

From Numerical Sensor Data to Semantic Representations:
A Data-driven Approach for Generating Linguistic Descriptions

To Sepideh
For her kindness and devotion,
and for her endless support.

Örebro Studies in Technology 78



HADI BANAEE

**From Numerical Sensor Data to Semantic Representations:
A Data-driven Approach for Generating Linguistic Descriptions**

© Hadi Banaee, 2018

Title: From Numerical Sensor Data to Semantic Representations:
A Data-driven Approach for Generating Linguistic Descriptions
Publisher: Örebro University 2018
www.publications.oru.se

Print: Örebro University, Repro 03/2018

ISSN 1650-8580
ISBN 978-91-7529-240-3

Abstract

Hadi Banaee (2018): From Numerical Sensor Data to Semantic Representations: A Data-driven Approach for Generating Linguistic Descriptions. Örebro Studies in Technology 78.

In our daily lives, sensors recordings are becoming more and more ubiquitous. With the increased availability of data comes the increased need of systems that can represent the data in human interpretable concepts. In order to describe unknown observations in natural language, an artificial intelligence system must deal with several issues involving perception, concept formation, and linguistic description. These issues cover various subfields within artificial intelligence, such as machine learning, cognitive science, and natural language generation.

The aim of this thesis is to address the problem of semantically modelling and describing numerical observations from sensor data. This thesis introduces data-driven approaches to perform the tasks of mining numerical data and creating semantic representations of the derived information in order to describe unseen but interesting observations in natural language.

The research considers creating a semantic representation using the theory of conceptual spaces. In particular, the central contribution of this thesis is to present a data-driven approach that automatically constructs conceptual spaces from labelled numerical data sets. This constructed conceptual space then utilises semantic inference techniques to derive linguistic interpretations for novel unknown observations. Another contribution of this thesis is to explore an instantiation of the proposed approach in a real-world application. Specifically, this research investigates a case study where the proposed approach is used to describe unknown time series patterns that emerge from physiological sensor data. This instantiation first presents automatic data analysis methods to extract time series patterns and temporal rules from multiple channels of physiological sensor data, and then applies various linguistic description approaches (including the proposed semantic representation based on conceptual spaces) to generate human-readable natural language descriptions for such time series patterns and temporal rules.

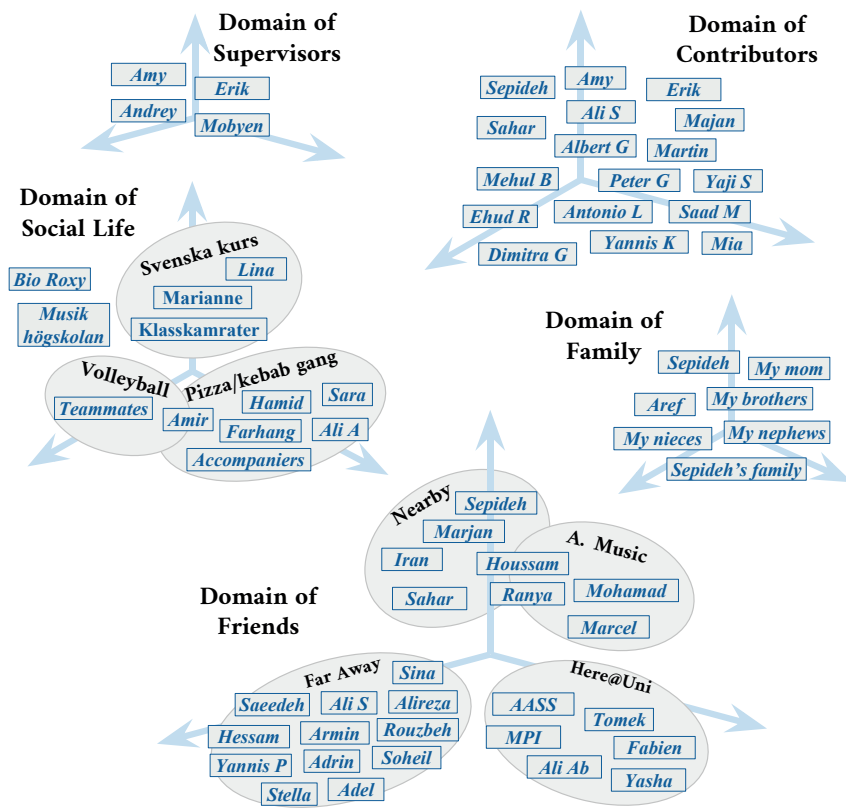
The main outcome of this thesis is the use of data-driven strategies that enable the system to reveal and explain aspects of sensor data which may otherwise be difficult to capture by knowledge-driven techniques alone. Briefly put, the thesis aims to automate the process whereby unknown observations of data can be 1) numerically analysed, 2) semantically represented, and eventually 3) linguistically described.

Keywords: Semantic representations, Conceptual spaces, Natural language generation, Temporal rule mining, Physiological sensors, Health monitoring system.

Hadi Banaee, School of Science and Technology
Örebro University, SE-701 82 Örebro, Sweden, hadi.banaee@oru.se

Acknowledgements

Conceptual Space of Acknowledgement¹:



¹ There are no linguistic descriptions generated for this conceptual space.

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Problem Statement	5
1.3	Research Question	6
1.4	Contributions	7
1.5	Thesis Outline	9
1.6	Publications	12
I	Creating Semantic Representations for Numerical Data	17
2	Background and Related Work	19
2.1	Semantic Representation	19
2.2	On the Theory of Conceptual Spaces	20
2.2.1	Identifying Quality Dimensions	24
2.2.2	Related Work on Conceptual Spaces and AI	25
2.3	Generating Linguistic Descriptions	27
2.3.1	Linguistic Descriptions of Data (LDD)	27
2.3.2	Natural Language Generation (NLG)	29
2.4	Conclusions	33
3	Data-Driven Construction of Conceptual Spaces	35
3.1	Domain and Quality Dimension Specification	37
3.1.1	Feature Subset Ranking	39
3.1.2	Feature Subset Grouping	41
3.2	Concept Representation	44
3.2.1	Convex Regions of Concepts	46
3.2.2	Context-dependent Weights of Concepts	49
3.3	Discussion	49

4	Semantic Inference in Conceptual Spaces	53
4.1	Symbol Space Definition	54
4.2	Inferring Linguistic Descriptions	56
4.2.1	Phase A: Inference in Conceptual Space	57
4.2.2	Phase B: Inference in Symbol Space	65
4.3	Discussion	70
5	Results and Evaluation: A Case Study on Leaf Data Set	73
5.1	Constructing a Conceptual Space of Leaves	74
5.1.1	Domain Specification for Leaf Data set	75
5.1.2	Concept Representation for Leaf Concepts	77
5.2	Semantic Inference for Unknown Leaf Samples	77
5.2.1	Inference in Conceptual Space of Leaves	78
5.2.2	Inference in Symbol Space of Leaves	78
5.3	Empirical Evaluation for Leaf Samples	78
5.3.1	Survey: Design and Procedure	80
5.3.2	Identifying Leaf Observations from Linguistic Descriptions	82
5.3.3	Rating Linguistic Descriptions of Leaf Observations . . .	83
5.4	Discussion	85

II Physiological Sensor Data: From Data Analysis to Linguistic Descriptions **89**

6	An Overview of Health Monitoring with Mining Physiological Sensor Data	91
6.1	Data Mining Tasks in Health Monitoring Systems	92
6.2	Data Mining in Health Monitoring Systems	95
6.2.1	Preprocessing	96
6.2.2	Feature Extraction/Selection	96
6.2.3	Modelling and Learning Methods	97
6.3	Data Sets: Acquisition and Properties	99
6.3.1	Sensor Data Acquisition	99
6.3.2	Sensor Data Properties	100
6.4	Discussion and Challenges	101
7	Physiological Time Series Data: Preparation and Processing	105
7.1	Input Time Series Sensor Data: Collection and Acquisition . . .	106
7.1.1	Wearable Sensors, Non-clinical Data	106
7.1.2	Clinical Physiological Data	107
7.2	Trend Detection in Physiological Time Series	109
7.3	Pattern Abstraction in Physiological Data	113
7.3.1	Background on Pattern Abstraction	113
7.3.2	Prototypical Pattern Abstraction	115

7.4	Discussion and Summary	117
8	Mining and Describing Physiological Time Series Data	119
8.1	Mining Temporal Rules in Physiological Sensor Data	120
8.1.1	Background on Temporal Rule Mining	120
8.1.2	A New Approach for Temporal Rule Mining	122
8.1.3	Temporal Rule Set Similarity	124
8.1.4	Results: Distinctive Rules in Clinical Settings	126
8.1.5	Evaluation of Rule Set Similarity in Clinical Conditions	128
8.2	Linguistic Descriptions for Patterns and Temporal Rules	131
8.2.1	Trend and Pattern Description	132
8.2.2	Temporal Rule Representation	133
8.3	Discussion and Summary	136
9	Linguistic Descriptions for Patterns using Conceptual Spaces	139
9.1	Constructing Conceptual Space of Patterns	140
9.1.1	Domain Specification for Time Series Pattern Data Set	141
9.1.2	Concept Representation for Pattern Concepts	143
9.2	Semantic Inference for Unknown Patterns	144
9.2.1	Inference in Conceptual Space of Patterns	144
9.2.2	Inference in Symbol Space of Patterns	145
9.3	Evaluation: Descriptions from Conceptual Spaces	147
9.3.1	Survey: Design and Procedure for Pattern Data Set	147
9.3.2	Identifying Pattern Observations from Linguistic Descriptions	147
9.3.3	Rating Linguistic Descriptions of Pattern Observations	149
9.4	Discussion and Summary	150
10	Conclusions	153
10.1	Summary of Contributions	153
10.1.1	Construction of Conceptual Spaces (C1)	153
10.1.2	Semantic Inference in Conceptual Spaces (C2)	154
10.1.3	Mining Physiological Sensor Data (C3)	155
10.1.4	Linguistic Description by Semantic Representations (C4)	155
10.2	Limitations	156
10.3	Societal and Ethical Impacts	158
10.4	Future Research Directions	159
10.5	Final Words	163
	References	165

List of Figures

1.1	A part of the Rumi’s poem in Persian, together with an illustration of the story, adapted from [7].	1
1.2	A schematic overview of the tasks to be performed in both theoretical (inner box) and application (outer box) focuses of the thesis.	6
1.3	Thesis MindMap, illustrating the appearance of the research tasks in the chapters, together with the contributions of this thesis.	14
2.1	A schematic presentation of a conceptual space of fruits.	24
2.2	The architecture of data-to-text systems, proposed by Reiter. . . .	31
3.1	Illustration of the main steps for constructing a conceptual space from a set of numeric data.	37
3.2	Two phases of the domain and quality dimension specification, with input and output parameters of each phase.	39
3.3	A weighted bipartite graph with two sets of vertices from the labels \mathcal{Y} and the selected features \mathcal{F}'	43
3.4	A bigraph graph and one selected biclique (blue edges) for the leaf example (explained in Example 3.5).	45
3.5	A concept representation example in a conceptual space with domains δ_a and δ_b	48
4.1	Illustration of the steps of inferring linguistic descriptions for an unknown observation via the constructed conceptual spaces and its corresponding symbol space.	54
4.2	Schematic of a conceptual space and the coupled symbol space. . . .	55
4.3	Two phases of the semantic inference for generating linguistic descriptions, with the input and output parameters of each phase. . . .	57
4.4	An illustration of four different cases with respect to the various positions of an instance points within domains.	60

4.5	Example of the membership functions of the linguistic terms <i>circular</i> , <i>elliptical</i> , and <i>elongated</i> , describing the <i>elongation</i> quality dimension.	65
4.6	Annotation and characterisation messages in AVM format	69
5.1	Six species as the known leaves in the leaf data set.	74
5.2	The bipartite graph presenting the relevance of features and labels in leaf data set.	75
5.3	The conceptual space of leaf data set.	76
5.4	A set of unknown leaf samples.	79
5.5	Screenshots from two types of questions designed for the survey.	81
5.6	The description of <i>leaf (a)</i> has been shown 31 times to the participants.	82
5.7	The box plot of the rating scores received for each of the models, deriving descriptions in <i>leaf</i> data set.	84
6.1	A schematic overview of the position of the main data mining tasks concerning the different aspects of the health monitoring systems.	94
6.2	A generic architecture of the primary data mining approach for wearable sensor data.	96
7.1	The wearable sensor, Bioharness3 [5], worn on the chest is able to locally store the measured data or wirelessly transmit it via Bluetooth.	106
7.2	An example of non-clinical measurements depicting thirteen hours of heart rate (HR, top) and respiration rate (RR, down) during sequential activities.	107
7.3	An example of clinical sensor data from MIMIC data set with variables heart rate (HR), blood pressure (BP), and respiration rate (RR).	109
7.4	An example of physiological time series data preprocessing and segmentation.	111
7.5	The output of partial trend detection algorithm for two segmented time series (HR on top and RR on bottom).	114
7.6	An example of trend detection outputs for two different resolutions of heart rate data.	115
7.7	An example of prototypical patterns for <i>HR</i> data in CHF condition.	117
8.1	Three temporal relations between one pattern P_1 in $P(t_{HR})$ and at most two patterns P_2 in $P(t_{BP})$	123
8.2	Result of cross-validation approach on a selected iteration for MI condition.	127

8.3	The number of rules for clinical conditions in $TRM-\rho^3$ and $TRM-\rho^1$ methods, in relation to the multivariate time series.	128
8.4	A selection of distinct temporal rules generated from physiological data in clinical conditions using $TRM-\rho^3$ approach.	129
8.5	Boxplot diagram of the occurrence ratios between one clinical condition's rule set and the other conditions.	131
8.6	The output of the partial trends for two segmented time series (HR on top and RR on bottom), shown in Figure 7.5.	134
9.1	Four sets of time series patterns, presenting the known classes of patterns in the data set.	141
9.2	The bipartite graph presenting the relevance of the features and the labels in data set of time series patterns.	142
9.3	The conceptual space of time series pattern data set: a graphical presentation of the determined domains with the corresponding quality dimensions and concepts.	143
9.4	A set of unknown samples of time series patterns.	144
9.5	Pie charts, showing the expertise level of the participants by measuring how familiar they are with the introduced terminology for class labels and features.	148
9.6	The box plot of the rating scores received for each of the models deriving descriptions in <i>pattern</i> data set.	149
10.1	Human-specified semantic features for <i>animal</i> domain.	160
10.2	Yet another illustration for the story of <i>The Elephant in the Dark</i> , adapted from [194].	163

List of Tables

2.1	Typical modules and tasks of an NLG system	29
5.1	The linguistic descriptions derived for the unknown leaf samples in Figure 5.4.	79
5.2	The overall scores calculated from rating responses to the different models in <i>leaf</i> data set.	84
5.3	Summary of the one-way ANOVA and Wilcoxon tests for the rating scores with respect to the models deriving descriptions. .	85
6.1	The summarisation of the most commonly used features of each wearable sensor data in the literature.	97
7.1	Clinical conditions and their subjects in mimic database, after removing unreliable measurements.	108
8.1	Occurrence ratios of rule sets for each pair of clinical conditions in multivariate time series $HR \& RR$, using $TRM\text{-}\rho^3$	130
8.2	Subjects with the same condition in their nearest rule sets. . . .	132
8.3	The instances of linguistic terms used for describing trends. . . .	133
8.4	A template-based textual representation of rules with temporal relations.	135
8.5	Textual representation of the acquired rules in Figure 8.4. . . .	136
9.1	The linguistic descriptions derived for the unknown samples of time series patterns in Figure 9.4.	146
9.2	The overall scores calculated from the rating responses to the different models in <i>pattern</i> data set.	149
9.3	Summary of the one-way ANOVA and Wilcoxon tests for the rating scores with respect to the models deriving descriptions. .	150

List of Algorithms

3.1	Feature Subset Ranking	41
3.2	Feature Subset Grouping	44
3.3	Concept Representation	46
4.1	Inference in Conceptual Space	61
4.2	Inference in Symbol Space	66
7.1	Partial Trend Detection	113
8.1	$\text{RuleMatch}(r, R, I_R)$	125
8.2	$\text{Occurrence}(R_1, R_2, I_{R_2})$	125

Chapter 1

Introduction

RUMI, the 13th century Persian poet and teacher of Sufism, has a story called *The Elephant in the Dark*¹ in his extensive book of poetry, *Masnavi*. In his retelling:

Some Hindus bring an elephant to be exhibited in a dark room. A number of men touch and feel the elephant in the dark and, depending upon where they touch it, they believe the elephant to be like a water spout (trunk), a fan (ear), a pillar (leg) and a throne (back)... [191].

عرشد را آورده بودندش بنود	بیل اندر خانه تاریک بود
اندر آن غلطی نمی شد حرکتی	از برای دیدنش مردم بسی
اندر آن تاریکی اش کف می رسود	دیدنش با چشم چون ممکن نبود
گفت: همچون ناودان است این نهاد	آن یکی را کف به خرطوم اوقعا
آن بره چون بادبزن شد پدید	آن یکی را دست بر گوشش رسید
گفت خود این چهل چون تختی بدست	آن یکی بر پشت او بنهاد دست
فهم آن می کرد در جایی شنید	بچنین حرکتی که جزوی که رسید
آن یکی داشت لقب داد این الف	از نظر که، گفتشان شد مختلف

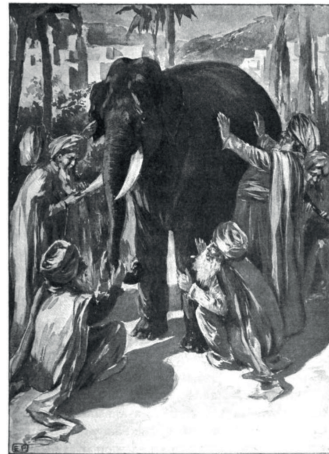


Figure 1.1: A part of the Rumi's poem in Persian, together with an illustration of the story, adapted from [7].

¹This parable originated on the Indian subcontinent and is known as “Blind men and an elephant” (See Figure 1.1).

Rumi uses this parable to demonstrate the problem of perception limitations. The individuals have their own perceptions of the elephant (an unknown concept for them) and therefore use their own *inference* to explain it. This problem is closely related to the problem of describing a concept based on the perceived information. The behaviour of the men in this story relied on a general cognitive process for learning concepts. They first perceived what they could observe (or more generally, could sense in the case of touching the elephant) and they then sought to map or categorise the perceived information according to similar concepts that were known to them. This process is known as *exemplar theory* in cognitive science. However, their failure to successfully characterise or describe the concept of *Elephant* was due to the limitations of their sensory perceptions, which led to overreaching misinterpretations [1].

Describing unknown observations in natural language appears to be an easy task for humans. Both speakers and hearers have a great deal of common sense understanding of the concepts and properties that they have encountered in life, and that enable them to describe such observations. For example, when J. K. Rowling presents the legendary creature of “Hippogriff” in her book *Harry Potter and the Prisoner of Azkaban*, she describes it as follows:

Hippogriffs have the bodies, hind legs, and tails of horses, but the front legs, wings, and heads of giant eagles, with cruel, steel-coloured beaks and large, brilliantly orange eyes. The talons on their front legs were half a foot long and deadly-looking... [190].

Rowling uses known or familiar concepts that are most similar (eagle and horse), together with perceivable features (orange, long, large, etc.) to explain the novel creature that she introduces². In doing so, she uses linguistic terms that are cognitively understandable for humans.

However, still, deriving descriptions for unknown concepts is no trivial task in *artificial intelligence* (AI). The challenge is how an artificial intelligence system can perform the task of semantically describing unknown observations by relying on a set of perceived information. This task becomes more crucial if the information given to the system is in the form of numeric or non-symbolic measurements (e.g., sensor data).

1.1 Motivation

Deriving semantics from real-world numerical observations such as sensor data has become increasingly important for creating a common understanding of information with humans. Artificial intelligence systems that can augment observations from sensor data in order to create conceptual representations are needed for applications that require interaction with humans in natural language. The motivation for the problem of the semantic description of numeri-

²The hippogriff was first mentioned by the Roman poet Virgil in his *Eclogues*, but the word hippogriff is derived from the ancient Greek.

cal data has its origin in the topic of knowledge representation in AI research. However, the initial motivation in this thesis comes from a real-world application for analysing sensor data in healthcare monitoring systems. The rest of this section explains both the theoretical and the application-based motivations that underlie the problem to be considered.

One goal of cognitive science is to construct artificial systems that can understand and model the cognitive activities of humans, such as concept learning and semantic inference [14]. However, a key issue is how the given information is to be modelled in knowledge representation frameworks [85, 87]. The two paradigms of *symbolic* and *sub-symbolic* representations have been the two dominant (and sometimes competing) approaches to addressing the issue of representation in AI [85, 219]. *Symbolic* approaches use explicit symbols as primitives when performing symbol manipulation in order to model high-level abstract concepts [155]. *Sub-symbolic* (connectionist) approaches often focus on the categorisation tasks per se. They process the activation patterns of input concepts at the perceptual level, using internally connected units of artificial neural networks [219].

With regard to the task of the semantic description of concepts by means of perceived data (henceforth, *concept description*), this thesis highlights two AI problems. The first is the problem of *induction* or, more generally, the issue of *learning*. Inductive inference performs a generalisation from a limited number of observations, which infers the characteristics of the concepts. Induction is highly related to the task of *concept learning* or *concept formation* in cognitive science [84]. The second problem is related to *semantics* or, in general terms, the issue of *explainability* or *interpretability* in AI. Semantic inference is the process of inferring meaningful descriptions or truth conditions from a set of (semantically enriched) information, which is usually represented in the form of logical or natural sentences. Neither of the representational approaches (symbolic and sub-symbolic) satisfactorily addresses the AI problems noted above (concept learning and semantic inference) simultaneously. Concept learning is a difficult task for symbolic approaches, since the symbolic AI has been formalised on the basis of rule-based representations of the logical source of knowledge, rather than having intrinsically learnt rules from observations [219]. Semantic inference cannot be addressed by the sub-symbolic approaches since there is no interpretable transformation from the low-level information of the model to the organised high-level symbols. In other words, the sub-symbolic approaches are unable to explain *what* the emerging learnt model represents [89].

Consequently, the theory of *conceptual spaces* was introduced by Gärdenfors [86] as a mid-level representation between the symbolic and the sub-symbolic approaches to addressing both the concept learning and the semantic inference problems [14, 116]. The theory of conceptual spaces presents a framework that consists of a set of *quality dimensions* in various *domains*. These are placed within a geometrical structure in order to model, categorise, and represent the

concepts in a multi-dimensional space [86]. In the literature, conceptual spaces are principally derived in a *knowledge-driven* manner. These spaces operate on the assumption that there is prior knowledge from perceptual mechanisms or experts that manually initialises the elements of the conceptual space (i.e., domains, quality dimensions, and concepts' regions) [10,189]. However, since this thesis relies on observed input data, the motivational challenge arises of how to automatically construct a conceptual space from the given information [131] in order to perform concept learning and semantic inference tasks. Performing these task are the required steps to address the problem of concept description (especially to describe unknown observations). This is an important motivation, due to a growing class of problems that involves more complicated input observations. These problems deal with raw sensor data that have little or no prior knowledge concerning their semantic significance [188]. Therefore, specifying the interpretable elements of a representational model for such problems is no trivial task.

New applications in several domains are increasingly required to rely on automatic methods for forming concepts directly from sensor data. One area that sensor data is crucially used in is the field of health monitoring, particularly in clinical conditions. Monitoring the vital signs of the subjects (whether healthy or unhealthy) is essential if the medical domain is to identify the different behaviours of the health parameters as symptoms of various medical diseases [28]. Such physiological parameters include *heart rate*, *respiration rate*, *blood pressure*, etc. Various sensors have been developed to measure these vital signs in both clinical conditions and home healthcare systems, and these accumulate a massive amount of data. With regard to *wearable sensors*, the continuous parameters in the form of time series signals are of particular interest in this research (in contrast to discrete ones), since it is more critical to consider the high resolution sequence of information (with very small time-stamps).

Critical challenges in the field of health monitoring include not only statistically mining massive amount of physiological time series data, but also discovering understandable interpretations for the extracted information [32]. A new aspect of analysing sensor data will involve going beyond expert knowledge and recognising information (e.g., events, patterns, anomalies) that is not pre-defined by the system. This aspect will lead to the analysis phase being conducted in a *data-driven* way in order to reveal information that was not previously seen, but is worth analysing and interpreting. A motivational example is the exploitation of interesting trends and patterns in sensor time series data. A knowledge-based system may look for the known pre-defined trends asked by the experts (e.g., peaks of heart rate signals). However, one can also analyse the data itself in order to detect interesting and meaningful patterns found throughout the data (e.g., repeated sudden drops of heart rate while sleeping). These are not necessarily requested by or even visible to the experts, but they are worth being reported. Nevertheless, the primary challenge remains finding

a suitable way to describe such data-driven extracted information. Here is the point where this thesis argues that a semantic representation can play the role of modelling physiological time series in order to describe unknown observations or patterns in natural language.

1.2 Problem Statement

The problem of semantically modelling and describing the unknown observation, which is termed the general semantic representation problem, forms the core of this research. This thesis first considers the task of semantic representation in describing the numerical observations (i.e., the theoretical focus of this thesis). The semantic representation task investigates representational models in order to be able to bind perceived numerical data as input into a set of linguistic characterisations as output.

In addition to considering this problem at the theoretical level, its instantiation is also presented with regard to a practical problem in a real-world application. This research proposes a specific use of semantic representation for the task of linguistically describing the unknown time series patterns derived from physiological sensor data (i.e., the application focus of this thesis). The proposed approach investigates data mining methods for analysing a set of raw physiological time series data as input, and it considers linguistic description approaches to generate a set of human-readable natural language text as output. Figure 1.2 presents a schematic overview of the tasks to be performed in both the theoretical focus and the application framework of this research.

The primary assumption for the stated problem is that there is no adequate expert or domain knowledge to be fed to the model in order to describe unseen but interesting observations extracted from a data set. Therefore, the proposed approach relies on the observed data and its own features in order to provide linguistic descriptions. This means that the objective is not to describe the observations on the basis of a set of pre-defined knowledge within a given knowledge representation (such as describing an *eagle* or a *horse* using an ontology of animals that perfectly defines the concepts of eagle and horse). Rather, the aim is to build a semantic representation of the domain using the known data set and its properties. Moreover, it is to extract and characterise the unseen but interesting observations within this representation (such as describing a *hippogriff* within a semantic representation of animals that includes the known concepts of eagle and horse, but does not necessarily include the new concept of hippogriff).

All the tasks defined within this solution have some assumptions regarding their inputs and outputs. For the semantic representation task, it is assumed that the input is a set of perceived or processed numerical information. This information is in the form of human understandable attributes (known as *semantic features*). Depending on the format of the raw input data, the semantic features can be either directly captured or be computed from the given data. *Skin colour* is an example of the former type of features that can be included

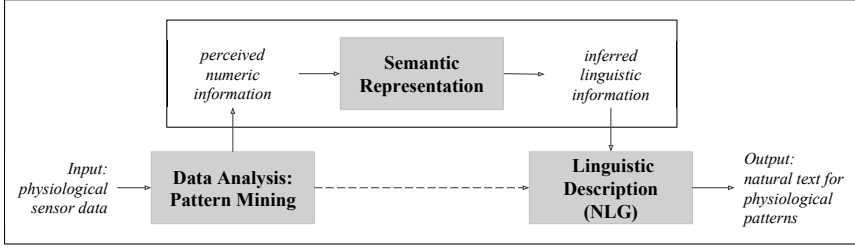


Figure 1.2: A schematic overview of the tasks to be performed in both theoretical (inner box) and application (outer box) focuses of the thesis.

in a data set of animals. However, the *fluctuation* of a signal is an example of the latter features that can be computed for a time series in a data set of sensor records. Therefore, the considered application for the physiological sensor data involves data analysis phase in order to extract the meaningful features that to be supplied to the semantic representation. Nevertheless, the proposed model for the task of semantic representation can be utilised on different types of data sets that include semantic features in their observations.

The output of the semantic representation is assumed to be a set of inferred linguistic information that characterises the input numerical observations. These linguistic characterisations (i.e., words) further need to be planned and realised into natural language messages (i.e., sentences) by means of *natural language generation* (NLG) systems. Although the text generation phase is particularised as a task in the framework for physiological sensor data, it can generally be applied to any linguistic characterisation that emerges from the semantic representation model. It is worth to mention that although the focus of this research is on the specific set of physiological time series data (which was the initial motivation for this thesis from the field of healthcare monitoring), the development of the framework is applicable as a *data-to-text* system in a variety of applications that deal with the problem of the linguistic description of data.

1.3 Research Question

The overall *objective* that this dissertation aims to achieve can be expressed as follows:

- O: “*Proposing data-driven approaches beyond expert knowledge to perform the tasks of (1) mining numerical observation (e.g., sensor data), and (2) the semantic representation of the derived information, in order to (3) describe unseen but interesting observations in natural language.*”

As noted in the problem statement, the semantic representation problem deals with a set of derived information (perceived or processed features). Therefore, one research question related to the theoretical part of the research focuses on the role of *semantic representation* and *linguistic description*, which is related to tasks (2) and (3) in objective O. Generally speaking, the question is

R1: “*How can perceived numerical information be semantically represented and described in a linguistic form?*”

This can be divided into more specific research questions as follows:

- *How can a semantic representation of a numerical data set automatically be created in a data-driven manner, based on the observed exemplars and their semantic features?*
- *How can these observations be conceptualised within the representation constructed in order to form the concepts of the data set’s domain?*
- *How can this representation be utilised in order to infer linguistic characterisations for a set of new unknown observations in natural language?*

Likewise, with regard to the problem statement about the application part of this thesis, the framework is dedicated to turning raw numerical data into natural language text. More specifically, however, it is concerned with considering real-world data sets from the field of healthcare (i.e., time series patterns derived from physiological sensor data). Therefore, the research question in the application part focuses on the role of *data analysis* and the *linguistic description* of patterns, which is related to tasks (1) and (3) in objective O. Generally speaking, the question is

R2: “*How can unseen but interesting patterns in physiological sensor data be extracted, represented, and linguistically described?*”

This can also be divided into more specific research questions as follows:

- *How can data-driven approaches be applied in order 1) to extract distinctive time series patterns in different clinical settings, and 2) to mine the temporal relations between multi-channels of physiological sensor data?*
- *How can a data-driven conceptual space be built as a semantic representation of the time series patterns in order to infer linguistic characterisations for the patterns?*
- *How can meaningful, interesting, and useful messages be generated in natural language for numerical patterns and their temporal relations using the proposed semantic model?*

1.4 Contributions

Given the general objective O, the core of the contributions in this thesis is the notion of *data-driven*. This applies to both the research questions, which

are concerned with semantically representing the perceived information, and numerically extracting interesting sensor patterns.

The contributions of this thesis are organised into four items. The first two contributions C1 and C2 (which address the research question R1), present an automatic construction of a conceptual space from numerical data, which can then be used to infer the semantic interpretation of novel unknown observations. The other two contributions C3 and C4 (which address the research question R2), present the automatic ways of extracting prototypical time series patterns from various physiological sensor data using temporal rule mining. In addition, they apply various linguistic description methods (including the proposed conceptual spaces) in order to generate text for such time series patterns.

C1: *Data-driven construction of conceptual spaces:*

In the theoretical part of this work, a data-driven approach is proposed in order to automatically *construct* conceptual spaces and perform *concept formation* based on the input exemplars of a numerical data set. Instead of initialising the domains and dimensions of the conceptual space from a priori knowledge or forming concepts in a rule-based manner, the proposed data-driven process automatically determines the relevant domains and dimensions on the basis of its ability to distinguish between exemplars of different concepts. This determination is performed using *machine learning* (ML) techniques in order to identify the relevant features of the observed data with the primitive concepts. Furthermore, an instance-based approach is presented for concept formation in which a concept is formed within the conceptual space on the basis of the spatial representation of its observed instances.

C2: *Semantic inference in the conceptual spaces:*

This thesis further proposes a semantic inference process for the conceptual space, which is built in order to provide explainability for the novel unknown observations in natural language descriptions. By grounding a semantic representation, the conceptual space is employed in order to provide a link between the numerical data and the semantic characterisation of the concepts. For a new observation, its representation within the conceptual space specifies either its *inclusion* in the known concepts or its values with regard the quality dimensions. This set of information is then mapped to a *symbol space* that infers the linguistic characterisation of the observation based on a set of concepts and features involved. This work will demonstrate the utility of conceptual spaces as a solution to the task of *content determination* within a natural language generation (NLG) system.

C3: *Data-driven mining of prototypical patterns and temporal rules in physiological sensor data:*

With regard to the task of data analysis for raw sensor data, this work focuses on data-driven approaches in order 1) to extract the prototypical patterns that frequently occur in the recorded signals of various health parameters such as heart rate, respiration rate, blood pressure, etc., and 2) to mine the temporal rules of the extracted patterns in order to find interesting relations between them. This data analysis is performed using unsupervised data mining techniques to discover unseen (but worthy of extraction) information beyond the expert knowledge. Moreover, the investigation of the temporal rule mining leads to an interesting outcome regarding the uniqueness of the extracted rules for different clinical conditions. This study shows the distinctive co-occurrence of prototypical patterns for each clinical condition.

C4: *Linguistic description of time series patterns using semantic representations:*

Given that the aim of the framework is to generate natural language text for physiological data, the processed time series data in the form of patterns are then used as input for the linguistic description approaches. In addition to the template-based methods for directly generating text for a pattern using stored features, one contribution of the work is to apply the proposed semantic representation in order to automatically build the conceptual space of the time series pattern into a physiological sensor data set. This conceptual space is then utilised to infer linguistic characterisations for such numerical data. An empirical evaluation of the output text for the time series pattern demonstrates the advantages of developing such a conceptual space as a content determination solution for an NLG system.

The intersection of the contributions is the idea of proposing *data-driven* approaches that are able to present new aspects of the sensor data (both numerically and symbolically) in a data set in which the observations are not catchable by knowledge-driven approaches, are nevertheless worth being 1) numerically extracted, 2) semantically represented, and 3) linguistically described.

1.5 Thesis Outline

This dissertation is composed of two parts, which consider the division of the problem statement according to the theoretical and the application focuses. The first part presents the theoretical focus which is the task of semantic representation, while the second part presents the application focus which is the tasks of numerically analysing and generating linguistic descriptions for time series patterns.

Part I: Creating semantic representations for numerical data

The first part of this thesis presents the notion of data-driven conceptual spaces as a semantic representation tool for linguistically describing the numerical data. After an overview of the background and the related work (Chapter 2), this part explains how a conceptual space can be constructed using the information observed (Chapter 3), before proceeding to explain how it can be utilised to infer semantics for the unseen observations (Chapter 4). This process is then be assessed by applying the approach to a data set of *leaf* examples (Chapter 5).

Chapter 2 begins with a discussion about the notion of semantic representations in the literature. It then presents the background to the theory of conceptual spaces, together with related work on the role of this theory in AI. It also includes background and related work on the linguistic description and natural language generation approaches. Finally, a brief discussion summarises the need to apply conceptual spaces in order to perform some tasks in existing NLG solutions.

Chapter 3 proposes a data-driven approach in order to automatically construct conceptual spaces. It starts by defining the parameters of a numerical data set. It then explains the steps involved in determining the quality dimensions and the domains of a conceptual space based on the most relevant features of the data. Furthermore, the chapter includes an instance-based algorithm for concept representation in the domains of the derived space.

Chapter 4 presents the design of an inference approach in order to provide a semantic characterisation of the novel unknown observations within the proposed conceptual space in Chapter 3. This chapter demonstrates the steps involved in checking the inclusion of a new instance within the domains. It then assigns the related linguistic terms using a defined symbol space based on the associated concepts and quality dimensions of the instance. Finally, the chapter presents the steps for performing the realisation task in order to turn linguistic terms into natural language messages.

Chapter 5 presents a case study that demonstrates the applicability of the approach proposed in Chapters 3 and 4 using a data set of leaf images. This chapter first describes the algorithms 1) to construct a conceptual space of leaves, 2) to represent each concept of leaves within the domains, and 3) to generate linguistic descriptions for a set of unknown leaf examples. Finally, it describes an empirical evaluation method for measuring the appropriateness of the developed space by using the worthiness of the messages generated.

Part II: Physiological sensor data: From data analysis to linguistic descriptions

The second part of the thesis focuses on the entire framework of describing the numerical data using the semantic model proposed in Part I, but specifically for physiological sensor data. A series of data analysis approaches are investigated in order to prepare suitable information (i.e., time series patterns) that is fed to the semantic representation. Chapters 6 to 8 are largely dedicated to this task. First, The current state of the art for the data mining approaches on sensor data is discussed (Chapter 6). Then, after exacting unseen but interesting time series patterns (Chapter 7) and the temporal rules between the patterns (Chapter 8), it is demonstrated how the proposed semantic model in Part I can be applied to linguistically describe these patterns in such a framework (Chapter 9).

Chapter 6 provides a survey of work on existing data mining approaches in order to analyse wearable sensors in health monitoring systems. It also considers the new trends in the field and the current challenges to data analysis in the healthcare domain.

Chapter 7 introduces the processes of collecting and acquiring the input physiological sensor data sets that were used in this work. It then describes the various unsupervised approaches first used to detect partial trends and to extract prototypical patterns in different channels of physiological time series data.

Chapter 8 presents a novel modified temporal rule mining approach to discovering the co-occurrence of prototypical patterns among various time series data of vital signs in a clinical condition. This chapter further describes the algorithms for comparing the extracted rules in order to show the uniqueness of the rules for different clinical conditions. Another central aspect of this chapter is the presentation of a template-based natural language generation method for describing temporal rules of patterns in natural language.

Chapter 9 presents the process of applying the proposed semantic representation in Part I in order to automatically construct the conceptual space of the abstracted physiological patterns. It then describes the process of inferring semantic descriptions for a set of unknown patterns. Furthermore, this chapter presents an evaluation to demonstrate the appropriateness of the conceptual space of the patterns used for text generation, together with a comparison between the output text of the physiological pattern from the semantic model and the output text of the template-based approaches.

See Figure 1.3 for the MindMap of the thesis, which illustrates the appearance of the research tasks in the chapters, together with the thesis' contributions. After presenting parts I and II,

Chapter 10 concludes the thesis by presenting a summary of the contributions, a discussion of the limitations of the proposed semantic representation and application framework, and finally an overview of the possible directions for the future work.

1.6 Publications

The contributions presented in this thesis have been published in the following journal and conference papers:

- Hadi Banaee, Erik Schaffernicht, and Amy Loutfi. Data-driven conceptual spaces: Creating semantic representations for linguistic descriptions of numerical data. *Journal of Artificial Intelligence Research*, submitted, 2018.
- Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. *Sensors*, 13(12):17472–17500, 2013.
- Hadi Banaee and Amy Loutfi. Data-driven rule mining and representation of temporal patterns in physiological sensor data. *IEEE journal of biomedical and health informatics*, 19(5):1557–1566, 2015.
- Hadi Banaee and Amy Loutfi. Using conceptual spaces to model domain knowledge in data-to-text systems. In *8th International Natural Language Generation (INLG) Conference, Philadelphia, USA*, pages 11–15, Association for Computational Linguistics, 2014.
- Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. Descriptive modelling of clinical conditions with data-driven rule mining in physiological data. In *8th International Conference on Health Informatics (HEALTH-INF)*, Lisbon, Portugal, pages 103–113, 2015.
- Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. A framework for automatic text generation of trends in physiological time series data. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Manchester, UK, pages 3876–3881. IEEE, 2013.
- Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. Towards NLG for physiological data monitoring with body area networks. *14th European Workshop on Natural Language Generation (ENLG)*, Sofia, Bulgaria, pages 193–197, Association for Computational Linguistics, 2013.

The following publications appeared as additional results of this research, but are not included in this thesis:

- Mobyen Uddin Ahmed, Hadi Banaee, and Amy Loutfi. Health monitoring for elderly: An application using case-based reasoning and cluster analysis. *ISRN Artificial Intelligence*, vol. 2013, 2013.
- Hadi Banaee and Amy Loutfi. What I talk about when I talk about data: Descriptive modelling of data in data-to-text systems. In *1st International Workshop on Data-to-Text Generation, Edinburgh, UK*, 2015.
- Mobyen Uddin Ahmed, Jesica Rivero Espinosa, Alenka Reissner, Àlex Domingo, Hadi Banaee, Amy Loutfi, and Xavier Rafael-Palou. Self-serve ict-based health monitoring to support active ageing. In *8th International Conference on Health Informatics (HEALTHINF), Lisbon, Portugal*, 2015.
- Mobyen Uddin Ahmed, Hadi Banaee, Xavier Rafael-Palou, and Amy Loutfi. Intelligent healthcare services to support health monitoring of elderly. In *Internet of things. user-centric IoT*, pp. 178-186. Springer, Cham, 2015.

Thesis MindMap

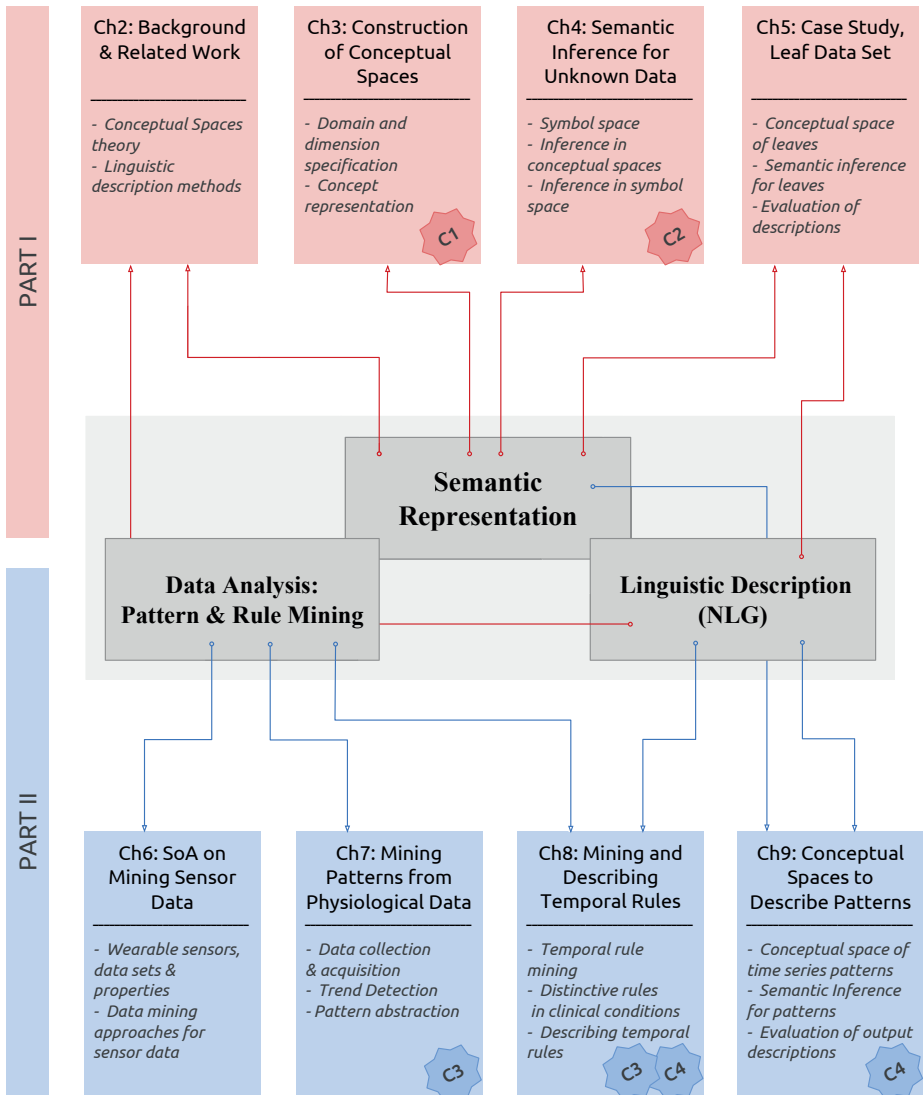
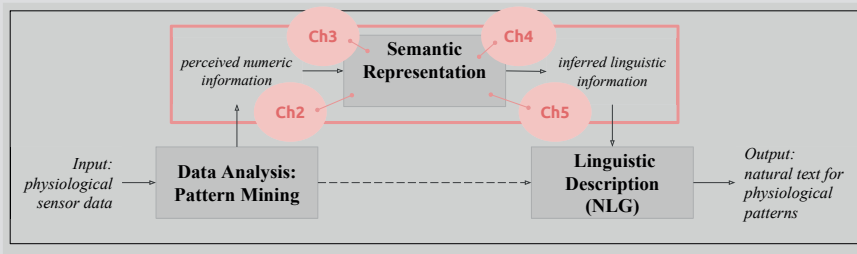


Figure 1.3: Thesis MindMap, illustrating the appearance of the research tasks in the chapters, together with the contributions of this thesis.

Part I

Creating Semantic Representations for Numerical Data

Part I of this thesis presents the notion of data-driven conceptual spaces as a semantic representation tool for linguistically describing numerical data. After an overview of the background and the related work (Chapter 2), this part explains how a conceptual space can be constructed using observed information (Chapter 3), before proceeding to explain how it can be utilised to infer semantics for unseen observations (Chapter 4). This process is then exemplified by applying the approach to a data set of *leaf* examples, and performing an empirical evaluation on the goodness of the derived descriptions for unknown samples (Chapter 5).



Chapter 2

Background and Related Work

“The world is my world: this is manifest in the fact that the limits of language (of that language which alone I understand) mean the limits of my world.”

— Ludwig Wittgenstein (1889–1951)

IN this chapter, the background and related work for this thesis are presented. This chapter focusses in particular on two fields: concept formation and concept acquisition using conceptual spaces, and approaches for generating linguistic descriptions from data. The chapter begins with a discussion about the term of *semantic representation* and the way that it has been used throughout this thesis. The theory of conceptual spaces then is introduced, together with a discussion about its use in artificial intelligence research. The background and the related work of linguistic descriptions of data (LDD) and natural language generation (NLG) are presented, together with the role of knowledge acquisition (KA) within the field of NLG. Moreover, a brief discussion summarises the advantage of using conceptual spaces as a semantic representation that can be later used for to perform NLG tasks.

2.1 Semantic Representation

The notion of a *semantic representation* has been used in a variety of ways in different areas such as knowledge representation in AI, cognitive science, and philosophy of language. Two prominent traditions for semantic representations exist [108]. One is to study the semantics of words by representing the relations of the words in natural language. For such representations, also called *amodal* approaches [78], input is linguistic information. Some computational models such as *semantic space* and *semantic networks* are examples of this kind of semantic representations in the field of linguistics [78, 108]. Another tradition

focuses on conceptual structures for the representation of meanings, which considers the relations between concepts and percepts or actions to model the semantics. In this case of semantic representations, also called *experiential* [225], the input is a set of perceptual information such as sensory data, memory, etc. The origin of this kind of semantic representation is the study of cognitive semantics, wherein the focus is on the meaning of the concepts as a cognitive phenomenon [19]. Cognitive semantics considers the meaning of linguistic expressions as mental entities coming from our perceptions. This perceptual information later is formed as concepts in our mind. This point of view is against the *realist* approaches that define semantics as something out in the world [86]. From the realist point of view, semantics can be represented using e.g., abstract propositions and description logic, and can be modelled and verified by truth conditions. This definition is highly related to the way that semantics has been used and modelled in the field of knowledge representation in AI [164]. The realist approach, also called *truth-conditional semantics*, seeks to connect language with statements about the real-world in the form of meta-language statements (e.g., in propositional or predicate calculus) [68, 164]. Within Cognitive semantics, however, meaning is a conceptual structure that comes before the truth [86]. Gärdenfors in his recent book, *The geometry of meaning* [90], includes another principle to the cognitive semantics as: ‘*a semantic theory shall account for the relation between perceptual processes and meaning*’¹. In other words, semantic representations, from the cognitive point of view, should be a conceptual structure which represents both perceptual and linguistic information.

The notion of a semantic representation in this thesis follows the latter mentioned definition, by first constructing a conceptual representation using perceptual information (numerical sensor data) and linguistic information (symbolic annotations), and then inferring semantically enriched descriptions. Therefore, in this thesis, a semantic representation of knowledge, as a core issue in the field of cognitive science, provides a conceptual structure for the meaning of perceived concepts [108]. This kind of representation eases the task of semantic reasoning of the perceived information, i.e., inferring the meaning of the concepts or words in language [14].

2.2 On the Theory of Conceptual Spaces

One approach to create semantic representations from perception is to use the *theory of conceptual spaces* introduced by Gärdenfors [86] as a knowledge representation framework at the conceptual level which relies on the paradigm of cognitive semantics.

¹Azzouni likewise discusses this principle in his book, *Semantic Perception* [25], where the contents of human perception sometimes involve semantic properties (e.g., meaningfulness). Thus, he argued that meaning is perceived, not inferred.

In cognitive science, both *explanatory* goals and *constructive* goals are considered. Explanatory goals aim to formulate the theories of cognition by studying the cognitive behaviour of humans or animals. Constructive goals aim to construct systems to fulfil cognitive tasks by building artefacts such as chess-playing programs and robots [87]. Both types of goals are dealing with the problem of *representations* in cognitive science to model how humans understand concepts.

From the AI point of view, two prominent approaches have been studied for the problem of modelling representations. *Symbolic* approaches are top-down representations that aim to model the high-level concepts using symbol manipulation. Abstract concepts are labelled by symbols, and the relations between the concepts are defined in a rule-based manner. Thus, inferences are logical and often are the result of first-order operations between the symbols and concepts [9, 116]. *Associationism* approaches aim to model cognition in a similar way that the neural structure of the brain associates the properties into an assumed concept. *Connectionism* as a particular case of this approach attempts to model the brain. This approach represents the interactions between the simple units as artificial neural networks in order to model or generate complex behaviours.

In both symbolic and associationism approaches, the main drawback is the lack of modelling various tasks of cognition such as concept learning, semantic similarity, and concept combination, simultaneously [9]. In the literature, conceptual spaces have been introduced as a mid-level representation model in cognitive systems between the high-level symbolic representation and low-level associationistic representations [14]. The aim of representing knowledge in a conceptual space is to develop an intuitive interpretation of the relationship between symbolic and sub-symbolic information [14, 87]. This theory explores how various types of information can be conceptually represented, both from an explanatory perspective and for constructing an artificial system [87]. Such a conceptual representation is the most important cognitive function, that, according to Hampton [111], “*stands at the centre of the information processing flow, with input from perceptual modules of differing kinds, and is centrally involved in memory, planning, decision-making, inductive inferences and much more besides*”. The ability to perceived information on a conceptual level relates the theory of conceptual spaces to the considering semantic representations.

Conceptual spaces are the geometric representations of how humans perceive, understand and learn concepts. Mainly, conceptual spaces are defined as geometric or topological structures that model, categorise and represent concepts in a set of multi-dimensional domains [87, 116]. The following is the description of the elements of a conceptual space as proposed by Gärdenfors [86]. The formal reformulation of these elements is presented later in Chapter 3.

- **Quality Dimensions:** A conceptual space consists of a set of *quality dimensions* (i.e., cognitively meaningful attributes). The quality dimensions present the quality attributes of objects in a metric space based on their measured quality values. Some examples of quality dimensions can be notions like *height*, *width*, and *depth*. Quality dimensions can be either interpreted as *phenomenal* (psychological) or *scientific* (theoretical) dimensions. The psychological interpretation of quality dimensions represents the phenomenal human responses in a semantically meaningful way, which are coming from human perceptions. *Colour* perception is a phenomenal example that can be described by three quality dimensions: *hue*, *saturation*, and *brightness*. The scientific interpretation is defined based on scientific theories which measure the values associated with e.g., sensors that measure wavelengths [86]. This distinction is tightly related to the mentioned goals of cognitive science. When the target is explanatory, the phenomenal interpretations of dimensions are in focus, and when the goal is constructive, the scientifically modelled dimensions are considered. Different scales of measurements including nominal, ordinal, interval, and ratio are used to assign values to the observations [86]. The values of these measurements on quality dimensions can be categorical (e.g., blood type) or continuous (e.g., size). Thus, these measurements enable the quality dimensions to calculate *distance* value between each two measured values. Depending on the type of the features (categorical or continuous) for the perceived information, different distance measures can be defined for each domain, separately.
- **Domains:** A *domain* in a conceptual space is represented as a set of interdependent quality dimensions which are logically integrated. A typical example of a domain is *colour* which is presented as a three-dimensional space in Figure 2.1. *Shape*, *taste*, *size*, and *weight* are other examples of perceptual domains. Some of the domains, like *weight* can be presented by a single dimension. The main reason to decompose the structure of conceptual spaces into domains, as Gärdenfors proposed, is to assign concepts to different quality attributes independently. As an example, an object can be independently described by its *colour*, without any need to consider its *weight*. According to the original definition of conceptual spaces, quality dimensions that depend on each other in forming a domain are considered to be *integral* dimensions, as opposed to *separable* ones [86]. Thus, within a domain, one cannot logically assign a value to one dimension without assigning values to the other dimensions. For example, a point within the colour domain cannot be defined with brightness and hue but without saturation.
- **Properties:** A property is a region in a single domain. As an example, *green* is a property corresponding to a region in the colour domain. Natural *Properties* are the convex regions expressing a particular attribute

of the domain. In natural language, properties are often associated with *adjectives*. Grasping the properties of a domain is not necessarily an intuitive task unless specified by the domain knowledge [86].

- **Concepts:** *Concepts* in a conceptual space are represented as a set of regions through multiple domains, and are modelled as n-dimensional areas in the space. A concept is described based on its properties in various domains. For instance, the concept of *Apple* in a conceptual space of fruits can be represented as a set of regions in the colour, shape, size, taste, and weight domains respectively. In natural language, the concepts often correspond to *nouns*. Some domains may be more *salient* while representing specific concepts. For example, to distinguish the concept *Apple* from other concepts like *Pear*, the colour and taste domains will be more salient than the weight domain.
- **Instances (Objects):** *Instances* of concepts are the sets of points in the conceptual space. These points are located within the domains by taking the values based on the quality dimensions.

Example 2.1. Assume a conceptual space for “fruits”. To define the Fruit space, one can introduce different **domains** such as ‘colour’, ‘taste’, ‘size’, ‘shape’, ‘nutrition’, etc., where these domains are defined by **quality dimensions** (e.g., ‘brightness’, ‘hue’, and ‘saturation’ dimensions for the colour domain). Now, a fruit **concept** like apple can be presented in this space by a set of regions (i.e., **properties**) within the domains. Verbally, the concept of an apple involves a ‘green’ property (a region) in the colour domain, ‘sweet-sour’ property in the taste domain, ‘roundish’ property in the shape domain, and so forth [89]. Now, one **instance** (or an object) of the concept apple can be presented by a set of points belonging to the regions of the concept. A ‘sweet apple’ object in the Fruit space has a multi-point presentation that contains a point in the e.g., sweet-sour region in the taste domain.

Following the above example, Figure 2.1 shows a schematic presentation of a conceptual space of fruits, together with presenting the regions of the concept of apple within the defined domains and quality dimensions.

The metric definition of domains allows one to depict the notion of *semantic similarity* in a conceptual space. Measuring the similarity robustly eases the consideration of cognitive tasks such as concept formation, semantic inferences, induction, and concept learning [116]. *Context* is another notion that has been considered in the theory of conceptual spaces. Since the semantics of concepts are conceptual structures in individual minds, the meaning of elements differs in various contexts. Thus, to formulate the context within the concept representation, it is possible to assign weights to the domains or dimensions to distinguish between similar concepts in different contexts [87].

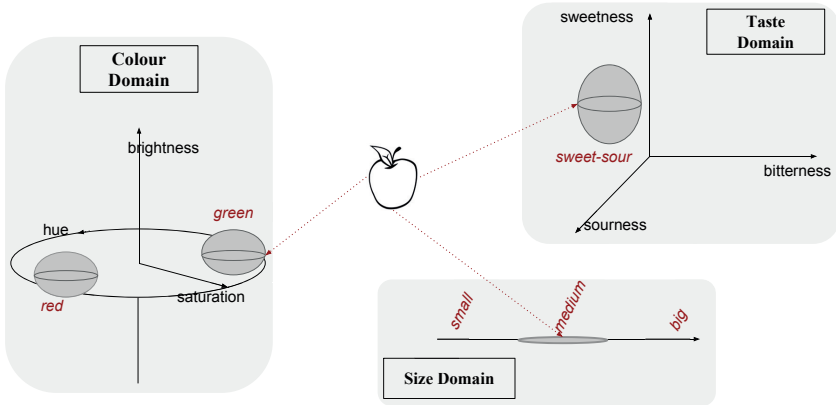


Figure 2.1: A schematic presentation of a conceptual space of fruits. It also shows the representation the concept of apple as a set of regions within the defined domains and quality dimensions as $\text{apple} = \langle \text{'sweet-sour'}, \text{'green'}, \text{'medium'} \rangle$.

Various formalisations of the conceptual spaces theory have been proposed in the literature, including [14, 116, 181, 189], which attempted to mathematically formalise how to construct and how to perform induction within conceptual spaces.

2.2.1 Identifying Quality Dimensions

The quality dimensions of human perceptions are revealed from the judgements about the similarity of concepts in a cognitive process [87]. The origin of quality dimensions is still an open question in the field of conceptual spaces [86]. Once the process of constructing a conceptual space starts, as Quine noted in [176], some innate quality dimensions are needed to make *concept learning* possible. However, there is no unique way to specify which set of dimensions is sufficient to characterise the concepts to be learned. There are two paradigms to identify quality dimensions: being chosen or being inferred. For a conceptual space that aims to model a scientific theory or an artificial cognitive task, quality dimensions are usually *chosen* by the developers of the theories (experts or scientist). In contrast, for phenomenal conceptual spaces, the quality dimensions need to be *inferred* from the perceived behaviours of the subjects. In many developed examples of conceptual spaces, determining the set of quality dimensions relies on the background knowledge, which comes from phenomenal (human perceptual) or scientific (sensory) quality dimensions [86]. Thus, in a real-world application, there is a need for knowledge engineering from experts within the application [188].

However, the issue of identifying the quality dimensions is more challenging when dealing with such systems where there is no prior knowledge to explain the semantics of dimensions, or there is a lack of knowledge about relevant quality dimensions [87]. This specific challenge motivates the investigations on how to derive the domains and quality dimensions in a data-driven manner.

An agreed point in the literature of conceptual spaces is that it is almost impossible to provide a complete list of human perceptual quality dimensions [89]. This statement emerges from the task of concept learning, wherein learning a new concept is often applicable by expanding a conceptual space with new quality dimensions [209]. To initialise a conceptual space, it would be challenging to realise which set of input features from the data can be used as the quality dimensions in order to form the target concepts. The work in this thesis attempts to identify a meaningful set of features out of predefined features in a numeric data set, relying on the hypothesis that the most discriminative features of concepts (classes) are the most representative quality dimensions for building a conceptual space.

2.2.2 Related Work on Conceptual Spaces and AI

As mentioned in the introduction, from the AI point of view, the aim of representing knowledge in a conceptual space is to develop an intuitive interpretation of the relationship between symbolic and sub-symbolic information [14, 86]. Gärdenfors has discussed thoroughly the role of conceptual spaces as a knowledge representation framework in AI systems [87], focusing on the tasks of *induction* and *reasoning* [88, 92]. Recently, Lieto et al. [147] have detailed the need of a conceptual representation as a mid-level of knowledge representation in-between the symbolic and the sub-symbolic one. This offers cognitive architectures a common language enabling the interaction between different types of representations. Schockaert and Prade [198] have focused on the problem of interpolative and extrapolative inference for different properties and concepts with the help of conceptual spaces. In addition to the theoretical AI problems, the feasibility of using conceptual spaces has been studied in various application domains of AI, such as geographical measurement [8, 199], cognitive robotics [57, 66, 138], object recognition [46], and visual perception [56]. A recent review [242] discusses further applications in diverse research areas (semantic spaces, computing meanings, and philosophical perspectives).

Concept formation tightly connects the theory of conceptual spaces to the induction (and particularly *learning*) problem. The aim of many *learning* systems is a general description of a category of observations as *concepts* [151]. If the input of a learning algorithm takes the form of instances, attributes, and concepts, then the process of learning is called *concept description* [230]. Instance-based learning refers to a class of learning algorithms which predicate the labels for unseen instances based on their similarity to the nearest training instances [129]. This model requires a similarity function to perform the

task of concept descriptions. However, in instance-based learning, the similarity functions are usually applied within a single-domain feature space [12]. A comparison of the practicality and effectiveness in instance-based learning and conceptual spaces was presented in [140]. However, in this work, the authors did not include the model construction process in their discussion. In this thesis, an instance-based approach is proposed for concept formation that considers the role of the features involved to derive a multi-domain space and represent the concepts in such a space.

Using data mining approaches in the process of deriving conceptual spaces has been studied in a few isolated works. Keßler [131] outlined the idea of using conceptual spaces to describe data, with some discussions on the possibility of automatically generating such spaces from databases. Lee [139] proposed a data mining method coupled with conceptual spaces, which addresses cognitive tasks such as concept formation using clustering techniques. The main drawback of these approaches is that they rely on knowing about the semantics of a domain (i.e., an application area) beforehand in order to directly determine the domains and the quality dimensions of a conceptual space. However, an essential challenge is to automatically find the most related features as integrated quality dimensions, without importing extra knowledge to the model. More precisely, the data-driven side of this thesis relies on the automatically finding the best subset of features and then grouping them into the domains, without expert knowledge.

The proposed approach in this thesis holds for certain classes of problems. It explores applications wherein the input data is complicated to be interpreted at first glance. Within such applications, the task of specifying the interpretable domains and dimensions based on human perceptions is no trivial. These classes of problems usually deal with raw sensor data (sometimes multivariate data) with little or no prior knowledge about their semantics [188]. The process of learning for these problems are typically performed by *connectionist* approaches (i.e., *neural network architectures* [219]) as solution for representing the relation of instances on a perceptual level [14]. But the main drawback of such solutions is that it neglects the explainability of the involved concepts or the interpretability of the learned model (i.e., features) from a semantic perspective.

This thesis aims to enable domain formation of a conceptual space that is highly data-driven. The motivation is to create a semantic model able to preserve induction and semantic inference. The motivation of the proposed approach is to deal with a class of learning problems that need a clear interpretation of the overall model (not only interpretability of the decisions made), but there is no prior knowledge to specify the relations within the model along with a vast number of input data. Thus, this thesis presents a method to create a semantic representation of sensor data and its interpretation using conceptual spaces, which facilitates the task of explaining concepts. This facilitation is performed in this work by applying approaches to generate linguistic descriptions

for unknown instances of concepts, using the constructed semantic representations as conceptual spaces. For this reason, the next section will provide an overview of linguistic description and natural language generation approaches that have been addressed in the literature.

2.3 Generating Linguistic Descriptions

In general, the task of generating linguistic descriptions is the process of generating understandable information in the form of natural linguistic expressions for the target user. This task is addressed in the literature in two research fields: *linguistic descriptions of data* (LDD), and *natural language generation* (NLG) [178]. LDD has its roots in fuzzy set theory and deals with summarising perception-based attributes of numeric data set using linguistic characterisations defined by fuzzy sets that are able to deal with the imprecision of human language. NLG, on the other hand, aims to automatically generate natural language text in the form of sentences that are as close as possible to human created text. Both of LDD and NLG fields contain elements which are relevant to this thesis. Herein, the definitions and the usage of these fields are presented.

2.3.1 Linguistic Descriptions of Data (LDD)

The field of *linguistic descriptions of data* has emerged from the use of *fuzzy set theory* and soft computing to perform linguistic computations on data. This task studies the necessity of automatically describing numeric data sets by employing a set of linguistic terms. Fuzzy set theory is a well-studied approach to bridge between numeric and linguistic information, specifically in perception-based systems [35, 126]. The basic idea of linguistic descriptions comes from the works of Zadeh [239] and Yager [233] on developing the paradigms of *computing with words* [239], and later, the *computational theory of perception* [238, 241], which express the ability of computing systems in a linguistic manner [73]. Within these paradigms, developed approaches are often based on fuzzy quantification models [70] to generate simple linguistic summaries on the variables, such as “*most of the apples are red*”. Although there is no high-level abstraction to present a generic model of linguistic descriptions, Ramos-Soto et al., [178] have listed the essential elements of linguistic description approaches as follows:

- **Input data**, usually including numerical and sequential data, which represents temporal or spatial properties. *Temperature, height, weight, and size* are some examples of input data. The input data is also called numeric variables or numeric properties of a domain.
- **Linguistic variables**, defining the fuzzy sets on the provided input data to categorise and annotate the input variable. Linguistic variables are tightly

related to the definition of *fuzzy granulation* of input variables. An example of such fuzzy granulation to map the numerical values of an input variable like *size* is a set of linguistic labels (i.e., words) such as *small*, *medium*, or *large*. These linguistic variables are characterised by fuzzy intervals with non-determined boundaries using fuzzy membership functions. (More details will be presented in Section 4.2.2).

- **Fuzzy quantifiers**, are the fuzzy granulations which lead to provide propositions like *low*, *increasing*, and *significant* for the input variables [240]. These quantifiers are also characterised by fuzzy membership functions.
- **Evaluation criteria**, defined to assess the appropriation of the generated descriptions based on several criteria such as data coverage degree, fulfilment degree, relevance and the length of the descriptions.

The algorithms that are employed in linguistic descriptions of data are influenced by fuzzy techniques to construct quantified sentences. LDD methods produce all the possible combinations of the sentences using the provided quantifiers and linguistic variables for a set of input data. Depending on the complexity of the domain, type-I or type-II of quantified sentences (which are the statements with simple or complex relations between quantifiers and variables, respectively [44]) may be employed [70, 163]. Afterwards, the generated sentences are ranked, added or removed from the output text based on the defined criteria. The generated linguistic sentences are simple from the natural language point of view with usually template-based messages to handle the sentences.

The theoretical works on developing linguistic descriptions started with the topic of fuzzy quantification as the task of computing with words [239]. The work of Kacprzyk [125, 126] introduced a way to relate the computation of words in fuzzy logic to an implementable linguistic summarisation of data. According to [126], a linguistic summarisation of a data set consists of a summariser *S* (e.g., *young*), a quantity agreement *Q* (e.g., *most*), and a truth degree *T*. Then, an abstract prototype of a linguistic summary can be in the form “*Q Y's are S*”, where *Y* is a set of observed objects. As an example, “*most of the employees are young*” is the result of a linguistic summarisation using fuzzy logic [126]. Practically, linguistic descriptions are applied in several applications. The granular linguistic model of phenomena (GLMP) [224] is a general framework that works based on applying fuzzy rules on a set of computational perceptions as inputs. This solution is employed for a verity of applications [178]. In [49], the concept of linguistic summarisation is employed to fulfil the precision and brevity on data related to the patient inflow, where the descriptions are for example: “*most of the days of June, patient inflow is medium*”. The work in [179] presents the use of LDD in meteorology to generate monthly reports emphasising related contrast descriptions, e.g.: “*The temperature was high for October, ... with very cold temperatures during the fourth week*”.

2.3.2 Natural Language Generation (NLG)

Studies on generating linguistic descriptions also encompass the field of *natural language generation* (NLG). An NLG system aims to generate human-readable natural language from non-linguistic information [185]. Typically, NLG systems generate text based on acquired knowledge about both language and the application domain. Based on the requirements of the system, NLG solutions follow two different goals: 1) Automatically generating *useful* text from non-textual inputs to comply with a specific set of needs, and/or 2) Automatically producing *human-like* text from non-textual inputs, to simulate an already known corpus of human-written text.

While a variety of NLG architectures and implementations are proposed in the literature, a generic architecture for an NLG system has been formulated and presented by Reiter and Dale [184, 185] based on the fact that the primary task of a natural language generation system is to convert acquired knowledge from underlying non-linguistic data into an understandable set of messages as output text. The proposed architecture consists of three main modules: *document planning*, *microplanning*, and *realisation*. Each of these modules performs a set of tasks related to the goal of the system. Table 2.1 depicts the modules and their corresponding tasks in a typical NLG system. The main content tasks of an NLG system are shortly described as follows. Note that in this thesis, the use of NLG systems is mostly limited to the task of *content determination*, where its related work will be further explained.

- ***Content determination*** is the task to decide what information will appear in the output text, based on the provided input information. Content determination is the most relevant aspect of NLG regarding the linguistic characterisation of numerical data [187, 237]. This task determines whether or not an acquired set of numerical information is able to be represented linguistically and be semantically labelled to appear in the final text. Thus, the content determination is the core of bridging from numerical data to the semantic representations in NLG systems. A recent survey on the task of content determination and content selection can be found in [103].

Table 2.1: Typical modules and tasks of an NLG system [185].

<i>Modules</i>	<i>Content task</i>	<i>Structure task</i>
Document Planning	Content determination	Document structuring
Microplanning	Lixicalisation; Referring expression generation	Aggregation
Realisation	Linguistic realisation	Structure realisation

- **Lexicalisation** is the task to decide what specific words should be selected to express the determined content. For example, the actual nouns, verbs, adjectives and adverbs to appear in the text are chosen from a lexicon. Lexicalisation determines the particular linguistic terms to be used to explain the domain concepts and their relations. In an NLG system, the role of lexicalisation is essential since the mapping between extracted numerical information to the predefined lexicon is not a trivial task.
- **Referring expression generation** is the task to decide what expressions should be used to refer to domain concepts and entities, in a way that the reader of the system recognises what the message refers to. So, referring expression generation is about selecting proper names and reasonable pronouns, or providing a sufficient set of descriptions for an entity or object in order to distinguish that entity from the rest.
- **Linguistic realisation** is the task to apply grammatical rules of the target language considering both morphology (the study of word forms) and syntax (the study of sentence structure). This task converts the provided abstract symbolic representations of the messages into the actual final sentences in natural language.

An important class of NLG frameworks is *data-to-text* systems, wherein a linguistic summarisation of numeric data is produced with the help of data mining and AI algorithms. The main architecture of data-to-text systems has been introduced by Reiter [182] that includes the following stages: *signal analysis*, *data interpretation*, *document planning*, *microplanning* and *realisation* (Figure 2.2). These systems identify and abstract the patterns of the numeric data, determine the most useful and relevant information, and generate a natural language text for the acquired knowledge in an understandable form [119].

A complete survey on reviewing NLG tasks, architectures, and applications has recently been provided by Gatt and Krahmer [93]. According to [93], the developed NLG systems can be divided into three approaches: rule-based (modular), planning-based, and data-driven approaches. Rule-based approaches consider a crisp division among the NLG tasks. These methods usually follow a set of pre-defined rules in order to perform each of sub-tasks within the NLG architecture. Planning approaches aim to look at the goal of generating text from data as the process of determining a sequence of actions. These approaches combine slightly the sub-tasks of NLG to generate text. Data-driven approaches are the new dominant trend in NLG, which consider the goal of text generation as an integrated task. Data-driven approaches perform this goal usually by applying statistical learning on the alignment of the input non-linguistic information and the output texts. This automatic learning of the correspondences between data and text make these approaches data-driven. Several pieces of research have been done from this perspective, especially by focusing on the neural network and deep learning methods (a wide review can

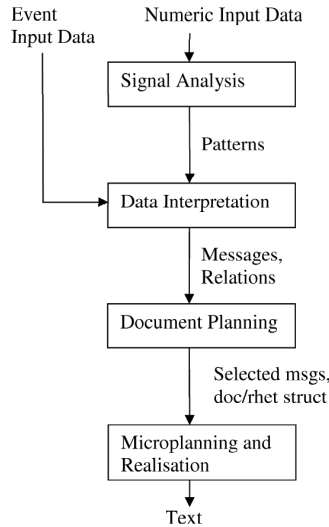


Figure 2.2: The architecture of data-to-text systems, proposed by Reiter [182].

be found in [93]). However, the remaining question is how much the learning process is dependent on the goodness of provided set of training information (aligned data and text).

Recently, there is a growing demand for NLG and data-to-text systems in real-world applications. Examples of well established NLG applications include the generation of weather forecasting reports from meteorological data [104, 186, 215], communicating financial and statistical information [79, 120]. BT-Nurse [119] and BabyTalk [94, 174] are the recent examples of data-to-text systems in order to generate documents in medical domains which produce summaries of data sets about the state of neonatal babies from intensive care data. Most of these applications use small amounts of homogeneous data and are supported by a significant amount of predefined knowledge.

Knowledge Acquisition for Content Determination

Knowledge acquisition (KA) is an essential part of building natural language generation systems. Two types of KA techniques including *corpus-based* KA and structured *expert-oriented* KA have been previously studied for NLG systems in [187]. These techniques aim to enrich the similarities between generated texts and natural human-written texts. Both of the techniques use rule-based methods to improve the quality of the acquired knowledge for such systems.

Expert-oriented techniques use experts of the domain to acquire knowledge in a structured manner (e.g., direct interviews, questionnaires). As an example

in the medical area, one can ask doctors or nurses about the necessary information that are needed to be captured and presented in the final generated text. Corpus-based techniques rely on the analysis of available corpus, and learn from provided sources of information such as data sets themselves. For example, one can use previous corpora of human-written reports (e.g., clinical reports) to acquire the necessary elements and information, to reproduce the same knowledge in the final generated text.

In data-to-text systems, the acquired knowledge, also called *domain knowledge*, is usually organised in the form of taxonomies or ontologies of information. All the stages of a data-to-text architecture (shown in Figure 2.2) then use these provided taxonomies of knowledge to perform their tasks. In particular, signal analysis stage extracts information that is determined in taxonomies such as simple patterns, events, and trends. Also, the data interpretation stage abstracts information into symbolic messages using the represented taxonomies. The most recent data-to-text frameworks developed by Reiter's architecture [182] have acquired the taxonomies or ontologies corresponding to the domain knowledge. For instance, the work on summarising gas turbine time series [237] has used expert knowledge to provide a taxonomy of the primitive patterns (i.e., spikes, steps, oscillations). Similarly, systems related to the BabyTalk project [94, 174] have stored medically known observation (e.g., bradycardia) in local ontologies. The core of such systems is based on the richness of the domain knowledge in the provided taxonomies which are usually bounded by expert rules. This organised domain knowledge is usually an inflexible input to the framework which restricts the output of the stages in data-to-text architectures. For instance, the taxonomy in [237] does not allow the system to represent unexpected observations (e.g., wave or burst) out of the predefined domain knowledge. Likewise, in the medical domain, an unknown physiological pattern will be ignored if it does not have a corresponding entity in the provided ontology by the domain experts.

Determining suitable content is mostly addressed using knowledge-driven and rule-based approaches to comply with the domain or user requirements [102, 237]. However, there is a new trend of data-driven approaches to performing such task without relying on a set of pre-defined knowledge to map data to lexicons. Instead, these approaches aim to use learning techniques, such as hidden Markov models [33] and optimisation methods [133], in order to automatically ground language acquisition and align numerical observation to their proper descriptions [93]. These approaches are independent of rules, however, they are still dependent on a set of well-defined correspondences between input data and the output text in a supervised manner. The way that this thesis aims to do the data-driven content determination is to rely on the observed data and its behaviours that can be beyond the user requirements, though still meaningful to represent [31].

Although some studies in NLG, like SumTime [186], claim that building a complete data-to-text system is good enough to be used for producing text as

good as a human does, there is still the lack of modelling cognitive task in NLG systems to study e.g., how humans model observations in their mind, and then explain them by linguistic terms. This issue in rule-based data-to-text systems reveals the necessity of reorganising the way of modelling domain knowledge in order to also cover explaining unseen information across the data.

2.4 Conclusions

A semantic representation of numerical data is the key point to bridge between conceptual spaces theory and linguistic description approaches by modelling *descriptive features* of data. On one hand, concept formation based on perceived information in a conceptual space relies on the descriptive features in order to determine the domains and quality dimensions of the spaces. This data-driven manner of modelling features leads to a semantic representation of non-linguistic information in order to infer meaningful descriptions. Importing descriptive features to computational systems includes the possibility of operating with linguistic information [35]. On the other hand, deriving linguistic descriptions for a set of numerical perceived data is dependent on the semantic level of its descriptive features. This set of information can be obtained from various sources such as observations, sensor measurements, mathematical analysis or visual perceptions [35]. Since conceptual spaces have been developed to model the descriptive features of concepts for further reasoning, it can be employed as a robust framework to perform content determination task in linguistic descriptions using semantic inferences.

Regarding the relation of these two fields in the literature, the problem of modelling natural language using conceptual spaces has been investigated in a few isolated works [10, 72]. Evidently, in most of the proposed conceptual representations, the primary interest is not the relation of concepts and natural language [108]. Aisbett et al., [15] recently investigated the integration of conceptual spaces theory with the topic of computing with words by introducing a fuzzy representation of conceptual spaces' elements. Domains and dimensions in their work, however, are crisp elements with no role concerning the qualification of objects within the space. Also, [72] attempted to derive the semantic relations within conceptual spaces built upon text documents. However, to the best of our knowledge, there is no study on the conceptual spaces to derive natural language descriptions for the numeric inputs through a conceptual representation.

As the final point to conclude this chapter, it worth mentioning that the proposed approach is called semantic representation (and not only conceptual representation), because it goes beyond representing the concepts as structures in mind. The proposed approach links the perceived information to natural language by "linking concepts to meaning" [111] using semantic inferences.

Chapter 3

Data-Driven Construction of Conceptual Spaces

“Solving a problem simply means representing it so as to make the solution transparent.”

— Herb Simon (1916–2001)

THIS chapter presents one of the leading contributions of this thesis, that is how to automatically construct a conceptual space in a data-driven manner from a numeric data set. The approach to constructing conceptual spaces is considered to be data-driven as it is automatically constructed by processing the data matrix of the observations based on the variable values and class labels. This is in contrast with knowledge-driven approaches that have to be manually constructed using psychologically or scientifically pre-defined knowledge about the relations between quality dimensions, domains and the concepts’ regions [10, 87]. Practically, the process of constructing a conceptual space is about determining its essential elements. According to [14, 181], the definition of a conceptual space is as follows:

Definition 3.1. *A conceptual space S is defined as a 4-tuple $\langle \mathcal{Q}, \Delta, \mathcal{C}, \Gamma \rangle$, where \mathcal{Q} is a set of quality dimensions, Δ is a set of domains, \mathcal{C} is a set of concepts in the space S , and Γ is a set of instances representing the concepts.*

The representations of the elements are rigorously explained in further definitions (from 3.2 to 3.5). To automate the process of constructing conceptual spaces, the definitions of the conceptual spaces’ elements are modified slightly compared to previous formulations (e.g., [9, 189]). These modifications are necessary to use the constructed conceptual space as a model of a semantic representation for the inference of unknown observations.

To begin, it is assumed that a given data set \mathcal{M} contains a set of possible class labels, a set of predefined features, and the input observations with known class labels, which are characterised by feature values. Given a set of class labels $\mathcal{Y} = \{y_1, \dots, y_m\}$ and a set of features $\mathcal{F} = \{X_1, \dots, X_n\}$, let \mathcal{D} be the set of known observations, denoted by $\mathcal{D} = \{o_i : (x_{o_i}, y_{o_i})\}$, where o_i consists of a n -dimensional feature vector $x_{o_i} = [x_1, \dots, x_n]$, and an output label $y_{o_i} \in \mathcal{Y}$. The component x_j ($j = 1, \dots, n$) in the vector x_{o_i} is the measured value of the corresponding feature $X_j \in \mathcal{F}$.

Here, each feature X is defined as a couple of values $X : \langle H_X, I_X \rangle$, where H_X indicates the linguistic name of the feature, and I_X is either a numeric interval or a categorical set that presents the possible range of values for X .

Example 3.1. *Consider the leaf data set [206] which is a set of photographed leaf samples (observation set \mathcal{D}^l) from various plant species (classes) such as: $\mathcal{Y} = \{y_{qr} : \text{'Quercus Robur'}, y_{ap} : \text{'Acer Palmatum'}, y_{no} : \text{'Nerium Oleander'}, y_{tt} : \text{'Tilia Tomentosa'}, \dots\}$. This data set includes a set of measurable features to characterise the features of each leaf sample, such as: $\mathcal{F} = \{X_{el} : \text{elongation}, X_{lo} : \text{lobedness}, X_{co} : \text{convexity}, X_{ro} : \text{roundness}, X_{so} : \text{solidity}, X_{in} : \text{indentation}, \dots\}$. An observed leaf such as $o_i \in \mathcal{D}^l$ that is labelled by y_{tt} takes the feature values as: $o_i : (x_{o_i}, y_{tt})$, where $x_{o_i} = [x_{el}, x_{lo}, x_{co}, x_{ro}, x_{so}, x_{in}]$.*

The goal described in this chapter is to find a mapping from the elements of a data set \mathcal{M} to various components needed to define a conceptual space S . In short, this mapping is achieved by performing the following steps:

- Initialise the primitive known concepts using the class labels. Consequently, the conceptual space S , which models the data set \mathcal{M} , will consist of a set of concepts $\mathcal{C} = \{C_1, \dots, C_m\}$, where $|\mathcal{C}| = |\mathcal{Y}|$. Thus, the notation C_y indicates the concept which corresponds to the class label $y \in \mathcal{Y}$.¹
- Specify the quality dimensions \mathcal{Q} and domains Δ . The quality dimensions are specified via selecting a subset of the features such that $\mathcal{Q} \subset \mathcal{F}$, and the domains are determined based on ranking and grouping the set of selected features as the quality dimensions (Section 3.1).
- Form the representation of each concept C_y within the domains Δ , based on the known corresponding instances (Section 3.2).

The following sections will outline how the two latter steps are achieved in detail. Figure 3.1 illustrates the steps of constructing a conceptual space from a set of numeric data, which are explained in the following sections.

¹This approach follows the assumption that any semantic modelling needs an innate set of knowledge [176]. Learning concepts without any initialised knowledge is not the scope of this work.

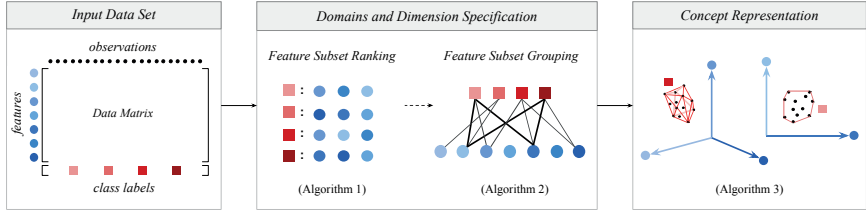


Figure 3.1: Illustration of the main steps for constructing a conceptual space from a set of numeric data. The domain and dimension specification is explained in Section 3.1, and the concept representation is described in Section 3.2.

3.1 Domain and Quality Dimension Specification: A Feature Selection Approach

A data-driven approach to build a conceptual space makes no prior assumption about the domains. Preferably, the known labelled observations and features are the inputs from which the quality dimensions and domains will be extracted. This approach aims to propose a set of observation-based associations between classes of data as concepts and the grouped subsets of features as domains. As a domain is an integrated subset of quality dimensions, and the quality dimensions are the subset of the initialised features, the first step is to determine a subset of informative features. This determination is performed by applying *feature selection* methods. Before explaining these methods, the formal definitions of a quality dimension and a domain is recalled. As mentioned before, these definitions are reshaped in a novel manner to be utilised in the task of semantic inference.

Definition 3.2. A quality dimension $q_X \in \mathcal{Q}$ is a triple $\langle H_q, I_q, \mu_q \rangle$, which corresponds to a selected feature $X \in \mathcal{F}$. H_q is the linguistic name of the quality dimension q_X , which is equal to H_X and I_q indicates the range of possible values for the quality dimension q_X , which is equal to I_X . μ_q is defined as a family of fuzzy membership functions² to map the subintervals of I_q onto a set of linguistic terms.

Definition 3.3. A domain δ is a triple $\langle \mathcal{Q}(\delta), \mathcal{C}(\delta), \omega_\delta \rangle$, where $\mathcal{Q}(\delta) \subset \mathcal{Q}$ is the set of integral quality dimensions involved in δ , $\mathcal{C}(\delta) \subset \mathcal{C}$ is the set of concepts that are represented in δ , and $\omega(\delta)$ is a weight function³ presenting the assigned salient weight between a concept and a quality dimension in δ .

²More details on the fuzzy membership functions μ , which quantify the changes in the values of a feature by assigning linguistic labels to the subintervals of dimension, will be given in Section 4.2.1.

³The weight function $\omega(\delta)$ is further explained in Section 3.1.2.

Example 3.2. Consider the leaf data set from Example 3.1, suppose that a quality dimension is elongation, which is defined as $q_{e1} = \langle \text{'elongation'}, [0, 1], \mu_{e1} \rangle$, and another one is lobedness, defined as $q_{l10} = \langle \text{'lobedness'}, (0, \text{inf}), \mu_{l10} \rangle$. One can conceptualise the leaves in various domains such as Shape, Texture, Colour, etc. Then, both the elongation and lobedness quality dimensions can belong to the shape domain. Moreover, μ_{e1} can return the linguistic labels for elongation as: 'circular', 'elliptical', 'elongated'.

Since the domains are constructed in a data-driven way without involving prior knowledge, it can be difficult to assign a semantic interpretation to the constructed domains that reflect human perception. However, the provided space still constitutes a conceptual one because of its ability to represent the concept formation and the semantic similarities between concepts and instances across the domains [87]. With quality dimensions in place, domain is then an integral subset of quality dimensions which are relevant to each other.

Identifying the most characteristic features of the data from the initialised set of features is an essential task, which is performed by *feature extraction* approaches [42]. There are two principal ways to extract informative features: *feature transformation* and *feature selection* [110]. The first approach finds a projection from the original feature space into a lower dimensional feature space. Transforming the original features into this lower dimensional space alters typically any associated descriptive attributes connected to the features. Therefore, the semantic meaning of the resulting features is often difficult, if not impossible, to assess [109]. The second approach selects a subset of original features by keeping relevant features and discarding the irrelevant ones. The retained features are not altered and the original semantic meaning of those features stays intact. Since the goal is to exploit external knowledge of the original features, *feature subset selection* techniques are applied.

Both *relevance*, and *redundancy* are important criteria to consider in feature selection. A subset of features is optimal if the relevance between selected features and the target classes is maximal, and the redundancy among the selected features is minimal. These two criteria guarantee that the selected features are adequate to distinguish the classes of data with the smallest number of features [75].

The proposed approach employs feature selection methods to specify the quality dimensions and domains using two phases: *feature subset ranking* and *feature subset grouping*. The feature subset ranking phase determines which features are most representative for every single target class, independent from other classes. Since a concept can rather be represented by one or several groups of features as domains, the feature subset grouping phase categorises the ranked features in a way to recognise what subset of features are most related to each other based on their relevancy to the concepts.

Figure 3.2 shows the two phases of (1) feature subset ranking and (2) feature subset grouping, along with the input and output parameters, to specify

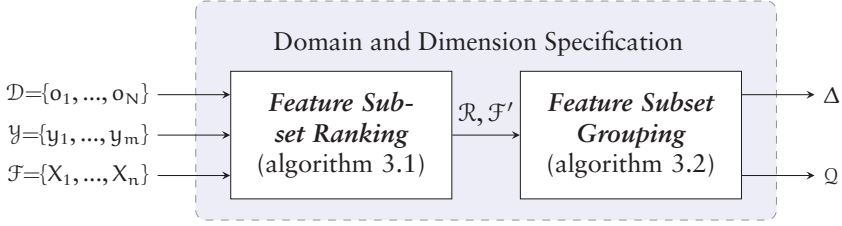


Figure 3.2: Two phases of the domain and quality dimension specification, with input and output parameters of each phase.

the suitable quality dimensions within a set of domains (the middle block in Figure 3.1). The following sections explain these two phases in details.

3.1.1 Feature Subset Ranking

Feature subset selection algorithms are categorized as either *filter* methods or *wrapper* methods [230]. *Filter* methods determine the subset of features based on the statistical characteristics of the input data set without referring to the used classifier. *Wrapper* methods are dependent on the learning algorithm (i.e., target classifier) that evaluates the selected subset of features based on the performance of the used learning algorithm. In the present work, the aim is to identify the meaningful set of understandable attributes out of predefined features, but not to classify the input data. *Filter* methods are chosen to be used for feature selection since this category of methods is independent of the final classifier approach, and it derives an informative subset of features concerning the input data set labels. It is worth to note that although the filter methods are more computationally efficient, the evaluation of such methods is not a trivial task [230] since there is no universally accepted relevance measure between a subset of selected features and the target class labels.

Filter methods rank the features using a scoring function, usually by employing a statistical measure or the measures from information theory to quantify relevance and redundancy. The top scored features are kept as selected features (or low scored ones are removed from the resulting subset). In this work, *mutual information*, one of the commonly employed scoring functions, is used [47]. One such technique is *MIFS* (mutual information-based feature selection) [34]. For an input set of features \mathcal{F} and 2-class labelled data \mathcal{D} , MIFS adds the feature $X_i \in \mathcal{F}$ to the already chosen subset of features \mathcal{F}' , to maximise

$$I(\mathcal{D}, X_i) - \beta \sum_{X_j \in \mathcal{F}'} I(X_i, X_j), \quad (3.1)$$

where $I(Y, X)$ is the mutual information between the variables Y and X [223]. This mutual information is defined based on the probability density functions [65], denoted by:

$$I(X, Y) = \int_x \int_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dy dx. \quad (3.2)$$

The first term in Equation 3.1 attempts to maximise the relevance of feature X_i to target labelled data, and the second term tries to minimise the redundancy between X_i and the already selected features in \mathcal{F}' (using a balancing parameter β). In this work, the term $I(Y, X)$ is estimated via histograms, but other estimation methods are applicable as well [196]. The MIFS technique is a heuristic approximation, since there is no independent assessment of the *joint mutual information* to determine when a feature is relevant to the class labels [223]. The proposed method is not dependent on the use of the MIFS algorithm. It can be substituted by other approximations of the joint mutual information [48, 80, 171]. It is notable that different filter methods do not necessarily produce the same ranking of the features. However, the focus here is to reach to a good enough set of features representing the data classes with high relevance and low redundancy [228].

The proposed method for feature subset ranking starts with defining a new set of input data for each target label y , wherein the data set of known observations \mathcal{D} is split into two classes: class y including all the observations labelled by class y , denoted by \mathcal{D}_y , and class $\bar{y} = \{Y \setminus y\}$ including the rest of observations labelled by other classes than class y , denoted by $\mathcal{D}_{\bar{y}}$. Then the MIFS procedure is applied to the feature set \mathcal{F} considering the generated 2-class data set $\mathcal{D}_{y\bar{y}} = \{\mathcal{D}_y \cup \mathcal{D}_{\bar{y}}\}$.

Example 3.3. *In the leaf data set from Example 3.1, for the class of the leaf species Tilia (y_{tt}), the set of $\mathcal{D}_{y_{tt}}$ is the leaf samples in \mathcal{D}^l that are labelled with y_{tt} , and $\mathcal{D}_{\bar{y}_{tt}}$ is the set of samples in \mathcal{D}^l that are not labelled with y_{tt} , or consequently, are labelled with one of the labels y_{qr} , y_{ap} , or y_{ia} .*

By separating one class (concept) of data from the other classes, the output of the feature ranking algorithm will return the features that individually characterise the observations of this concept and separate it from the rest. The output of the filter method for each label is a sorted list of features with a score for each feature. Formally, the output for a class y is a ranked list of features with the highest scores according to y , as

$$\mathcal{R}(y) = \{ (X, w_{y,X}) \mid X \in \mathcal{F}, w_{y,X} \in [0, 1] \} \quad (3.3)$$

where $w_{y,X}$ is the normalised weight (or the score) that is assigned to the relation of feature X and label y . The features in $\mathcal{R}(y)$ are the k most relevant features of the class label y . From a conceptual point of view, these k features

Algorithm 3.1: Feature Subset Ranking

```

Function FeatureRanking( $\mathcal{D}, \mathcal{Y}, \mathcal{F}$ )
   $\mathcal{R}, \mathcal{F}' \leftarrow \emptyset$ 
  foreach  $y \in \mathcal{Y}$  do
    // Define 2-class data set
     $\mathcal{D}_y = \{\mathbf{o}_i : (\mathbf{x}_i, y_i) \in \mathcal{D} \mid y_i = y\}$ 
     $\mathcal{D}_{\bar{y}} = \{\mathbf{o}_i : (\mathbf{x}_i, y_i) \in \mathcal{D} \mid y_i \neq y\}$ 
     $\mathcal{D}_{y\bar{y}} \leftarrow \mathcal{D}_y \cup \mathcal{D}_{\bar{y}}$ 
    // Find the list of top scored features ( $X, w_{y,X}$ )
     $\mathcal{R}(y) \leftarrow \text{MIFS}(\mathcal{D}_{y\bar{y}}, \mathcal{F})$ 
     $\mathcal{F}' \leftarrow \mathcal{F}' \cup \{X \mid \exists (X, w_{y,X}) \in \mathcal{R}(y) \wedge X \in \mathcal{F}, w \in [0, 1]\}$ 
  return  $\mathcal{R}, \mathcal{F}'$ 

```

of $\mathcal{R}(y)$ are the suitable candidates to be the quality dimensions that distinguishes the concept C_y from the other concepts. The score $w_{y,X}$ determines the importance of the selected feature X to represent the class label y . From the conceptual space point of view, the scores indicating the weights show the significance of the chosen quality dimensions for C_y .

Algorithm 3.1 shows the steps for finding the ranked scored list of features for each label y . The output of the algorithm is then a set of filter method results for all the class labels, denoted by $\mathcal{R} = \{\mathcal{R}(y_1), \dots, \mathcal{R}(y_m)\}$. In this algorithm, \mathcal{F}' is the set of all features that appeared (at least once) in the ranked features:

$$\mathcal{F}' = \bigcup_{y \in \mathcal{Y}} \{X \mid \exists (X, w_{y,X}) \in \mathcal{R}(y) \wedge X \in \mathcal{F}, w_{y,X} \in [0, 1]\}, \quad (3.4)$$

where $\mathcal{F}' \subset \mathcal{F}$. The set of features \mathcal{F}' is the set of potential features to become quality dimensions. However, in feature grouping, some of these features may be filtered out from the target set of quality dimensions.

Example 3.4. *Continuing of the leaf data set in Example 3.1, suppose that after applying the MIFS method, elongation, lobedness, and roundness are selected as the top features for y_{tt} , $\mathcal{R}(y_{tt}) = \{(X_{e1}, w_{y_{tt}, X_{e1}}), (X_{l1o}, w_{y_{tt}, X_{l1o}}), (X_{r1o}, w_{y_{tt}, X_{r1o}})\}$. Also, elongation, roundness, and indentation are selected for y_{no} , $\mathcal{R}(y_{no}) = \{(X_{e1}, w_{y_{no}, X_{e1}}), (X_{r1o}, w_{y_{no}, X_{r1o}}), (X_{i1n}, w_{y_{no}, X_{i1n}})\}$. Then, $\mathcal{F}' = \{X_{e1}, X_{l1o}, X_{r1o}, X_{i1n}\}$.*

3.1.2 Feature Subset Grouping

In order to specify a set of domains as subsets of integral quality dimensions out of selected features, the method should partition features into a number of subsets regarding their relevancy to the defined class labels. Based on the

definitions in conceptual space theory, a quality dimension usually appears in a single domain along other relevant dimensions to represent a specific aspect of conceptualised observations [86, 242]. It might be possible to have the same dimension in various domains, but this requires a priori knowledge [31]. Moreover, repeating dimensions in a multi-domain space increases the redundancy, and consequently decreases the accuracy of learning concepts. Therefore, the selected features are divided into distinct partitions of features as target domains, to avoid either creating a single domain of full features or repeating features in all of the constructed domains.

Here, a heuristic method is proposed to detect distinct subsets of features, where the features in each subset are highly representative of the most relevant classes. The output set \mathcal{R} in Algorithm 3.1 is a set of ranked features for each label. It is obvious that some features might be repeated in the ranked set of different class labels in \mathcal{R} . From the information in the set \mathcal{R} , the goal is to extract those subsets of features that are associated with each other based on their co-appearance in the ranked features of each class. This section first introduces a graph representation of the label-feature relation, and then it explains how to derive the correlated features using a greedy search on this graph. More specifically, this approach proposes to build up a bipartite graph and search for the bicliques that identify the most associated feature subsets (i.e., domains).

Let $G = (V_Y \cup V_{\mathcal{F}'}, E, w)$ be a bipartite graph with two sets of vertices V_Y and $V_{\mathcal{F}'}$, a set of edges E , and $w : V_Y \times V_{\mathcal{F}'} \rightarrow \mathbb{R}$ as a weight function for the edges. The vertex set V_Y denotes the class labels in \mathcal{Y} . The vertex set $V_{\mathcal{F}'}$ denotes the top-ranked features in \mathcal{F}' . A vertex $v_Y \in V_Y$ is connected to a vertex $v_X \in V_{\mathcal{F}'}$ if $X \in \mathcal{F}'$ has been selected for $y \in \mathcal{Y}$ in Algorithm 3.1. In other words, for each pair $(X, w_{y,X}) \in \mathcal{R}(y)$ a new edge $v_Y v_X$ is added to the edge set E of bipartite graph G between vertices v_Y and v_X , where the weight of the edge $v_Y v_X$ is denoted by $w(v_Y v_X) = w_{y,X}$. Figure 3.3 is an illustration of such weighted bipartite graph G .

The idea of grouping features is to find the maximal connected subgraphs in G . More precisely, a subset of features which are all connected to the same set of classes is a suitable subset of features for feature grouping. A biclique $\hat{G} \subset G$ is a special bipartite graph where every vertex in one part of vertices is connected to all the vertices in the other part of the vertices. The highlighted edges in Figure 3.3 depicts an example of a biclique in the given bipartite graph. Let \hat{G} be a biclique denoted by $\hat{G} = (\hat{V}_Y \cup \hat{V}_{\mathcal{F}'}, \hat{E}, w)$, where $\hat{V}_Y \subset V_Y$, $\hat{V}_{\mathcal{F}'} \subset V_{\mathcal{F}'}$. In this biclique, assume $|\hat{V}_Y| = \hat{m}$, $|\hat{V}_{\mathcal{F}'}| = \hat{n}$, thus $|\hat{E}| = \hat{m} \times \hat{n}$. The proposed approach is looking for a biclique with the highest score as \hat{G}_{\max} among all the bicliques in G . The score of a biclique \hat{G} is calculated using a scoring function $\text{Score}_{\hat{G}}$, based on the weights of its edges, as follows:

$$\text{Score}_{\hat{G}} = \sum_{v_Y \in \hat{V}_Y} \left(\prod_{v_X \in \hat{V}_{\mathcal{F}'}} w(v_Y v_X) \right) / \hat{n} \quad (3.5)$$

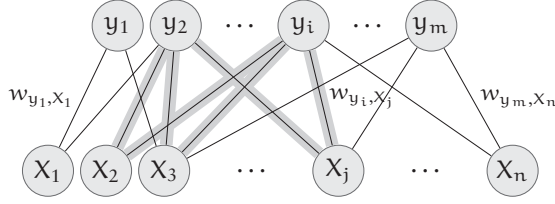


Figure 3.3: A weighted bipartite graph with two sets of vertices from the labels \mathcal{Y} and the selected features \mathcal{F}' . Also, an example of a biclique shown in the highlighted edges.

This scoring function calculates the average of the weights associated with the biclique. This scoring function will return higher values for the bicliques with the higher number of class labels, and lower number of features⁴.

In the selected biclique (\hat{G}_{\max}), the involved features $\hat{V}_{\mathcal{F}'}$ then will be the subset of features as the quality dimensions of a domain δ . To identify the next domain, the set of features $\hat{V}_{\mathcal{F}'}$ is eliminated from the graph G since these features are already assigned to a domain. After that, the process of finding the best biclique repeats on the updated graph G to find the next maximal biclique. Algorithm 3.2 shows the steps of determining the domains with feature subset grouping.

As stated in Definition 3.3, a domain δ is a triple $\langle Q(\delta), C(\delta), \omega_\delta \rangle$. The weight function $\omega_\delta = \mathcal{C}(\delta) \times \mathcal{Q}(\delta) \rightarrow \mathbb{R}$ is a function presenting the assigned salient weight between a concept $C_y \in \mathcal{Y}(\delta)$ and a quality dimension $q_x \in \mathcal{Q}(\delta)$. For a chosen biclique $\hat{G}_{\max} = (\hat{V}_y \cup \hat{V}_{\mathcal{F}'}, \hat{E}, w)$, a domain δ can be constructed as a triple $\langle \mathcal{Q}(\delta), \mathcal{C}(\delta), \omega_\delta \rangle = \langle \hat{V}_{\mathcal{F}'}, \hat{V}_y, w \rangle$. More specifically, for a constructed domain $\delta = \langle \mathcal{Q}(\delta), \mathcal{C}(\delta), \omega_\delta \rangle$, the set of quality dimensions is $\mathcal{Q}(\delta) = \{q_x \mid v_x \in \hat{V}_{\mathcal{F}'}\}$. Also, the set of concepts related to δ is defined as $\mathcal{C}(\delta) = \{C_y \mid v_y \in \hat{V}_y\}$. Then $\omega_\delta(C_y, q_x) = w(v_y v_x)$.

It is worth noting that for the next iteration of finding bicliques, the vertices with the labels \hat{V}_y of a chosen biclique are not eliminated while updating the bipartite graph, because a class label can be involved in other bicliques in further iterations. From a conceptual space point of view, it is also meaningful, since a concept can be represented in several domains.

Example 3.5. *The corresponding bipartite graph to the ranked features from Example 3.4 is illustrated in Figure 3.4. In this bipartite graph, one biclique is highlighted, which can potentially be the best biclique. If so, then the features elongation and roundness will become the quality dimensions of a new domain δ as: $\mathcal{Q}(\delta) = \{q_{el}, q_{ro}\}$ and $\mathcal{C}(\delta) = \{C_{tt}, C_{no}\}$.*

⁴Assume that all the weights of the edges in a biclique \hat{G} are equal to a constant weight w_c . Then $\text{Score}_{\hat{G}} = \frac{m}{n} (w_c)^{\frac{n}{m}}$.

Algorithm 3.2: Feature Subset Grouping

```

Function FeatureGrouping( $\mathcal{R}, \mathcal{F}', \mathcal{Y}$ )
   $\Delta, Q \leftarrow \emptyset$ 
  // Build bipartite graph
   $V_{\mathcal{Y}} \leftarrow \mathcal{Y}$ 
   $V_{\mathcal{F}'} \leftarrow \mathcal{F}'$ 
  foreach  $(X, w_{y,x}) \in \mathcal{R}(y) : y \in \mathcal{Y}$  do
     $E \leftarrow E \cup v_y v_x$ 
     $w(v_y v_x) \leftarrow w_{y,x}$ 
   $G = (V_{\mathcal{Y}} \cup V_{\mathcal{F}'}, E, w)$ 
  // Find max cliques as domains
  do
     $\hat{G}_{\max}(\hat{V}_{\mathcal{Y}} \cup \hat{V}_{\mathcal{F}'}, \hat{E}, w) \leftarrow \text{MaxBiclique}(G)$ 
    if  $\hat{G}_{\max} = \emptyset$  then
      break
     $\delta : \langle Q(\delta), \mathcal{C}(\delta), \omega_{\delta} \rangle \leftarrow \langle \hat{V}_{\mathcal{F}'}, \hat{V}_{\mathcal{Y}}, w \rangle$ 
     $\Delta \leftarrow \Delta \cup \delta$ 
     $Q \leftarrow Q \cup Q(\delta)$ 
    // Update bipartite graph
     $G \leftarrow (V_{\mathcal{Y}} \cup (V_{\mathcal{F}'} \setminus \hat{V}_{\mathcal{F}'}), (E \setminus \hat{E}), w)$ 
  until  $(Q = \mathcal{F}')$ 
return  $\Delta, Q$ 

```

So far, a methodology to extract domains including their distinct quality dimensions out of a given data set of observations and features is introduced. Now, each class or label of the observations needs to be represented as concepts in the proposed conceptual space.

3.2 Concept Representation: An Instance-based Approach

The vital concern to represent a concept in a conceptual space is to decide which are the most relevant quality dimensions and consequently most relevant domains. A concept may be represented in one domain or several domains. An important point is that a concept is not necessarily associated with a certain subset of domains⁵, but usually, one domain or a few numbers of domains are prominent to represent a concept [86]. In Section 3.1, the selected features are grouped into a set of domains out of the extracted bicliques. Using the fact

⁵For example, the concept of 'apple' is mainly represented by *colour*, *texture*, and *shape* domains. But it can be associated with further biological features and domains.

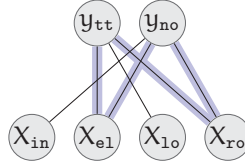


Figure 3.4: A bigraph graph and one selected biclique (blue edges) for the leaf example (explained in Example 3.5).

that each class label will be involved in at least one selected biclique, then the corresponding concept is assigned to (or associated with) a certain number of domains (at least one). With the output of Algorithm 3.2 it is already known which concepts are associated with which domains.

Example 3.6. For the selected biclique \hat{G} in Figure 3.4, the concepts C_{tt} and C_{no} are associated with a generated domain δ_i including the quality dimensions q_{e1} and q_{ro} .

It is possible that one concept appears in two bicliques, which means the concept is relevant to both specified domains. This fact is consistent with the conceptual spaces theory since a concept is not always represented within a single domain. The common example is the concept of ‘apple’ which is represented with more than a single domain, such as *colour*, *taste*, *size*, etc. A concept with merely one related domain is called *property* [86]. In fact, a property is a special form of a concept defined in a single domain [91]. For example, the colour ‘green’ is a property which is represented only in the *colour* domain. Thus, a concept can be specified as a single property within a single domain (e.g., green), or as a collection of properties within several domains (e.g., apple). Depending on the domain specification process, a class label in the input data set might be represented either as a property in one domain or as a concept in several domains. The problem of deriving the domains in a data-driven manner is that for a concept represented in several domains, there is no trivial interpretation for the meaning of its properties within the domains. This issue comes from the fact that the interpretation of the data-driven domains themselves is also tricky.

For a set of concepts $\mathcal{C} = \{C_{y_1}, \dots, C_{y_n}\}$, the problem is how to formulate the geometrical representation of concepts in the conceptual space with the extracted set of domains Δ . In general, a natural concept is a collection of regions across one or more domains along with a set of salient weights to the domains [86]. For a concept C_y , let $\Delta(y) \in \Delta$ be a subset of domains that contain concept C_y in their concept sets, as $\Delta(y) = \{\delta_i \mid \delta_i \in \Delta \wedge C_y \in \mathcal{C}(\delta_i)\}$, assuming that $|\Delta(y)| = k$. The concept C_y is presented by a collection of *sub-*

Algorithm 3.3: Concept Representation

Input: $\mathcal{D}_y \subset \mathcal{D}$: set of observations that are labelled by $y \in \mathcal{Y}$ and
 $\Delta(y) \subseteq \Delta$: domains that contain C_y in their concept sets.
Output: A Concept $C_y = \{c_y^1, \dots, c_y^k\}$, representing label y in the
conceptual space.

foreach $o \in \mathcal{D}_y$ **do**
 $\gamma_o \leftarrow \text{Vectorise}(o, \Delta(y), \mathcal{Q})$ // determine $p_\gamma^1, \dots, p_\gamma^k$ in $\delta_1, \dots, \delta_k$
 $\Gamma(y) \leftarrow \Gamma(y) \cup \gamma_o$

foreach $\delta_i \in \Delta(y)$ // $\delta_i = \langle Q(\delta_i), C(\delta_i), \omega_{\delta_i} \rangle$, $1 \leq i \leq k$
do
 $c_y^i \leftarrow \emptyset$
 foreach $\gamma \in \Gamma(y)$ **do**
 $p_y^i \leftarrow P_y^i \cup \{p_\gamma^i\}$
 $\eta \leftarrow \text{ConvexHull}(P_y^i)$
 $\phi \leftarrow \{\omega_{\delta_i}(C_y, q^i) | C_y \in \mathcal{C}(\delta_i), q^i \in \mathcal{Q}(\delta_i)\}$
 $c_y^i \leftarrow \langle \eta, \phi \rangle$
 $C_y \leftarrow C_y \cup c_y^i$

concepts, denoted: $C_y = \{c_y^1, \dots, c_y^k\}$, where each c_y^i is the representation of C_y within the domain $\delta_i \in \Delta(y)$ ⁶.

Definition 3.4. A sub-concept c_y^i , representing the concept C_y in the domain δ_i , is defined as a tuple $\langle \eta, \phi \rangle$, where η is the region representing the geometrical area of C_y in the domain δ_i , and ϕ is a set of weights indicating the assigned degrees of salience between C_y and each quality dimension $q \in \mathcal{Q}(\delta_i)$.

In order to represent a concept, the representation of its sub-concepts is defined. The following two sections describe the way to formally represent a concept, by defining its regions and its set of weights, respectively. Algorithm 3.3 shows the steps of the concept representation, with the required parameters to represent a concept C_y .

3.2.1 Convex Regions of Concepts

The identification of the geometrical regions of concepts is based on the location of the known observations. The concept $C_y \in \mathcal{C}$ is represented using the subset of observations $\mathcal{D}_y = \{o_1, o_2, \dots, o_{n_y}\}$ which are labelled with $y \in \mathcal{Y}$. The set of *instances* $\Gamma(y)$ is defined related to the observations in \mathcal{D}_y , denoted

⁶ If a concept is associated with only one domain, then the representation of the concept is in fact equivalent to the representation of its sub-concept in that domain.

by $\Gamma(y) = \{\gamma_1, \gamma_2, \dots, \gamma_{n_y}\}$. These instances then specify the geometrical representation of the concept C_y as a set of regions within the domains. The set of all instances Γ in a conceptual space S is defined as: $\Gamma = \bigcup_{y \in Y} \Gamma(y)$.

Definition 3.5. *An instance γ related to the concept C_y is a finite set of n -dimensional points $\gamma = \{p_\gamma^1, \dots, p_\gamma^k\}$ with a one-to-one mapping from the instance points to the domains $\Delta(y)$, where $|\Delta(y)| = |C_y| = k$.*

An instance $\gamma_o \in \Gamma(y)$ is the representation of the observation $o \in \mathcal{D}_y$. The points of γ_o are basically the values of the associated quality dimensions, which are stored in the feature vector \mathbf{x}_o . Formally, each point $p_\gamma^i \in \gamma_o$ in a domain $\delta_i \in \Delta(y)$ is a numeric vector of the values of the quality dimensions in δ_i , denoted: $p_\gamma^i = \langle q_1(\gamma_o), \dots, q_{|Q(\delta_i)|}(\gamma_o) \rangle$, which is a sub-vector of the feature vector \mathbf{x}_o that includes the features associated with the quality dimensions in $Q(\delta_i)$. This process of determining the points of an instance γ_o using the feature vector of the observation o is called *vectorisation*.

Since all the instances with the label y have a point p^i in domain $\delta_i \in \Delta(y)$, to identify the convex region η of a sub-concept c_y^i , it is necessary to know the location of all these points in the domain. Let P_y^i be the collection of all the points which their corresponding instances are labelled by y , and these points are located in domain δ_i . So, $P_y^i = \{p_{\gamma_1}^i, p_{\gamma_2}^i, \dots, p_{\gamma_{n_y}}^i\}$, where $p_{\gamma_j}^i \in \gamma_j$, $\gamma_j \in \Gamma(y)$, and $j = 1 \dots n_y$.

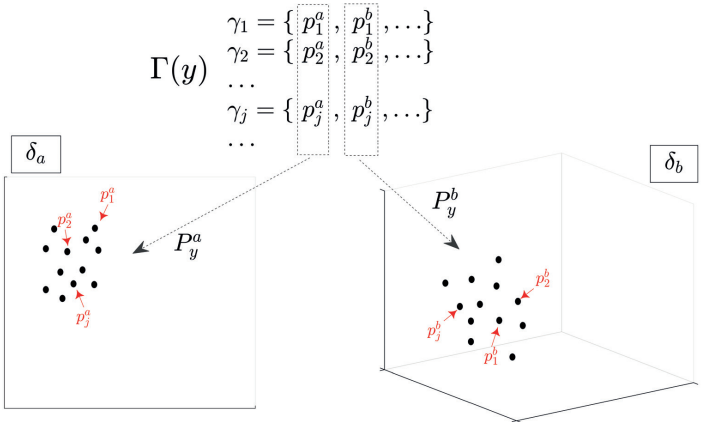
Example 3.7. *Figure 3.5a consists of two domains δ_a and δ_b with two and three quality dimensions, respectively. Assume a concept C_y has two sub-concepts c_y^a and c_y^b within these domains. So, each instance $\gamma_j \in \Gamma(y)$ includes two points p_j^a and p_j^b located in the domains. Figure 3.5a depicts the set of points P_y^a and P_y^b , which will be used to represent sub-concepts c_y^a and c_y^b , respectively.*

The *convexity*, *connectedness*, and *betweenness* are the geometrical criteria required to define a region for a concept in the theory of conceptual spaces [86]. The convexity of concepts is crucial to facilitate the *learnability* of concepts through the instances [90]. Also, with the convexity, this approach satisfies the notion of *betweenness* notion as an essential relation of observations in a same region. So, if two observations of a concept are located at points p_1 and p_2 , then any instance located between p_1 and p_2 also belongs to the same concept [86]. A convex region is a geometric structure within a multidimensional domain which satisfies convexity and connectedness criteria. There are various approaches to identify the convex region covering a set of giving points, such as *convex hull* and *Voronoi tessellations* algorithms, or defining an ellipsoid around the points [9, 86]. For the purpose of this work, the convex hull (CH) is a more convenient choice among others, because it also satisfies the *betweenness* criterion. Since the concepts' regions are formed by its instances, both

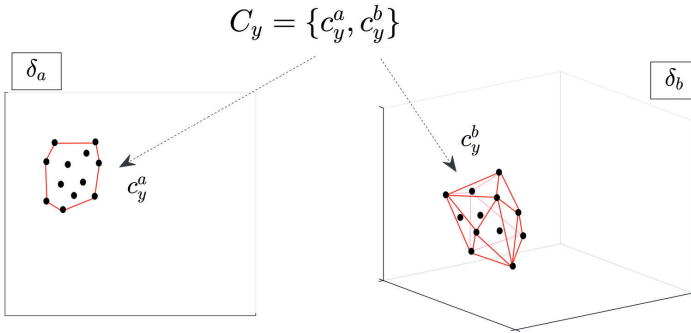
ellipsoid and Voronoi regions assign some points of the space to a concept's region, which are not necessarily between the concept's known instances.

For a sub-concept c_y^i , the convex region η is defined as the convex hull (i.e., convex polytope) of the points in P_y^i , as: $\eta(c_y^i) = \text{CH}(P_y^i)$.

Example 3.8. Figure 3.5b shows the convex regions of the sub-concepts c_y^a and c_y^b . The convex hull $\eta(c_y^a)$ is a 2D polygon in δ_a , and $\eta(c_y^b)$ is a 3D polytope in δ_b . These convex hulls are calculated based on the points P_y^a and P_y^b from the instances in $\Gamma(y)$ in Figure 3.5a.



(a) Instances in $\Gamma(y)$ and their corresponding set of points, P_y^a and P_y^b .



(b) Convex regions of the sub-concepts c_y^a and c_y^b .

Figure 3.5: A concept representation example in a conceptual space with domains δ_a and δ_b .

3.2.2 Context-dependent Weights of Concepts

Depending on the *context*, the *salience* given to various aspects of a concept may vary [86]. In the example of the *apple* concept, in one context the *taste* domain might be more prominent, but in another context, *shape* domain can be salient. In contrast, in such examples of concepts that there is no common knowledge about the salience of the domains in various concepts, the data itself determines the salience of domains and quality dimensions and defines the context-based weights for the concepts. In other words, the observations from different contexts define which domain and quality dimensions are more important to represent the given concepts.

Example 3.9. *For the example of leaf data set, suppose that shape and colour are the domains, and suppose that one wants to differ between the contexts of Swedish leaves and Japanese leaves. Knowing the common-sense knowledge about these contexts might be useless to determine the weights of the domains and dimensions. However, based on the observed data in each of these contexts, one can realise that e.g., the quality dimensions in the shape domain are more salient rather than the colour domain to represent the Swedish leaves, but this inference may not be necessarily valid for Japanese leaves.*

These relative degrees of salience assigned to the dimensions of the domains implicitly represent the notion of context. Here, the context-dependent weights are already embedded in the representation by calculating the relevance of quality dimensions to the concepts (i.e., the weights of the bipartite graph) while specifying the domains. The salient weights ϕ for a sub-concept c_y^i come from the assigned weights ω_{δ_i} in δ_i between $C_y \in \mathcal{C}(\delta_i)$ and the any quality dimension in $\mathcal{Q}(\delta_i)$. Formally:

$$\phi(c_y^i) = \{\omega_{\delta_i}(C_y, q^i) \mid C_y \in \mathcal{C}(\delta_i), q^i \in \mathcal{Q}(\delta_i)\}. \quad (3.6)$$

So, each sub-concept has its own set of weights in relation to the domain's quality dimensions. This point individualises the definition of context-dependent weights from the definition of context weights in other developed conceptual spaces. In other conceptual spaces, a set of overall weights is assigned to the domains without considering the role. But here, for two independent sub-concepts, the assigned weights in the same domain might vary.

3.3 Discussion

This chapter has introduced a data-driven *construction* of conceptual spaces from known observations in a given numeric data set. The framework proposed here has employed machine learning algorithms for the task of identifying relevant features and concepts in a numerical data set, to shape the domains and quality dimensions of a conceptual space. It has been argued that a set of selected and grouped features that provide discrimination between concepts are

adequate to specify the domains and dimensions while preserving the semantic interpretation of the features and concepts. Then, an instance-based approach has been shown for the task of concept formation within the derived conceptual space. A key finding from the data-driven construction of conceptual space is that it provides a generalisation for concept representation, where the model can be constructed or extend by different types of input instances.

This work has demonstrated how to generate a model with which it is possible to create semantic interpretations of new observations. Keßler [131] states that any data-driven approach to generating conceptual spaces cannot be fully automated and require at some point external (symbolic) information. Following his statement, the presented approach in this thesis also relies to some degree on pre-defined concepts related to the input data set, which make it a supervised process. However, the processes of domain/quality dimension specification and concept formation perform automatically based on the data. In other words, performing these processes to create conceptual spaces is not dependent on the symbolic information of the provided knowledge about concepts and features.

In the following, some issues inherent to the data-driven approach are discussed.

Interpretation of features: One important issue to address is to determine how interpretable the selected features are for representing the derived concepts. Inherently, the quality dimensions capture the attributes that can cognitively categorise the concepts [92]. Thus, it makes sense that the feature selection considers the separability between concepts when generating conceptual spaces in a data-driven manner. At the same time, a data-driven approach cannot be separated entirely from meaningful semantics. Hence, a further implicit selection criterion has been to select features that can be expressed in natural language, or as stated by [86], features which can be given a meaningful perceptual interpretation. The interpretability of features is certainly context-dependent and occurs at different levels of feature abstractness [214]. As an example, for a given leaf sample, a *large* or *small area* of a leaf is not informative whereas knowing the *elongation* or *wideness* enables the model to depict a more meaningful description.

Semantics of domains: Another issue is how to form the domains in a conceptual space without human perception (Section 3.1). While the quality dimensions can be mapped to the feature selection methods, the domains which are formed by grouping the features should be semantically meaningful. In this approach, grouping the features is based on how well a subset of the features distinctly represents the various concepts. However, there still exists the problem of verifying the semantic dependency of the quality dimensions within a domain, to realise what quality dimensions are *integral* and what are *separable* without necessarily involving background knowledge. While no solution is

presented to this problem by this study, the problem itself has been discussed in the literature. For example, Gärdenfors in [86] suggests that the verification of deciding whether two quality dimensions are integral or not can be done by empirical testing based on the subject judgements as of the domain experts, and not necessarily using statistical or analytical techniques. It is seemingly difficult, if not impossible, to realise the semantic dependency of the features analytically. For example, by looking at the values of RGB as the dimensions of the colour domain for a set of observations, there is no indication to realise their semantic relations. At the very least in the proposed approach, with relying on the associations between the observations and features, the method considers quality dimensions to be integral if they have high relevance to each other that measured by their high relevance to the concepts. Indeed, solving the issue of how to derive a grouping of features for domain specification, can lead to forming a general solution to the problem of determining an evaluation criterion to choose between competing conceptual spaces, an issue raised by Gärdenfors in [88].

Quality and quantity of known instances: On concept representation (Section 3.2), convex regions and the salient weights are induced in a data-driven way by considering the observations as the instances of the concepts. Still, these representations suppose that there are sufficient known instances that are representative of the concepts to determine geometric regions and the weights. This is a crucial point which is inherent in all data-driven methods. In the literature of conceptual spaces, the knowledge-driven approaches have used simple thresholds or cutoffs to determine the regions, as illustrated in the definition of the regions for the *mountain* and *hill* concepts described by [8]. However, in such systems like the leaf data set, it is not trivial to initialise in advance the specific geometric boundaries for the categories of the leaf species as the concepts, or the precise salient weights the provided domains. So, in a data-driven approach it is essential to have an adequate set of the known instances of the concepts, and thus an area that is worth to study is to determine on how to dynamically adjust a conceptual space when a set of new observations emerge.

In sum, this chapter has explained the process of deriving the elements of a conceptual space in a data-driven manner in order to represent concepts. This data-driven conceptual space will then be utilised for the task of semantic inference, which is explained in Chapter 4.

Chapter 4

Semantic Inference in Conceptual Spaces

“There are no facts, only interpretations.”

— Friedrich Nietzsche (1844–1900)

THE aim of this chapter is to design an approach for inference in order to provide a semantic characterisation for unknown observations. The focus of this chapter is on solving two central questions (1) how a new (unknown) observation is represented in a conceptual space, and (2) how this representation enables the inference of semantic descriptions for the observation.

The first question refers to the problem of *induction* in conceptual spaces theory [88]. To develop such inductions in a conceptual space, it is important to realise which concepts represent a new observed instance. Due to the geometrical representation of the conceptual space, the *similarity* between the instances in the space enables the model to define the notion of *inclusion* as an operator to measure the similarity of new observations to the specified concepts within metric domains.

The second question refers to the problem of *symbol grounding* in conceptual spaces theory [14]. The inference of a semantic representation for any input observation in natural language is enabled by defining a *symbol space* in the conceptual space. With the use of the symbol space, the inference can be done by semantic reasoning in which new observations are assigned to a set of linguistic symbols in the symbol space.

This chapter first introduces the notion of a symbol space. Then, it proposes a process for semantic inference to provide linguistic descriptions for the new observations based on the symbol space. In general, the proposed inference in a conceptual space consists of the following steps:

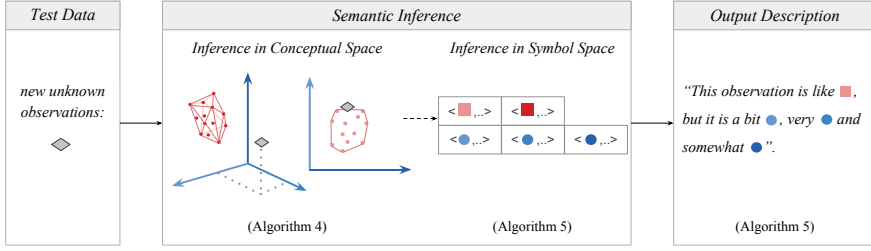


Figure 4.1: Illustration of the steps of inferring linguistic descriptions for an unknown observation via the constructed conceptual spaces and its corresponding symbol space. The details of the semantic inference step is explained in Section 4.2.

- **Defining the *symbol space***, based on the prior knowledge for linguistic characterisation of the concepts and quality dimensions,
- **Inferring *linguistic descriptions***, for each new unknown observation, based on the *inclusion* of its corresponding *instance* in the concepts:
 - **Inference in *conceptual space***: specifying the geometrical location of a new instance within the conceptual space, examining the *inclusion* of the instance, and determining the linguistic labels in the symbol space based on the associated concepts and dimensions,
 - **Inference in the *symbol space***: annotating and characterising the instance from the provided set of symbolic terms, and generating linguistic descriptions.

Example 4.1. Consider the concepts and quality dimensions of the leaf conceptual space in Example 3.5. A new observed leaf can be either linguistically represented by a known concept (e.g., C_{no}) where “The new observation is a *Nerium* leaf.”, or by a set of related quality dimensions (e.g., q_{el} and q_{ro}), such as “The new observation is an **elongated** and **lance-shaped** leaf.”

Figure 4.1 illustrates the step of inferring linguistic descriptions for an unknown observation through the constructed conceptual spaces and its corresponding symbol space. The details of each component are explained in Section 4.2, after formally defining the symbol space in Section 4.1.

4.1 Symbol Space Definition

According to a general formulation proposed by Aisbett and Gibbon [14], a conceptual space can be augmented with a *symbol space*. This extension provides an internal mapping between geometrical elements in conceptual space

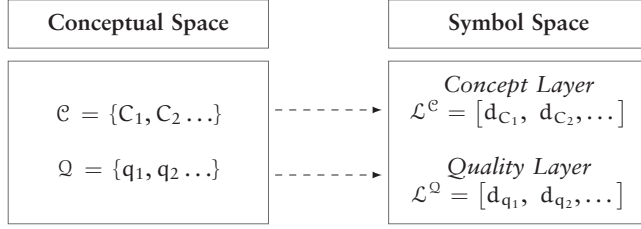


Figure 4.2: Schematic of a conceptual space and the coupled symbol space.

(such as concepts, dimensions, domains, etc.) and the symbolic labels (typically words) in symbol space.

Definition 4.1. A symbol space \mathcal{S} of size n is a space containing n symbol dimensions \mathcal{L}^S , wherein each concept and quality dimension in the conceptual space is linked to a symbol dimension. Symbol dimensions are isomorphic to the real number interval $[0, 1]$.

Each concept and quality dimension in the conceptual space is linked to a symbol dimension in the symbolic space. Based on the definition of Aisbett and Gibbon, the symbol dimensions need to be named by the primitive input labels of the associated concepts (classes) and quality dimensions (features). The construction of the symbol space is a knowledge-based process, wherein the prior knowledge is encoded [14]. The prior knowledge specifies the symbolic expressions in natural language form, related to the elements of conceptual space \mathcal{S} .

This study proposes a two-layer symbol space containing the symbol dimensions of the concepts, $\mathcal{L}^C = [d_{C_1}, d_{C_2}, \dots]$, as *concept layer*, and the symbol dimensions of quality dimensions, $\mathcal{L}^Q = [d_{q_1}, d_{q_2}, \dots]$, as *quality layer*. The symbol dimensions \mathcal{L}^C and \mathcal{L}^Q are acquired from the set of input labels \mathcal{Y} , and set of selected features \mathcal{F}' , respectively. So, for every concept $C_y \in \mathcal{C}$, there is a symbol dimension in the concept layer, and for each quality dimension $q \in \mathcal{Q}$, there is a symbol dimension in the quality layer. Figure 4.2 shows a schematic presentation of the associations between the elements in a conceptual space and the two-layer symbol dimensions in a symbol space.

Example 4.2. Consider the conceptual space of leaves \mathcal{S}^l in Example 3.5. The associated symbol space \mathcal{S}^l is defined with two-layer symbol dimensions, denoted by $\mathcal{L}^C = [d_{tt} : \text{label}(C_{tt}), d_{no} : \text{label}(C_{no})]$ in the concept layer, and $\mathcal{L}^Q = [d_{e1} : \text{label}(q_{e1}), d_{ro} : \text{label}(q_{ro})]$ in the quality layer.

Any instance in a conceptual space is then associated with a point in the symbol space, namely a *symbol vector*. For a given instance γ , the associated symbol vector \mathcal{V}_γ in \mathcal{S} specifies the applicability of the symbol dimensions for γ in the range 0 to 1 for each dimension [14]. The symbol vector \mathcal{V}_γ is a

concatenation of two vectors $\mathcal{V}_\gamma :< \mathcal{V}_\gamma^c, \mathcal{V}_\gamma^q >$, one vector in the concept layer and one vector in quality layer, respectively. Thus, $|\mathcal{V}_\gamma^c| = |\mathcal{L}^c|$, and $|\mathcal{V}_\gamma^q| = |\mathcal{L}^q|$.

Example 4.3. Consider the conceptual space of leaves S^l in Example 3.5. For a new leaf instance γ , the symbol vector \mathcal{V}_γ is defined as a 4-dimensional vector with concatenation of $\mathcal{V}_\gamma^c = < v_{d_{tt}}, v_{d_{no}} >$, and $\mathcal{V}_\gamma^q = < v_{d_{e1}}, v_{d_{ro}} >$.

The symbol vector is defined as a two-element vector, wherein each $v_i \in \mathcal{V}_\gamma$ consists of a pair of values $v_i = (\text{label}, \text{value})$. The *label* shows the related symbolic term and the *value* shows how representative is the instance to the dimension d_i (either how similar to its concepts or how related to the quality dimensions). The notion $\mathcal{V}_\gamma^c(C_y) = v_{d_{cy}}$ indicates the value of the symbol vector in the concept layer for the dimension related to the concept C_y , and similarly, $\mathcal{V}_\gamma^q(q_j) = v_{d_{qj}}$ indicates the value of the symbol vector in the quality layer for the dimension related to the quality dimension q_j . The further sections explain how the elements of a symbol vector for a new instance are assigned values based on the inclusion of the instance within the domains.

4.2 Inferring Linguistic Descriptions for Unknown Observations

For any given unknown observation, the goal is to infer a semantic description in natural language form. The core of the inference process is to cope with the notion of *similarity* in conceptual spaces. To place the new instances in the space and choose the best concepts that include (or are similar enough to) an observation [189]. The metric structure of a geometrical conceptual space enables the model to measure the semantic similarity of concepts and instances in the space [9]. The proposed construction of the conceptual space in Chapter 3 facilitates these measurements since the representations of concepts and instances span across domains using the geometric elements, i.e., convex regions and points. From the point of view of NLG, inferring linguistic descriptions for unknown observations covers the main tasks of an NLG pipeline for generating natural language text out of non-linguistic data: *Content determination*, *Microplanning* (including *lexicalisation*), and *Realisation* [185]. This phase employs various developed methods for linguistic descriptions (i.e., fuzzy set theory [177]) to ease the process of quantifying the location of unknown samples within a conceptual space, and infer the proper linguistic terms.

The process of inferring linguistic descriptions for an instance γ' is presented in two following phases.

- *Phase A: Inference in Conceptual Space*, that first determines the inclusion of the new instance γ' in the concepts within the domains $\Delta(\gamma')$ using semantic similarity, and then sets the values of symbol vector $\mathcal{V}_{\gamma'}$ (performing the content determination task).

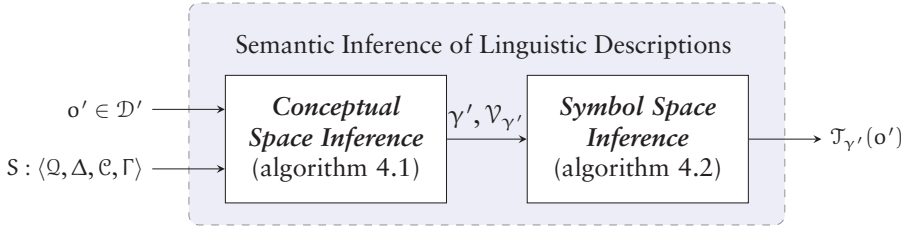


Figure 4.3: Two phases of the semantic inference for generating linguistic descriptions, with the input and output parameters of each phase.

- *Phase B: Inference in Symbol Space*, that verbalises the symbol vector $\nu_{\gamma'}$ into a set of lexical items which are human-readable descriptions. (performing the lexicalisation and realisation tasks).

For a given set of new observations $\mathcal{D}' = \{o'_i : (\mathbf{x}_{o'_i})\}$, let γ' be the corresponding instance to the unknown observation $o' \in \mathcal{D}'$ which is not assigned to any of the known class labels of \mathcal{Y} . Also, let $\Delta(\gamma') \subseteq \Delta$ be a set of domains that γ' has corresponding points in each of them¹, where $|\Delta(\gamma')| = k'$.

Figure 4.3 illustrates the phases of inferring a linguistic description for a new observation, with their input and output parameters. The details of the phases A and B are explained in the following sections.

4.2.1 Phase A: Inference in Conceptual Space (Content Determination)

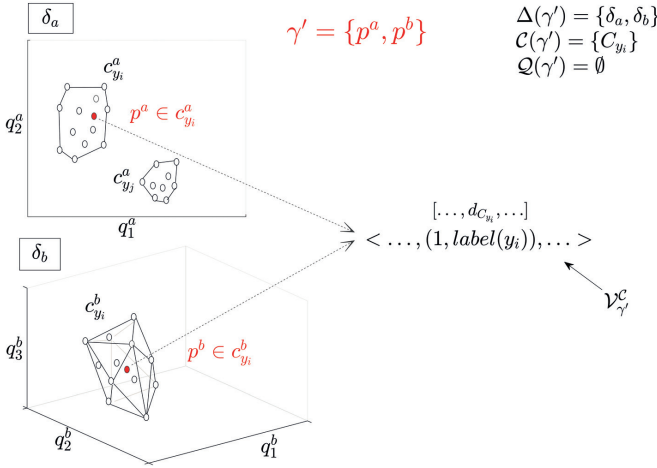
This phase first presents the comparison of the new instance with each of the concepts using the similarity measure to check whether it is included within concept regions or not. Then, based on the result of this inclusion, it describes how to initialise a symbol vector and set its values in both concept layer and quality layer. From the NLG perspective, this performs the task of *content determination* [182], that decides which set of information is required to characterise a new observation in the final description. The process of checking the inclusion is a simple fuzzy extension of an *instance-based* method that measures the membership of the new instance to the nearest labelled region of instances. Although any classification approach can do it, the process is formulated here with respect to the definitions of the introduced conceptual space.

For an instance γ' , the symbol vector $\nu_{\gamma'}$ is calculated based on the inclusion of the instance points $p_{\gamma'}$ in different regions within the domains. As defined before, γ' is represented with a set of points $\gamma' = \{p_{\gamma'}^1, \dots, p_{\gamma'}^{k'}\}$. In

¹It is notable that a new observation is not necessarily defined in all domains, since there might be no calculated values for some of the features/quality dimensions. So, the corresponding instance may not have points in all the provided domains.

general, with placing a new instance in a conceptual space, four cases can occur. Without losing generality, assume γ' consists of two points in two domains δ_a and δ_b , denoted by $\gamma' = \{p^a, p^b\}$. Also assume that there are two concepts C_{y_1} and C_{y_2} that have been represented in one or both of these two domains. Figure 4.4 shows the four different cases with respect to various positions of the points p^a and p^b , and their relations to the sub-concepts' regions within the domains. One instance can be located in the space differently as follows:

1. Totally included in a concept within all the domains (case one, Figure 4.4a),
2. Partially included in just a concept (case two, Figure 4.4b),
3. Partially included in two distinct concepts (case three, Figure 4.4c),
4. Not included in any concept (case four, Figure 4.4d).



(a) Case one: γ' is totally included in concept C_{y_i} (in all the domains).

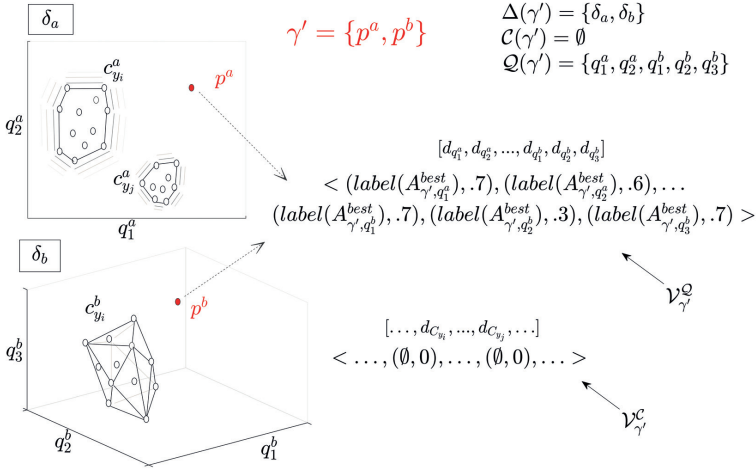
(d) Case four: γ' is not included in any concept.

Figure 4.4: An illustration of four different cases with respect to the various positions of an instance points $\gamma' = \{p^a, p^b\}$ within two domains δ_a and δ_b , together with the assigned values to symbol vector $\mathcal{V}_{\gamma'}$, according to the inclusion of the points in the presented sub-concepts.

To assign a concept to γ' , it is necessary to check the inclusion of the instance points in the concept's regions within $\Delta(\gamma')$ using a similarity measure. Within a single domain, two states need to be considered: 1) If the instance point is included in a region, then the region's concept will be assigned to γ' . So, the corresponding symbol dimension of the concept will be activated in the symbol vector of γ' (in the concept layer). 2) If there is no region that the instance belongs to, then no concept will be assigned to γ' within that domain. The symbol dimensions related to the quality dimensions of the domain will be activated in the symbol vector of γ' (in the quality layer). Formally, the symbol vector $\mathcal{V}_{\gamma'}$ gets the values in each domain $\delta_i \in \Delta(\gamma')$ as follows: First a function called *Graded Membership* function, $\mathcal{G}(p_{\gamma'}^i, c^i)$, is defined as an inclusion operator to determine the similarity degree of a point $p_{\gamma'}^i$, to the region of a sub-concept $c^i \in \delta_i$ within δ_i . If $p_{\gamma'}^i$ is similar enough to the sub-concept's convex region with a certain membership degree, then the value of $\mathcal{G}(p_{\gamma'}^i, c^i)$ is set to $\mathcal{V}_{\gamma'}^C(C_y)$, where $C_y \ni c^i$. Moreover, another function called *Graded Quality* function, $\mathcal{H}(p_{\gamma'}^i, q^i)$, is defined to measure to what degree $p_{\gamma'}^i$ belongs to a quality dimension $q^i \in \mathcal{Q}(\delta_i)$. If $p_{\gamma'}^i$ is not included in any of the sub-concepts, then the value of $\mathcal{H}(p_{\gamma'}^i, q^i)$ is set to $\mathcal{V}_{\gamma'}^Q(q^i)$.

Algorithm 4.1: Inference in Conceptual Space

```

Function ConceptualSpaceInference( $\gamma', \Omega, \Delta, \mathcal{C}$ )
  foreach  $\delta_i \in \Delta(\gamma')$  do
     $p \leftarrow p_{\gamma'}^i \in \gamma'$ 
    // Set the symbol vector in concept layer
    foreach  $c \in \delta_i$  do
       $\text{sim}_{\gamma', c_y} = \max(\text{sim}_{\gamma', c_y}, \mathcal{G}(p, c))$  //  $C_y \ni c$ 
      if  $\text{sim}_{\gamma', c_y} \neq 0$  // cases 1 and 3 (and partially 2)
      then
         $\text{label}_{\gamma', c_y} = \text{label}(C_y)$ 
      else
         $\text{label}_{\gamma', c_y} = \emptyset$ 
       $\mathcal{V}_{\gamma'}^c(C_y) = (\text{label}_{\gamma', c_y}, \text{sim}_{\gamma', c_y})$ 
    // Set the symbol vector in quality layer
    if  $\mathcal{V}_{\gamma'}^c(C_y) == (\emptyset, 0)$  // case 4 (and partially 2)
    then
      foreach  $q \in \delta_i$  do
         $\text{degree}_{\gamma', q} = \mathcal{H}(p, q)$ 
         $\text{label}_{\gamma', q} = \text{label}(A_{\gamma', q}^{\text{best}})$ 
         $\mathcal{V}_{\gamma'}^q(q^i) \leftarrow (\text{label}_{\gamma', q}, \text{degree}_{\gamma', q})$ 
  return  $\mathcal{V}_{\gamma'} : \langle \mathcal{V}_{\gamma'}^c, \mathcal{V}_{\gamma'}^q \rangle$ 

```

Example 4.4. Consider the instance γ' in Figure 4.4b. Point p^b is included to the region of $c_{y_i}^b$ within δ_b . So, $\mathcal{V}_{\gamma'}^c(C_{y_i})$ in concept layer will get the graded membership value between p^b and $c_{y_i}^b$. But, point p^a is not included in any of the regions within δ_a . So, $\mathcal{V}_{\gamma'}^q(q_1^a)$ and $\mathcal{V}_{\gamma'}^q(q_2^a)$ in quality layer will get the graded quality values between p^a and two quality dimensions q_1^a and q_2^a , respectively. but p^a is not included in any of the regions within δ_a . So, the symbol vector $\mathcal{V}_{\gamma'}$ will get values at three indices: $\mathcal{V}_{\gamma'}^c(C_{y_i})$ in concept layer, and $\mathcal{V}_{\gamma'}^q(q_1^a)$ and $\mathcal{V}_{\gamma'}^q(q_2^a)$ in quality layer.

Algorithm 4.1 shows the steps to set the symbol vector values while iterating through all the involved domains $\Delta(\gamma')$. Both graded membership function and graded quality function are formally defined in the following sections.

Graded Membership function

In Section 3.2, a sub-concept c is defined as a pair $c : \langle \eta_c, \phi_c \rangle$, where η_c is the convex region representing the geometrical area of the corresponding concept in domain δ , and ϕ_c is a set of weights showing the assigned degrees of

salience between sub-concept and each quality dimension within the domain. The problem of *inclusion* has been studied in the literature of the conceptual spaces theory with various definitions such as inclusion operator [9], graded similarity [92], and graded membership [69, 112]. These definitions calculate the similarity of the instance points to the regions based on their geometrical distances, with or without considering the gradedness of membership. Here, since the convex regions of concepts are constructed based on the observed instances, it does not make sense to adhere to the *crisp* boundaries of the calculated regions rigidly. So, a point which is not certainly inside the calculated boundaries a region, but is similar enough to the region, can be counted as a member of the region's concept.

The *Graded Membership function* $\mathcal{G}(p, c)$ is defined as an inclusion operator between a given point p and a defined sub-concept $c : \langle \eta_c, \phi_c \rangle$. This function shows how similar is p to the convex region η_c of c with the certain set of weights ϕ_c within a metric domain δ . The graded membership is calculated by applying geometrical algorithms that consider whether an n -dimensional point is included in the n -dimensional convex hull or not. If point p is certainly included in the region η_c , then $\mathcal{G}(p, c)$ is equal to one. But if p is outside the region, then the similarity of point p to the region is defined as a monotonic decreasing function [69] which is measured using a fuzzy membership function of distance $\text{sim}(p, c) = f[d(p, c)]$. This similarity takes values between $[0, 1]$ and expresses the graded degree of inclusion of p in c . From the experiments on similarity cognition, the similarity can be measured as an exponential decay function of the distance: $\text{sim}(d) = e^{-r \cdot d}$ [202] (where r is a constant factor). Using a fuzzy membership function to measure the similarity has the advantage of using the notions from the fuzzy set theory that provide linguistic descriptions for the output fuzzy degrees [188].

Several methods have been proposed to compute the distance of a point p to a convex region η_c . The *Hausdorff distance* $d^H(p, \eta_c)$ [14] is a proper choice, which relies on the definition of a distance measure between two n -dimensional points (namely *Weighted Minkowski Metric*). As a function of similarity measure, a graded membership function is defined inspired by Hampton's definition [112] in which a determinate boundary region of membership is assumed. For the points inside the region, the membership value is one. For the points out of the region, with a given lower-bound threshold θ_L , if $d(p, \eta_c) \leq \theta_L$, then p is similar enough to be counted as a member of c , and if $d(p, \eta_c) > \theta_L$, then p is far to be counted as an instance of c [188]².

²The definition of graded membership function in [69, 112] is slightly different, where it is based on three thresholds to define the lower, upper, and the middle level of the boundary regions.

Definition 4.2. *The graded membership function $\mathcal{G} : \delta \rightarrow [0, 1]$ is the similarity measure between a point p and a sub-concept $c \in C_y$ as:*

$$\mathcal{G}(p, c) = \begin{cases} 1 & \text{if } p \in \eta_c \\ e^{-r^{d^H(p, \eta_c)}} & \text{if } p \notin \eta_c \text{ \& } d^H(p, \eta_c) \leq \theta_L \\ 0 & \text{if } p \notin \eta_c \text{ \& } d^H(p, \eta_c) > \theta_L \end{cases} \quad (4.1)$$

The symbol vector of a new instance in the concept layer ($\mathcal{V}_{\gamma'}^c$) is set by the graded membership function by measuring the similarity of an instance point to each region of the sub-concept within the domains (Algorithm 4.1). As mentioned in Section 4.1, each $v_i \in \mathcal{V}_\gamma$ consists of a pair of values $v_i = (\text{label}, \text{value})$. The similarity values greater than zero will lead to assign a non-empty label to the corresponding concept's index in symbol vector. Formally, for a given γ' and C_y , two elements (label, value) are calculated as: $\mathcal{V}_{\gamma'}^c(C_y) = (\text{label}_{\gamma', C_y}, \text{sim}_{\gamma', C_y})$, where

$$\text{sim}_{\gamma', C_y} = \max_{p_i \in \gamma', c_j \in C_y} (\mathcal{G}(p_i, c_j)), \quad (4.2)$$

and

$$\text{label}_{\gamma', C_y} = \begin{cases} \text{label}(C_y) & \text{if } \text{sim}_{\gamma', C_y} > 0 \\ \emptyset & \text{o.w} \end{cases} \quad (4.3)$$

Example 4.5. *Figures 4.4a, 4.4b, and 4.4c show the example values of the graded membership function (\mathcal{G}) calculated for the points p^a and p^b based on their positions and distances to the convex regions of the sub-concepts in the space. For example in Figure 4.4c, suppose in δ_a , $\mathcal{G}(p_{\gamma'}^a, c_{y_j}^a) = 0.9$. Then, the elements of $\mathcal{V}_{\gamma'}^c(C_{y_j})$ are set to $(\text{label}(C_{y_j}), 0.9)$.*

Graded Quality Function

According to Algorithm 4.1, if a point p in a domain δ is not similar enough to any sub-concept within δ , then the values of the symbol vector in the quality layer will be set based on the graded value of p for each quality dimension of δ . Recall from Chapter 3, a quality dimension $q : \langle H_q, I_q, \mu_q \rangle$ contains a family of membership functions μ_q , representing the linguistic terms related to graded values of q . In particular, this function is defined as a fuzzy granulation to exploit the linguistic characterisation of feature values, which are identified by prior knowledge. To formalise μ_q , fuzzy membership functions are defined for a set of pre-defined label classes which forms a *fuzzy partition* of the interval I_q [165]. Suppose A_i is a class (label) acquired for q (e.g., linguistic label *tall* for feature *height*). The corresponding fuzzy set is defined as: $A_i = \{(x, \mu_{A_i}) \mid x \in I_q\}$, where μ_{A_i} is a sigmoidal membership function with

certain parameters to define the lower and upper boundaries of the function³. Then, $\mu_q = \{\mu_{\Lambda_1}, \dots, \mu_{\Lambda_n}\}$.

Example 4.6. Consider the elongation, described in Example 3.2. A set of label classes to describe the elongation is $\{A_{\text{cir}} = \text{'circular'}, A_{\text{elp}} = \text{'elliptical'}, \text{and } A_{\text{eld}} = \text{'elongated'}\}$. Then the family of membership functions for q_{el} is $\mu_{q_{\text{el}}} = \{\mu_{\Lambda_{\text{cir}}}, \mu_{\Lambda_{\text{elp}}}, \mu_{\Lambda_{\text{eld}}}\}$. Figure 4.5 depicts the defined membership functions for the elongation quality dimension.

This linguistic mapping provides a symbolic representation for numeric interval values of the quality dimensions. The graded quality value of an instance γ' for a quality dimension q is calculated based on the quality dimension value of the instance point as $p_q = p_{\gamma'}(q)$, where $p_q \in I_q$. Using the defined fuzzy membership functions, p_q is mapped into the fuzzy set best matching to the given value. Using the functions in μ , the values of the symbol vector can be set in the quality layer. Recall $p = \langle p_{q_1}, \dots, p_{q_{|Q(\delta)|}} \rangle$ as the vector of quality dimension values for the point p in δ .

Definition 4.3. Graded quality function $\mathcal{H} : I_q \rightarrow [0, 1]$ is the degree of membership, wherein for a quality dimension q , it returns the maximum degree of membership of p_q using the membership functions in μ_q , as:

$$\mathcal{H}(p, q) = \max_{\mu_{\Lambda_i} \in \mu_q} \mu_{\Lambda_i}(p_q) \quad (4.4)$$

The symbol vector of a new instance in the quality layer ($\mathcal{V}_{\gamma'}^Q$) is filled with the values of the graded quality function (Algorithm 4.1). Similar to the concept layer, each $v_i \in \mathcal{V}_{\gamma'}$ consists of a pair of values $v_i = (\text{label}, \text{value})$. The value of the graded quality function, which is the maximum degree of membership of v_i assigns the best match fuzzy subset (i.e., symbolic label) to the corresponding quality dimension's index in symbol vector. Formally, for a given γ' and C_y , two elements (label, value) of are calculated as: $\mathcal{V}_{\gamma'}^Q(q) = (\text{label}_{\gamma',q}, \text{degree}_{\gamma',q})$, where

$$\text{degree}_{\gamma',q} = \mathcal{H}(p, q), \quad (4.5)$$

and

$$\text{label}_{\gamma',q} = \text{label}(A_{\gamma',q}^{\text{best}}). \quad (4.6)$$

Here, $A_{\gamma',q}^{\text{best}}$ is the fuzzy subset with the maximum degree of membership such that $\mu_{\Lambda_{\gamma',q}^{\text{best}}} \in \mu_q$ and $\forall(\mu_{\Lambda_i} \in \mu_q) : \mu_{\Lambda_{\gamma',q}^{\text{best}}}(p_q) \geq \mu_{\Lambda_i}(p_q)$.

Example 4.7. Considering the functions $\mu_{q_{\text{el}}}$ in Example 4.6 (Figure 4.5), assume that for a given instance point $p_{\gamma'}$, its elongation value is $p_q = 0.75$. So,

³Sigmoidal membership function is defined as: $f(x, a, c) = \frac{1}{1 + e^{-a(x-c)}}$

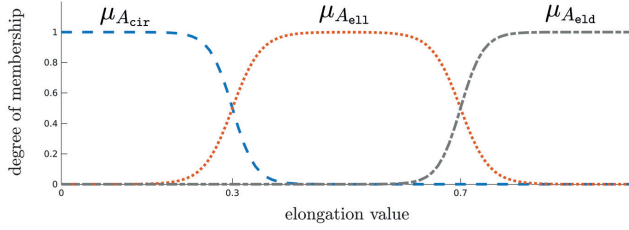


Figure 4.5: Example of the membership functions of the linguistic terms *circular*, *elliptical*, and *elongated*, describing the *elongation* quality dimension.

$\mu_{A_{\text{cir}}}(\mathbf{p}_q) = 0$, $\mu_{A_{\text{elp}}}(\mathbf{v}_q) = 0.15$, and $\mu_{A_{\text{eld}}}(\mathbf{p}_q) = 0.9$. Then, $\mathcal{H}(\mathbf{p}_{\gamma'}, \mathbf{q}_{\text{el}}) = 0.9$, and $A_{\gamma', \mathbf{q}_{\text{el}}}^{\text{best}} = A_{\text{eld}}$. Thus, one can say that “the given instance is elongated to a degree 0.9”.

Example 4.8. Figures 4.4b and 4.4d show the example values of graded quality function (\mathcal{H}) calculated for the points \mathbf{p}^a and \mathbf{p}^b based on their values related to the quality dimensions of two domains. For example in Figure 4.4b, suppose \mathbf{p}^a is not included in any region in δ_a . Then, two vector indices of the symbol vector in the quality layer get values as: $\mathcal{V}_{\gamma'}^Q(\mathbf{q}_1^a) = (\text{label}(A_{\gamma', \mathbf{q}_1^a}^{\text{best}}), 0.7)$ and $\mathcal{V}_{\gamma'}^Q(\mathbf{q}_2^a) = (\text{label}(A_{\gamma', \mathbf{q}_2^a}^{\text{best}}), 0.6)$.

4.2.2 Phase B: Inference in Symbol Space (Lexicalisation and Realisation)

In this phase, the aim is to infer a linguistic lexicon from the symbol vector of an unknown instance, to generate descriptions in natural language form. From the NLG perspective, this is the task of *lexicalisation*, that decides which linguistic terms (i.e., natural words) should be selected from the determined content [182]. This can be done by verbalising the linguistic labels that are calculated and stored in the symbol vector. This verbalisation is done either with *annotating* a new instance via the concept labels, or with *characterising* the instance via the quality dimension labels. The tasks of *annotation* and *characterisation* will assign a set of lexical items to an unknown observation. This collection of linguistic terms is then turned to the natural language phrases (i.e., sentences) using the *realisation* tools in NLG systems. Algorithm 4.2 shows the steps of the tasks in phase B.

Annotation in the Concept Layer

Annotation for an instance γ' is to annotate a set of linguistic labels which are derived from the associated concepts in the concept layer of symbol vec-

Algorithm 4.2: Inference in Symbol Space

```

Function SymbolSpaceInference( $\mathcal{V}_{\gamma'}$ )
    // Annotation in the concept layer
    foreach  $C_y \in \mathcal{C}$  do
        if  $\mathcal{V}_{\gamma',e}(C_y) \neq (\emptyset, 0)$  then
             $\mathcal{C}(\gamma') \leftarrow \mathcal{C}(\gamma') \cup \{C_y\}$ 
     $\mathcal{O}_\mathcal{C}(\gamma') \leftarrow \text{OrderConceptLabels}(\mathcal{C}(\gamma'))$ 
     $\mathcal{T}_\mathcal{C}(\gamma') \leftarrow \text{Annotate}(\mathcal{O}_\mathcal{C}(\gamma'), \mathcal{V}_{\gamma',e})$  // in concept layer
    // Characterisation in the quality layer
    foreach  $\delta \notin \Delta(\mathcal{C}(\gamma'))$  do
         $\mathcal{Q}(\gamma') \leftarrow \mathcal{Q}(\gamma') \cup \{\mathcal{Q}(\delta)\}$ 
     $\mathcal{O}_\mathcal{Q}(\gamma') \leftarrow \text{OrderQualityLabels}(\mathcal{Q}(\gamma'))$ 
     $\mathcal{T}_\mathcal{Q}(\gamma') \leftarrow \text{Characterise}(\mathcal{Q}(\gamma'), \mathcal{V}_{\gamma',\mathcal{Q}})$  // in quality layer
    // Linguistic Realisation
     $\mathcal{T}_{\gamma'} \leftarrow \text{Realise}(\mathcal{T}_\mathcal{C}(\gamma'), \mathcal{T}_\mathcal{Q}(\gamma'))$ 
return  $\mathcal{T}_{\gamma'}$ 

```

tor $\mathcal{V}_{\gamma',e}$. Each C_y is included in the set of associated concepts of the instance, $\mathcal{C}(\gamma')$, if the corresponding element in $\mathcal{V}_{\gamma',e}$ is not empty. Formally, $C_y \in \mathcal{C}(\gamma') \iff \mathcal{V}_{\gamma',e}(C_y) \neq (\emptyset, 0)$.

Example 4.9. *Considering Figures 4.4a and 4.4b, γ' is associated with only one concept C_{y_i} . So, $\mathcal{C}(\gamma') = \{C_{y_i}\}$. In Figure 4.4c, γ' is associated with two concepts C_{y_i} and C_{y_j} . So, $\mathcal{C}(\gamma') = \{C_{y_i}, C_{y_j}\}$. Finally, in Figure 4.4d, there are no associated concepts. So, $\mathcal{C}(\gamma') = \emptyset$.*

After determining $\mathcal{C}(\gamma')$, if γ' is associated with two or more distinct concepts as $\mathcal{C}(\gamma') = \{C_{y_i}, C_{y_j}, \dots\}$, then γ' is an instance of all the associated concepts. In this case, an extra process is needed to sort and combine the concept labels to annotate γ' with a new set of linguistic labels.

The task of *concept combination* is discussed widely in the literature of conceptual space theory [86, 142, 188]. What is important for the inference is to distinguish which concept labels are the *modifiers* and which are the *modified* concepts. This distinction leads to order the labels in the final linguistic expression [9]. In particular, an *ordered set*⁴ of the associated concepts, $\mathcal{O}_\mathcal{C}(\gamma') = \{C'_1, C'_2, \dots\}$ is defined to prioritise the modifier concepts over modified ones. Since there is no background knowledge to define the semantic order of the associated concepts, the ordering process relies based on the *graded membership* values that can be retrieved from $\mathcal{V}_{\gamma',e}$.

⁴In the set theory, an ordered set is defined as a set of elements, plus a relation \leq between each pair of the elements that presents the order of them [229].

Example 4.10. *Considering the case three in Figure 4.4c, since γ' is located in sub-concept $c_{y_i}^b$ with graded membership 1 and in sub-concept $c_{y_j}^a$ with graded membership 0.9, then the corresponding ordered set of concepts for γ' will be: $\mathcal{O}_C(\gamma') = \{C_{y_i}, C_{y_j}\}$. Suppose there is an unknown instance γ' in a conceptual space including Colour and Taste domains. If γ' is located in both 'red' sub-concept with grading degree 0.95 and 'sweet' sub-concept with grading degree 0.7 (in colour and taste domains, respectively), then the corresponding ordered set of concepts is: $\mathcal{O}_C(\gamma') = \{C_{\text{red}}, C_{\text{sweet}}\}$.*

The set of annotations for γ' then is defined as an ordered set of lexical items $\mathcal{T}_C(\gamma') = \{\text{label}(C_1'), \text{label}(C_2'), \dots\}$. These annotations are the corresponding linguistic terms to the ordered concepts in $\mathcal{O}_C(\gamma')$, that are retrieved from the labels in $\mathcal{V}_{\gamma',C}$ for the concepts in $\mathcal{C}(\gamma')$.

Example 4.11. *For the conceptual space of leaves presented in Example 3.5, suppose an unknown leaf sample γ' is associated with both known concepts Tilia and Nerium. Assume that $\mathcal{V}_{\gamma',C}(C_{\text{tt}}) = (\text{label}(C_{\text{tt}}), 0.5)$ and $\mathcal{V}_{\gamma',C}(C_{\text{no}}) = (\text{label}(C_{\text{no}}), 0.9)$. Then, $\mathcal{T}_C(\gamma') = \{\text{label}(C_{\text{no}}) = \text{'Nerium'}, \text{label}(C_{\text{tt}}) = \text{'somewhat Tilia'}\}$.*

Characterisation in the Quality Layer

Characterisation for γ' is to assign the linguistic descriptions of the associated quality dimension based on the values in the quality layer of the symbol vector $\mathcal{V}_{\gamma',Q}$. The motivation behind the characterisation comes from the lack of the concept annotation in the cases with no associated concept within the domains (like case four and partially case two in Figure 4.4). This is especially important for those instances that are completely unknown for the systems and are not representable by any of the defined concepts, but still are explainable with their quality dimensions' values.

According to Algorithm 4.1, if γ' within a domain does not belong to any sub-concept, then $\mathcal{V}_{\gamma',Q}$ gets values from the domain's quality dimensions. Obviously, if γ' has even one associated concept within the domain, there is no need to involve the quality dimensions of that domain in the characterisation process. For the calculated $\mathcal{V}_{\gamma',Q}$, each quality dimension q is included in the set of associated quality dimensions $\mathcal{Q}(\gamma')$, if the corresponding elements in the $\mathcal{V}_{\gamma',Q}$ are not empty. Formally, $q \in \mathcal{Q}(\gamma') \iff \mathcal{V}_{\gamma',Q}(q) \neq \{\emptyset, 0\}$.

Example 4.12. *Considering Figure 4.4b, γ' is not associated with any concepts within δ_a . So, $\mathcal{Q}(\gamma') = \{q_1^a, q_2^a\}$. Also, in Figure 4.4d, there are no associated concepts in any of the domains. So, $\mathcal{Q}(\gamma') = \{q_1^a, q_2^a, q_1^b, q_2^b, q_3^b\}$. In Figures 4.4a and 4.4c, $\mathcal{Q}(\gamma') = \emptyset$.*

Based on Equation 4.6, $\text{label}_{\gamma',q} = \text{label}(A_{\gamma',q}^{\text{best}})$ is the linguistic term related to the best interval of the quality dimension with the maximum degree of membership.⁵

Similar to the process of annotation, a sorting operation is needed to derive the order of quality dimension labels in the final linguistic descriptions. Relying on the *graded quality* values in $\mathcal{V}_{\gamma',\Omega}$, an ordered set of the associated quality dimensions, $\mathcal{O}_{\Omega}(\gamma') = \{q'_1, q'_2, \dots\}$, is defined. The characterisation set for γ' then is simply defined as an ordered set of lexical items $\mathcal{T}_{\Omega}(\gamma') = \{\text{label}(A_{\gamma',q'_1}^{\text{best}}), \text{label}(A_{\gamma',q'_2}^{\text{best}}), \dots\}$. These characterisations are retrieved from the corresponding labels in $\mathcal{V}_{\gamma',\Omega}$, for the quality dimensions in $\Omega(\gamma')$.

Example 4.13. *Considering Example 4.12, suppose γ' is not associated with any known concepts. Assume that for the quality dimensions elongation and roundness, $\mathcal{V}_{\gamma',\Omega}(q_{\text{el}}) = (\text{label}(A_{\text{elid}}), 0.9)$ and $\mathcal{V}_{\gamma',\Omega}(q_{\text{ro}}) = (\text{label}(A_{\text{ro}}), 0.7)$ (referring to Example 4.6). Then, $\mathcal{T}_{\Omega}(\gamma') = \{\text{label}(A_{\text{elid}}) = \text{'elongated'}, \text{label}(A_{\text{ro}}) = \text{'lanced shape'}\}$. For the above example, if the values of dimensions (with the value range $[0\ 1]$) are $v_{\text{weight}} = 0.9$ and $v_{\text{height}} = 0.2$, then the corresponding characterisation set will be $\mathcal{T}_{\Omega}(\gamma') = \{\text{label}(\alpha_{q_{\text{weight}}}) = \text{'heavy'}, \text{label}(\alpha_{q_{\text{length}}}) = \text{'short length'}\}$.*

Phrase Specification and Linguistic Realisation

Before applying the linguistic realisation, there is a need to specify an abstract representation of the provided set of lexicons. According to [185], *messages* are the abstractions that mediate between the set of lexicons and eventual text. Here, based on the annotation and characterisation lexicons, two types of messages can be defined as: *AnnotationMsg* and *CharacterisationMsg*. *Phrase specification* is a structure to specify the elements of a *message* for a single sentence, which is the proper representation of the output for microplanning and the input for realisation. Various levels of abstraction have been proposed for phrase specification such as *Syntactic structure*, *Canned text*, *Case frame*, etc. [185]. Since the aim was to describe an observation with its attribute labels, the complexity of the final sentence is limited to a reasonably straightforward template of abstract representation. Here the *syntactic structure* is employed, which describes the linguistic elements to be used as uninflected words, plus a set of features to determine how to realise the final text. Figure 4.6 illustrates the simple template for the final text in an *attribute value matrix* (AVM) format [185, 237].

⁵The linguistic name of a quality dimension $\text{label}(q)$ is H_q , as defined in Definition 3.2, but this name can be also added to the linguistic label of the fuzzy intervals on quality dimension. As an example, for two dimensions 'time' and 'length', the linguistic labels of the first intervals in both can be 'short', but to be precise, the labels can consist of the dimension names to be defined as 'short time' and 'short length', respectively.

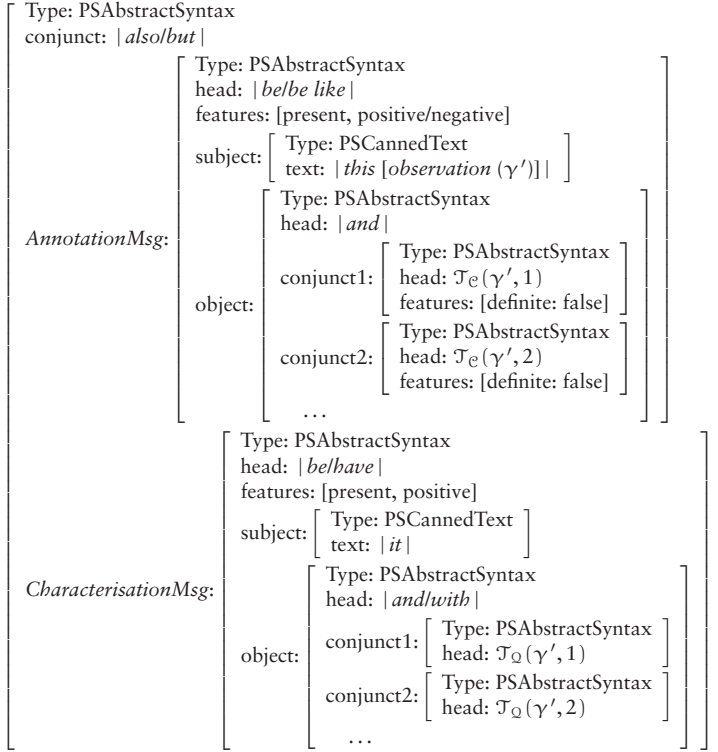


Figure 4.6: An abstract representation for the annotation and characterisation messages in the AVM format.

Linguistic realisation is the process of applying a set of rules to abstract representations of the lexical items in order to specify well-formed sentences in natural language, which are syntactically and morphologically correct. More specifically, the realisation maps the acquired abstract phrase specifications into the surface text [185]. Some of the linguistic realisations represent sentences by *template-like structures* when only limited syntactic variability is needed in the output description [184]. One instance of an output format for the sentences to describe the observations is as follows:

“This [Obs.] [be/be not/be like] [a *con.label1*] [and [a *con.label2*] and ...], [but/also] it [be/have] [*dim.label1*] [and/with [*dim.label2*] and/with ...].”

As this descriptive sentence is linguistically formed in a simple format, applying any realisation technique (e.g., *SimpleNLG* engine) will produce grammatically correct sentences as the output text. The standard architecture of

NLG systems provides a well-defined realisation for the abstracted representations, which the details of the architecture can be found in [185] and the implementation details are available in [95].

Example 4.14. *Consider the set of annotations and characterisations from Examples 4.11 and 4.13. Then, the output of realisation will be a message like: “This unknown leaf observation is like **Nerium** leaves and somewhat **Tilia** leaves, but it is an **elongated** and a **lance-shaped** leaf.”*

4.3 Discussion

This chapter has presented the utility of extending conceptual spaces as semantic *inference* models to generate linguistic descriptions for unknown observations in natural language. It has been shown that linking a symbol space to the conceptual space facilitates the process of inferring linguistic characterisations for the unknown observations. One advantage of this inference model is that the generated descriptions include adequate set of interpretable features that are derived from the inclusion of unknown observations within the conceptual space. following, a number of issues related to the semantic inference approach are discussed.

Lexicalisation: On the inference process, one point related to the lexicalisation task is to determine the most descriptive terms to use in order to linguistically represent a new observation. As seen in the Leaf example, the most obvious psychological description could be regarding a leaf’s similarity or dissimilarity to the known concepts. Thus, either the linguistic labels of the known concepts or the linguistic terms of quality dimensions which are stored in symbol space can be used. So, the semantic interpretability of such labels will affect directly on the descriptive quality of the natural output text.

Concepts as nouns or adjectives: One point related to the lexicalisation and realisation is that from the natural language point of view, the concepts in a conceptual space typically represent the *nouns* while the sub-concepts or properties are related to the *adjectives*. The approach to determine linguistic descriptions does not make the distinction between adjectives and nouns, and consequently, between the properties and concepts for the class labels from the input learning data set⁶. Preferably, one region within a domain is considered as a concept or sub-concept that can be either nouns or adjectives, and be used to

⁶The reason is that from machine learning point of view, there is no linguistic distinction between the different types of the classes, in a sense that the learning problem is to classify whether the *noun* concepts, *adjective* concepts, or even *verb* concepts. For example, in *Iris* data set the classes are nouns, as each class refers to a name of iris plant [146], however, in *Wine quality* data set the classes are adjectives, as each class refers to the quality level of wines from excellent to poor quality [64].

describe an observation. For example, descriptions of an object in a conceptual space of *fruits* can be either “*the object is an apple*” or “*the object is red*”. In the former, the label refers to the concept *apple* as a noun, and in the latter one, the label refers to the sub-concept *red* as an adjective. It is assumed that all the linguistic terms of the quality dimensions are considered as the adjectives. For instance, if *weight* is involved in the description of an object, the term *heavy* is considered as the adjective in such descriptions like “*the object is heavy*”.

In sum, this chapter together with chapter 3 has shown the process of creating data-driven conceptual spaces and inferring linguistic description from such spaces. The presented approaches in these two chapters will be exemplified in chapter 5 to show the goodness of the developed semantic representation for real-world data samples.

Chapter 5

Results and Evaluation: A Case Study on Leaf Data Set

“It doesn’t matter how beautiful your theory is, it doesn’t matter how smart you are. If it doesn’t agree with experiment, it’s wrong.”

— Richard Feynman (1918–1988)

THIS chapter presents an assessment of the formal methods described in the chapters 3 and 4. A case study from a data set of leaves is investigated to show the feasibility and the plausibility of the proposed approach in real-world applications. This chapter also describes an empirical evaluation that is designed to assess the goodness of the developed methods for describing unknown observations using the conceptual spaces. Finally, the results of this empirical assessment for the leaf data set are presented.

The leaf data set [206] is a set of photographed leaf samples from different plant species. Here, six species are selected as the labelled data set (See Figure 5.1), and the rest of the species are used as unlabelled data. The leaf data set is a good first example, as it provides a tangible example of physical objects while the vocabulary used to describe the leaves is not necessarily familiar to non-specialists. For the case study on the leaf data set, the primitive set of features is initialised by expert-oriented questionnaires or domain-oriented background knowledge. These semantically interpretable features are describable in natural language and are able to distinguish the known classes from each other perceptually.

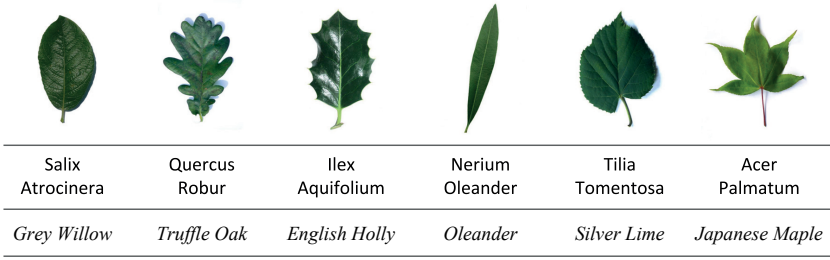


Figure 5.1: Six species as the known leaves in the leaf data set. The first row of labels shows the scientific names [206], and the second row of labels shows the common names [231].

5.1 Constructing a Conceptual Space of Leaves

The leaf data set includes 72 known leaf samples that are categorised in six species. Formally, $\mathcal{D}^l = \{o_1, \dots, o_{72}\}$ and $\mathcal{Y}^l = \{y_{sa} : \text{'Salix Atrocinera'}, y_{qr} : \text{'Quercus Robur'}, y_{ia} : \text{'Ilex Aquifolium'}, y_{no} : \text{'Nerium Oleander'}, y_{tt} : \text{'Tilia Tomentosa'}, y_{ap} : \text{'Acer Palmatum'}\}$. Figure 5.1 shows the prototypical samples of leaf species for the leaf labels in \mathcal{Y} , along with their popular names.¹ According to the set of class labels \mathcal{Y}^l , the set of concepts is defined as $\mathcal{C}^l = \{C_{ia}, C_{tt}, C_{no}, C_{qr}, C_{ap}, C_{sa}\}$. Here, the concepts are initiated, but the representation of each concept will be formally presented later.

The first step to build a conceptual space of leaves is to specify the initial set of features that characterises the leaves. The primary criterion while initialising the features is how descriptive or interpretable the chosen features are in the linguistic form. In other words, this approach is looking for such features that are representable in natural language with a perceptual interpretation. For example, the values of *area* and *perimeter* features might be useful for statistical analysis or classification tasks, but these features do not carry meaningful information to describe and distinguish the leaf observations. In contrast, a feature like *elongation* meaningfully describes a perceptual feature of a leaf observation.

Besides the leaf samples in different species, Silva et al. [206] have provided a set of attributes that describe the shape and texture features of leaves. Among the semantic attributes, the following features and used in the model as the initial set of features $\mathcal{F}^l = \{X_i : \langle H_{X_i}, I_{X_i} \rangle\}$:

¹Note that in the model, the scientific names of the leaves have been applied as labels used in the original data set [206]. However, in the final descriptions for the evaluation, the common names of leaves are used [231] which were more familiar to the general users.

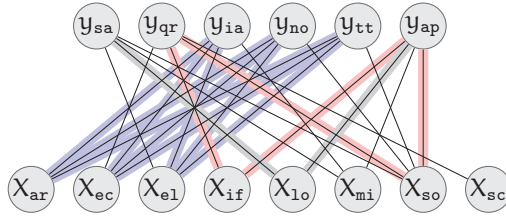


Figure 5.2: The bipartite graph presenting the relevance of features and labels in leaf data set. Also, three chosen bicliques (as the domains) are highlighted with blue, red, and grey edges.

X_{ec} : ‘Eccentricity’, $[0, 1]$ (eccentricity of the ellipse),
 X_{ar} : ‘Aspect Ratio’, $[1, \text{inf}]$ (values close to 1 indicate an elongated shape),
 X_{el} : ‘Elongation’, $[0, 1]$ (minimum is achieved for a circular region),
 X_{so} : ‘Solidity’, $(0, 1]$, (how well the leaf fits a convex shape),
 X_{if} : ‘Isoperimetric Factor’, $[1, \text{inf}]$ (curvy intertwined contours yield low values),
 X_{lo} : ‘Lobedness’, $(0, \text{inf}]$ (characterises how lobed a leaf is),
 X_{mi} : ‘Maximal Indentation Depth’, $(0, 1]$, (how deep are the indentations),
 X_{sc} : ‘Stochastic Convexity’, $[0, 1]$, (probability of a random segment in a leaf to be fully contained).

5.1.1 Domain Specification for Leaf Data set

The values of these features for every observation are acquired from [206]. After that, the construction of the conceptual space is performed with the inputs of the labelled observations \mathcal{D}^l , label set \mathcal{Y}^l , and feature set \mathcal{F}^l . The algorithm first applies the feature filtering approach, i.e. MIFS (Algorithm 3.1) to provide a ranking matrix which shows the mutual relations of features and labels. Then, using Algorithm 3.2, a feature subset grouping is performed. Figure 5.2 illustrates the created bipartite graph, which leads to determining the domains and quality dimensions.

The chosen bicliques (with the highest scores) determine the three domains $\Delta^l = \{\delta_1, \delta_2, \delta_3\}$, where each domain is specified as follows:

- Domain $\delta_1 = \langle \mathcal{Q}(\delta_1), \mathcal{C}(\delta_1), \omega_{\delta_1} \rangle$, wherein
 $\mathcal{Q}(\delta_1) = \{q_{ar}, q_{el}, q_{ec}\}$,
 $\mathcal{C}(\delta_1) = \{c_{ia}, c_{tt}, c_{no}\}$.
- Domain $\delta_2 = \langle \mathcal{Q}(\delta_2), \mathcal{C}(\delta_2), \omega_{\delta_2} \rangle$, wherein
 $\mathcal{Q}(\delta_2) = \{q_{so}, q_{if}\}$,
 $\mathcal{C}(\delta_2) = \{c_{qr}, c_{ap}\}$.

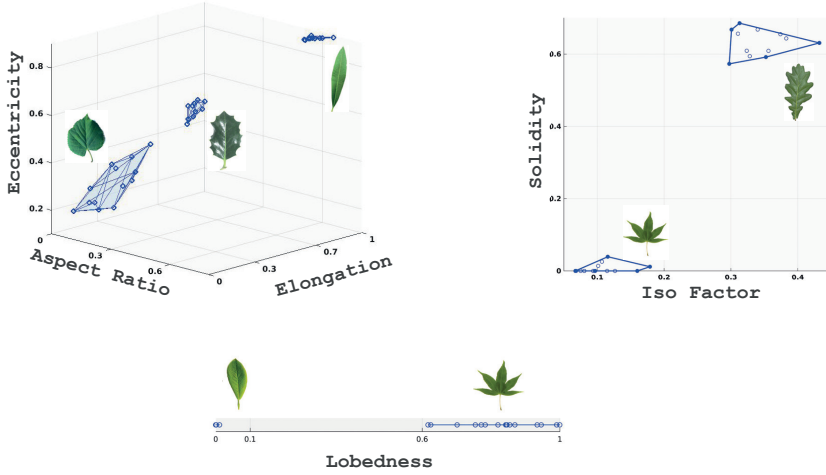


Figure 5.3: The conceptual space of leaf data set: a graphical presentation of the determined domains with the corresponding quality dimensions and concepts.

- Domain $\delta_3 = \langle Q(\delta_3), C(\delta_3), \omega_{\delta_3} \rangle$, wherein
 $Q(\delta_3) = \{q_{1o}\}$,
 $C(\delta_3) = \{C_{sa}, C_{ap}\}$.

Figure 5.3 depicts a graphical presentation of the determined domains with the corresponding quality dimensions and concepts. As an example, domain δ_2 is specified by two quality dimensions ‘solidity’ and ‘isoperimetric factor’, and is associated with two concepts ‘Quercus’ and ‘Acer’. An example of calculated weights in a domain is $\omega_{\delta_2}(C_{ap}, q_{so}) = 0.61$,² which shows the salience of the relation between leaf concept ‘Acer’ and quality dimension ‘solidity’ within δ_2 . Although the process of deriving the domains is data-driven, there may be an interpretation of each specified domain. For instance, one can say that δ_1 illustrates the *convexity* of the known leaves, while δ_2 shows the *indentation* of the known leaves (see Figure 5.3).

As an output of the domain specification phase for the conceptual space of leaves, the set of quality dimensions is $\mathcal{Q}^l = \{q_{ar}, q_{el}, q_{ec}, q_{so}, q_{if}, q_{1o}\}$, and the set of instances is $\Gamma^l = \cup_{y \in \mathcal{Y}} \Gamma(y)$, where $|\Gamma^l| = |\mathcal{D}^l|$. Each $\gamma \in \Gamma^l$ corresponds to a known leaf sample $o \in \mathcal{D}^l$, and consists of three points (one in each domain).

²It is notable that the values of weights, calculating with filter method (MIFS), are not interpretable individually. However, they are involved in the model, since they are helpful to compare and measure the similarities between the concepts and instances.

5.1.2 Concept Representation for Leaf Concepts

According to the output of concept representation, each leaf concept in \mathcal{C}^l appears in only one domain and thus has exactly one sub-concept, except concept C_{ap} which has two sub-concepts in two different domains. Using Algorithm 3.3, the elements of the sub-concepts for each known leaf concept in \mathcal{C} is calculated as follows.

- Leaf concepts ‘Ilex’, ‘Tilia’, and ‘Nerium’ are represented in δ_1 as, respectively:
 $C_{ia} = \{c_{ia}^1 : \langle \eta_{ia}^1, \phi_{ia}^1 \rangle\},$
 $C_{tt} = \{c_{tt}^1 : \langle \eta_{tt}^1, \phi_{tt}^1 \rangle\},$
 $C_{no} = \{c_{no}^1 : \langle \eta_{no}^1, \phi_{no}^1 \rangle\}.$
- Leaf concept ‘Acer’ is represented in two domains δ_2 and δ_3 as:
 $C_{ap} = \{c_{ap}^2 : \langle \eta_{ap}^2, \phi_{ap}^2 \rangle, c_{ap}^3 : \langle \eta_{ap}^3, \phi_{ap}^3 \rangle\}.$
- Leaf concept ‘Quercus’ is represented in δ_2 as:
 $C_{qr} = \{c_{qr}^2 : \langle \eta_{qr}^2, \phi_{qr}^2 \rangle\}.$
- Leaf concept ‘Salix’ is represented in δ_3 as:
 $C_{sa} = \{c_{sa}^3 : \langle \eta_{sa}^1, \phi_{sa}^1 \rangle\}.$

In these representations, for example, η_{qr}^2 shows the 2D convex polygon of leaf concept ‘Quercus’ within δ_2 (see Figure 5.3). Also, as an example for the weights, $\phi_{qr}^2 = \{\omega_{\delta_2}(C_{qr}, q_{so}), \omega_{\delta_2}(C_{qr}, q_{if})\}$ shows the salience between leaf concept ‘Quercus’ and two quality dimensions ‘solidity’ and ‘isoperimetric factor’ within δ_2 . In Figure 5.3, the graphical presentation of leaf concepts is shown by illustrating the convex hulls of their corresponding sub-concepts.

Now, with the provided elements, the conceptual space of the leaf data set is presented as: $S^{leaf} = \langle \mathcal{Q}^l, \Delta^l, \mathcal{C}^l, \Gamma^l \rangle.$

5.2 Semantic Inference for Unknown Leaf Samples

Inference step aims to derive a semantic description for unknown observations using the developed conceptual space. The utility of the conceptual space of leaves is presented using a set of unknown leaf samples from the plant species. Figure 5.4 presents the selected set of unknown leaf samples to be represented. Here, It is shown how to apply the inference approach on one the samples (e.g., leaf (a) in Figure 5.4). According to the inference process, an unknown observation like (a) is firstly vectorised to an instance γ_a . Then a linguistic description for a is inferred in two phases: setting symbol vectors by inferring in conceptual space and setting the lexical items by inferring in symbol space. Instance γ_a is a set of points within the domains $\Delta^l(\gamma_a)$ denoted by $\gamma_a = \{p_{\gamma_a}^1, p_{\gamma_a}^2, p_{\gamma_a}^3\}$, where the numeric values of each point are

the feature values of (a) for each quality dimension in S^{leaf} . For example in δ_2 , $p_{\gamma_a}^2 = (q_{so}(a), q_{if}(a)) = (0.86, 0.45)$.

5.2.1 Inference in Conceptual Space of Leaves

Here, it is determined whether γ_a is included in any defined concept's regions, and then infer the semantic labels based on the inclusion of the instance to the regions. Considering the leaf sample (a) in Figure 5.4, γ_a belongs to the sub-concept c_{tt}^1 in δ_1 , however, it does not belong to any sub-concept in δ_2 and δ_3 . Thus, based on Algorithm 4.1, the symbol vector for γ_a is set as follows: In the concept layer, using the graded membership function (defined in Definition 4.2): $\mathcal{V}_{\gamma_a, c}(C_{tt}) = ('Tilia Tomentosa', 0.85)$. In the quality layer, for the quality dimensions of δ_2 and δ_3 , using the graded quality function (defined in Definition 4.3): $\mathcal{V}_{\gamma_a, \Omega}(q_{if}^2) = ('tipped/toothed', 0.75)$, $\mathcal{V}_{\gamma_a, \Omega}(q_{so}^2) = ('smooth edges/entire', 0.86)$, and $\mathcal{V}_{\gamma_a, \Omega}(q_{lo}^3) = ('low lobedness', 0.6)$.

5.2.2 Inference in Symbol Space of Leaves

By retrieving the information of the symbol vector $\mathcal{V}(\gamma_a)$, it is possible to verbalise the elements of symbol vectors into a set of natural language descriptions. As mentioned in Section 4.2.2, γ_a is annotated using the values of $\mathcal{V}_{\gamma_a, c}$, and characterised by the values of $\mathcal{V}_{\gamma_a, \Omega}$. In particular, the annotation set is $\mathcal{T}_c(\gamma_a) = \{'Tilia Tomentosa'\}$, and the characterisation set will be $\mathcal{T}_\Omega(\gamma_a) = \{'tipped', 'smooth edges', 'low lobedness'\}$. Then the realisation for γ_a is as follows: $\mathcal{T}_{\gamma_a} = \text{'like Tilia Tomentosa(Silver Lime), also with smooth and tipped edges, and low lobedness'}$.

The same approach is applicable for other unknown leaf samples (shown in Figure 5.4) to describe them in natural language. Table 5.1 presents the generated descriptions derived from the semantic inference.

5.3 Empirical Evaluation: Design and Procedure for Leaf Samples

Finding a direct solution for assessing the applicability of the proposed conceptual space representation is not a trivial problem. Instead, the usefulness of the constructed conceptual spaces for the case study of leaves is evaluated via the linguistic descriptions derived from such spaces. The experiment evaluates the following aims: (1) to measure the *feasibility* of deriving *accurate* descriptions to *distinguish* unknown observations, and (2) to assess the *goodness* of the descriptions derived from conceptual spaces in comparison to the descriptions derived from other base-line models. To these ends, a survey has been conducted in which the participants were asked to

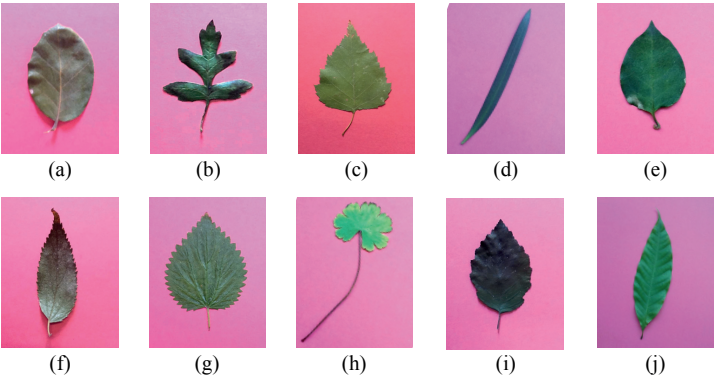


Figure 5.4: A set of unknown leaf samples.

Leaves	Linguistic Description
Fig. 5.4(a)	<i>This leaf is like Grey Willow, but it is round with a slightly serrated margin.</i>
Fig. 5.4(b)	<i>This leaf is like Japanese Maple, but it is oval with lobed margin.</i>
Fig. 5.4(c)	<i>This leaf is like Silver Lime, but it is tipped with a slightly toothed margin.</i>
Fig. 5.4(d)	<i>This leaf is not like any known leaf species, but it is linear and elongated with entire margin.</i>
Fig. 5.4(e)	<i>This leaf is like Grey Willow, but it is round and tipped.</i>
Fig. 5.4(f)	<i>This leaf is not like any known leaf species, but it is oval and tipped with toothed margin.</i>
Fig. 5.4(g)	<i>This leaf is like Silver Lime, also it is tipped with low lobed and toothed margin.</i>
Fig. 5.4(h)	<i>This leaf is not like any known leaf species, but it is round with toothed and lobed margin.</i>
Fig. 5.4(i)	<i>This leaf is like English Holly, but it is tipped with serrated margin.</i>
Fig. 5.4(j)	<i>This leaf is like Oleander, and it is also tipped.</i>

Table 5.1: The linguistic descriptions derived for the unknown leaf samples in Figure 5.4.

1. identify specific leaf based on their linguistic description derived from the conceptual space, and
2. rate the goodness of descriptions produced by different models on a Likert scale.

5.3.1 Survey: Design and Procedure

The survey was conducted online³. After an introduction to the used vocabulary, the participants self-evaluated their familiarity with the terminology. The main body of the survey was composed of two parts.

The first part is designed as a set of 4 *multiple choices* questions wherein the participants have been asked to read the conceptual space description of a randomly selected sample (among 15 leaves) and to choose the corresponding image of the leaf from four shown options. The three incorrect options were also randomly selected from a pool of unknown examples. This task-based (or extrinsic) evaluation [183] allowed evaluation process not only to measure how well the participants can connect a description of an unknown sample to its correct image, but to investigate the incorrectly identified examples and their relation in the conceptual space.


For the second part, a set of *rating scale* questions has been designed, again using four questions per participant. In each question, an image of one unknown observation is randomly selected and shown to the participants, along with the three generated descriptions for that observation from three different models. Participants then are asked to rate each text from 1 to 5 (Likert-scale scoring) concerning how well each of the descriptions helps them to refer to the image. This simultaneous human-rating (or intrinsic) evaluation [183] of descriptions enables the evaluation process to compare the approaches relative to each other, as well as to evaluate the absolute goodness of each approach.


In total 207 responses have been received⁴, out of which 102 valid responses have been considered for the *leaf* data set. The survey was publicly distributed online to anyone who was interested in participating. However, the outcome for both data sets showed that most of the participants were in the range of 25-44 years old and they are mostly educated in computer science or equivalent. Besides, most of the participants were fluent in English speaking. About the expertise level of the participants, The participants were not quite familiar with the terminology that has been used for the *leaf* data. From the responses, 20% of the participants knew none of the lexical items, 70% knew few or some of them, and just 10% knew almost all of them.


³The survey can be accessed at <https://survey.bana.ee>


⁴The exact number of individuals is unknown since each participant could decide to perform one of the data sets each time or even redo it with a new set of random questions.

Select the leaf which is like **Grey Willow**, but it is round with a slightly serrated margin.

☐ 

☐ 

☐ 

☐ 

PREV 1 / 4 NEXT

(a) A multi choice question

How well do the following descriptions help you refer to the leaf in the image?

not at all

This leaf has some similarity to **English Holly**, also to **Grey Willow**. ☐ ☐ ☐ ☐ ☐

This leaf is round, wide, connected and entire, with no indentation and smooth margin. ☐ ☐ ☐ ☐ ☐

This leaf is like **Grey Willow**, but it is round with a slightly serrated margin. ☐ ☐ ☐ ☐ ☐

How would you describe the leaf in this image? (optional)

Your description for this leaf (optional)

very well

PREV 1 / 4 NEXT

(b) A rating scale question

Figure 5.5: The screenshots from two types of questions designed for the survey: (a) An example of *multi choice* question that shows a generated description, and it asks the participants to identify which given image of leaf (from the choices) is related to the description, (b) An example of *rating scale* question that shows three different descriptions generated for a single leaf image, and it asks the participant to rate the goodness of each given descriptions.

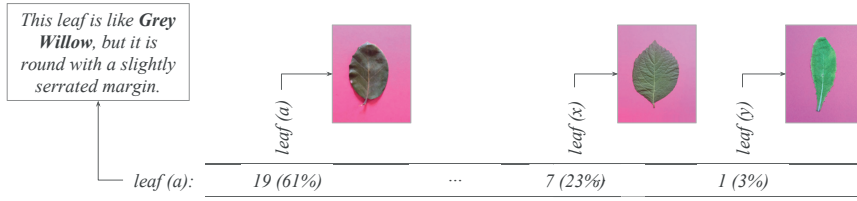


Figure 5.6: The description of *leaf (a)* has been shown 31 times to the participants. In 19 responses, the correct image of *leaf (a)* is chosen by the participants (61%). The most common misidentified example (7 responses, 23%) was the image of *leaf (x)*, which subjectively is quite similar, and interestingly, it is the closest instance to the *leaf (a)* in the conceptual space. However, the closest instance to *leaf (a)* in the full feature space is *leaf (y)* that its image is rarely misidentified by the participants (only 1 response, 3%).

5.3.2 Identifying Leaf Observations from Linguistic Descriptions

Participants were able to successfully identify all the unknown observations (15 leaves) with the help of the corresponding conceptual space descriptions. The success rate to identify the correct image for each description in the *leaf* data set was $73\% \pm 13\%$.

For further investigation of the incorrectly identified (i.e., misidentified) examples, the geometrical similarity of these answers was calculated in order to the correct one in the conceptual space (multi-domain). According to [86], the similarity in conceptual spaces can be calculated by applying *Euclidean* distance with the domains and *city-block* distance between them. To assess of the similarities in conceptual space, the geometrical similarity of the same instances is also calculated, but in a full feature space (single-domain) by applying *Euclidean* distance. Two interesting results have been obtained: First, the misidentified examples are not uniformly distributed between all possible choices, but instead, participants tended to make similar mistakes. Second, the common misidentified examples are most of the times (73% leaves) the closest instance to the correct one in the conceptual space. In the full feature space, this was only occasionally true (46% leaves). This shows that the confused examples with each other are commonly the nearest instances within the multi-domain conceptual space, which is mostly not true in the full feature space. One example regarding this outcome is illustrated in Figure 5.6.

The results from the first part of the survey show that the proposed conceptual space representation a) is applicable to derive semantic descriptions for unknown observations, and b) is suitable to represent the cognitively similar observations among the multiple domains.

5.3.3 Rating Various Linguistic Descriptions of a Leaf Observation

In the results of the rating scale questions, the description derived from the conceptual space model is compared with the descriptions derived from the two models of concept formation within the full feature space, one using a generative model and the other a discriminative model [122, 162]. The *generative* model forms the concepts by modelling the distribution of individual classes (i.e., bound each of them with a convex region). Then a new observation either belongs to an existing class or none of them. On the other hand, the *discriminative* model forms the concepts by learning the (hard or soft) boundary between classes (i.e., divide them into Voronoi regions). Hence, a new observation always belongs to at least one class label.

Concerning these two models of concept formation, two base-line approaches have been developed to generate linguistic descriptions. Inspired by the idea of *fuzzy sets* for the *linguistic description of data* [177, 180], the generative and discriminative models have been extended to quantify the inclusion of new observations within the full feature space. This will allow the model to verbalise the numeric output of the models with linguistic descriptions. In a generative model, fuzzy sets are employed to quantify the numeric values of the features within the full-feature space. the same semantic inference algorithm is applied (Chapter 4) on this model to derive such descriptions that most likely involves only the quantified terms of the characteristic features. In a discriminative model, with the help of fuzzy sets, the model is extended to multi-label classification [211], which quantifies the membership of the instances to the concepts. The inference algorithm is applied to this model to derive descriptions that involve only the assigned labels of concepts with a quantification of their membership degrees.

Example 5.1. For leaf (a) in Figure 5.4, here are the output descriptions from three various approaches:

- *Conceptual*: “This leaf is like **Grey Willow**, but it is round with slightly serrated margin.”
- *Generative*: “This leaf is round, wide, connected and entire, with smooth margin and no indentation.”
- *Discriminative*: “This leaf is similar to **English Holly**, also has some similarity to **Grey Willow**.”

Table 5.2 shows the statistical summary of the rating scores received for the descriptions derived from each of the approaches in *leaf* data set. Also, these scores are depicted in the form of boxplots in Figure 5.6.

An ANOVA test has been applied to show that the conceptual space description (*Conceptual*) is significantly preferable rated than the two alternatives (*Generative* and *Discriminative*). Here, the Likert-scale scores were used

Table 5.2: The overall scores calculated from rating responses to the different models in *leaf* data set. The numbers show average scores (and standard deviations) in the range of 1 to 5.

	Mean (SD)
<i>Conceptual</i>	3.62 (1.19)
<i>Generative</i>	3.27 (1.33)
<i>Discriminative</i>	2.72 (1.24)

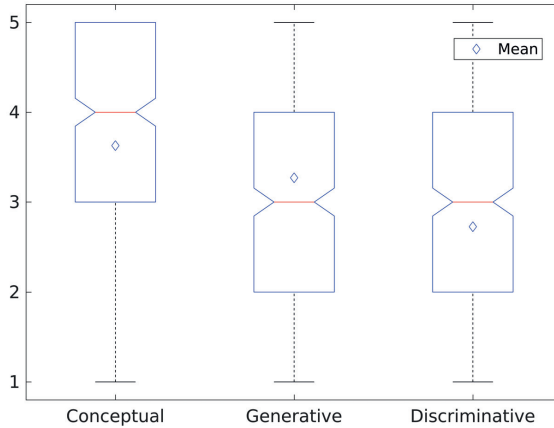


Figure 5.7: The box plot of the rating scores received for each of the models, deriving descriptions in *leaf* data set.

as the dependent variable, and the models were used as the independent variables (groups). Then, the Post-hoc Tukey test was employed to identify the significant difference between the models. The one-way ANOVA test showed a significant effect of the models on the scores. For *leaf* data set, *Conceptual* has the mean significantly different from *Generative* and *Discriminative*, $p < .0001$ (two-tailed). The details of the test has been shown in Table 5.3.

Moreover, since the ratings are ordinal, also a non-parametric test (i.e., *Wilcoxon Test*) has been carried out to identify the significant differences between ratings by comparing each pair of the scores. Table 5.3 shows the p-values of this method for each pair of models. The output showed that *Conceptual* model is significantly different from *Generative* and *Discriminative* models ($p < .0001$).

Overall, the results from this part of the survey show that the proposed conceptual space representation a) is an appropriate semantic inference model

Table 5.3: Summary of the one-way ANOVA and Wilcoxon tests for the rating scores with respect to the models deriving descriptions.

ANOVA Test	<i>Conceptual vs. Generative & Discriminative</i>	leaf data set
		$F(2, 1221)=52.82,$ $p<10^{-21}$
Wilcoxon Test	<i>Conceptual vs. Generative</i>	$p<10^{-4}$
	<i>Conceptual vs. Discriminative</i>	$p<10^{-23}$
	<i>Generative vs. Discriminative</i>	$p<10^{-07}$

to derive linguistic descriptions for unknown observations, and b) successfully derives descriptions (from multi-domain space) that are naturally preferred by participants, in comparison to the other alternative models (from single-domain space).

5.4 Discussion

In this chapter, the feasibility of the approach has been demonstrated in a case study of the real-world numerical data set. But it is notable that the framework proposed in this work is introduced for general use in any numeric data sets wherein the known features and categories are available, but the perceptual domains and the concept formation need to be determined. By performing an empirical evaluation, it has been assessed how well linguistic descriptions that are generated based on the derived conceptual space enable human users first to identify the unknown observations. This followed by a comparison between different semantic models of generating linguistic descriptions to show how well human users prefer the descriptions from the conceptual space model to refer to unknown observations. Following, a number of issues related to the inference approach are discussed.

Multi-domain representation: The evaluation results indicate that a multi-domain representation of concepts (i.e., conceptual spaces) can lead to a better presentation of output descriptions in comparison to a single-domain representation, since the multi-domain spaces preserve the various semantic aspects of the attributes for a concept, while the others combine all the attributes into a single space.

Generalisation: While the proposed approach has been tested on two case studies to verify its plausibility, it would be necessary to identify a general class of problems that the proposed approach can address. Many AI problems dealing with numeric data as the input of learning systems require semantic interpretation for these data, which is needed for interaction with humans. However, in most of the cases, there is no a priori or expert knowledge to explain the

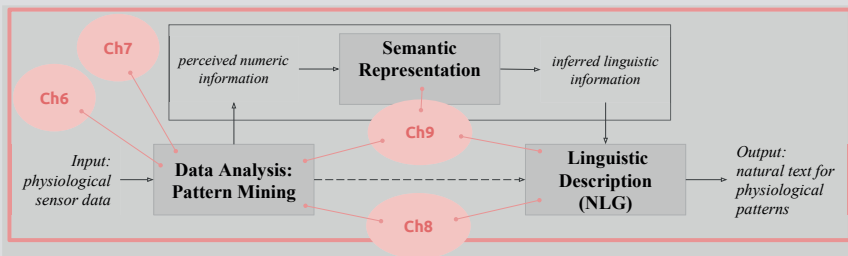
aspects of the input observations. This is more problematic when *connectionist* approaches are applied since they cannot explain what the learnt emerging model represents. Hence, the introduced framework to construct and utilise conceptual spaces is generally applicable for those AI applications wherein: (1) there is a need for *concept learning* and *concept description* only based on the known available observations, and (2) there is a lack of *interpretability* while creating a learning model, as well as the lack of *explainability* while testing the model by completely unknown observations.

In sum, Part I of this thesis is concluded with the investigation on how the semantic representations proposed in Chapter 3 and Chapter 4 can be applied to a real-world data set, and the utility of this representation for describing new observations.

Part II

Physiological Sensor Data: From Data Analysis to Linguistic Descriptions

Part II of this thesis focuses on the full framework of describing the numerical data, specifically considering physiological sensor data. This part shows how to mine physiological patterns from sensor data, and how to use the semantic representations proposed in Part I to linguistically describe such patterns. First, the state of the art on data mining approaches for sensor data is discussed (Chapter 6). After mining unseen but interesting time series patterns (Chapter 7) and temporal rules between the patterns (Chapter 8), it is shown how the proposed semantic model in Part I can be applied to linguistically describe these patterns (Chapter 9).



Chapter 6

An Overview of Health Monitoring with Mining Physiological Sensor Data

“Without data you’re just another person with an opinion.”

— W. Edwards Deming (1900–1993)

As noted in chapter 1, a framework to deal with the problem of representing and describing numerical data needs first to acquire a proper set of raw data. Subsequently, the framework needs to turn this data into information that is suitable to be fed to a semantic representation or any other data to text system. The second part of the thesis presents the task of mining and analysing the raw measured physiological data in order to provide a set of valuable information for the semantic model. For this reason, before going through the proposed methods for analysing physiological data and extracting valuable information, this chapter presents an overview of current health monitoring systems within mining physiological sensor data. This overview gives a better understanding of the existing methods for analysing such data, along with investigating their strengths and weaknesses to be used for the goal of data explanation in natural language.

The past few years have witnessed an increase in the development of wearable sensors for health monitoring systems. With the increase of healthcare services in non-clinical environments using vital signs provided by wearable sensors, the need to mine and process physiological measurements is growing significantly. This increase has been due to several factors such as development in sensor technology as well as directed efforts on political and stakeholder lev-

els to promote projects which address the need for providing new methods for care given challenges with an ageing population. An important aspect in such system is how the data is treated and processed. This chapter provides a review of the recent methods and algorithms used to analyse data from wearable sensors, which are used for monitoring of physiological vital signs in healthcare services. More precisely, it outlines the common data mining tasks that have been recently applied for health monitoring, such as anomaly detection, prediction and decision making when considering continuous time series measurements. Moreover, the chapter details the suitability of particular data mining and machine learning methods used to process physiological data.

In health monitoring systems the focus has been recently shifting from the obtaining data to developing intelligent algorithms to perform a variety of tasks [23]. Such tasks not only include traditional pattern recognition and anomaly detection, but also consider decision-based systems which can handle context awareness, subject-specific models, and personalisation. As the literature in this field is vast, the scope of this thesis has been limited to only cover wearable sensors that measure health parameters such as vital signs for disease management and prevention. Specifically, this review is concentrated on the following vital sign parameters: *electrocardiogram* (ECG), *oxygen saturation* (SpO₂), *heart rate* (HR), *Photoplethysmography* (PPG), *blood glucose* (BG), *respiratory rate* (RR), and *blood pressure* (BP).

Works such as [51] and [27] have focused on the needs of involving wearable sensors and overcoming essential bottlenecks for the use of wearable sensors such as the clinical acceptability and interoperability in health records. Most of the recent review articles on data mining of physiological sensor data are related to general studies for healthcare, i.e., well-known problems in healthcare with simple and routine data mining approaches [159]. Recently, Sow [212] categorised the main challenges of sensor data mining in five following stages: acquisition, preprocessing, transformation, modelling and evaluation. The authors in [235] and [37] have used the data mining algorithms mainly in two categories 1) *descriptive* or unsupervised learning (i.e. clustering, association, summarisation) and 2) *predictive* or supervised learning (i.e. classification, regression). However, they are lacking deeper insight into the suitability of the algorithms for handling the special characteristics of the sensor data in health monitoring systems.

6.1 Data Mining Tasks in Health Monitoring Systems

Recently, the research area of health monitoring systems has shifted from the pure reasoning of wearable sensor readings (like calculating the sleep hours or the number of steps per day) to the higher level of data processing to give more valuable information to the end users. Therefore, healthcare services have

been focusing on more in-depth data mining tasks to be achieved. Based on the selected literature, three types of data mining tasks as the objectives of healthcare systems are predominant. These three tasks are: 1) *prediction*, 2) *anomaly detection* which include the subtask of raising *alarms*, and 3) *diagnosis* as a *decision making* process.

Figure 6.1 provides a depiction of each task concerning three dimensions or aspects. The first dimension involves the *setting* in which the health monitoring occurs (home/remote or clinical settings). Most monitoring applications which consider remote settings deal predominantly with *prediction* and *anomaly detection* tasks, whereas the applications in clinical settings are typically focused on *diagnosis* task [27, 216]. This fact is explained by the growing desire to have a more preventative approach (prediction) via wearable sensors and to consider the possibility to facilitate independent living in home environments by increasing the sense of security (alarm). Similarly, in clinical settings much more information is available to provide diagnosis and assist in decision making [37]. The second dimension in the figure shows the type of *subjects* that are under observation (healthy or patient). For patients with known disease and medical records, both *diagnosis* and precisely the possibility to raise *alarms* are the essential tasks. For health monitoring which typically include healthy individuals who want to ensure the maintenance of good health, *prediction* and *anomaly detection* are the considered tasks in the literature [158]. The final dimension depicted in the figure considers how the acquired data sets have been processed (online/offline). For all three mentioned tasks, the data sets have been addressed both online and offline manners. However, most of the alarm-related tasks are naturally investigated in the context of online and continuous monitoring [212].

Anomaly Detection

Anomaly detection is the task of identifying unusual patterns which do not conform to the expected behaviour of the data [52]. Detected unusual patterns in health parameters, especially for home monitoring systems, enables the clinicians to make accurate decisions in short time [53]. Anomaly detection techniques are often developed based on classification methods to distinguish the data set into the regular classes and the outliers [97]. For example, support vector machines [141], Markov models [243] and Wavelet analysis [100] are used in healthcare systems for anomaly detection. Most of the related works using the anomaly detection approach usually deal with short-term [118] and multivariate data sets [61] to characterise the entire the data to find discords. Some of the studies considered finding irregular patterns in vital signs time series such as abnormal episodes in ECG [61, 222] and SpO₂ [54], which mostly discover unusual temporal patterns in continuous data. In online healthcare systems, alarms as soon as detecting any anomaly in vital signs will be triggered to have an instant reaction. Such alarm system is designed for monitoring patients in

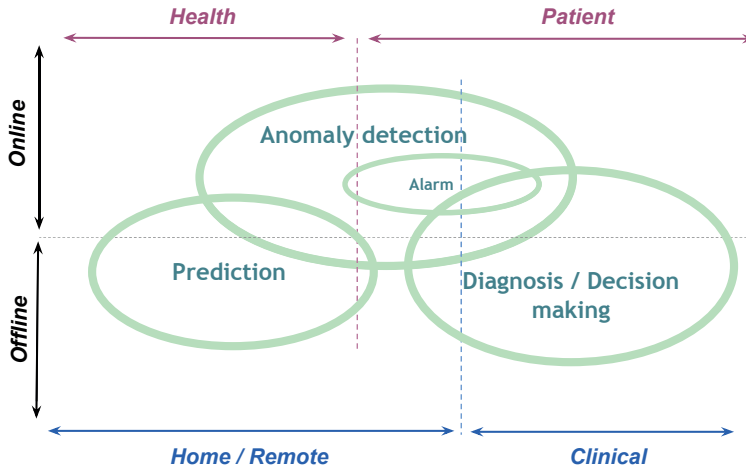


Figure 6.1: A schematic overview of the position of the main data mining tasks (anomaly detection, prediction, and diagnosis/decision making) concerning the different aspects of wearable sensing in the health monitoring systems.

clinical units [54]. However, anomaly detection task can be obtained by offline techniques in a sense to detect abnormal readings for subjects based on their historical measurements [243].

Prediction

Prediction task has been widely considered in data mining field that identifies the upcoming events based on the previously recorded information. This approach is getting more interesting for the healthcare providers in the medical domain since it prevents further chronic problems [37] and leads to make decisions about prognosis [38]. The role of the predictive data mining considering wearable sensors is non-trivial due to the requirement of modelling sequential patterns acquired from vital signs. This approach is also known as a supervised learning model [235]. As the typical examples of the predictive models, authors in [205, 218] presented a method which predicts the further stress levels of a subject. A similar case of using predictive models in healthcare are blood glucose level prediction [55], mortality prediction by clustering electronic health data [153], and a predictive decision making system for dialysis patients [234]. Recently, there are new studies concerning prediction tasks, which have used experimental wearable sensor data to perform in non-clinical settings [60, 97].

Diagnosis/Decision Making

Decision making in diagnosis is one of the main tasks of clinical monitoring systems which is often based on retrieved knowledge using vital signs, and also other information such as electronic health records and metadata [210]. Some examples of recent works on diagnosis/decision making tasks consist of: estimating the severity of health episodes of patients suffering chronic disease [40, 41, 101], sleep issues such as *polysomnography* and *apnea* [43, 232], estimation and classification of health conditions [169, 226], and emotion recognition [81]. Most of these studies have used online databases with annotated episodes to have sufficient and trustable real-world disease labels to evaluate the decision making process. Considering the complexity of the data to infer diagnosis, some researchers frequently used classification methods such as neural network [128] and decision trees [81] on short-term clinical data sets.

6.2 Data Mining in Health Monitoring Systems

In health monitoring systems, the role of data analysis is to extract information from the low-level sensor data and bridge them to the high-level knowledge representation. For this reason, recent health monitoring systems have given more attention to the data processing phase to catch more valuable information based on the expert user requirements. Data mining techniques that have been applied to wearable sensor data in health monitoring systems have varied, and it is also not uncommon that several techniques are used within the same architecture.

Regardless of the data mining technique used, the most standard and widely used approaches to mining information from sensor data sets are given in Figure 6.2. Acquiring the data sets as the input of the architecture is discussed in Section 6.3, and the data mining tasks as the goals of the architecture are presented in Section 6.1. The main steps of the data mining approach consist of 1) *data preprocessing*, 2) *feature extraction*, and 3) *modelling and learning* the information (considering expert knowledge and metadata) to perform the defined tasks.

It should be noted that parameters such as expert knowledge, historical data measurements, electronic health records, and stable parameters (e.g. sex, age) are essential in a data mining method. These parameters (metadata) provide contextual analysis and improve the process of knowledge extraction [100, 227]. One example is that every healthcare system, that uses HR sensor data, needs to investigate the effect of metadata such as age, sex, weight, and medicine in order to have meaningful reasoning in order to find i.e., basic abnormal heart rates or to personalise the critical pulses based on the mentioned metadata [98, 115].

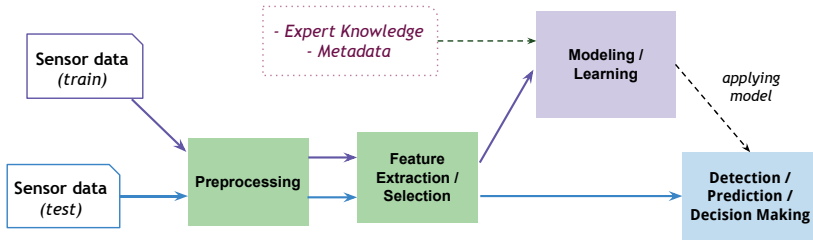


Figure 6.2: A generic architecture of the primary data mining approach for wearable sensor data.

6.2.1 Preprocessing

Due to the occurrence of noise, motion artefacts, and sensor errors in any wearable sensor networks, preprocessing of the raw data is necessary. Preprocessing in the healthcare domain involves 1) filter unusual data to remove artefacts [22, 152, 208], and 2) remove high frequency noise [81, 117, 137]. The main challenges of the preprocessing phase in healthcare systems are addressed in [212] which includes data formatting, data normalisation, and data synchronisation. Since the gathered sensor data is often unreliable and massive, studies working with large-scale and continuous data have necessarily employed a pre-processing step [101].

6.2.2 Feature Extraction/Selection

According to the magnitude and complexity of the raw data, feature extraction provides a representation of the sensor data which can formulate the relation of raw data with the expected knowledge for decision making [39]. Moreover, reducing the amount of sensor data is another task in feature extraction and selection phases which leads to having an arranged vector of features as an input of data mining techniques like classifier methods [101, 152].

Since Wearable sensor data that provide monitoring of vital sign parameters tend to be continuous time series readings, most of the considered features are related to the properties of time series signals [60]. Two main aspects of analysing signals are time domain and spectral domain [24]. In physiological data, the time-domain features are common, because the traditional decision making frameworks on vital signs have relied on the observable trends in the signal [54]. However, for extra knowledge about the periodic behaviour of time series, research in the medical field has given more attention to the features acquired from frequency-domains [100, 101]. Table 6.1 summarises the most appeared features in the literature that extracted from wearable sensor data.

	Time Domain	Spectral Domain	Other Features
ECG	mean R-R, std R-R, mean HR, std HR [218], number of R-R interval [141], mean R-R, std R-R interval [39]	spectral energy [117, 141], power spectral density [222], low-pass filter [101], low/high frequency [39, 218]	-
SpO ₂	mean, zero crossing counts, entropy [232], mean, slope [22], self-similarity [152]	energy, low frequency [152]	drift from normality range [22], entropy [152]
HR	mean, slope [22], mean, self-similarity, std [152]	energy, low/high frequency [152], low/high frequency [208], wavelet coefficients of data segments [101], low/high frequency, power spectral density [60]	Drift from normality range [22], Entropy, Co-occurrence coefficients [152]
PPG	rise times, max, min, mean [208]	low/high frequency [208]	-
BP	Mean, Slope [22]	-	rule based [13]
RR	max, min, mean [39].	-	residual and tidal volume [39].
Other Sensor Data	zero crossings count, peak value, rise time (EMG) [175], mean, duration (GSR) [208], pick value, min, max (SCR) [81], total magnitude, duration (GSR) [218].	spectral energy (EEG) [141], median and mean frequency, spectral energy (EMG) [175], energy (GSR) [208].	bandwidth, peaks count (GSR) [208].

Table 6.1: The summarisation of the most commonly used features of each wearable sensor data in the literature.

6.2.3 Modelling and Learning Methods

Appropriate data processing techniques are essential, in order to make sense of the data [212]. This section briefly outlines the most common data mining algorithms used for modelling and learning sensor data.

Neural networks *Neural Network* (NN) is an artificial intelligence approach which is widely used for classification and prediction [168]. Due to the predictive performance of NN, it is presently the most popular data modelling method used in the medical domain [38]. The NN is able to profoundly model nonlinear systems such as physiological records where the correlation of the input parameters is not easily detectable [21]. A wide range of the *diagnosis* and *decision making* tasks has been done by NN in the healthcare sys-

tems [143, 167]. Since the progress of learning in NN is to some extent complicated, the method is commonly used in clinical conditions with large and complicated data sets [101]. Also, as the modelling in NN is a black box progress, NN methods need to be adjusted for different input data [55].

Decision trees *Decision tree* is a vital learning technique which provides an efficient representation of rule classification [81]. The decision tree is a reliable technique to use in different areas of the medical domain in order to make a right decision [150, 173]. Nowadays, upon dealing with complex and noisy sensor data, the C4.5 algorithm is the used characteristic version of this method [234]. More attention has been given to decision trees in the medical domain when short-term data with few numbers of subjects have been used [40, 97, 100]. This method is also suitable for handling multivariate sensors due to the construction of independent levels in the decision tree [218].

Support vector machines *Support vector machine* (SVM) is a statistical learning method to classify unseen information by deriving a high dimensional hyperplane for the features in order to make a decision model [63]. Common health parameters considered by SVM methods are ECG, HR, and SpO₂ which are mostly used in the short-term and annotated form [117, 141, 222]. In general, SVM techniques are often proposed for *anomaly detection* and *decision making* tasks in healthcare services. However, SVM is not an appropriate method to integrate domain knowledge in order to use metadata or symbolic knowledge seamlessly with the sensor measurements [141].

Gaussian mixture models *Gaussian mixture model* (GMM) is a statistical approach that used for classification and pattern recognition [227]. Studies using GMM usually deal with annotated medical data in order to assess the performance of the model [61, 227]. Despite the fact that the GMM method can detect unseen information in physiological data, it has rarely been used for prediction tasks because the computation time of the constructing models is relatively high [101].

Other methods Out of the considered methods, there are other data mining techniques, which are roughly used in physiological data analysis. Some examples are: *Hidden Markov Models* for anomaly detection task [26, 243], *Bayesian networks* for prediction task [97, 143], *Rule-based reasoning* for anomaly (event) recognition [16, 127], *Fourier and wavelet transforms* for mostly feature selection [100, 128] and noise reduction in physiological data [74, 195], and finally *Association rule mining* for prediction and diagnosis tasks [114, 234]. More details can be found in a literature review article [28] that have been done by the author of this thesis.

6.3 Data Sets: Acquisition and Properties

In any health monitoring system, having a robust data processing stage requires adequate information about the data itself. Knowing about the way of collecting data and its properties while recording process will allow the data analysis system to perform the tasks such as selecting the proper data mining approach, designing new methods, and tuning the parameters. This section examines the types of data and the methods for acquiring data that have been used in different experimental validations in the literature. This information gives the opportunity to distinguish the data processing methods that are applied based on the type of sensor data.

6.3.1 Sensor Data Acquisition

Several input sources and data acquisition methods have been considered in the literature for wearable sensor data in health monitoring systems. Here, three major data gathering approaches have been identified:

Experimental wearable sensor data: Studies developed health monitoring systems have mostly used their own data gathering experiments to design, model and test the data analysis step [97,200,205]. In this case, the gathered data are usually obtained based on the predefined scenarios due to the test and evaluate the performed results [208]. But usually, these studies do not provide the precise annotations and meaningful labels on physiological signals.

Clinical or online databases of sensor data: Despite the attention of the literature is on the role of data mining on vital signs in health monitoring, several studies in this area have used the stored clinical data sets [61,195]. The most of the works used categorised and complex multivariate data sets with formal definitions and annotations by domain expert [101,118,153]. One common online database is the *PhysioNet* [4,105] that consists a wide range of physiological data sets with categorised and robust annotations for complex clinical signals. Several studies in the literature have used two main data sets in *PhysioNet* bank, *MIMIC* data sets (e.g. [22,152,161]) and *MIT* data sets (e.g. [74,141,170]) that contain the time series of patients vital signs obtained from hospital medical information systems.

Simulated sensor data: For the sake of having a comprehensive controlled analysis system, few works have designed and tested their data mining methods through simulated physiological data [226]. Data simulation would be useful when the focus of data processing method is on the efficiency and the robustness of the information extraction [227,243], rather than on handling real-world data including artefact and errors.

Another reason to create and use simulated data is the lack of long-term and large-scale data sets [243], where simulated data helps the proposed data mining systems to work with huge amount of data.

6.3.2 Sensor Data Properties

In addition to data acquisition methods, the following properties of wearable sensor data have been collected from the literature. These properties present which kind of data sets are usually used in healthcare systems.

Time horizon (long-term/short-term): The length of time for considering data set measurements is a particular challenge for wearable sensor data in order to orientate data mining techniques and the manner of data interpretation. Here, the time horizon of considered sensor data is categorised to short-term and long-term data. Some data analysis systems in healthcare were designed to process short signals such as a few minutes of ECG data [123, 170], a few hours of heart rate or oxygen saturation [22, 152] or the measurement of blood pressures over a day [13]. On the other hand, dealing with long-term data is a significant portion of some data mining methods for handling and processing. This period could be longer than a number of days or even a year of measurements. Blood glucose monitoring is an example of long-term data analysis for the sake of right decision making [55, 243].

Scale (large/small): A big challenge of data analysis in any health monitoring system is the examination of the proposed method on more than an individual. Depending on the design of sensor network, data gathering, and the goal of decision making, the scale of subjects in the frameworks would differ. Here, the studies considering 50 or more subjects (patient or healthy) are counted as large-scale studies [118], which can handle the same data modelling for large-scale of monitoring [61].

Labelling (annotated/unlabelled): Health monitoring systems need to evaluate their results to show the correctness of the decision making process. Due to having significant data analysis step, the attention of the most research is given to annotated data. By considering the behaviour of vital signs the domain expert is able to mark the data with several annotations such as arrhythmia disease [141], sleep discords [43], the severity of health [22], stress levels [208], and abnormal pulse in ECG [222]. These annotations also acquired using another source of knowledge like electronic health record (EHR), coronary syndromes, and also the history of vital signs [153]. On the other hand, working with unlabelled data leads to having unsupervised learning methods to extract unseen knowledge among raw sensor data. Some studies have been done on unlabelled

data to consider uncontrolled situations for especially *experimental data sets* [41, 200, 218].

Single Sensor/Multiple Sensors: Commonly, single sensor data have been used for specific analysis on individual physiological data such as ECG signal analysis [74, 170, 222] or blood glucose monitoring [55, 243]. Besides, some of the researchers have used several sensors [41, 118, 144, 153] to have global reasoning in health monitoring. Although using several wearable sensors in health monitoring frameworks are common, but few pieces of research have performed the multivariate data analysis to extract useful information through multi sensor data [118].

6.4 Discussion and Challenges

This chapter has presented an overview of the data mining approaches used for analysing the measured vital signs from the wearable sensing devices. For each approach, a reflection of its suitability for health monitoring was provided. From these reflections, the following guidelines for applying data mining methods were extracted:

- The selected data mining technique is highly dependent on which data mining task is in focus. According to the data mining tasks mentioned in Section 6.1, for *anomaly detection* task, SVM, HMM, statistical tools and frequency analysis are more commonly applied. *Prediction tasks* have often used decision tree methods as well as other supervised techniques. But rule-based methods, GMM, and frequency analysis have not been employed for prediction due to the shortcoming in modelling the data behaviours. Finally, any *decision making* task needs a modelling and inferring system with considering the contextual information. So, non-statistical methods like SVM, NN, and decision tree techniques have usually applied with success.
- The requirements for a real-time system should guide the selection of the data mining methods. To design a real-time health monitoring system, such methods like NN, GMM, and frequency analysis are not efficient for the sake of their computational complexities. But simple methods such as rule-based, decision tree and statistical techniques can quickly handle the online data processing requirements.
- The properties of the data set and experimental condition also influence the choice of method. Data mining methods (e.g., rule-based, decision tree) have been used in the clinical situation with the controlled conditions and clear data sets, but the efficiency of them are not tested in real experiments of healthcare services. In contrast, NN, HMM, and frequency techniques have been used to handle complex physiological data and discover the unexpected patterns in real-world situations.

- The level of supervision and labelled data is a crucial factor. Methods like SVM, NN, and GMM have been designed and justified to model long-term data. However, they could not deal with the unlabelled data to model the raw data in an unsupervised manner. For multivariate analysis of sensor data, the methods such as rule-based, decision tree, and statistical tools are more usable, while, e.g., GMM and HMM cannot play this role in healthcare systems.

While these guidelines can assist implementing technical systems to select appropriate methods for data analysis, the field is still challenged by some factors which have been discussed below. These are general challenges in the area of health monitoring and have been emerged from the literature studies investigated in this chapter. They include:

Need for large-scale monitoring: One challenge is still many applications using more massive data sets, and also still consider the monitoring task in the clinical contexts. This challenge will become more important for applications which examine target groups such as elderly, healthy persons etc. to make the significant effort in collecting reliable data sets for processing.

Dealing with annotated data sets: Data mining approaches gain increasing attention in this field, open data sets, as well as benchmark data sets, become essential to validate different approaches. Still however in this area, few benchmark data sets are available. The last mentioned point raises a second challenge of how data annotation (labelling) can be best done for such target groups. The process of annotating data is expensive, time-consuming and non-trivial considering long-term continuous data. To confront this challenge, an exciting avenue of study will be the efficacy of data mining in unsupervised contexts using unlabelled data sets. This type of data mining applies both to the modelling as well as eventual preprocessing of data, where for example, unsupervised feature learning techniques [136] for time series data could show promise.

Multiple measurements: Another challenge in this field is to exploit the multiple measurements of vital signs simultaneously. In particular, sensor fusion techniques which are able to consider dependencies and correlations between different vital sign parameters could assist in performing the primary data mining tasks of prediction, decision making and anomaly detection. Some attention to this issue has been given in the literature such as [118].

Contextual information: The usage of contextual information to assist in data mining is of ever increasing importance. Such contextual information could include meta information about subjects such as weight, height, age, sex, history of vital signs, as well as the history of previous decisions. It is also pos-

sible to automate the retrieval of high-level information via available ontologies [17, 132] and link this information to the data.

Discovering of unseen features: Still, important features of the data which may be unintuitive, e.g., frequency domain features may be needed for providing proper analysis and uncovering essential characteristics from the data which cannot be obtained by hand-engineered features. It is also worth noting that in real-world system such as home monitoring, it would be difficult to model the unexpected features with straightforward techniques.

Post-processing and representation: As a result of healthcare systems, an upcoming approach could use classical data mining techniques together with methods such as natural language generation which uncover trends in the data but also explain the process to both expert and non-expert users. Works such as [29, 119] have demonstrated the possible uses of such systems in both clinical and experimental contexts.

In sum, this chapter has provided an overview of the current trends and challenges of mining physiological sensor data within health monitoring systems. The chapter investigated 1) the main data mining tasks in health monitoring systems, 2) the dominant data mining approaches that have been used, and 3) various data types and their properties.

In relation to the rest of this thesis, this chapter has shown the need of considering new data analysis approaches to extract valuable information from the data beyond expert knowledge. This need of data-driven mining of sensor data is then performed in chapters 7 and 8, as the first step within the framework of describing physiological sensor data.

Chapter 7

Physiological Time Series Data: Preparation and Processing

“We are drowning in information but starved for knowledge.”

— John Naisbitt (1929–)

THIS chapter presents the first steps of data analysis to process the raw data of physiological sensors and exploit a set of meaningful and interesting information. This processing includes data collection/acquisition and mining in order to extract trends and patterns. This kind of information is then used for the tasks of creating semantic representations and linguistic descriptions for physiological sensor data (chapter 9).

With the increase of wearable sensor technology in both clinical and at home settings, the accumulation of physiological sensor data requires a concentrated effort on the analysis and modelling of this data [58]. Via sensor data analysis and modelling, it is possible to achieve a deeper understanding of the correlations between long-term measurements of physiological parameters and medical conditions. Typically, this process involves diverse data mining techniques on sensor data to acquire patient-specific models [28, 213]. In general, such approaches are either knowledge-driven or data-driven. Using a knowledge-driven approach leads to a supervised model of information extraction, but information is restricted to expert domain knowledge [235]. On the other hand, data-driven methods enable a system to discover hidden and potentially useful information through the physiological sensor data and to build models based on the experimental data [28]. In order to leverage from data-driven approaches, a solution whereby hidden patterns can be captured and made explicit in human consumable terms, i.e., semantics, is beneficial. Such an approach would not only facilitate automatic monitoring but contribute to



Figure 7.1: The wearable sensor, Bioharness3 [5], worn on the chest is able to locally store the measured data or wirelessly transmit it via Bluetooth.

a deepening and betterment of our knowledge in understanding the relationship between specific ailments and large-scale physiological time series and continuous data. This chapter aims to present some of data mining models to find such patterns in physiological sensor measurements.

7.1 Input Time Series Sensor Data: Collection and Acquisition

This section describes the process of collecting and acquiring sensor data in non-clinical and clinical conditions. In this thesis, the collected non-clinical data is used for the trend detection and trend descriptions. The acquired clinical data though is used wider for pattern abstraction, and later on in chapters 8 and 9 for linguistic description of such patterns.

7.1.1 Wearable Sensors, Non-clinical Data

The proposed approach here is designed to consider several continuous health parameters which are collected by wearable sensors or clinical records of physiological data. In this work, a wearable sensing device called Bioharness3 [5] is used which records various vital signs of the body including heart rate, respiration rate, skin temperature, activity, and electrocardiogram (ECG).¹ This sensor is worn on the chest and is able to locally store data or wirelessly transmitting it via Bluetooth (Figure 7.1). In this process of data analysis, the input data is continuous measurements which include physiological time series signals for a specific period.

Within the non-clinical condition, focus is primarily given to health parameters heart rate (HR) and respiration rate (RR), which are common vital signs in the health monitoring domain [28]. In this study, each health parameter is

¹ This specific product is now available on the provider's web page with another commercial name: *Zephyr™ Strap* [6].

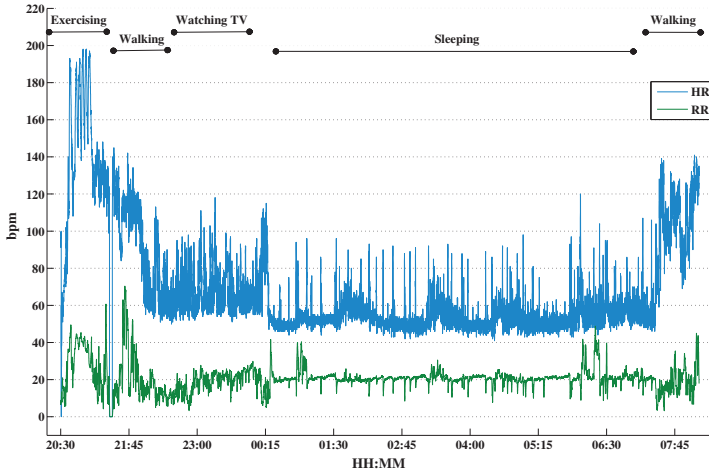


Figure 7.2: An example of non-clinical measurements depicting thirteen hours of heart rate (HR, top) and respiration rate (RR, down) during sequential activities. The unit *bpm* is used for both signals (beats per minute for heart rate and breaths per minute for respiration rate).

considered independently. An example of the input measurements is shown in Figure 7.2. The data in this figure has been recorded for thirteen continuous hours during the sequential activities such as exercising, walking, watching TV, and sleeping.

7.1.2 Clinical Physiological Data

A challenge in the evaluation is to find reliable data sets consisting of long-term measurements of physiological parameters (vital signs) where such sensor data is annotated with ground truth information about patients' conditions. Although the proposed approach is applicable to a variety of settings (Intensive care unit or ICU, ambulatory, and at-home monitoring), established benchmarks are more readily available from sensor data sets in a clinical setting. Thus, a data set of physiological sensor data from the online *PhysioNet* database is used [105]. In particular, a numeric data set within this database called *MIMIC* (multi parameter intelligent monitoring for intensive care) database [3] is considered. This data set contains periodic numeric measurements of physiological variables, such as heart rate, blood pressure, respiration rate, and oxygen saturation, obtained from bedside ICU monitors [156]. The entire database includes multiple recordings with various lengths of measurements (from 1 hour to 77 hours) which are acquired from 90 subjects (pa-

Table 7.1: Clinical conditions and their subjects in mimic database, after removing unreliable measurements.

Clinical conditions	No. of subjects (records)	No. of male/female	Age: [min,max], average	Average length of records
Resp. failure	10 (17)	7/3	[38,90], 67	32h25m
Bleed	2 (4)	1/1	[45,70], 57	44h45m
CHF	13 (17)	6/7	[54,92], 75	33h15m
Brain injury	2 (3)	1/1	[68,75], 70	21h30m
Sepsis	4 (5)	3/1	[27,88], 64	31h20m
MI	6 (8)	2/4	[63,80], 68	42h35m
Angina	2 (4)	1/1	[67,68], 67	41h10m
Post-op Valve	2 (5)	0/2	[49,67], 58	40h45m
Post-op CABG	3 (3)	1/2	[49,80], 66	40h20m

tients) with different ages and genders. The subjects in this database have been manually labelled different clinical categories related to their medical problems. In this work, the numeric records of the subjects from nine major clinical conditions have been selected to be analysed and modelled. The considered clinical conditions include *Respiratory failure*, *Bleed* (loss of blood from the circulatory system), *CHF* (chronic heart failure), *Brain injury*, *Sepsis*, *MI* (myocardial infarction, i.e. heart attack), *Angina*, *Post-op Valve* (heart valve surgery), and *Post-op CABG* (coronary artery bypass grafting surgery). General properties of the nine clinical conditions and the information about the selected subjects are listed in Table 7.1.

Here, the subjects with records consisting of at least 12 hours of continuous readings are considered to facilitate the identification of patterns over longer time horizons. Three physiological variables have been chosen to be processed: heart rate (*HR*), means of blood pressure (*BP*), and respiration rate (*RR*). Each measurement consists of long-term sequential data (i.e. time series) with a resolution of 1Hz. As an example, Figure 7.3 shows seven hours of sensor readings from the raw sensor data of a patient suffering from the CHF condition for variables *HR*, *BP*, and *RR*.

Working with the clinical measurements in the MIMIC database is challenging, due to dealing with incomplete, sparse, non-uniform, and irregular raw data [154]. Before analysing the records of the subjects, the measurements are cleaned in the following steps:

1. All records that include three mentioned physiological variables are picked for analysis,

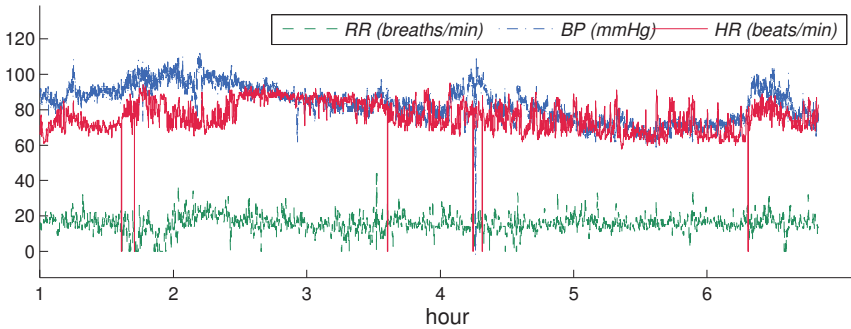


Figure 7.3: An example of clinical sensor data from MIMIC data set with variables heart rate (HR), blood pressure (BP), and respiration rate (RR).

2. Measurements with short episodes (less than 12 hours) were discarded, because finding meaningful patterns in a short period of data is not reasonable,
3. Since the data is collected in a clinical environment with wearable sensors, the signal readings involve plenty of artefacts and noise. To avoid processing incorrect information
 - (a) Sensor readings with unreliable values (e.g. zero value for heart rate) are discarded²,
 - (b) A *local regression* method (LOESS) as a smoothing function [99] is applied on readings to reduce the amount of noisy data.

After cleaning the data set, 45 subjects in nine clinical conditions are chosen to be analysed, which include reliable measurements with all three variables.

7.2 Partial Trend Detection in Physiological Time Series Data

Mining of physiological time series data is significant not only to model but also to detect specific health-related vital signs. One of the main challenges in the healthcare area is how to analyse physiological data such that valuable information can help the end user (physician or layman). The data analysis module aims to detect and represent the principal events and significant trends which are relevant for the end user. The proposed data processing method is unsupervised i.e., without expert knowledge or pre-defined rules, and can discover information which is not necessarily recognisable by an expert at first glance.

²The ranges of accepted values for three health parameters are: $20 < \text{HR} < 200$, $30 < \text{BP} < 220$, and $5 < \text{RR} < 100$.

The data analysis module includes preprocessing and segmentation steps which help the system to perform statistical information and trend detection components.

Preprocessing

In comparison with clinical data, noise and artefacts (i.e., disturbances or abnormalities within the signals) are more predominant in the signal from wearable sensors. Thus, the preprocessing of signals is a necessary task. In this work, artefacts are eliminated heuristically by setting some thresholds for each health parameter. Then, the local regression method (LOESS) has been applied to reduce the noise in signals. This method is a non-parametric regression method which is commonly used as smoothing function [149]. Figure 7.4b shows an example of the LOESS smoothing model for heart rate and respiration rate for the raw data presented in Figure 7.4a. The bandwidth parameter in this method is adapted depending on the requirements of the output for resolution of information. The output of this step is a prepared time series data for further analysis.

Signal Segmentation

After the smoothed signals are generated, a representation for each time series that captures temporal changes in the data is generated. Several methods have been introduced such as Fourier and Wavelet transforms, Symbolic Mappings, and Piecewise Linear models etc. [148] to represent the primary apparent attributes of the signals. Here, piecewise linear approximation (PLA) [130] as a segmentation method is selected for this system which can make a significant representation of the time series simply and efficiently. The output of the PLA method on a time series with length n is a set of linear segments with size m ($m \ll n$). The most popular approach to calculate the PLA is Bottom-up method. This approach starts with $n/2$ segments and merges the two next segments which have minimum distance error after merging. This process repeats till some stopping criteria are satisfied. The criteria could be setting a threshold on the maximum distance error and on the number of segments.

There are several methods to find the optimal number of segments [82]. In this work, the threshold parameter is heuristically tuned based on the requested resolution of the output trends.

Figure 7.4c and Figure 7.4d show the examples of output of PLA method for the smoothed heart rate and respiration rate time series presented in Figure 7.4b with $m=10$ and $m=25$, respectively. Based on the request from the end user, it is possible to provide all the details of events during the measurements or just reporting the major trends and changes with tuning the number of segments.

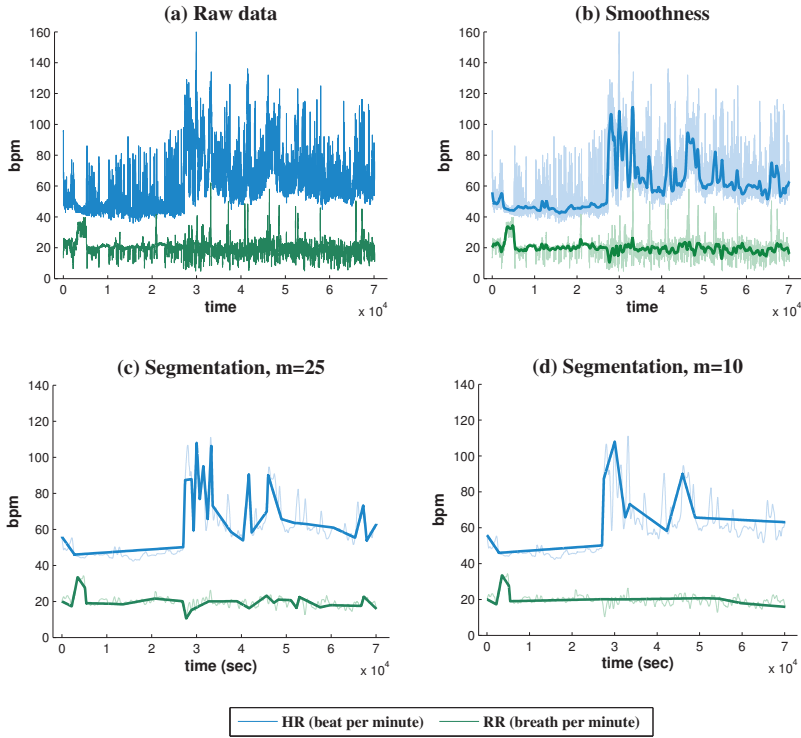


Figure 7.4: An example of physiological time series data preprocessing and segmentation: (a) the raw data of heart rate (HR, top) and respiration rate (RR, down) for 22 hours, (b) an instance of smoothing raw signals using LOESS method, (c) and (d) examples of the segmentation method (PLA) of the smoothed time series in Figure b with $m=25$ and $m=10$, respectively.

Partial Trends

Trends are the essential features to detect in physiological time series data as it can provide indications at an early stage of potential health issues and facilitate prevention. The acquired segments from the PLA approach are used to detect trends by applying a trend detection algorithm. The aim of this method is finding specific trends such as *dropping*, *rising*, or *unstable changes* in the measurements. Several studies have previously examined partial trends on time series segmentation [45, 124]. However, the challenge here is to determine which number of segments corresponds to significant events in the data. In other words, if the number of generated segments is high and each corresponds to a human understandable event, then the approach will produce too many

events and overburden the user. In contrast, if the number of segments is low, then valuable information about meaningful events could be lost.

To find a proper solution to represent the trends, a partial trend is considered to be a subset of segments, which has a similar tendency as the segments and relates to the orientation of data. Each time series is therefore represented as a collection of *partial* trends. The preliminary step of the algorithm is to normalise both axes of data by scaling them between 0 and 1. With this normalisation, the features of detected trends will be independent of the *duration* and *range* of the data (either long-term or short-term data).

The set of segments obtained from the previous step is considered as the input of the proposed trend detection method. After normalising, this method characterises the main attributes of the segments, which are longitude and gradient. For each segment s_i , the length of segment (len_{s_i}) and its gradient (grad_{s_i}), the trend detection algorithm starts with a set of segments, $S = \{s_1 \dots s_m\}$. Based on the defined parameters, the algorithm decides for the segment s_i : keep it and concatenate it with the current trend, keep it as the first segment of a new trend, or ignore it. The following function has been defined to make a balance between the features of s_i :

$$f(\text{grad}_{s_i}, \text{len}_{s_i}) = (\alpha - \text{grad}_{s_i}) - 1/(\lambda - \text{len}_{s_i}) \times k \quad (7.1)$$

where the α and λ are the heuristic thresholds for gradient and length, respectively and k is a coefficient to adjust the dependency of features. In this function, if $f(\text{grad}_{s_i}, \text{len}_{s_i})$ is more than zero then s_i is kept, otherwise it will be eliminated (except some conditions related to length of s_i and the gradients of its adjacent). Algorithm 7.1 illustrates the trend detection method with showing in which cases the algorithm makes a new trend or merges the segments in the current trend. Figure 7.5 presents an output of trend detection algorithm for the segmented heart rate and respiration signals. The annotations on the detected trends will be described in Section 8.2.

Depending on the end user requirements, the proposed system supports multi-resolution processing of the input signal and is able to summarise both long and short-term measurements. Figure 7.6 shows the output of the trend detection algorithm for two different resolutions of one measurement on heart rate. The first one is long-term data in 22 hours (Figure 7.6, top) and the second one is short-term data in 4.5 hours (Figure 7.6, bottom). The algorithm has detected several partial trends in the second diagram within 4.5 hours. However, the same portion of data has been identified as a single partial trend within 20 hours of time series data in the first diagram.

Algorithm 7.1: Partial Trend Detection

Input: Set of segments, $S = \{s_1 \dots s_m\}$
Output: Set of trends, $A = \{a_1 \dots a_l\}$, $l \leq m$
 $A \leftarrow \emptyset$
new trend a
foreach $s_i \in S$ **do**
 if $f(\text{grad}_{s_i}, \text{len}_{s_i}) > 0$ **then**
 if s_i and s_{i-1} *are in different gradient* **then**
 add a to A
 new trend a
 add s_i to a
 else
 if $\text{len}_{s_i} < \lambda$ **then**
 if s_i and s_{i-1} *are in different gradient* **then**
 if s_i and s_{i+1} *are in same gradient* **then**
 add a to A
 new trend a
 add s_i to a
 else
 add s_i to a
 else if $\text{grad}_{s_i} < \alpha$ **then**
 add a to A
 new trend a

7.3 Prototypical Pattern Abstraction in Physiological Time Series Data

7.3.1 Background on Pattern Abstraction

Dealing with large time series with high granularities is typically a challenge [134]. One of the objectives of this section is to find prototypical patterns in sequential data. This is mainly related to the general task of *pattern abstraction* [82]. The main goal of prototypical pattern abstraction is to provide a set of representative patterns from raw time series data, which includes two phases: 1) *discretisation* and 2) *clustering*.

Discretisation or segmentation is a solution to transform a time series $t = (t_1, \dots, t_n)$ with n time points into a discrete sequence of segments $S(t) : s_1 s_2 \dots s_m$, where generally $m \ll n$. Within different approaches for time series discretisation [82], a *sliding window* method is the most commonly used algorithm. In a sliding window approach, a time series t is discretised to a set of segments $S(t)$ by sliding a window of size w with a given overlap on two

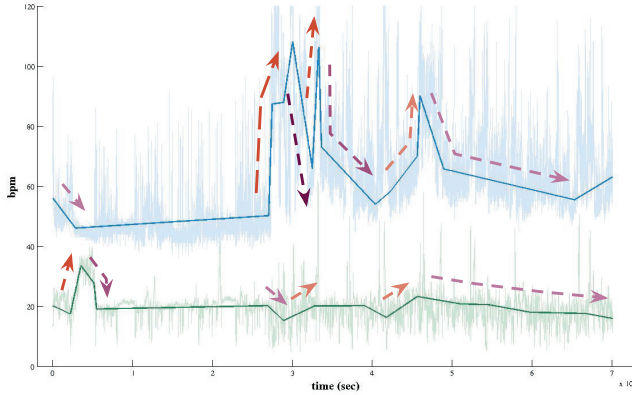


Figure 7.5: The output of partial trend detection algorithm for two segmented time series (HR on top and RR on bottom).

consecutive windows. Each segment $s_i = (t_{i_1}, \dots, t_{i_w})$ is a subsequence of the time series t , ($1 \leq i \leq m$). The provided segments are potentially the candidate to describe the unique attributes of the input data.

Clustering techniques are used for categorising the subsequences of time series, in order to exploit a reasonable number of representative patterns from numerous segments. The advantage of using a clustering algorithm is that the prototypical patterns are provided in a data-driven way without involving any domain knowledge to customise the typical patterns. Applying a mean normalisation on each segment is a part of clustering progress to minimise the effect of amplitudes of segments. Clustering subsequences of time series is a challenging matter discussed in the literature [71] and dependent on the discretisation algorithm and distance method. Several clustering algorithms (e.g. k-means, hierarchical, DBscan, etc.) along the various distance measures [145] can be applied on the data to cluster all the subsequences $s_i \in S(t)$ of time series t into a specific number of clusters (k). Cluster centres are considered as the prototypical patterns of the time series. In other words, these patterns are captured by averaging on the subsequences of clusters. Suppose $C_t = \{c_1, \dots, c_k\}$ is the set of prototypical patterns (clustering centres) of time series t , in which a prototypical pattern $c_j = (t'_{j_1}, \dots, t'_{j_w})$ is not necessarily an exact subsequence of time series t . Thus, in the sequence of segments $S(t)$, by replacing each segment s_i with its cluster centre, the corresponding sequence of prototypical patterns $P(t)$ is generated as: $P(t) : p_1 \dots p_m$, where $p_i \in C_t$.

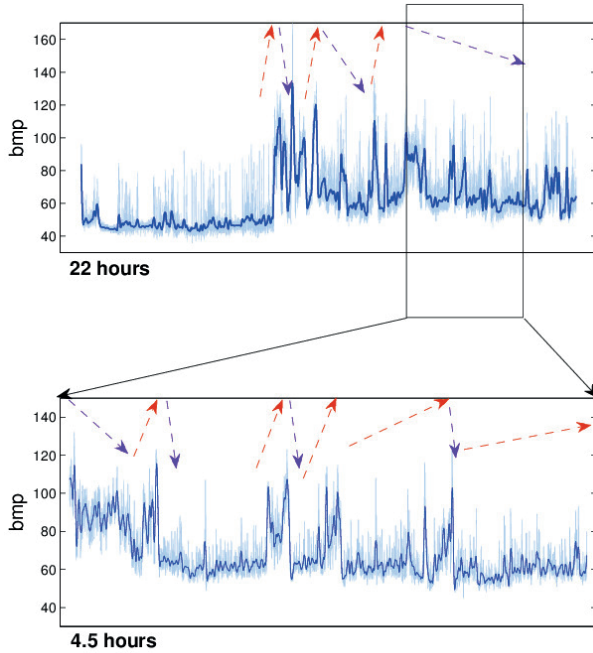


Figure 7.6: An example of trend detection outputs for two different resolutions of heart rate data. Top: 22 hours data with detected trends. Bottom: 4.5 hours data captured from the last trend in the first diagram.

7.3.2 Prototypical Pattern Abstraction

The considered measurement for each condition includes three mentioned variables HR , BP , and RR . Suppose three time series t_{HR} , t_{BP} , and t_{RR} , with the same length of n corresponds to the measurements in HR , BP , and RR , respectively.

Discretisation: In order to provide the sequence of prototypical patterns for each time series, the first step is a discretisation method, introduced in Section 7.3.1. Discretisation of time series data needs first to determine the size of sliding window (w). Since this approach aims to provide a set of descriptive rules based on the patterns, a meaningful range of values for the size of the sliding window (w) is tested. The length of overlap of two consecutive windows is set to half of the window's size, in order to avoid certain breaks in the signals that might lead to losing a segment involving prototypical patterns. By applying the discretisation method to the time series t_{HR} , t_{BP} , and t_{RR} , the sequences of

segments will be obtained for physiological variables as $S(t_{HR})$, $S(t_{BP})$, and $S(t_{RR})$, where $|S(t_v)| = 2 \times (n/w) - 1$, and $v \in \{HR, BP, RR\}$.

Clustering: The next phase is to extract the prototypical patterns of each time series using clustering methods. Here, a k-means clustering is applied to each set of segments, to categorise the segments into a set of clusters (k). In this algorithm, k segments are selected as initial centres. Then other segments are assigned to these centres based on their similarity, and the centre of each cluster is updated. This process is repeated until the centres do not change [106]. To optimise the number of clusters [121], a range of values is validated by considering the modelling results while the clustering approach is used for temporal rule mining task (this will be described in Section 8.1.4).

Before applying clustering, each subsequence is prepared as follows: If there are several artefacts in a subsequence, then this subsequence is not considered for further processing. The maximum number of allowed artefacts in a subsequence should be less than half of the length of the subsequence. If the number of artefacts in the segment's values exceeds a defined threshold, the segment s_i is removed from $S(t_v)$. Otherwise, the artefacts will be replaced with the values given by an interpolation method (i.e. cubic interpolation). After that, each segment $s_i \in S(t_v)$ (with the average value μ_{s_i}) is normalised to get zero mean by subtracting the μ_{s_i} from all values of s_i . This normalisation will invalidate the amplitude of segment values. So, the focus will be given to the trends of segments while clustering applies. The normalisation is crucial, because the segments with the same shape and trend need to be categorised in the same cluster, rather than the segments with a similar range of values. The k-means algorithm classifies the pre-processed segments of $S(t_v)$ into k clusters, with the set of centres C_{t_v} . Then the corresponding sequence of the prototypical patterns $P(t_v)$ is defined as: $P(t_v) : p_1 \cdots p_{|S(t_v)|}$, where $p_i \in C_{t_v}$ and $1 \leq i \leq |S(t_v)|$.

It is worth noting that applying various distance functions implies the various amount of computational effort. The Euclidean distance is employed in the clustering algorithm since it provides an efficient computation of the distances between segments. Moreover, different clustering algorithms with various distance functions (e.g. Euclidean distance, dynamic time warping, etc.) will provide distinct clusters and construct multiple prototypical patterns [172]. However, due to the time complexity of the other algorithms on big data, the k-means algorithm with Euclidean distance is employed in the clustering progress, which is not guaranteed to be the optimal, but sufficient for the goal of finding meaningful patterns.

An example of abstracted prototypical patterns from a sensor reading is shown in Figure 7.7a, which depicts the cluster centres obtained from *HR* sensor data in CHF condition ($w = 180$, $k = 7$). Figure 7.7b presents the sequence of prototypical patterns ($P_{t_{HR}}$) for the first two hours of *HR* data shown in Figure 7.3.

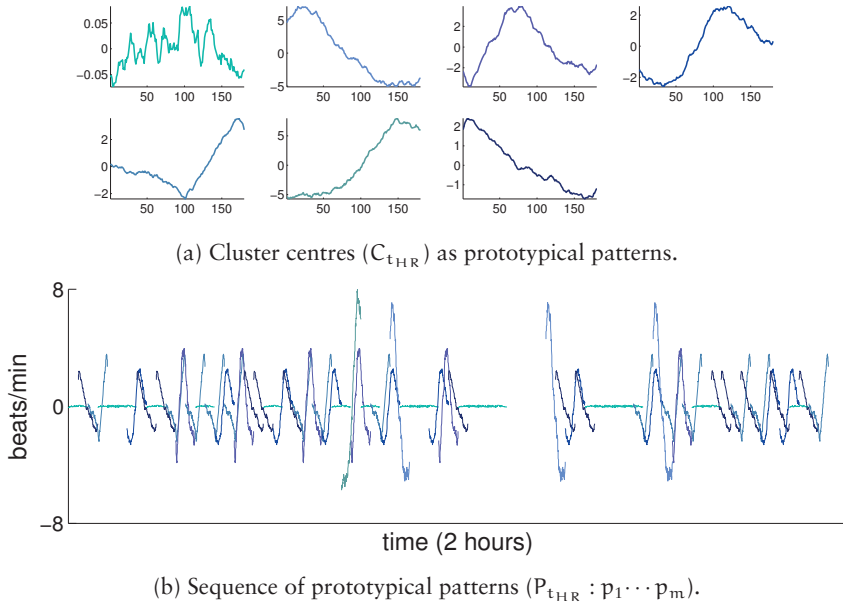


Figure 7.7: An example of prototypical patterns for *HR* data in CHF condition.

7.4 Discussion and Summary

This chapter has first presented the process of collecting and acquiring the input physiological sensor data sets. This process involved collecting data from wearable sensors as well as acquiring proper physiological sensor data from clinical conditions. Afterwards, this chapter has provided two main processes to analyse time series sensor data to catch the crucial behaviours of data. First, a trend detection method has been presented to capture the partial trends among the time series. Then, a pattern abstraction approach is introduced to extract the prototypical patterns throughout the time series. These prototypical patterns are the representative abstractions of the all possible time series subsequences within the data. The importance of extracting such prototypical patterns is to be able to determine which behaviours repeatedly occur in the time series. This information is data-driven in a sense that there is no prior knowledge about the patterns or no expert knowledge to define them beforehand. Also, detecting such prototypical patterns is not a trivial task for an expert such as clinicians or doctors by just exploring the recorded time series data. Therefore, these prototypical patterns are beneficial for further investigations.

There are some limitations related to the data preprocessing and data cleaning methods. For both trend detection and pattern abstraction algorithms, sev-

eral parameters have been set (either learned or just selected heuristically) that are dependent on the input recorded data. For example, in the pattern abstraction approach, an automatic approach has been suggested to select the parameters of *window size* and *number of clusters* by using the given set of data. However, providing a general method to automatically perform the parameter selection task for any input data set is not investigated in this study. Besides, the computational cost of the proposed algorithms (especially for pattern abstraction) is not optimised. There is a lack studying the suitable mechanism to minimise the computational cost of the approaches. This point will be critical if the size of input data grows.

In sum, this chapter shows how the raw numerical sensor data can be prepared and processes in order to either 1) be used for more complex data minings such as temporal rule mining among the patterns (Section 8.1), 2) be directly mapped to linguistic descriptions using template-based approaches (Section 8.2), or 3) be fed as perceived information to a semantic representation to be turned to linguistic descriptions using the semantic inference (Chapter 9).

Chapter 8

Mining and Describing Physiological Time Series Data

“Nothing in the world is more exciting than a moment of sudden discovery or invention, and many more people are capable of experiencing such moments than is sometimes thought.”

— Bertrand Russell (1872–1970)

FOLLOWING the structure of the thesis (Figure 1.2), a data preparation approach has been presented in Chapter 7 to provide information for semantic representations. But before showing how to apply the semantic representation approaches on physiological patterns, this chapter considers, mining more complex information, and using template-based NLG approaches to map the numerical data to the linguistic descriptions directly.

The output trends and patterns from data analysis can be considered as input for different aspects of the work on the application side. Two sections of this chapter present two aspects of data processing and interpreting such prepared patterns and trends. 1) The first section introduces a data processing upon the abstracted patterns by automatically mining temporal rules from the physiological sensor data in clinical conditions. This mining of temporal rules will enrich the processed information into more useful knowledge. 2) The second section proposes simple linguistic description approaches to interoperate the mind information in natural language, for all the processed information: detected trends, abstracted patterns, and temporal rules.

8.1 Mining Temporal Rules in Physiological Sensor Data

This section is dedicated to present temporal rule mining as a data-driven analysis of the abstracted patterns. The generated set of rules captures the temporal relationships between patterns of physiological data. This rule mining approach is presented in this thesis to be used later in a template-based linguistic description approach (explained in Section 8.2). This section first presents an overview of association rule mining methods in general and mining temporal rules in the medical domain. Afterwards, it gives details on the proposed approach to extract temporal rules from multi-channels of physiological time series data. The output rules for clinical conditions are then compared using a rule similarity measure that highlights the uniqueness of the rules in each condition.

8.1.1 Background on Temporal Rule Mining

Temporal rule mining is a promising approach to generate meaningful association rules from sequential data [197]. This section first describes the standard *association rule mining* method. Then it reviews rule mining approaches for temporal data in the medical domain.

Association Rule Discovery

Suppose $I = \{i_1, \dots, i_d\}$ is a set of items (e.g. all the products in a store), and $D = \{d_1, \dots, d_N\}$ is a transactional database with N transactions (e.g. all the shopping lists in a year). The support of an *itemset* $A \subset I$ is the frequency of the occurrence of A in all the transactions of D . The standard association rule mining provides a set of rules in form of $A \Rightarrow B$. In this rule, A is antecedent and B is the consequent, which are disjoint itemsets. Generally, a rule like $A \Rightarrow B$ means if the items of A occur in a transaction d_i , then the items of B also will plausibly appear in d_i . Typical measures to show the strength of a rule are support (*sup*) and confidence (*conf*). Support of a rule shows how often the rule itemsets occur in the database. Further, the confidence of rule $A \Rightarrow B$ determines how frequently the itemset B occurs in transactions which contain itemset A . Let $P_D(A)$ be the probability of the occurrence of an itemset A in D . Then, support and confidence of the rule $A \Rightarrow B$ are defined as [220]:

$$\text{sup}(A \Rightarrow B) = P_D(A \cup B), \quad (8.1)$$

$$\text{conf}(A \Rightarrow B) = p_D(A|B) = \text{sup}(A \Rightarrow B)/p_D(A) \quad (8.2)$$

The rules with sufficient support and confidence are typically known as strong rules. The values of *minsup* and *minconf* are specified as the thresholds for strong and meaningful rules. Association rules with low support may have

occurred accidentally that will not be interesting as significant rules. Similarly, a rule with low confidence cannot represent the frequent relations. Thus, the thresholds *minsup* and *minconf* given by the user can avoid involving the ineffective rules in the result.

Temporal Relations in Association Rules

Several versions of association rule mining algorithms have been introduced to deal with non-transactional data which consist sequential items (i.e., time series) to give temporal rules [134]. These algorithms adapt the definition of elements in association rules based on the time-stamped data to involve temporal constraints between the antecedent and the consequent of a rule. As in the case of $A \xRightarrow{T} B$, which intends “If *A* happens, *B* will happen within time *T*” [67]. Defining the temporal rules needs a reasonably good understanding of time-dependent relations between the temporal observations (items) [193].

Based on the *Allen’s temporal logic* [18], 13 possible relationships between each pair of temporal patterns can be specified. For association rule mining of temporal data, the abstracted patterns from time series are defined as the items. Then the set of transactions in rule mining method is constructed by all the combinations of Allen’s operations between temporal patterns. Then all these combinations of related temporal patterns from single or multivariate time series build the set of transactions. For instance, suppose two time series t_1 and t_2 , with the prototypical patterns C_{t_1} and C_{t_2} , also sequences of prototypical patterns $P(t_1) : p_1 \cdots p_m$ and $P(t_2) : q_1 \cdots q_m$. To find the coincident rules between t_1 and t_2 , the set of items is $I = C_{t_1} \cup C_{t_2}$, and the set of transactions D is constructed with all pairs of $d_i : (p_i, q_i)$ according to the ‘equal’ operation ($1 \leq i \leq m$). The next step would be to apply the described association rule mining algorithm to the provided transactions D and items I . The output of rule mining is a set of temporal rules $R = \{r_1, r_2, \dots\}$, where each rule $r_i : A \xRightarrow{\rho} B$ represents the repetitive relation of itemsets A and B along the operation ρ , where $A, B \subset I$ and $\rho \in \{\text{‘equal’}, \text{‘start’}, \text{‘meet’}, \dots\}$. While these approaches have been applied to physiological data [62, 166], they lack comparing the provided rule sets in various medical conditions. This comparison can reveal valuable insights about the data. The approach presented in this chapter attempts to address this problem.

Temporal Rule Mining in Clinical Settings

Recently, temporal association rule mining methods have been applied on the clinical data stream to identify complex relationships of the physiological sensor observations. Sacchi et al. [192] presented a knowledge-based approach for rule mining from labelled temporal patterns in biomedical data. In [62], the authors present temporal rule extraction for physiological data and address the problem of visually analysing this kind of data. The study in [113] proposes a

novel multivariate association rule mining based on change detection for complex data sets including numerical data streams. The authors in [157] introduce an approach to generate the rules automatically from the linguistic data of coronary heart disease using subtractive clustering and fuzzy inference to determine the diagnosis. In [20], a temporal technique for discovering frequent temporal patterns is proposed to extract well-known patterns of sleep apnea-hypopnea syndrome.

Although these systems have used rule mining techniques for health monitoring, none of them has focused on modelling the individual behaviours, along with a descriptive approach to represent the output of the system (i.e. generated rules). Also, those are mostly dependent on the initial knowledge provided by the user.

8.1.2 A New Approach for Temporal Rule Mining

To apply an association rule mining approach on each clinical condition from the MIMIC data set, all the selected records of the subjects with the same condition are accumulated to be analysed together. In this way, a more significant amount of data is involved in the process of rule mining, which leads to having more robust rules for each clinical condition. Recall from Chapter 7, the considered measurements of physiological data for each condition includes three variables HR , BP , and RR . Suppose three time series t_{HR} , t_{BP} , and t_{RR} , with the same length of n corresponded to the measurements in HR , BP , and RR , respectively.

The sequences of patterns $P(t_{HR})$, $P(t_{BP})$, and $P(t_{RR})$, with size of m , are obtained from the prototypical pattern abstraction, explained in Section 7.3. Association rule mining is a suitable approach to discover the coherence relations between the patterns occurred among the multi-variables. In this work, the focus is on the association rules between two pairs of physiological time series, i.e. heart rate with blood pressure ($HR \& BP$), and heart rate with respiration rate ($HR \& RR$), although, more compound relations are also applicable by applying complex temporal abstraction techniques [201]. As discussed in Section 8.1, the main requirement for association rule mining is to identify the set of items I and the set of transactions D . While considering the relation of HR and BP patterns, the set of items I includes all the prototypical patterns in both $C_{t_{HR}}$ and $C_{t_{BP}}$.

Different temporal relations can be defined on the discovered patterns to specify the transactions, but in this study, a modified set of relations between the patterns in physiological data is specified. Consider two multivariate signals HR and BP with the sequences of patterns $P(t_{HR})$ and $P(t_{BP})$, respectively. Let P_1 represents one pattern in $P(t_{HR})$ and P_2 represents at most two patterns in $P(t_{BP})$. Three temporal relations between P_1 and P_2 are considered: ' P_1 equals P_2 ', ' P_1 before P_2 ', and ' P_1 after P_2 '. It is worth noting that further temporal relations such as *meets* and *overlaps* are only slightly different with *before* and

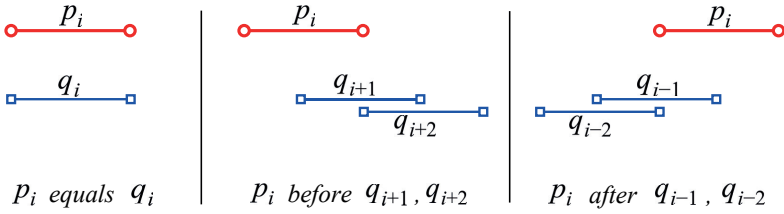


Figure 8.1: Three temporal relations between one pattern P_1 in $P(t_{HR})$ and at most two patterns P_2 in $P(t_{BP})$.

after relations. Likewise, the relations *during*, *starts*, and *finishes* are mostly covered by *equals* relation. Thus, for all the patterns in the sequences $P(t_{HR})$ and $P(t_{BP})$, three corresponding transactions for ‘*equals*’, ‘*before*’, and ‘*after*’ relations are defined as follows:

$$p_i \text{ equals } q_i : \quad d_i^e = (p_i, q_i), \quad (8.3)$$

$$p_i \text{ before } q_{i+1}, q_{i+2} : \quad d_i^b = (p_i, q_{i+1}), (p_i, q_{i+2}), \quad (8.4)$$

$$p_i \text{ after } q_{i-1}, q_{i-2} : \quad d_i^a = (p_i, q_{i-1}), (p_i, q_{i-2}), \quad (8.5)$$

where $p_i \in P(t_{HR})$ and $q_i \in P(t_{BP})$. Note that the relation ‘ P_1 *before* P_2 ’ is equivalent to the relation ‘ P_2 *after* P_1 ’, which means the opposite relations between *BP* and *HR* are also covered by this definition. Figure 8.1 shows the relational positions of patterns p_i , and $q_{i-2} \cdots q_{i+2}$ in their corresponding pattern sequences for three defined temporal relations.

The Apriori algorithm introduced in [11] is an efficient algorithm for association rule mining from a set of transactions D , which initialises all possible itemsets from the items I and then generates a set of sufficient rules like $A \Rightarrow B$ based on the co-occurrence of A and B in the transactions. This algorithm is based on the symbolic order of items, which can destroy the temporal relations in sequential data. However, in this approach, the temporal relations of the patterns are implicitly specified in the definition of the introduced transactions. In other words, the Apriori algorithm is applied, but with an adapted set of transactions D^ρ including defined temporal relations as their items: $D^\rho = \{d_i^e, d_i^b, d_i^a \mid 1 \leq i \leq |P(t_{HR})|\}$. Using the defined set of temporal transactions and the set of items (prototypical patterns), the generated rule is formulated as: $r : A \stackrel{\rho}{\Rightarrow} B$, where the antecedent (A) and consequent (B) can be any of two subsequences P_1 and P_2 that are co-occurred in D^ρ . The rule r also expresses additional information about the temporal relation ρ between p and q , such that $\rho \in \{\text{‘equal’}, \text{‘before’}, \text{‘after’}\}$. Applying the Apriori algorithm

with accurate values for *minsup* and *minconf* leads to have a set of temporal rules $R = \{r_1, r_2, \dots, r_n\}$ as a result. This rule set consists of the main repetitive relations of physiological data in sensor observations.

8.1.3 Temporal Rule Set Similarity

In this work, a similarity function is proposed to compute a ratio between the number of rules from one rule set which occur in another rule set. The provided temporal rules can be extended to represent the individual behaviour of vital signs in a given condition. This similarity function can evaluate the distinction between temporal rule sets. Suppose there are two rule sets $R_1 = \{r_1, \dots, r_{n_1}\}$ and $R_2 = \{r_1, \dots, r_{n_2}\}$ including m and n rules, respectively. The overlapping ratio of rule sets is a primary measure to investigate the characteristic properties of rule sets with the same sets of items [76]. The overlapping ratio as a similarity function between a pair of rule sets is typically defined as:

$$\text{Overlap}(R_1, R_2) = |R_1 \cap R_2| / |R_1 \cup R_2|. \quad (8.6)$$

In a standard rule association mining with a constant database of items, counting the intersection of the rules in R_1 and R_2 is straightforward, since it is easy to check the equivalence of rules. Two rules $r_i : A \Rightarrow B$ and $r_j : C \Rightarrow D$ are equivalent if their corresponding itemsets are equal: $A = C$ and $B = D$. However, the main issue with temporal rule sets produced by this approach is that the sets of items in different rule sets are entirely distinct. In other words, for different cases, there are different sets of prototypical patterns (items), and consequently different itemsets in the final rules. Thus, finding the overlap of temporal rule sets using Equation 8.6 utilising this function is not informative. Here, an alternative solution is proposed.

Occurrence Ratio

Suppose I_{R_1} and I_{R_2} are the most likely distinct sets of items (patterns) for the temporal rule sets R_1 and R_2 , respectively. To find the equivalent rule to $r_i : A \xRightarrow{\rho} B \in R_1$ in rule set R_2 (if exists), the approach searches for the most analogous rule $r'_j : A' \xRightarrow{\rho'} B' \in R_2$ which is sufficiently similar to r_i . Here, the similarity of itemsets is measured through the use of the pattern matching algorithms to find the best-matched patterns [59]. If r'_j exists, then one overlap is found between R_1 and R_2 , which means $A \approx A'$, $B \approx B'$, and $\rho = \rho'$. It is notable that the corresponding itemsets need to be approximately equal, whereas, the temporal relations have to be the same.

Searching for the occurrence of one rule in a rule set is presented in Algorithm 8.1. This algorithm shows how to find the most analogous rule from R to an input rule r (with the assumption of $r \notin R$). If such a matched rule r_m can be detected in R , it derives that the rule r most likely appears in R as well.

Algorithm 8.1: RuleMatch(r, R, I_R)

Data: $r : A \xrightarrow{\rho} B$, $R = \{r_1, \dots, r_n\}$ with set of items I_R , ($r \notin R$).
Result: $r_m : A_m \xrightarrow{\rho} B_m$, where $r_m \in R$ and $A_m, B_m \subset I_R$.
 $A_m \leftarrow$ best patterns matched to A from I_R ;
 $B_m \leftarrow$ best patterns matched to B from I_R ;
 $r_m \leftarrow A_m \xrightarrow{\rho} B_m$;
foreach $r_i : A_i \xrightarrow{\rho_i} B_i \in R$ **do**
 if $r_i = r_m$ ($A_i = A_m \ \& \ B_i = B_m \ \& \ \rho_i = \rho$) **then**
 return r_m ;
return \emptyset ; //rule not found

Algorithm 8.2: Occurrence(R_1, R_2, I_{R_2})

Data: Rule set R_1 , and rule set R_2 with set of items I_{R_2} .
Result: $\text{ratio}_{R_1 \text{ in } R_2}$: Occurrence ratio of R_1 in R_2 .
 $\text{weight}_{R_1 \text{ in } R_2} \leftarrow 0$;
 $\text{Weight}_{R_2} \leftarrow 0$;
foreach $r_i \in R_1$ **do**
 $r' \leftarrow \text{RuleMatch}(r_i, R_2, I_{R_2})$;
 if $r' \neq \emptyset$ **then**
 $\text{weight}_{R_1 \text{ in } R_2} \leftarrow \text{weight}_{R_1 \text{ in } R_2} + \text{sup}_{R_2}(r') \times \text{conf}_{R_2}(r')$;
foreach $r_j \in R_2$ **do**
 $\text{Weight}_{R_2} \leftarrow \text{Weight}_{R_2} + \text{sup}_{R_2}(r_j) \times \text{conf}_{R_2}(r_j)$;
 $\text{ratio}_{R_1 \text{ in } R_2} \leftarrow \text{weight}_{R_1 \text{ in } R_2} / \text{Weight}_{R_2}$;
return $\text{ratio}_{R_1 \text{ in } R_2}$

The next step is to measure how strong a rule occurs in another rule set. The method for checking the occurrence of a rule in another rule set leads to define a non-symmetric similarity measure, called $\text{Occurrence}_{R_1}(R_2)$, the occurrence ratio of R_1 in R_2 (previously called *Appearance ratio* in [30]). This measure represents how often rules with high support and confidence that appear in R_1 also occur in R_2 , considering their strength in R_2 . It means that while finding the closest rules of R_2 to the rules in R_1 , the values of support and confidence of matched rules are also considered in the occurrence ratio.

Algorithm 8.2 presents computing the occurrence ratio measure, which is scaled by the summation on the support and confidence of the rules in R_2 . Evidently, if the occurrence ratio of a rule set in another is considerably high, it shows these two rule sets are meaningfully associated. In contrast, if the ratio is considerably low, it indicates of few connections between rule sets, in a sense that two rule sets are distinct.

8.1.4 Results: Distinctive Rules in Clinical Settings

This section presents the experimental results of the rule sets in clinical settings from MIMIC database records. As discussed in Section 7.1.2, the raw data is fetched from the online MIMIC database. This data set includes a set of physiological measure nets that each belongs to a subject (i.e., patient) with a certain type of clinical condition. In this data set, only records that include three health parameters heart rate (*HR*), blood pressure (*BP*) and respiration rate (*RR*) are considered. The average length of all the measurements for a clinical condition in the data set is about 250 hours, and this average for a subject is around 50 hours. For each of the nine clinical conditions described in Table 7.1, temporal rule mining approach is applied to two pairs of sensor data: *HR* & *BP* and *HR* & *RR*. The output model of the rule mining approach is thus a collection of rule sets for clinical conditions.

Parameter Selection

To select the optimal values during pattern abstraction and rule mining phases, a voting approach is used considering the strength of the generated rules. Four parameters are optimised: window size, number of clusters, and the best thresholds for support and confidence. The window size is applied between 1 to 5 minutes, and the number of clusters is set between 5 to 9. Due to the small number of patients in the data set, optimising the parameters by applying rule mining on the entire measurements will lead to overfitting the model, and it does not indicate how well the method will generalise to unknown data sets. So, to avoid overfitting the model, a leave-one-out cross-validation approach [50] is used for finding the best parameters. For each clinical condition, this approach leaves out a patient's records in each fold of the validation as the hold-out set and applies the modelling on the rest of the patients as the training set.

After preparing the data for both training and hold-out sets for each combination of parameters, the validation of parameters are examined with two measures: *Interest* and *J-measure*. These measures indicate the quality of rules in different aspects [221].

By voting between the top rules with the highest values in these measures on the hold-out sets in all the iterations (folds), this approach is able to find the parameters that achieve the best average results. More precisely, with different values of w and k , all the models (on the train and test data sets) are generated in each clinical condition. Then, with different thresholds on support and confidence, the measures *Interest* and *J-measure* are calculated.

By voting again on the highest results of measures in the entire models, the best values of parameters are selected as $w = 180$ and $k = 7$. These values are obtained with the best cut-off values $\text{minsup} = 0.05$ and $\text{minconf} = 0.45$. Figure 8.2 depicts an example of applying the cross-validation approach for the MI clinical condition. This run includes six iterations (folds) for six

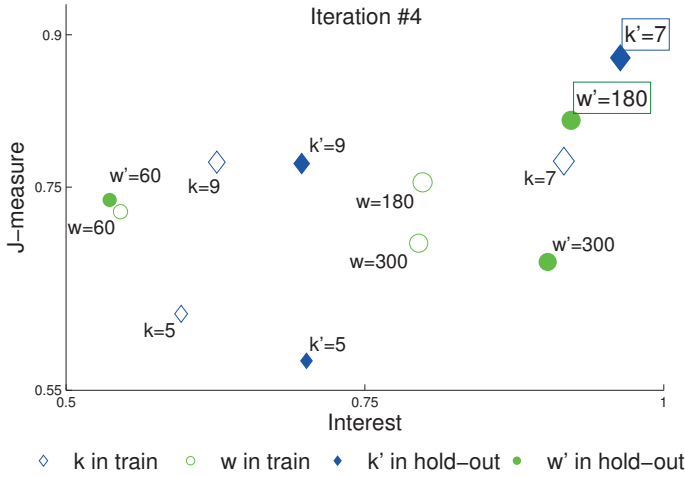


Figure 8.2: Result of cross-validation approach on a selected iteration for MI condition. The diagram shows the values of measures Interest and J-measure for various values of parameters w and k in both training set and holding set.

patients with the MI condition. But just to illustrate the values of two measures for each parameter, the figure shows the results of one iteration. As shown in Figure 8.2, the best values of measures Interest and J-measure are achieved by the mentioned values of w and k for a selected hold-out set.

Output Temporal Rule Sets

The results of rule sets in clinical conditions have been published in two versions by the author of this thesis. In the first version, only the *after* relation in patterns of *HR* and the other two variables has been considered. The results of this version are published in [30]), but not detailed in this chapter. The explained approach in this chapter (published in [32]) is the extension of the early version to consider more temporal relations. It is worth mentioning that the number of temporal relations is just one aspect of the distinctions between these two approaches. To compare the results of the extended version with the previous one, let's call the first approach $TRM-\rho^1$ (temporal rule mining with one relation), and the extended approach $TRM-\rho^3$ (temporal rule mining with three relations). Figure 8.3 shows the number of rules provided in both $TRM-\rho^1$ and $TRM-\rho^3$ approaches concerning the multivariate time series $HR \odot BP$ and $HR \odot RR$ in each clinical condition. The output sets of temporal rules indicate a collection of data-driven features which are independently able to describe their corresponding clinical conditions. To illustrate the variation of prototype-

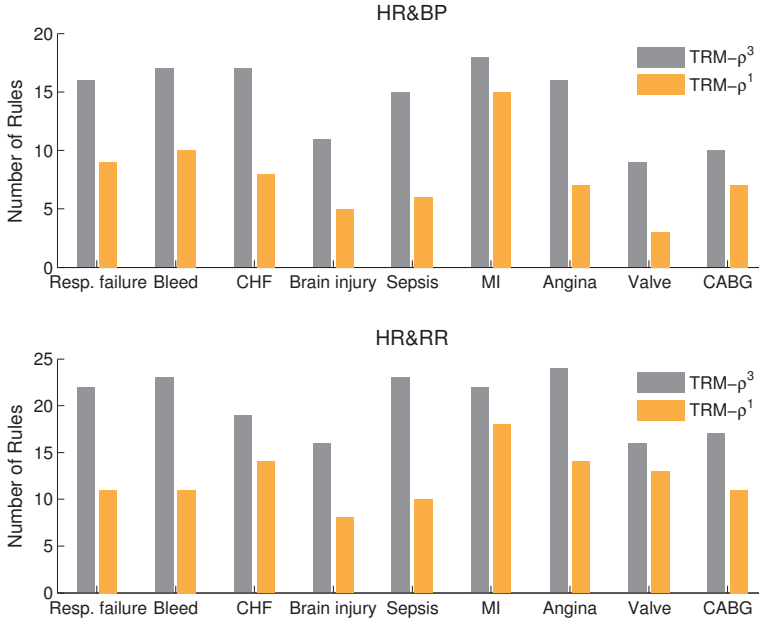


Figure 8.3: The number of rules for clinical conditions in $TRM-\rho^3$ and $TRM-\rho^1$ methods, in relation to the multivariate time series $HR\&BP$ and $HR\&RR$.

ical patterns among the temporal rules, a selection of distinct temporal rules from different rule sets with different temporal relations is represented in Figure 8.4.

8.1.5 Evaluation of Rule Set Similarity in Clinical Conditions

This section evaluates the uniqueness of rule sets for clinical conditions. For this reason, the new evaluation method based on the similarity function proposed in Section 8.1.3 is applied to measure the occurrence ratio of rules in other rule sets. This evaluation is first applied to the rule sets of clinical conditions to show the distinctness of rule sets. Besides, the rule sets for a selection of subjects in the same data set are compared with all the clinical conditions to consider the closeness of subjects to their corresponding conditions.

Occurrence Ratio of Rule Sets in Clinical Conditions

Based on the rule sets achieved from the $TRM-\rho^3$ method for clinical conditions, the evaluation approach is applied to each pair of rule sets. For nine clinical categories, the occurrence ratios for temporal rule sets are calculated.

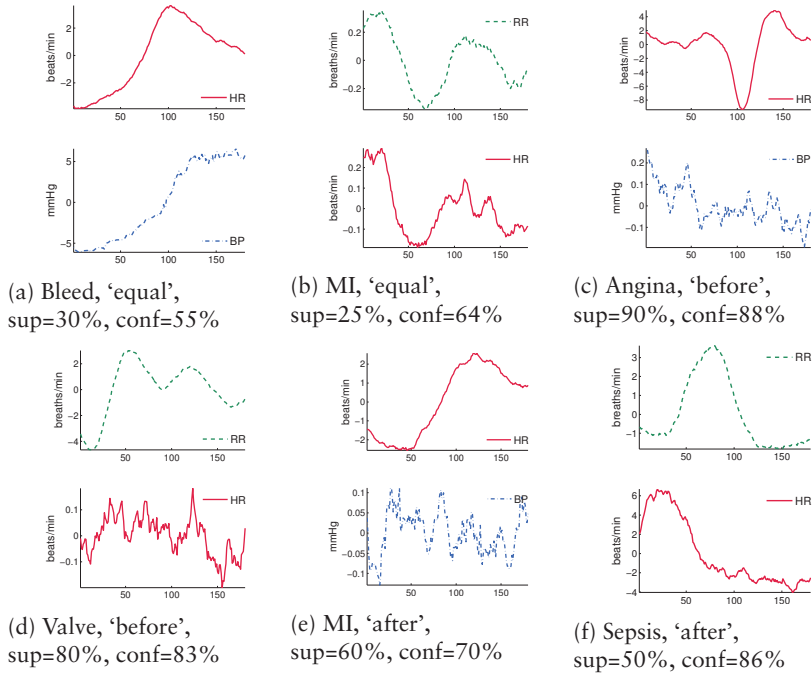


Figure 8.4: A selection of distinct temporal rules generated from physiological data in clinical conditions using $TRM-\rho^3$ approach.

As an example of output ratios, the matrix in Table 8.1 shows the obtained values of the occurrence ratio for temporal rule sets in $HR \& RR$ time series. Since the occurrence ratio is a non-symmetric similarity function, the values in Table 8.1 are not symmetric. For instance, the $Occurrence_{R_{MI}}(R_{CHF})$ is 27%, whereas $Occurrence_{R_{CHF}}(R_{MI})$ is 16%. The main reason for this variation is that the occurrence ratio is a weighted function which is calculated based on the support and confidence of the rules in only the second rule set. So, a subset of temporal rules with strong support and confidence values in their own rule set may appear in another rule set with weak corresponding support and confidence in the second one. The results in the matrix show the low occurrence ratios between the rule sets of clinical conditions.

In comparison with the former method $TRM-\rho^1$, the proposed approach for temporal rule mining $TRM-\rho^3$ is also providing much lower values of occurrence ratios between clinical conditions since the temporal rules are more specialised using the extra temporal relations. Figure 8.5 depicts a graphical comparison between the results of the occurrence ratios based on two approaches $TRM-\rho^1$ and $TRM-\rho^3$ in a box plot diagram. This figure shows most of the ratio

Table 8.1: Occurrence ratios of rule sets for each pair of clinical conditions in multivariate time series $HR \& RR$, using $TRM-\rho^3$.

	Resp. failure	Bleed	CHF	Brain injury	Sepsis	MI	Angina	Post-op Valve	Post-op CABG
Resp. failure	-	67%	0.2%	0%	3%	3%	2%	3%	7%
Bleed	61%	-	1%	6%	8%	2%	8%	3%	0.5%
CHF	8%	4%	-	1%	3%	16%	3%	0%	38%
Brain injury	11%	8%	0%	-	0%	1%	0.5%	2%	13%
Sepsis	3%	3%	3%	0%	-	0%	1%	0%	0%
MI	2%	14%	27%	4%	0%	-	3%	32%	15%
Angina	1%	49%	5%	1%	0.1%	1%	-	3%	27%
Post-op Valve	12%	2%	1%	10%	1%	16%	9%	-	55%
Post-op CABG	1%	0.5%	32%	41%	0.5%	6%	14%	65%	-

values are close to the zero in both versions, but much closer to the zero for the $TRM-\rho^3$ method. More precisely, for the occurrence ratios in clinical conditions for time series, more than 90% of all occurrence ratios are lower than 30%. Also, 83% of them are lower than 15% (in $TRM-\rho^1$ it was 70% lower than 15%). So, this evaluation to some extent can guarantee that the temporal rule mining methods generate relatively distinctive rule sets, which the rules in one category of the clinical condition can sufficiently provide an individual behaviour of its vital signs.

Occurrence Ratio of Subjects in Clinical Conditions

Another aspect of validating the performance of rule set similarity measure is to show the robustness of this measure for analogous rule sets. For this reason, some individual subjects with specific clinical condition labels are considered through the use of $TRM-\rho^3$. The temporal rule set of each subject is then compared with the rule sets of each clinical category via measuring their occurrence ratio in the rule sets of clinical conditions, using a leave-one-out method to avoid overfitting the modelling and the comparison of rule sets. If the occurrence ratio of a subject is higher in its corresponding clinical condition, rather than other conditions, then it shows the closeness of the subject's rule set and its corresponding condition's model. In other words, every provided rule set can represent a descriptive model of specific features which is recognisable from the other models.

Since the number of subjects in some clinical conditions is not enough, to avoid having biased results, the subjects with four major clinical labels, Respiratory failure, CHF, MI, Sepsis (33 subjects in total) are tested. The other clinical conditions do not have enough subjects to perform this evaluation. The

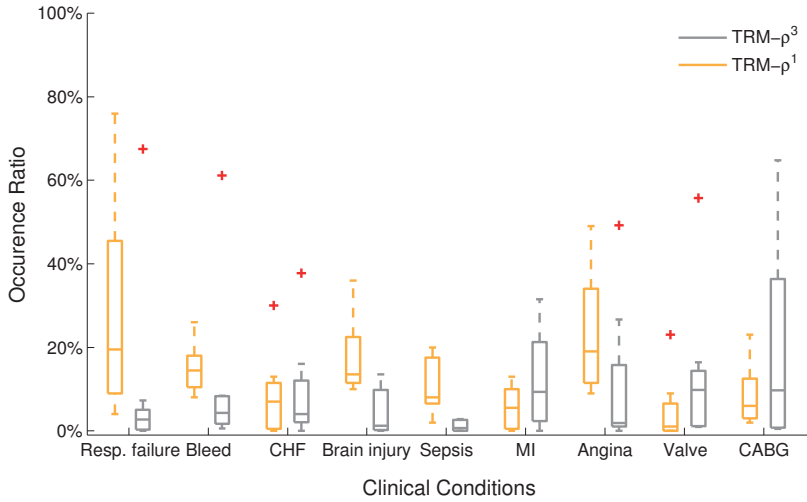


Figure 8.5: Boxplot diagram of the occurrence ratios between one clinical condition’s rule set and the other conditions with $TRM-\rho^3$ (each row in Table 8.1), in comparison with the results of $TRM-\rho^1$.

rule set of a subject sbj is compared with every rule set achieved from nine clinical conditions C_i through the calculation of $Occurrence_{R_{sbj}}(R_{C_i})$, where $1 \leq i \leq 9$. For each subject, two clinical conditions with top occurrence ratios have been selected as nearest conditions to the subject. If the closest conditions to a subject involve the same clinical label as the subject’s label, it shows a rich similarity between the rule sets of the subject and its corresponding clinical condition. The number of subjects with the same clinical label in their nearest conditions can indicate the correctness of the proposed evaluation.

Table 8.2 shows the three significant clinical conditions, with the number of subjects in each of them, in which their clinical labels have been revealed as one of their nearest conditions. For the subjects in CHF, MI, and Sepsis conditions, most of the rule sets were significantly close to their clinical model. However, the behaviour of rule sets for the Respiratory failure condition and its subjects are not adequately similar.

8.2 Linguistic Descriptions for Patterns and Temporal Rules

This section presents how a set of data-driven information can be turned into a set of natural language sentences that are humanly understandable. In particular, it explains template-based approaches which directly map the output

Table 8.2: Subjects with the same condition in their nearest rule sets.

Selected clinical condition	No. of Subjects	No. of nearest conditions with same label	percentage (%)
CHF	13	11	85%
MI	6	5	83%
Resp. failure	10	6	60%
Sepsis	4	3	75%

numerical information to linguistic characterisations. These approaches are applied to three sets of information derived from data analysis methods: partial trends, prototypical patterns, and temporal rules of patterns. It is notable that the proposed approaches here are heuristic and are the first attempt to generate natural language text for derived numerical information. It will be shown in Chapter 9 that how semantic representations can be involved in the process of text generation in order to enrich the final description of such information.

8.2.1 Trend and Pattern Description

A text generation method proposed in [29] provides a framework to detect and represent partial trends in sequential patterns. The method first detects the partial trends of an input time series based on their numeric features such as slope and duration. Then it characterises the partial trends in a textual form using a mapping function between numeric and symbolic terms such as sudden increase, steady decay, much fluctuated, and so on. By employing this method, the patterns in a temporal rule can be described based on their partial trends. The benefit of using natural language generation to represent the trends is that all the temporal events from a set of physiological time series data could be summarised in a textual output, which helps the end user to get a global perspective of the repetitive patterns and their temporal correlations in a massive amount of measurements.

The linguistic description approach applied for trend characterisation is inspired by the NLG architecture proposed by Reiter and Dale [185]. As presented in Chapter 2, this architecture includes the steps of data interpretation, document planning, microplanning and realisation (See Section 2.3.2 for more details). This section describes the linguistic characterisation of the detected trends only in the data analysis phase which is a part of *microplanning* module. For other tasks in NLG system, the framework follows the developed NLG methods in [185] or more recently in [119].

While extracting partial trends from time series data to represent them in natural language, the orientation of detected trends is interpreted in linguistic

Table 8.3: The instances of linguistic terms used for describing trends.

Range \ Duration	Short	Medium	Long	
Small	<i>steadily, nor- mally</i>	<i>slowly gradually</i>	<i>Adverb</i>	
Medium				
Big	<i>suddenly</i>	<i>sharply</i>	<i>regularly</i>	
-	<i>rise, drop</i>	<i>increase, decrease, recover</i>		<i>Verb</i>

terms. For this reason, two following features of each trend are considered: (1) the duration of trend and (2) the range of values that trend belongs to. To meet the requirements of the end user and domain specificity, the system uses a fuzzy granulation. A heuristic method is used to map between the mentioned features and the linguistic terms considering the following behaviours of trends: the duration of the trend to be represented (short, medium, long), and the range of trend would be represented (small, medium, big). Note that depending on some criteria like the goal of the system, the end user's needs and the type of input health parameter, the function of identifying these terms may vary.

With this categorisation, the system can fetch the linguistic terms to describe each trend (with specified duration and range) in natural language sentences. These sentences include particular portions such as subject, verb, adverb etc. which have to be clarified by the system. An example of defined lexicons for the trend's behaviour is illustrated in Table 8.3 which includes a set of suggestions for the proper verbs and adverbs in each combination of specified duration and range. Figure 8.6 shows some instances of linguistic terms for extracted trends in HR (top) and RR (down) signals.

8.2.2 Temporal Rule Representation

One descriptive way of representing the rules is to generate a textual representation of them for the end user of the system. A simple representation of a typical rule, $r : A \Rightarrow B$ in natural language text is to put the relation and the definition of the itemsets as the antecedent and consequent in a textual format such as: "When (If/while) A occurs (happens), then (after that, at the same time) B will occur". For instance, in the example of market basket [207], a rule could be explained as: "Customers who buy bread and cheese are likely to buy milk." The main challenge in the textual representation of the temporal rules $r : A \xRightarrow{\rho} B$ is to involve the temporal relation (ρ) into the rule representation.

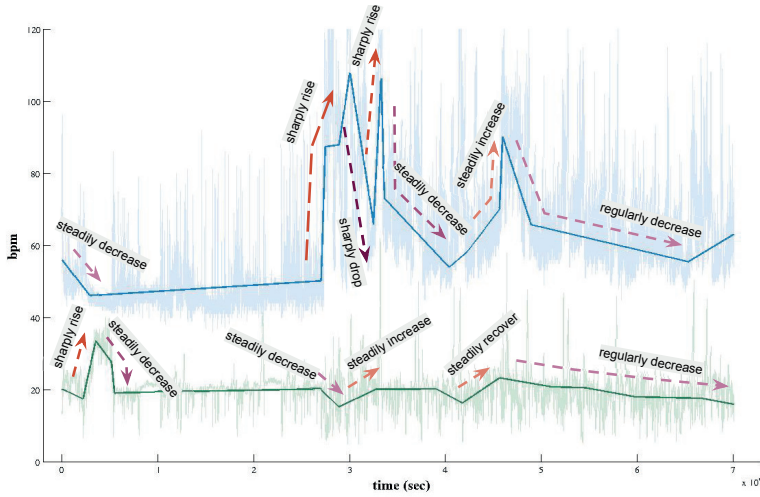


Figure 8.6: The output of the partial trends for two segmented time series (HR on top and RR on bottom), shown in Figure 7.5.

As mentioned before, for two subsequences of patterns P_1 and P_2 with the relation ρ , both temporal rules $P_1 \xrightarrow{\rho} P_2$ and $P_2 \xrightarrow{\rho} P_1$ can be generated through the association rule mining method. Although the temporal relation is same for these two rules, the meaning and interpretation of them are practically different, because the roles of antecedent and consequent have been swapped. For this reason, a linguistic mapping from temporal rules to their messages should be defined. Table 8.4 illustrates a mapping for these two rules with the itemsets P_1 and P_2 while considering the defined temporal relations $\rho \in \{\text{'equal'}, \text{'before'}, \text{'after'}\}$.

Another challenge of textual rule representation while dealing with patterns is how to explain the antecedent and consequent patterns as temporal subsequences in a meaningful way. Since a data-driven approach extracts the prototypical patterns, there is no predefined characterisation of them by the expert. Therefore, a linguistic description of the prototypical patterns of the subsequences P_1 and P_2 should be generated in the place-holders denoted by $[P_1]$ and $[P_2]$ in Table 8.4. For instance, an output text like “*After a gradual decrease in pattern P_1 , then pattern P_2 has a big rise and then a sharp drop*” is understandable, to interpret the behaviour of patterns in discovered rules for the end user.

Table 8.4: A template-based textual representation of rules with temporal relations.

	$P_1 \stackrel{g}{\Rightarrow} P_2$	$P_2 \stackrel{g}{\Rightarrow} P_1$
$P_1 \text{ equals } P_2$	when $[P_1]$, at the same time $[P_2]$.	when $[P_2]$, at the same time $[P_1]$.
$P_1 \text{ before } P_2$	when $[P_1]$, after that $[P_2]$.	when $[P_2]$, before that $[P_1]$.
$P_1 \text{ after } P_2$	when $[P_1]$, before that $[P_2]$.	when $[P_2]$, after that $[P_1]$.

Textual Representation of Temporal Rules

As described in Section 8.2.2, the significant tasks in temporal rule representation are 1) characterising the main trends in each of antecedent and consequent as patterns and 2) realising the form of temporal relation in a provided rule (Table 8.4). The partial trends in the patterns of a temporal rule are textually represented based on their numeric features and conduct, which is proposed in Section 7.2. The linguistic demonstrations of the temporal relation between the antecedent (here, *HR*) and consequent (here, *BP* and *RR*) of a rule are provided by a variety of words which are employed from expert knowledge. For instance, the *equal* relation is presented with the terms “*at the same time, simultaneously, concurrently, etc.*” or the relations *before* and *after* are shown with “*before that, earlier, just after that, later, afterwards, etc.*”. Moreover, the strength of a temporal rule based on its support and confidence values can be also represented in the corresponding sentence. It provides a meaningful impression on the rule strength for the reader of the textual messages. Various terms and phrases for the values of support and confidence can be used. As an example the sentence of a temporal rule with a high confidence value is started with the terms like: “*most of the time*” or “*commonly*”. In this work, since the rules are generated to show the sequential happenings in the entire data, the general conditional (if-then) sentence is implemented to characterise the antecedent and consequent of rules. It is worth to note that to make the final text more natural, different templates of conditional sentences have been applied (e.g. using “*when*” or “*after*”, instead of “*if*”).

Table 8.5 shows the generated textual representation of the acquired temporal rules in Figure 8.4. In these output examples, each sentence describes a discovered temporal rule to specify the temporal relation inside the rule with the partial trends in each of the appeared prototypical patterns, followed by the corresponding clinical condition. An advantage of generating final output in natural language is a textual description that is understandable and interpretable by the end user of the system.

Table 8.5: Textual representation of the acquired rules in Figure 8.4.

Rules	Linguistic Description
Fig. 8.4 (a)	<i>In the Bleed condition, occasionally when the heart rate normally rises (7 beats) and steadily decreases (4 beats), at the same time the blood pressure normally rises (10 units).</i>
Fig. 8.4 (b)	<i>In the MI condition, usually when the respiration rate decays and then rises in a very small range, simultaneously the heart rate decays and rises very slowly.</i>
Fig. 8.4 (c)	<i>In the Angina condition, most frequently if the heart rate sharply decreases (10 beats) and suddenly rises (13 beats), later, the blood pressure reduces very slowly.</i>
Fig. 8.4 (d)	<i>In the Post-op Valve condition, most of the time the respiration rate sharply increases (7 breaths) and steadily reduces (3 breaths, just before that, the heart rate decreases in a very small range.</i>
Fig. 8.4 (e)	<i>In the MI condition, usually after the heart rate steadily increases (5 beats) and normally reduces (2 beats), the blood pressure fluctuates in a very small range.</i>
Fig. 8.4 (f)	<i>In the Sepsis condition, most of the time before the respiration rate normally rises (2 breaths) and suddenly decreases (5 breaths), the heart rate steadily decreases (6 breaths).</i>

8.3 Discussion and Summary

The approach introduced in the first section of this chapter presents a descriptive model of temporal rule mining to generate meaningful rules for physiological sensor data in a clinical setting. This modelling also underlies the *uniqueness* of the rule sets for considering cases, which means each provided rule set contains distinct rules that are unique to their model. The advantage of providing distinctive rules for clinical conditions is to enable physicians to discover specific behaviours of vital signs, which are not necessarily recorded in medical ontologies. The proposed approach is able to exploit unseen and distinctive information per patient or condition. This information can assist the clinicians in individual decision making.

The second section of this chapter introduces the approaches to describe the mind information linguistically. First, it is shown how linguistic terms can annotate the partial trends and prototypical patterns (extracted in data analysis phase in Chapter 7). Then, with the use of the annotated trends and patterns, a

template-based approach is presented to generate linguistic descriptions for the mined temporal rules involving the temporal relation of the unknown but interesting patterns. Although the output results are reasonably human-readable text, the limitation of this approach is the richness of the provided annotations and labels for the partial trends and patterns. This method has been only relying on the shape and the trends of the time series patterns. So, an alternative pass to describe unknown patterns and temporal rules is to describe them by a model that is constructed on the basis of the known annotated patterns. Semantic representations that are presented in Part I can play this role in a linguistic description method.

In sum, this chapter provides a more in-depth analysis of the extracted patterns of physiological data to find temporal rules among those patterns. After, it presented linguistic description approaches to turn patterns and rules into natural language texts. Chapter 9 shows how a semantic representation can help the linguistic description approaches to enrich the final text, and how it can help to describe further (and possibly unknown) observations.

Chapter 9

Linguistic Descriptions for Time Series Patterns using Conceptual Spaces

“Long before worrying about how to convince others, you first have to understand what’s happening yourself.”

— Andrew Gelman (1965–)

THIS chapter presents the application of the proposed semantic representations in Part I in the area of physiological sensor data. As mentioned in Part I, the input of a semantic representation needs to be a set of perceived information with certain attributes (namely involving labels and features for observations). This thesis has employed data analysis approaches to extract meaningful and interesting information. These approaches have already been widely discussed in chapters 7 and 8. The abstracted patterns from sensor data are now used in a semantic representation, and then be utilised to infer linguistic descriptions, as will be presented in this chapter. The distinctive approach presented here is the ability of modelling and interpreting new observations that are could be unknown (i.e., not pre-defined) for the system.

Within the field of time series data mining, the perception-based analysis of patterns attempts to formalise knowledge and simulate human reasoning [35]. Linguistic descriptions can represent the perceptions (i.e., words such as low, increasing, most of the time, etc.) of time series patterns. A time series pattern is a subsequence of a univariate time series containing a meaningful behaviour or trend of the data. Many studies consider the problem of qualitative analysis of time series patterns and its manipulation with linguistic information [32, 36, 124, 165, 237]. However, in most of them, the required linguistic

information is limited by the expert or domain knowledge. In other words, the developed systems are designed to cover specific trends and shapes of patterns with a particular set of the requested linguistic characterisations as the general vocabulary. For example, Yu et al. [237] developed a natural language generation (NLG) framework to summarise the patterns in large time series in a few sentences. This framework uses an ontology of patterns as general vocabulary like a spike, a step, etc. to comply with the linguistic requirements. The major drawback of such a system is that any other observation (pattern) which is not matched with the provided vocabulary cannot be described and reported in the final summary.

Here, the conceptual spaces theory is used to represent the linguistic characterisation of time series patterns in a semantic model. According to the proposed approach in Part I, this model provides a symbolic representation of both known and unknown patterns. This chapter shows how to construct such conceptual space for a given set of time series patterns, also shows how to inference in the conceptual space for the linguistic characterisation of time series patterns.

9.1 Constructing a Conceptual Space of Time Series Patterns

Assume that there is a set of time series patterns with varying lengths, which are labelled with a set of linguistic terms (See Figure 9.1). Here, the same set of physiological patterns that are extracted from the MIMIC data set (explained in Chapter 7) are used. The time series patterns are exploited from heart rate and respiration rate recorded in several clinical conditions [32]. From this data set, 78 patterns are used as known observations with varying time durations, which are categorised (i.e., labelled by experts) into four known classes: *increasing*, *decreasing*, *spike*, and *oscillation*. These class labels of time series patterns are acquired from the common behavioural labels used in the literature of mining and linguistic characterisation of shapes and trends in time series data [35, 107, 165, 236].

Formally, the data set and the labels of the patterns are defined as: $\mathcal{D}^P = \{o_1, \dots, o_{78}\}$ and $\mathcal{Y}^P = \{y_{in} : \text{'Increasing'}, y_{de} : \text{'Decreasing'}, y_{sp} : \text{'Spike'}, y_{os} : \text{'Oscillation'}\}$. Although there many other types of behaviours for time series patterns, these four classes are primarily chosen to simplify the process of conceptualising the patterns. Figure 9.1 shows the typical examples of such patterns for each of the mentioned classes of the data set. Regarding the set of class labels \mathcal{Y}^P , the set of concepts is defined as: $\mathcal{C}^P = \{C_{in}, C_{de}, C_{sp}, C_{os}\}$.

Initialising the primitive set of characteristic features is another input to build the conceptual space of time series patterns. There are many characteristic features for analysing and modelling time series data, from simple statistical features to frequency related ones. As mentioned in the leaf data set, the criterion is how describable or interpretable the features are in lin-

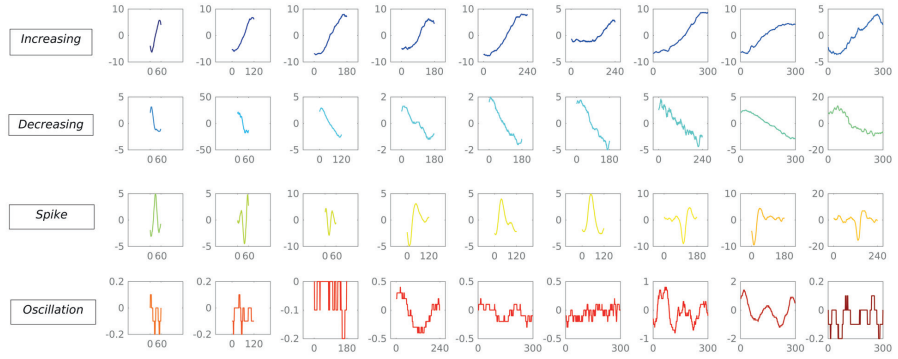


Figure 9.1: Four sets of time series patterns, presenting the known classes of patterns in the data set.

guistic form. For example, in time series pattern data set, the values of *integral* or mean features can be useful for analytical tasks in time series mining, but these values are meaningless to the end user of the system to visualise or distinguish it from other patterns. In contrast, a feature like *slope* of a pattern is perceptually interpretable for the user in natural language. Among the various features in the literature of feature-based time series data mining [83,107,160,203], the following features have been chosen as the initial set of features $\mathcal{F}^p = \{X_i = \langle H_{X_i}, I_{X_i} \rangle\}$:

- $X_\alpha : \langle \text{'Slope'}, (-\pi, \pi) \rangle$ (the slope of pattern),
- $X_{\Delta_{mm}} : \langle \text{'Min - Max Diff'}, [0, \text{inf}) \rangle$ (absolute difference between min and max values),
- $X_{\Delta_{se}} : \langle \text{'Start - End Diff'}, [0, \text{inf}) \rangle$ (difference between start and end values),
- $X_{\Delta t} : \langle \text{'Time interval'}, (0, \text{inf}) \rangle$ (time duration of pattern),
- $X_{en} : \langle \text{'Entropy'}, [0, \text{inf}) \rangle$ (how chaotic is the pattern),
- $X_{fft} : \langle \text{'Frequency'}, [0, 1] \rangle^1$,
- $X_{\partial x} : \langle \text{'First Derivative'}, [0, 1] \rangle$,
- $X_{\partial \partial x} : \langle \text{'Second Derivative'}, [0, 1] \rangle$,
- $X_\sigma : \langle \text{'Standard Deviation'}, [0, \text{inf}) \rangle$.

9.1.1 Domain Specification for Time Series Pattern Data Set

After calculating all these features for every known observation, the conceptual space construction approach has been applied with the inputs of known labelled observations \mathcal{D}^p , label set \mathcal{Y}^p , and feature set \mathcal{F}^p . The approach first

¹For some features, I_i needs to be set manually based on a mapping function from feature's values to one value in an interval. For example, the outputs of Fourier transform (fft) function of a pattern is mapped to the values in the interval $[0.1]$, likewise for first and second derivatives

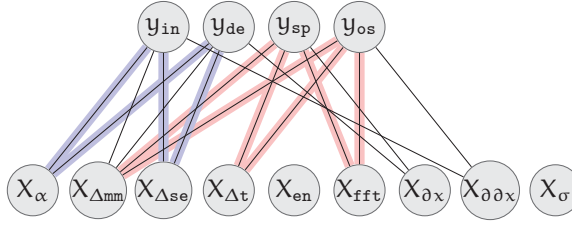


Figure 9.2: The bipartite graph presenting the relevance of the features and the labels in data set of time series patterns. Also, two chosen bicliques (as the domains) are highlighted with the blue and red edges.

utilises the feature filtering approach, i.e. MIFS (Algorithm 3.1) to provide a ranking matrix which shows the mutual correlation of features and labels. Then, the feature subset grouping determines which subsets of features as domains represent which labels as concepts using the Algorithm 3.2. Figure 9.2 illustrates the created bipartite graph, which presents the specified domains and quality dimensions. Two selected maximum bicliques determines two domains $\Delta = \{\delta_1, \delta_2\}$, where each domain is specified as follows:

- Domain $\delta_1 = \langle \mathcal{Q}(\delta_1), \mathcal{C}(\delta_1), \omega_{\delta_1} \rangle$, wherein
 $\mathcal{Q}(\delta_1) = \{q_{\alpha}, q_{\Delta se}\},$
 $\mathcal{C}(\delta_1) = \{C_{in}, C_{de}\}.$
- Domain $\delta_2 = \langle \mathcal{Q}(\delta_2), \mathcal{C}(\delta_2), \omega_{\delta_2} \rangle$, wherein
 $\mathcal{Q}(\delta_2) = \{q_{\Delta t}, q_{\Delta mm}, q_{fft}\},$
 $\mathcal{C}(\delta_2) = \{C_{sp}, C_{os}\}.$

Figure 9.3 depicts a graphical presentation of the determined domains with the corresponding quality dimensions and concepts for the known time series patterns. As an example, δ_1 is specified by two quality dimensions ‘start – end diff’ and ‘slope’, and is associated with two concepts ‘Increasing’ and ‘Decreasing’. An example of the calculated weights in a domain is $\omega_{\delta_1}(C_{in}, q_{\alpha}) = 0.62$, which shows the salience of the relation between pattern concept ‘Increasing’ and quality dimension ‘slope’ within δ_1 . Similar to the leaf conceptual space, although the process of specifying the domains is data-driven, there may be an interpretation for each determined domain. Here, the interpretation of perceived domains is more sensible. For instance, one can say that δ_1 illustrates the *trend direction* of the known patterns, while δ_2 shows the *shape* of the known patterns (see Figure 5.3). As the output of the domain specification phase for the conceptual space of patterns, the set of quality dimensions will be $\mathcal{Q}^p = \{q_{\alpha}, q_{\Delta se}, q_{\Delta t}, q_{\Delta mm}, q_{fft}\}$. Moreover, the set of instances is defined as: $\Gamma^p = \cup_{y \in \mathcal{Y}} \Gamma(y)$, where $|\Gamma^p| = |\mathcal{D}^p|$.

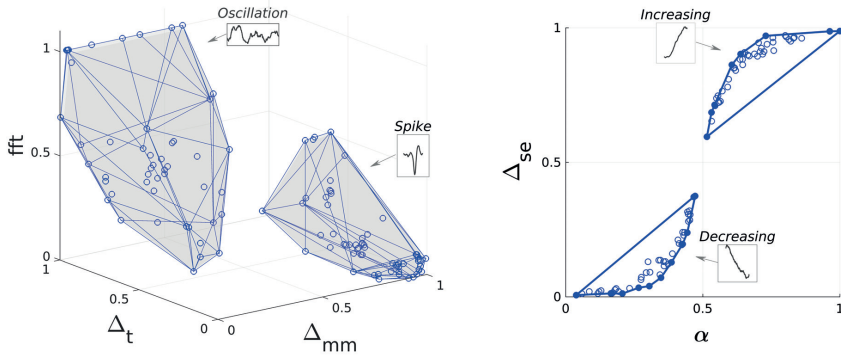


Figure 9.3: The conceptual space of time series pattern data set: a graphical presentation of the determined domains with the corresponding quality dimensions and concepts.

9.1.2 Concept Representation for Pattern Concepts

Regarding the output of the domain specification process, each concept in \mathcal{C}^P appears in only one domain (has precisely one sub-concept), as $C_y = \{c_y\}$. By applying Algorithm 3.3, the elements of the sub-concept for each concept in \mathcal{C}^P is derived as follows.

- Pattern concepts ‘Increasing’ and ‘Decreasing’ are represented in δ_1 as, respectively:

$$C_{in} = \{c_{in}^1 : \langle \eta_{in}^1, \phi_{in}^1 \rangle\},$$

$$C_{de} = \{c_{de}^1 : \langle \eta_{de}^1, \phi_{de}^1 \rangle\}.$$

- Pattern concepts ‘Spike’ and ‘Oscillation’ are represented in δ_1 as, respectively:

$$C_{sp} = \{c_{sp}^2 : \langle \eta_{sp}^2, \phi_{sp}^2 \rangle\},$$

$$C_{os} = \{c_{os}^2 : \langle \eta_{os}^2, \phi_{os}^2 \rangle\}.$$

In these representations, for example, η_{inc}^2 shows the 3D convex polytope of pattern concept ‘Spike’ within δ_2 (see Figure 9.3). Also, as an example for the weights, $\phi_{sp}^2 = \{\omega_{\delta_2}(C_{sp}, q_{\Delta_t}), \omega_{\delta_2}(C_{sp}, q_{\Delta_{mm}}), \omega_{\delta_2}(C_{sp}, q_{ffft})\}$ shows the salience between pattern concept ‘Spike’ and three quality dimensions ‘time interval’, ‘min – max diff’, and ‘frequency’ within δ_2 . In Figure 9.3, the graphical presentation of time series pattern concepts is shown by illustrating the convex hulls of their corresponding sub-concepts.

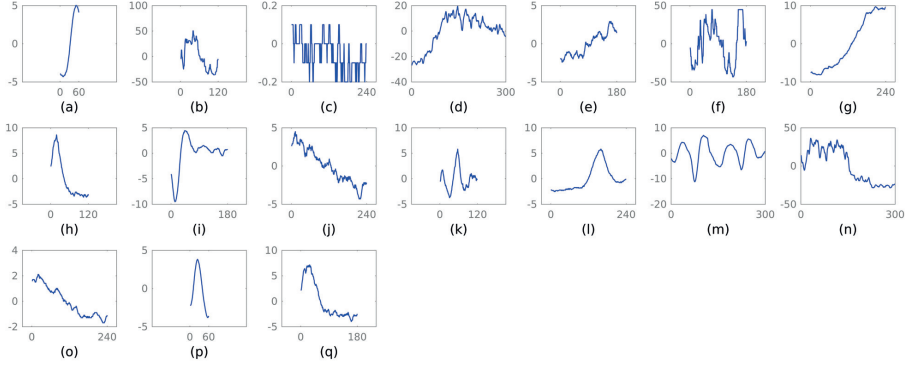


Figure 9.4: A set of unknown samples of time series patterns.

Now, with the provided elements, the conceptual space of the time series pattern data set is presented as: $S^{patterns} = \langle Q^p, \Delta^p, C^p, \Gamma^p \rangle$.

9.2 Semantic Inference for Unknown Patterns

The aim is to derive a linguistic description for unknown time series patterns. Figure 9.4 shows a number of examples of unknown patterns that are chosen to be considered. According to the proposed inference process, an unknown pattern sample (e.g., pattern (a) in Figure 9.4) is first vectorised to an instance γ_a . Then, a linguistic description for (a) is inferred in two phases: set the values of symbol vector and set the lexicons, by inferring in conceptual space and symbol space, respectively. γ_a is a set of points within $\Delta^p(\gamma_a)$ as: $\gamma_a = \{p_{\gamma_a}^1, p_{\gamma_a}^2\}$, where the numeric values of each point are the feature value of (a) for each quality dimension. For example, in δ_1 : $p_{\gamma_a}^2 = \langle q_{\alpha}(a), q_{\Delta se}(a) \rangle = \langle 0.34, 0.28 \rangle$.

9.2.1 Inference in Conceptual Space of Patterns

In task of this phase is basically to check whether or not the new instance γ_a is included in any defined concept's regions, and then infer semantic descriptions based on the closeness of its values to the regions. Considering the pattern sample (a) in Figure 9.1, γ_a belongs to the sub-concept c_{de}^1 in δ_1 , but it does not belong to any sub-concept in δ_2 . Based on Algorithm 4.1, the symbol vector for γ_a is set as follows: In the concept layer, using the graded membership function (defined in Definition 4.2): $\mathcal{V}_{\gamma_a, c}(C_{de}) = (\text{'Decreasing'}, 1)$. In the quality layer, for the quality dimensions of δ_2 , using the graded quality function (defined in Definition 4.3): $\mathcal{V}_{\gamma_a, \Omega}(q_{\Delta t}^2) = (\text{'long'}, 0.75)$, $\mathcal{V}_{\gamma_a, \Omega}(q_{\Delta mm}^2) = (\text{'medium range'}, 0.62)$, and $\mathcal{V}_{\gamma_a, \Omega}(q_{fft}^2) = (\text{'fluctuate'}, 0.7)$.

9.2.2 Inference in Symbol Space of Patterns

By retrieving the information of symbol vector $\mathcal{V}(\gamma_a)$, it is possible to verbalise the elements of symbol vectors into a set of natural language descriptions. As mentioned in Section 4.2.2, γ_a is annotated by the values of $\mathcal{V}_{\gamma_a, e}$, and characterised by the values of $\mathcal{V}_{\gamma_a, c}$. In particular, the annotation set is $\mathcal{T}_C(\gamma_a) = \text{'Decreasing'}$, and the characterisation set will be $\mathcal{T}_Q(\gamma_a) = \{ \text{'long'}, \text{'medium range'}, \text{'fluctuate'} \}$. Then the realisation for γ_a is as follows: $\mathcal{T}_{\gamma_a} = \text{'Decreasing, also fluctuates and it is long with medium range'}$. Table 9.1 present more results derived from the semantic inference in the conceptual space for the time series patterns shown in Figure 9.4.

Patterns	Linguistic Description
Fig. 9.4(a)	<i>This pattern is an Increasing pattern, but it is smooth and very short, within a medium range of values.</i>
Fig. 9.4(b)	<i>This pattern is like a Spike pattern, but it is noisy with a sharp decreasing trend in a high range of values.</i>
Fig. 9.4(c)	<i>This pattern is an Oscillation pattern, within a very short range of values.</i>
Fig. 9.4(d)	<i>This pattern is an Increasing pattern, but it fluctuates in a very long duration, within a large range of values.</i>
Fig. 9.4(e)	<i>This pattern is an Increasing pattern, but it fluctuates within a medium range of values.</i>
Fig. 9.4(f)	<i>This pattern is not like any known pattern, but it fluctuates within a high range of values.</i>
Fig. 9.4(g)	<i>This pattern is an Increasing pattern, but it is long, within a medium range of values.</i>
Fig. 9.4(h)	<i>This pattern is like Decreasing Spike pattern, in a short duration.</i>
Fig. 9.4(i)	<i>This pattern is like a Spike pattern, but it has a sharp rise and fluctuation, within a medium range of values.</i>
Fig. 9.4(j)	<i>This pattern is a Decreasing pattern, but it fluctuates in a long duration, within a medium range of values.</i>
Fig. 9.4(k)	<i>This pattern is a Spike pattern, with the same start and end values.</i>
Fig. 9.4(l)	<i>This pattern is like a Spike pattern, but it has a smooth and slow increasing trend, within a medium range of values.</i>
Fig. 9.4(m)	<i>This pattern is not like any known pattern, but it wavy within a medium range, and very long duration. Also, it has the same start and end values.</i>
Fig. 9.4(n)	<i>This pattern is not like any known pattern, but it has a normal decreasing trend in a very long duration. Also, it is very fluctuating in a large range of values.</i>
Fig. 9.4(o)	<i>This pattern is like Decreasing Oscillation pattern.</i>
Fig. 9.4(p)	<i>This pattern is like a Spike pattern, but it is very short and smooth, within a high range of values</i>
Fig. 9.4(q)	<i>This pattern is like Spike and Decreasing patterns.</i>

Table 9.1: The linguistic descriptions derived for the unknown samples of time series patterns in Figure 9.4.

9.3 Evaluation: Descriptions from Conceptual Spaces vs. Other Semantic Models

As described in Section 5.3 in Part I, assessing the benefits of the proposed conceptual space representation directly is not a trivial problem. Instead, the usefulness of the constructed conceptual space of the time series patterns has been evaluated via the linguistic descriptions derived from such space. Again, the experiment evaluates the following aims: (1) to measure the *feasibility* of deriving *accurate* descriptions to *distinguish* unknown pattern observations, and (2) to assess the *goodness* of the descriptions derived from conceptual spaces in comparison to the descriptions derived from other base-line models. To these ends, a survey was conducted in which participants were asked to

1. identify specific time series pattern based on their linguistic description derived from the conceptual space, and
2. rate the goodness of descriptions produced by different models on a Likert scale.

9.3.1 Survey: Design and Procedure for Pattern Data Set

Similar to the survey used for leaf data set, the main body of the survey for patterns was composed of two parts, with the same designs of questions explained in Section 5.3.1. Here, it is worth to mention that the evaluation considered the results of 17 unknown patterns from a pool of unknown pattern examples. Moreover, among all responses to the survey, 89 valid responses have been studied for the pattern data set. Most of the participants were in the range of 25-44 years old, and they were mostly educated in computer science or equivalent. Besides, most of the participants were fluent in English speaking.

About the expertise level of the participants, the results show that the participants were more familiar with the terminology that have been used for *pattern* data set, rather than *leaf*. As shown in Figure 9.5, for the *leaf* data set, 20% of the participants knew none of the lexical items, 70% knew few or some of them, and only 10% almost all of them (See Figure 9.5a). But for the *pattern* data set, 30% of the participants knew few, or some of the lexical items, more than 40% knew most of them, and more than 25% knew all the introduced terminology (See Figure 9.5b). Comparing the percentages of the expertise level shows that the lexicon used in the descriptions of the patterns was more familiar to the participants.

9.3.2 Identifying Pattern Observations from Linguistic Descriptions

Participants were able to successfully identify all the unknown observations (17 patterns) with the help of the corresponding conceptual space descriptions.

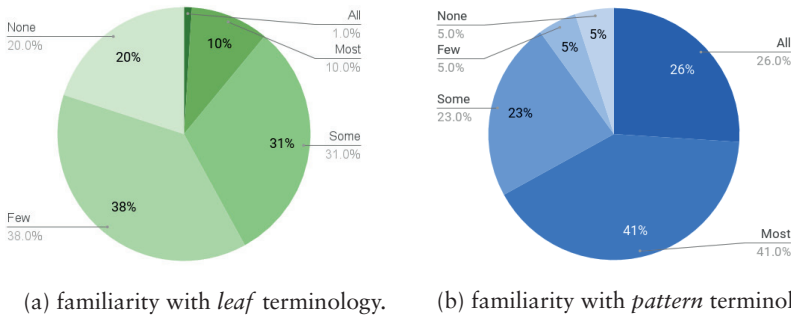


Figure 9.5: Pie charts, showing the expertise level of the participants by measuring how familiar they are with the introduced terminology for class labels and features in (a) *leaf* and (b) *pattern* data sets. Participants’ choices were: I am familiar with none/a few/some/most/all of the terminology.

The success rate to identify the correct image for each description in the *pattern* data set was $79\% \pm 11\%$. As mentioned in Section 5.3, this success rate for *leaf* data set was $73\% \pm 13\%$. The difference is reasonable to assume that the higher familiarity of participants with the pattern data set is a possible explanation for the better success rates.

For further investigation of the incorrectly identified (i.e., misidentified) examples, the geometrical similarity of these answers to the correct one is calculated in the conceptual space (multi-domain). According to [86], the similarity in conceptual spaces can be calculated by applying Euclidean distance with the domains and the city-block distance between them. To assess the similarities in conceptual space, also the geometrical similarity of the same instances is calculated but in a full feature space (single-domain) by applying Euclidean distance. Two interesting results have been obtained: First, the misidentified examples are not uniformly distributed between all possible choices, but instead, participants tended to make similar mistakes. Second, the common misidentified examples are most of the times (76% for patterns) the closest instance to the correct one in the conceptual space. In the full feature space this was only occasionally true (29% for patterns). This shows that the confused examples with each other are commonly the nearest instances within the multi-domain conceptual space, which is mostly not true in the full feature space.

The results from the first part of the survey show that the proposed conceptual space representation a) is applicable to derive semantic descriptions for unknown pattern observations, and b) is suitable to represent the cognitively similar pattern observations among the multiple domains.

Table 9.2: The overall scores calculated from the rating responses to the different models in *pattern* data set. The numbers show average scores (and standard deviations) in the range of 1 to 5.

	Mean (SD)
<i>Conceptual</i>	3.76 (1.07)
<i>Generative</i>	3.18 (1.29)
<i>Discriminative</i>	3.36 (1.22)

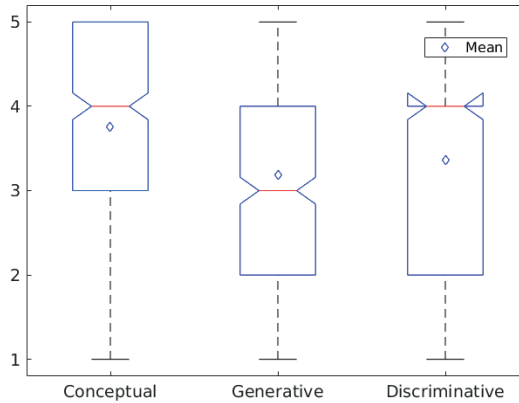


Figure 9.6: The box plot of the rating scores received for each of the models deriving descriptions in *pattern* data set.

9.3.3 Rating Various Linguistic Descriptions of a Pattern Observation

In the results of the rating scale questions, the description derived from the conceptual space model is compared with the descriptions derived from the two other semantic models that are explained in details in Section 5.3.3. Table 9.2 shows the statistical summary of the rating scores received for the descriptions derived from each of the approaches (*Conceptual*, *Generative* and *Discriminative*) in *pattern* data set. Also, these scores are depicted in the form of box plot in Figure 9.6.

Similar to the analysis for *leaf* data set, an ANOVA test has been applied to show that the conceptual space description (*Conceptual*) is significantly preferable rated than the two alternatives (*Generative* and *Discriminative*). The one-way ANOVA test showed a significant effect of the models on the scores. For the *pattern* data set, *Conceptual* has the mean significantly different from *Gen-*

Table 9.3: Summary of the one-way ANOVA and Wilcoxon tests for the rating scores with respect to the models deriving descriptions.

ANOVA Test	<i>Conceptual</i> vs. <i>Generative & Discriminative</i>	pattern data set
		$F(2, 1173) = 23.72,$ $p < 10^{-13}$
Wilcoxon Test	<i>Conceptual</i> vs. <i>Generative</i>	$p < 10^{-12}$
	<i>Conceptual</i> vs. <i>Discriminative</i>	$p < 10^{-07}$
	<i>Generative</i> vs. <i>Discriminative</i>	$p > 0.05$ (*)

erative and *Discriminative*, $p < .0001$ (two-tailed). The details of the test has been shown in Table 9.3.

Moreover, since the ratings are ordinal, also a non-parametric test (i.e., *Wilcoxon Test*) is carried out to identify the significant differences between ratings by comparing each pair of the scores. Table 9.3 shows the p-values of this method for each pair of models. The output showed that *Conceptual* model is significantly different from *Generative* and *Discriminative* models ($p < .0001$). Beside the significant difference of *conceptual* model, an interesting outcome of the tests is the scores of *Generative* and *Discriminative* for two data sets. In the *leaf* data set, participants have given higher scores to *Generative* than *Discriminative* ($p < .0001$). But in the *pattern* data set, there is no significant difference between these two models ($p > .05$). It can be interpreted that beside *Conceptual* which is the most preferred description, subjects preferred to see all the features as the description for leaf samples. But about the pattern samples, there is no preference between the descriptions including either the features or the concept labels related to the shown patterns.

Overall, the results from this part of the survey show that the proposed conceptual space representation a) is an appropriate semantic inference model to derive linguistic descriptions for unknown pattern observations, and b) successfully derives descriptions (from multi-domain space) that are naturally preferred by participants, in comparison to the other alternative models (from single-domain space).

9.4 Discussion and Summary

This chapter has presented the process of applying the semantic representation proposed in Part I on the physiological pattern data sets. This process automatically constructs the conceptual space of the abstracted physiological patterns, and then, it infers semantic descriptions for a set of unknown patterns. The constructed conceptual space of patterns involves two domains that include five quality dimensions in total. This space represents four class of patterns: Increasing, Decreasing, Spike, and Oscillation. As it was expected before applying the data-driven approach to construct the conceptual spaces, the concepts of increasing and decreasing have been represented in the same domain.

Beside the '*slope*' feature, the other selected quality dimension (i.e., '*start-end diff*') to represent these concepts was not the obvious choice. However, it seems that the values of this feature were dominant enough to play the distinguish the Increasing and Decreasing concepts from the rest of the concepts. The other domain that represents Spike and Oscillation concepts also contains interesting quality dimensions. The '*time interval*', '*min-max diff*', and '*frequency*' are the features that are more or less the obvious choices to distinguish or describe these two concepts. However, as one can see, some features like '*entropy*' or '*standard deviation*' have not been appeared in the conceptual space, meaning that their values were not good enough (in comparison with the other features) to separate any concept from the rest.

Furthermore, this chapter has performed an empirical evaluation, similar to the assessment presented in Chapter 5, to show the goodness of the conceptual space of the patterns used for text generation. This assessment has conducted a survey by asking the human subjects to identify the instances by reading the generated texts by conceptual spaces model, along with rating three different generated texts by different semantic models to compare the output text of the physiological patterns.

In sum, this chapter has shown the ability of the semantic representation approach proposed in Part I to infer linguistic descriptions for physiological sensor patterns, which are not necessarily known for the end user. An empirical evaluations was performed to evaluate the goodness of the generated texts.

Chapter 10

Conclusions

“And you may ask yourself: Well, how did I get here?”

— Talking Heads (1975–1991)

THIS chapter summarises the contributions of this thesis. It provides an overview of the limitations of different aspects presented in the thesis. Finally, the chapter concludes by presenting an outlook of the future research in different directions.

10.1 Summary of Contributions

The overall contribution of this thesis has been to provide different data-driven strategies for mining and representing numerical data into a semantic representation and then generate linguistic descriptions for such information. The process, in summary, has been presented in three steps: 1) extracting and mining numerical information (e.g., physiological sensor data), 2) modelling the information in a semantic representation, and 3) generating linguistic descriptions for such information.

To present the achievements of this thesis, the rest of this section revisits the introduced contributions (C1 to C4), given in Chapter 1 by explaining how each of the contributions has been accomplished using the proposed approaches.

10.1.1 Construction of Conceptual Spaces (C1)

This thesis has first introduced a data-driven approach to automatically construct conceptual spaces based on the input observations and their semantic attributes (Chapter 3). The following contributions have addressed the task of constructing a conceptual space:

1) The specification of domains and quality dimensions was presented using feature selection and feature grouping methods, which exploit the relevance of the selected features to the known concepts. The approach has applied a feature selection method based on mutual information for feature selection (MIFS) to rank the relevance of features and concepts. This ranking algorithm has been applied to each concept individually by applying the MIFS method on the input data set. After that, grouping the selected features has been introduced through first constructing a bipartite graph representation of the feature-concept associations, and then exploiting the most representative subsets of features as the domains by finding the best bicliques in such graph.

2) The concept representation was described in an instance-based manner. To form the concepts within the specified domains, the approach has calculated the convex regions of concepts and the salient weights of concept in relation to the quality dimensions of the domains. Note that this calculation was entirely formulated based on the associated observations to the concepts, without involving external knowledge.

A key finding in this contribution is that the proposed approach to construct conceptual spaces provides a generalisation for concept representation, where this representation can be derived from different types of input instances.

10.1.2 Semantic Inference in Conceptual Spaces (C2)

This thesis has introduced a semantic inference process to linguistically represent a new observation within the built conceptual space (Chapter 4). The following contributions have addressed the task of semantic inferring in a conceptual space:

1) A symbol space was introduced as the complementary space to the conceptual space, which includes the semantics of the corresponding concepts and quality dimensions. This space enables the approach to determine the relevant symbolic terms to represent a new observation.

2) The inference process to generate linguistic descriptions for an unknown observation was presented in two phases: To determine associated concepts and quality dimensions of an unknown instance, its location within the constructed conceptual space has been investigated. This determination has been done by considering the inclusion of the instance in the regions of the space and the use of similarity measures in such space. After that, the lexicalisation of the instance has been induced by extracting the semantic labels of the associated concepts and quality dimensions. Finally, microplanning and realisation techniques have been applied in order to generate natural language descriptions.

One advantage of such inference model is that the proposed approach concerns which features and concepts should be inferred as the most suitable set of interpretable contents to describe an unknown observation.

10.1.3 Mining Prototypical Patterns and Temporal Rules in Physiological Sensor Data (C3)

From the application point of view, this thesis has focused on the field of health-care monitoring to analyse the physiological sensor data. The data analysis phase has addressed the task of mining partial trends, prototypical patterns and distinctive temporal rules using data-driven approaches. The following contributions have addressed the task of mining physiological sensor data:

1) A literature review on mining physiological sensor data was presented. This review has considered several health monitoring systems that use wearable sensors to monitor the vital signs. Different data mining tasks for healthcare systems have been studied, as well as various machine learning techniques to address such tasks (Chapter 6).

2) This thesis has introduced an unsupervised approach to extract prototypical patterns from a physiological time series data. This approach has been designed based on desensitising the time series and clustering the sub-sequences of data to exploit the final patterns. Besides, a partial trend detection method has also been proposed to capture the partial behaviours of a time series concerning the shape and trend of the data (Chapter 7).

3) The central data analysis part of this work was the process of mining temporal rules from various channels of sensor data in clinical conditions. This process has introduced a new temporal rule mining method to extract the repeated co-occurrences of the physiological patterns in a set of long recorded sensor data. The significant output of such temporal rule mining was the fact that the extracted rules from the data of each clinical condition are distinguishable from the temporal rules that are derived from other conditions. A new approach proposed to compare temporal rules has confirmed this uniqueness of rules in each condition (Chapter 8).

The key outcome of the data analysis phase in this thesis is that the entire process relies on the measured data itself to express the valuable information. The proposed data-driven approaches have shown that there are some aspects of the sensor data (i.e., unseen patterns and temporal rules) that are not known or not readily observable by the domain experts, but they are still interesting to be extracted and moreover are valuable to be interpreted.

10.1.4 Linguistic Description of Time Series Patterns using Semantic Representations (C4)

After data analysis phase, this thesis considered the task of representing numerical information into linguistic descriptions. This representation has been done using both template-based approaches and the proposed semantic representation based on data-driven conceptual spaces. The following contributions have addressed the task of describing physiological sensor data:

1) A template-based linguistic description approach was presented to turn the partial trends and patterns into a set of natural language terms. Also, for the extracted temporal rules, this thesis has applied a fuzzy mapping method to annotate the co-occurrence of the patterns and also the frequency of their happening using linguistic terms, and then it has generated natural language sentences for the rules in each clinical condition (Chapter 8).

2) Applying the proposed semantic representation using data-driven conceptual spaces was one of the leading contributions for describing time series patterns. The proposed approach in the first part of this thesis has been applied to a collection of processed known time series patterns as input numerical information. Then, using the constructed conceptual space of time series patterns, the inference process has generated linguistic descriptions for a set of unknown time series patterns. The presented empirical evaluation has shown the goodness of the generated texts using the proposed semantic over the introduced generative and descriptive models (Chapter 9).

The advantage of applying the semantic model to the physiological data is to be able to interpret data-driven extracted patterns that are unknown by definition. Involving semantic representation helps the system to not be restricted to analyse and mine only the pre-defined patterns requested by the domain expert, but to search for any interesting information and be sure that the semantic model can generate a description for such information.

10.2 Limitations

The approaches proposed in both parts of this thesis have presented data-driven strategies (numerically, and semantically) for linking numerical information to linguistic descriptions. The discussions given at the end of each chapter have shown the key points and critical issues of the contributions. In this section, a list of more general issues and limitations related to the accomplished contributions are discussed.

Construction of conceptual spaces Regarding the proposed approach for construction of conceptual spaces, one limitation is the assumption of having *semantic features* as inputs. These features are assumed to be understandable or be interpretable by the human, which means they can be used as semantic terms in the final linguistic descriptions. Besides, another assumption for the input of constructing conceptual spaces is the set of labelled or annotated observations with known classes or concepts. With this assumption, the process of concept forming can be categorised as a supervised modelling, since the labelled data leads the process of the domain specification and concept representation.¹

¹Note that regardless of being supervised or not, still the approach is data-driven in a sense that there is no extra knowledge rather than the known input instances to influence the processes of constructing the space.

Another limitation of constructing conceptual spaces is related to the proposed algorithms. The algorithms for the domain and quality dimension specification have been introduced heuristic approaches to filter and to group the features, which are not the optimal ways to do so. Therefore, there is a lack of studying on how to measure the goodness of the specified domains in comparison with other non-greedy approaches. It is worth mentioning that these algorithms have not aimed to perform classification task to discriminate the classes of data with high accuracy. Instead, they have attempted to select the most distinctive features for each class of data to enrich the descriptivity of the model by presenting multi-domain space. Therefore, it might be meaningless to apply classification measures like recall or precision in order to measure the goodness of the model.

Semantic inference in conceptual spaces: Regarding the semantic inference in conceptual spaces, the structure of the introduced symbol space is dependent on domain knowledge, which is the set of linguistic annotations and symbols of the input features and concepts. An important point to mention is that there is no inference to exploit or generate a new semantic label or term by reasoning among the relation of the provided symbols or words based on any knowledge-based system (e.g., ontologies). Instead, the inference process chooses the most representative terms and labels for a new unknown observation among all the already provided information.

Another constraint in the inference process is related to the unknown observations as the inputs. The unknown observations should have the same properties that the known observations have. More precisely, the unknown observations should be able to be vectorised within the constructed conceptual space. Thus, any unknown observation with missing data or out-of-range values will be problematic in the current version of the proposed inference approach.

The linguistic description task in the inference process has also some limitations. The main one is that the semantic role of the concepts, sub-concepts, and quality dimensions are reduced or simplified to some specific linguistic roles in a sentence. The concepts are restricted to be the nouns and sub-concepts or properties are constrained to be the adjectives in the final realisation task. Regarding the quality of the inferred descriptions, one limitation is that the generated text is subject to questions related to some criteria such as conciseness versus verbosity of the descriptions, the target audience of the system, and the ultimate aim the text is expected to support.

Case study and evaluation: Regarding the evaluation of the semantic representation approach, it is not trivial to find a general solution to evaluate each component or step of the approach, separately. For the construction part, there is not yet a measure to compare the constructed conceptual spaces in term of applicability or sufficiency. Moreover, the inference process within the conceptual space cannot be isolated and be evaluated individually. Therefore, this thesis

has assessed the final output of the inference step in order to evaluate the goodness of multi-domain model in comparison with the single-domain models. This comparison accesses the idea of constructing *multi-domain* conceptual spaces out of numerical data, but not the individual components of the construction process.

Mining physiological sensor data: There are some limitations related to data analysis of physiological data. First of all, the source of data itself is a challenge. There are few available data sets including the vital signs recorded for a long time from different subjects. This limitation has made the constraint to only test the approach on open access data sets, rather than collecting data via wearable sensors (e.g., within a controlled environment). The main reason to rely on long-term data is that the proposed algorithms for pattern abstraction and temporal rule mining will return meaningful information when the input is a long stream of the sensor data with minimum interruptions or noises.

Regarding the temporal rule mining, the presented approach has considered the frequency of the co-occurrence of the patterns throughout the different channels of data. The current approach has only provided rules which are temporal correlations of the patterns from the co-occurrence aspects. Thus, there is a lack of considering other ways of analysing data to capture more meaningful information, such as the causes and effects of the patterns.

Linguistic description of physiological patterns: The main limitation of the proposed approach for describing time series patterns is the richness of the provided features. In general, there is a limited number of semantic features, representing the behaviour of the time series patterns, in the literature. In addition, most of the features to analyse the time series are complicated to be interpreted or are not relevant to the context. For example, wavelet coefficients are the useful features for numerical time series analysis, but they are complex to be translated into understandable natural language terms. Some features like the area under the signal might be explainable but are not interpretable or meaningful when e.g., the shape and the trend of the time series are in the focus. So, the proposed conceptual space of the patterns is limited to a small number of input features which are more or less understandable by the human subjects to show the behaviour of the time series.

10.3 Societal and Ethical Impacts

Nowadays sensors are everywhere to collect various kinds of information. Understanding and interpreting this information is a significant challenge, especially when the information is related to crucial aspects of people's life. This thesis has provided a system to process and describe unknown observations derived from sensor data. Within a society, this research might be applicable in different scenarios. In any situation that the perceived information is not explic-

itly recognisable, the proposed system can help to interpret those observations. An example of the usage of the system in society could be to help particular groups of people (e.g., children, persons with disabilities) in order to make sense about the observations around them. This goal can be achieved by coupling the approach with other perception or/and decision making systems. The impact of the proposed approach will be more critical in the medical domain since it can help the patients of clinicians to see new aspects of the perceived information that have been unknown beforehand.

Besides the societal impacts of this research, several ethical impacts must be considered. The ethical issues in this thesis can be seen from three perspectives: input data, proposed approach, and output text. Regarding the input data, the primary ethical challenge (especially in the medical domain) is the anonymity of the subjects who use the sensors. Systems that use the recorded data from users should ensure to protect the privacy of them under regulations such as the General Data Protection Regulation (GDPR) [2]. In this thesis, all the acquired physiological sensor data have been kept anonymous. The ethical issue related to the approach is the problem of blindly applying data-driven strategies. A potential risk within the data-driven approaches is that the constructed models ((e.g., learning or decision making models) can be easily biased if the input observations are biased. Thus, one might think about which observations are suitable to be fed to such blind models. Last but not least, ethical issues are related to the output linguistic descriptions. A generated text for a specific end-user might involve ambiguous, inadequate content, or even might misplace the provided contents in the sentence. These problems can lead to critical issues such as misinterpreting the content of the text, and consequently making inaccurate or incorrect decisions. This thesis has not directly addressed this ethical issue in its case studies, but it is worth to keep it in mind for further developments upon such a framework.

10.4 Future Research Directions

Besides the limitations of the proposed approaches mentioned above, there is a number of future research directions that require further investigation. This section presents the future work regarding each part of the thesis.

Conceptual Spaces: Construction and Inference

One possible direction of the future work is to focus on acquiring semantic features that are contextually understandable by the human and are able to differentiate distinct concepts in the domain [77, 96, 217]. There has been a preliminary ongoing study on this topic by the author of this thesis, which needs to be extended and be evaluated. Two possible approaches can be applied for the task of semantic feature acquisition. One way is to specify the attributes coming from the human perceptions of instances that bridges be-

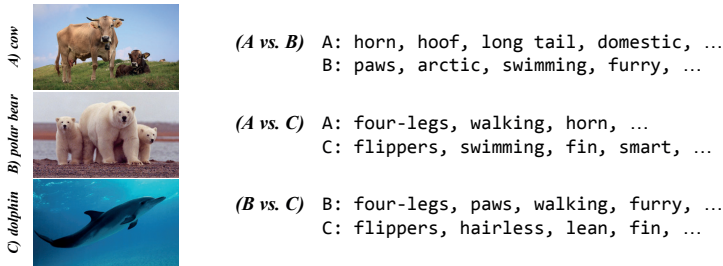


Figure 10.1: Human-specified semantic features for *animal* domain.

tween low-level features and linguistic words [204]. Directly asking the experts or users to specify the attributes is an obvious way to provide both understandable and discriminative features. As an example one can differ between different *animals* by specifying the semantic features like *hair type*, *domestic/wild*, or *having horns* (See Figure 10.1 for some examples of human-specified semantic features). Another way is to verify the attributes that are scientifically measurable and potentially interpretable in natural language but are not obvious to specify by humans in the first glance. Such examples of these features for *animal* domain are *agility* or *hibernation* which are discriminative features for categorising animal species [135]. These ways of acquiring semantic features then guarantee that the input features to the construction of conceptual spaces are human explainable information that can be later used for inferring linguistic descriptions of unknown observations as well.

Construction of conceptual spaces has the advantage of using labelled input observations to form the concepts. Another way to look at the problem of turning numerical data into the conceptual spaces is to deal with unlabelled data sets. In this case, there will be no pre-defined classes or concepts for the data. An unsupervised approach to specify the domains and quality dimensions can be a new way to build a data-driven conceptual space. This suggestion might need to use clustering methods and clustering index criteria to form the clusters of data as concepts that are not labelled by any specific term, but still are explainable by their quality dimensions.

Moreover, the proposed concept representation provides a set of geometrical domains, which has been used in the inference process. This representation has the potential to be utilised for other cognitive tasks such as concept combination, inductive inference, and property reasoning. This would be a promising research direction, especially when it comes to data-driven representations of the cognitive architectures.

While evaluating the proposed approach, a new hypothesis has been raised related to the topic of *referring expression generation* in the NLG area. The idea of comparing the descriptions from multi-domain spaces versus from single-

domain spaces brings up the question of referring to an unknown object by whether describing its features or mentioning its closest known concepts. From the cognitive point of view, one question is how humans prefer to use the features in order to refer to an object. One possible solution is to use a mixture of known associated concepts and the relevant descriptive features. For example, in the domain of animals, one can describe a Unicorn (an unknown animal let's say) as “*This animal is like a Horse (a known class of animals), but it has a single large horn (a feature for animals) on its head.*” Throughout the evaluation of this work, some textual results have been provided to support this hypothesis, along with some comparisons using an empirical evaluation. However, this hypothesis is not formalised as a new aspect of referring expression generation in NLG.

Another direction for the future work would be to assess the quality of the automatically derived domains and dimensions. As Gärdenfors discussed in [88], there is a need to determine evaluation criteria to choose among competing conceptual spaces. The proposed framework has the potential to address this need by defining statistical measures to compare the specified domains in a data-driven manner.

Physiological Sensor Data: Mining and Describing

Turning to the data analysis of the physiological sensor data, there are many future research directions to improve or extend the proposed approaches. One of the critical updates might be to extend the pattern abstraction approach to deal with newly recorded data in a streaming data set. In the current version, the prototypical patterns are abstracted by analysing the entire time series at once. It could be interesting to extend the algorithm to incrementally update the patterns using the new bunch of recorded data. This extension will also help to handle any large-scale recording sensor data that is crucial for rule mining approach.

Additionally, involving the causality to the process of rule mining is another direction of research, which impacts profoundly on the level of explainability of the derived information. One extension would be to use data-driven ways to identify causes and effects of the patterns' behaviours using causal inference approaches and then describing them in the form of e.g., *if-else* statements. This identification will help the system to obtain more enriched (still unseen, but more interesting) information.

The proposed approach for generating linguistic descriptions of the physiological patterns has the ability to be improved in future research. First, it would be worth to compare the generated text by conceptual spaces with the template-based generated text. Although these two approaches aim to capture the same behaviours of the pattern, still they might be differently accepted or used by the end user. Second, there is a lack of extrinsic evaluation of the framework in a real-world application within the medical domain to see the actual

impact of the generated text to experts such as clinicians, medical doctors, or even caregivers. This task-based evaluation is needed to assess the usability of the generated text for patterns in the real-world, meaning that how much this data-driven information (in the form of natural language) are interesting or/and helpful to the end users for any further decision making task. Here, the focus in the evaluation studies has largely been on identification of the observations, but adapting the proposed methods to other uses, such as decision support in a medical setting, or other audiences, such as experts versus non-experts, would be an interesting road for future work.

In a more general perspective, the approaches for generating linguistic descriptions and natural language generation can be more involved in the field of healthcare monitoring. Besides considering the wearable sensor data, there is much other information within a healthcare system that will be more acceptable by using the natural language to explain them. This information such as medical history, environmental sensors, activity records, personal reports, etc. (which are often ontological knowledge) can be merged with the data-driven information in order to analyse the health monitoring scenarios and to perform high-level reasoning for medical purposes. Then, it would be worth to use approaches to generate linguistic descriptions in order to explain such inferred or reasoned information in a natural language which is understandable by humans.

10.5 Final Words

To conclude this thesis, let's revisit the story of *The Elephant in the Dark* and the problem of perception limitations or, as called in this thesis, the problem of concept description. The proposed approach in this thesis made valuable contributions to address this problem by applying data-driven methods to link unknown perceived data to understandable linguistic descriptions. This thesis can be called a success if a reader of the thesis will be able to “adequately” describe an elephant to a person who has neither encountered the notion of an elephant previously or seen one in real life. Hopefully, such a description would be better than the one shown in Figure 10.2.

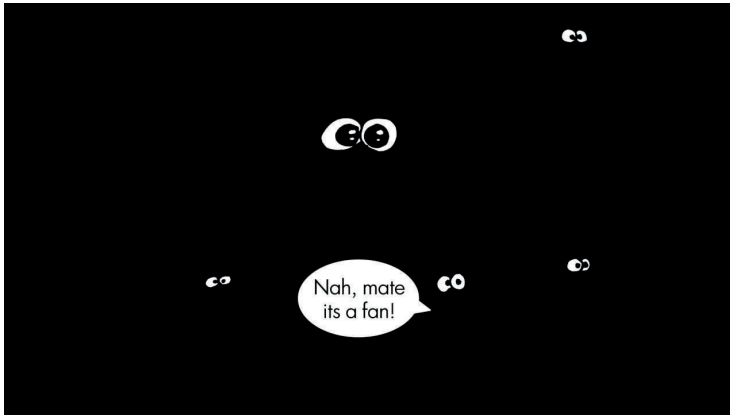


Figure 10.2: Yet another illustration for the story of *The Elephant in the Dark*, adapted from [194].

Bibliography

- [1] Blind men and the elephant. Available online: <http://www.allaboutphilosophy.org/blind-men-and-the-elephant.htm> (Accessed March 3, 2018). (Cited on page 2.)
- [2] Eu general data protection regulation (gdpr). Available online: <https://www.eugdpr.org/>. (Accessed March 5, 2018). (Cited on page 159.)
- [3] Mimic II data set. Available online: physionet.org/physiobank/database/mimicdb/numerics. (Accessed March 3, 2018). (Cited on page 107.)
- [4] Physiobank archive index. Available online: <http://www.physionet.org/physiobank/database/> (Accessed March 3, 2018). (Cited on page 99.)
- [5] Zephyr™ bioharness3. Available online: <http://www.zephyr-technology.nl/en/product/71/zephyr-bioharness.html>. (Accessed April 10, 2013). (Cited on pages x and 106.)
- [6] Zephyr™ strap. Available online: <https://www.zephyranywhere.com/online-store>. (Accessed November 29, 2016). (Cited on page 106.)
- [7] *The Heath Readers by Grades, Book Five*. (public domain image on page 69), Boston: D.C. Heath & co., Boston, USA, 1907. (Cited on pages ix and 1.)
- [8] Benjamin Adams and Martin Raubal. Conceptual space markup language (csml): Towards the cognitive semantic web. In *IEEE International Conference on Semantic Computing (ICSC) 2009*, pages 253–260, 2009. (Cited on pages 25 and 51.)
- [9] Benjamin Adams and Martin Raubal. A metric conceptual space algebra. In *Spatial information theory*, pages 51–68. Springer, 2009. (Cited on pages 21, 35, 47, 56, 62, and 66.)

- [10] Francesco Agostaro, Agnese Augello, Giovanni Pilato, Giorgio Vassallo, and Salvatore Gaglio. A conversational agent based on a conceptual interpretation of a data driven semantic space. In *Congress of the Italian Association for Artificial Intelligence*, pages 381–392. Springer, 2005. (Cited on pages 4, 33, and 35.)
- [11] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993. (Cited on page 123.)
- [12] David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991. (Cited on page 26.)
- [13] Nor Faizah Ahmad, Doan B. Hoang, and M. Hoang Phung. Robust preprocessing for health care monitoring framework. In *Proceedings of the 11th international conference on e-Health networking, applications and services*, pages 169–174, Piscataway, NJ, USA, 2009. IEEE Press. (Cited on pages 97 and 100.)
- [14] Janet Aisbett and Greg Gibbon. A general formulation of conceptual spaces as a meso level representation. *Artificial Intelligence*, 133(1):189–232, 2001. (Cited on pages 3, 20, 21, 24, 25, 26, 35, 53, 54, 55, and 62.)
- [15] Janet Aisbett, John T Rickard, and Greg Gibbon. Conceptual spaces and computing with words. In *Applications of Conceptual Spaces*, pages 123–139. Springer, 2015. (Cited on page 33.)
- [16] Ahmad A. Al-Hajji. Rule-based expert system for diagnosis and symptom of neurological disorders "neurologist expert system (nes)". In *Proceedings of the 1st Taibah University International Conference on Computing and Information Technology*, pages 67–72, Buraydah, March 2012. Qassim University. (Cited on page 98.)
- [17] Marjan Alirezaie and Amy Loutfi. Automatic annotation of sensor data streams using abductive reasoning. In *5th International Conference on Knowledge Engineering and Ontology Development*, September 2013. (Cited on page 103.)
- [18] James F Allen. Towards a general theory of action and time. *Artificial intelligence*, 23(2):123–154, 1984. (Cited on page 121.)
- [19] Jens S Allwood and Peter Gärdenfors. *Cognitive semantics: Meaning and cognition*, volume 55. John Benjamins Publishing, 1999. (Cited on page 20.)

- [20] Miguel R Álvarez, Paulo Félix, and PurificaciÓN CariñEna. Discovering metric temporal constraint networks on temporal databases. *Artificial intelligence in medicine*, 58(3):139–154, 2013. (Cited on page 122.)
- [21] Filippo Amato, Alberto López Rodríguez, Eladia María PeñaandMéndez, Petr Vaňhara, Aleš Hampl, and Josef Havel. Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, 11:47–58, 2013. (Cited on page 97.)
- [22] Daniele Apiletti, Elena Baralis, Giulia Bruno, and Tania Cerquitelli. Real-time analysis of physiological data to support medical applications. *IEEE Transactions on Information Technology in Biomedicine*, 13(3):313–321, May 2009. (Cited on pages 96, 97, 99, and 100.)
- [23] Louis Atallah, Benny Lo, and Guang-Zhong Yang. Can pervasive sensing address current challenges in global healthcare? *Journal of Epidemiology and Global Health*, 2(1):1–13, 2012. (Cited on page 92.)
- [24] Akin Avci, Stephan Bosch, Mihai Marin-Perianu, Raluca Marin-Perianu, and Paul Havinga. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In *Proceedings of the 23th International Conference on Architecture of Computing Systems*, pages 167–176, Berlin, February 2010. VDE Verlag. (Cited on page 96.)
- [25] Jody Azzouni. *Semantic perception: How the illusion of a common language arises and persists*. Oxford University Press, 2015. (Cited on page 20.)
- [26] Joonbum Bae and Masayoshi Tomizuka. Gait phase analysis based on a hidden markov model. *Mechatronics*, 21(6):961–970, 2011. (Cited on page 98.)
- [27] MirzaMansoor Baig and Hamid Gholamhosseini. Smart health monitoring systems: An overview of design and modeling. *Journal of Medical Systems*, 37(2):1–14, 2013. (Cited on pages 92 and 93.)
- [28] Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. *Sensors*, 13(12):17472–17500, 2013. (Cited on pages 4, 98, 105, and 106.)
- [29] Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. A framework for automatic text generation of trends in physiological time series data. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pages 3876–3881. IEEE, 2013. (Cited on pages 103 and 132.)

- [30] Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. Descriptive modelling of clinical conditions with data-driven rule mining in physiological data. In *HEALTHINF 2015: 8th International Conference on Health Informatics*, 12-15 january, Lisabon, Portugal, 2015. (Cited on pages 125 and 127.)
- [31] Hadi Banaee and Amy Loutfi. Using conceptual spaces to model domain knowledge in data-to-text systems. In *8th International Natural Language Generation (INLG) Conference*, 19-21 June, Philadelphia, Pennsylvania, USA, pages 11–15. Association for Computational Linguistics, 2014. (Cited on pages 32 and 42.)
- [32] Hadi Banaee and Amy Loutfi. Data-driven rule mining and representation of temporal patterns in physiological sensor data. *IEEE journal of biomedical and health informatics*, 19(5):1557–1566, 2015. (Cited on pages 4, 127, 139, and 140.)
- [33] Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of HLT-NAACL*, pages 113–120, 2004. (Cited on page 32.)
- [34] Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994. (Cited on page 39.)
- [35] Ildar Batyrshin, Leonid Sheremetov, and Raul Herrera-Avelar. Perception based patterns in time series data mining. In *Perception-based Data Mining and Decision Making in Economics and Finance*, pages 85–118. Springer, 2007. (Cited on pages 27, 33, 139, and 140.)
- [36] Ildar Z Batyrshin and LB Sheremetov. Perception-based approach to time series data mining. *Applied Soft Computing*, 8(3):1211–1221, 2008. (Cited on page 139.)
- [37] Riccardo Bellazzi, Fulvia Ferrazzi, and Lucia Sacchi. Predictive data mining in clinical medicine: a focus on selected methods and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5):416–430, 2011. (Cited on pages 92, 93, and 94.)
- [38] Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, 77(2):81–97, 2008. (Cited on pages 94 and 97.)
- [39] Christos C. Bellos, Athanasios Papadopoulos, Roberto Rosso, and Dimitrios I. Fotiadis. Extraction and analysis of features acquired by wearable sensors network. In *10th IEEE International Conference on Information Technology and Applications in Biomedicine*, pages 1–4, 2010. (Cited on pages 96 and 97.)

- [40] Christos C. Bellos, Athanasios Papadopoulos, Roberto Rosso, and Dimitrios I. Fotiadis. Categorization of patients' health status in copd disease using a wearable platform and random forests methodology. In *Proceedings of the IEEE International Conference on Biomedical and Health Informatics*, pages 404–407, 2012. (Cited on pages 95 and 98.)
- [41] Christos C. Bellos, Athanasios Papadopoulos, Roberto Rosso, and Dimitrios I. Fotiadis. A support vector machine approach for categorization of patients suffering from chronic diseases. In Konstantina S. Nikita, James C. Lin, Dimitrios I. Fotiadis, and Maria-Teresa Arredondo Waldmeyer, editors, *Wireless Mobile Communication and Healthcare*, volume 83, pages 264–267. Springer Berlin Heidelberg, 2012. (Cited on pages 95 and 101.)
- [42] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. (Cited on page 38.)
- [43] Anna Maria Bianchi, Martin Oswaldo Mendez, and Sergio Cerutti. Processing of signals recorded through smart devices: Sleep-quality assessment. *IEEE Transactions on Information Technology in Biomedicine*, 14(3):741–747, 2010. (Cited on pages 95 and 100.)
- [44] Fatih Emre Boran, Diyar Akay, and Ronald R. Yager. An overview of methods for linguistic summarization with fuzzy sets. *Expert Systems with Applications*, 61:356 – 377, 2016. (Cited on page 28.)
- [45] Sarah Boyd. Trend: a system for generating intelligent descriptions of time series data. In *IEEE International Conference on Intelligent Processing Systems (ICIPS1998)*. Citeseer, 1998. (Cited on page 111.)
- [46] João Mario Lopes Brezolin, Sandro Rama Fiorini, Marcia de Borba Campos, and Rafael H Bordini. Using conceptual spaces for object recognition in multi-agent systems. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 697–705. Springer, 2015. (Cited on page 25.)
- [47] Gavin Brown. A new perspective for information theoretic feature selection. In *International conference on artificial intelligence and statistics*, pages 49–56, 2009. (Cited on page 39.)
- [48] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(Jan):27–66, 2012. (Cited on page 40.)

- [49] Rita M Catillo-Ortega, Nicolás Marín, and Daniel Sánchez. A fuzzy approach to the linguistic summarization of time series. *Journal of Multiple-Valued Logic & Soft Computing*, 17, 2011. (Cited on page 28.)
- [50] Gavin C Cawley and Nicola LC Talbot. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition*, 36(11):2585–2592, 2003. (Cited on page 126.)
- [51] Marie Chan, Daniel Estéve, Jean-Yves Fourniols, Christophe Escriba, and Eric Campo. Smart wearable systems: Current status and future challenges. *Artificial Intelligence in Medicine*, 56(3):137–156, November 2012. (Cited on page 92.)
- [52] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15:1–15:58, July 2009. (Cited on page 93.)
- [53] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection for discrete sequences: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):823–839, May 2012. (Cited on page 93.)
- [54] Sylvie Charbonnier and Sylviane Gentil. On-line adaptive trend extraction of multiple physiological signals for alarm filtering in intensive care units. *International Journal of Adaptive Control and Signal Processing*, pages 382–408, 2009. (Cited on pages 93, 94, and 96.)
- [55] Samir Chatterjee, Kaushik Dutta, Harry (Qi) Xie, Jongbok Byun, Akshay Pottathil, and Miles Moore. Persuasive and pervasive sensing: A new frontier to monitor, track and assist older adults suffering from type-2 diabetes. In *Proceedings of the 2013 46th Hawaii International Conference on System Sciences*, pages 2636–2645, Washington, DC, USA, 2013. IEEE Computer Society. (Cited on pages 94, 98, 100, and 101.)
- [56] Antonio Chella, Marcello Frixione, and Salvatore Gaglio. A cognitive architecture for artificial vision. *Artificial Intelligence*, 89(1-2):73–111, 1997. (Cited on page 25.)
- [57] Antonio Chella, Marcello Frixione, and Salvatore Gaglio. Anchoring symbols to conceptual spaces: the case of dynamic scenarios. *Robotics and Autonomous Systems*, 43(2):175–188, 2003. (Cited on page 25.)
- [58] Hsinchun Chen, Sherrilynne S Fuller, Carol Friedman, and William Hersh. *Medical informatics: knowledge management and data mining in biomedicine*, volume 8. Springer Science & Business Media, 2006. (Cited on page 105.)

- [59] Yueguo Chen, Mario A Nascimento, Beng Chin Ooi, and Anthony KH Tung. Spade: On shape-based pattern detection in streaming time series. In *IEEE 23rd International Conference on Data Engineering (ICDE) 2007*, pages 786–795. IEEE, 2007. (Cited on page 124.)
- [60] Jongyoon Choi, Beena Ahmed, and Ricardo Gutierrez-Osuna. Development and evaluation of an ambulatory stress monitor based on wearable sensors. *IEEE transactions on information technology in biomedicine*, 16(2):279–286, 2012. (Cited on pages 94, 96, and 97.)
- [61] Lei Clifton, David A Clifton, Marco AF Pimentel, Peter J Watkinson, and Lionel Tarassenko. Gaussian processes for personalized e-health monitoring with wearable sensors. *IEEE Transactions on Biomedical Engineering*, 60(1):193–197, 2013. (Cited on pages 93, 98, 99, and 100.)
- [62] Carlo Combi and Alberto Sabaini. Extraction, analysis, and visualization of temporal association rules from interval-based clinical data. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 238–247. Springer, 2013. (Cited on page 121.)
- [63] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. (Cited on page 98.)
- [64] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. (Cited on page 70.)
- [65] Thomas M Cover and Joy A Thomas. Elements of information theory 2nd edition. 2006. (Cited on page 40.)
- [66] Richard Cubek, Wolfgang Ertel, and Günther Palm. High-level learning from demonstration with conceptual spaces and subspace clustering. In *IEEE International Conference on Robotics and Automation (ICRA) 2015*, pages 2592–2597. IEEE, 2015. (Cited on page 25.)
- [67] Gautam Das, King-Ip Lin, Heikki Mannila, Gopal Renganathan, and Padhraic Smyth. Rule discovery from time series. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining KDD’98*, pages 16–22, 1998. (Cited on page 121.)
- [68] Donald Davidson. Truth and meaning. *synthese*, 17(1):304–323, 1967. (Cited on page 20.)
- [69] Lieven Decock and Igor Douven. What is graded membership? *Noûs*, 48(4):653–682, 2014. (Cited on page 62.)

- [70] Miguel Delgado, M Dolores Ruiz, Daniel Sánchez, and M Amparo Vila. Fuzzy quantification: a state of the art. *Fuzzy Sets and Systems*, 242:1–30, 2014. (Cited on pages 27 and 28.)
- [71] Anne M Denton, Christopher A Besemann, and Dietmar H Dorr. Pattern-based time-series subsequence clustering using radial distribution functions. *Knowledge and Information Systems*, 18(1):1–27, 2009. (Cited on page 114.)
- [72] Joaquín Derrac and Steven Schockaert. Inducing semantic relations from conceptual spaces: a data-driven approach to plausible reasoning. *Artificial Intelligence*, 228:66–94, 2015. (Cited on page 33.)
- [73] Félix Díaz-Hermida, Martin Pereira-Fariña, Juan Carlos Vidal, and Alejandro Ramos-Soto. Characterizing quantifier fuzzification mechanisms: A behavioral guide for applications. *Fuzzy Sets and Systems*, 2017. (Cited on page 27.)
- [74] Hao Ding, Hong Sun, and Kun mean Hou. Abnormal ecg signal detection based on compressed sampling in wearable ecg sensor. In *International Conference on Wireless Communications and Signal Processing*, pages 1–5, 2011. (Cited on pages 98, 99, and 101.)
- [75] Włodzisław Duch. Filter methods. In *Feature Extraction*, pages 89–117. Springer, 2006. (Cited on page 38.)
- [76] Damian Dudek. Measures for comparing association rule sets. In *International Conference on Artificial Intelligence and Soft Computing*, pages 315–322. Springer, 2010. (Cited on page 124.)
- [77] Alireza Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition*, pages 1778–1785, June 2009. (Cited on page 159.)
- [78] Yansong Feng and Mirella Lapata. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99. Association for Computational Linguistics, 2010. (Cited on page 19.)
- [79] Leo Ferres, Avi Parush, Shelley Roberts, and Gitte Lindgaard. Helping people with visual impairments gain access to graphical information through natural language: The igrph system. In *International Conference on Computers for Handicapped Persons*, pages 1122–1130. Springer, 2006. (Cited on page 31.)

- [80] François Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5(Nov):1531–1555, 2004. (Cited on page 40.)
- [81] Christos A Frantzikidis, Charalampos Bratsas, Manousos A Klados, Evdokimos Konstantinidis, Chrysa D Lithari, Ana B Vivas, Christos L Padelis, Eleni Kaldoudi, Costas Pappas, and Panagiotis D Bamidis. On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):309–318, 2010. (Cited on pages 95, 96, 97, and 98.)
- [82] Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011. (Cited on pages 110 and 113.)
- [83] Ben D Fulcher and Nick S Jones. Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):3026–3037, 2014. (Cited on page 141.)
- [84] Peter Gärdenfors. Three levels of inductive inference. *Studies in Logic and the Foundations of Mathematics*, 134:427–449, 1995. (Cited on page 3.)
- [85] Peter Gärdenfors. Symbolic, conceptual and subconceptual representations. In *Human and Machine Perception*, pages 255–270. Springer, 1997. (Cited on page 3.)
- [86] Peter Gärdenfors. *Conceptual spaces: The geometry of thought*. MIT press, 2000. (Cited on pages 3, 4, 20, 21, 22, 23, 24, 25, 42, 44, 45, 47, 49, 50, 51, 66, 82, and 148.)
- [87] Peter Gardenfors. Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, 2(2):9–27, 2004. (Cited on pages 3, 21, 23, 24, 25, 35, and 38.)
- [88] Peter Gärdenfors. Induction, conceptual spaces and ai. *The Dynamics of Thought*, pages 109–124, 2005. (Cited on pages 25, 51, 53, and 161.)
- [89] Peter Gärdenfors. Semantics based on conceptual spaces. In *Indian Conference on Logic and Its Applications*, pages 1–11. Springer, 2011. (Cited on pages 3, 23, and 25.)
- [90] Peter Gärdenfors. *The geometry of meaning: Semantics based on conceptual spaces*. MIT Press, 2014. (Cited on pages 20 and 47.)
- [91] Peter Gärdenfors and Massimo Warglien. Using conceptual spaces to model actions and events. *Journal of Semantics*, page ffs007, 2012. (Cited on page 45.)

- [92] Peter Gärdenfors and Mary-Anne Williams. Reasoning about categories in conceptual spaces. In *Proceedings of the 17th international joint conference on Artificial, IJCAI'01*, pages 385–392, 2001. (Cited on pages 25, 50, and 62.)
- [93] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018. (Cited on pages 30, 31, and 32.)
- [94] Albert Gatt, Francois Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. From data to text in the neonatal intensive care unit: Using nlg technology for decision support and information management. *Ai Communications*, 22(3):153–186, 2009. (Cited on pages 31 and 32.)
- [95] Albert Gatt and Ehud Reiter. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93. Association for Computational Linguistics, 2009. (Cited on page 70.)
- [96] Albert Gatt, Roger PG van Gompel, Emiel Krahmer, and Kees van Deemter. Does domain size impact speech onset time during reference production? In *34th Annual Meeting of the Cognitive Science Society*, pages 1584–1589, 2012. (Cited on page 159.)
- [97] Elena Gaura, John Kemp, and James Brusey. Leveraging knowledge from physiological data: On-body heat stress risk prediction with sensor networks. *IEEE transactions on biomedical circuits and systems*, 7(6):861–870, 2013. (Cited on pages 93, 94, 98, and 99.)
- [98] Ronald L Gellish, Brian R Goslin, Ronald E Olson, AUDRY McDONALD, Gary D Russi, and Virinder K Moudgil. Longitudinal modeling of the relationship between age and maximal heart rate. *Medicine and science in sports and exercise*, 39(5):822–829, 2007. (Cited on page 95.)
- [99] James E Gentle, Wolfgang Karl Härdle, and Yuichi Mori. *Handbook of computational statistics: concepts and methods*. Springer Science & Business Media, 2012. (Cited on page 109.)
- [100] John Gialelis, Petros Chondros, Dimitrios Karadimas, Sofia Dima, and Dimitrios Serpanos. Identifying chronic disease complications utilizing state of the art data fusion methodologies and signal processing algorithms. In Konstantina S. Nikita, James C. Lin, Dimitrios I. Fotiadis,

- and Maria-Teresa Arredondo Waldmeyer, editors, *Wireless Mobile Communication and Healthcare*, volume 83, pages 256–263. Springer Berlin Heidelberg, 2012. (Cited on pages 93, 95, 96, and 98.)
- [101] Donna Giri, U Rajendra Acharya, Roshan Joy Martis, S Vinitha Sree, Teik-Cheng Lim, Thajudin Ahamed VI, and Jasjit S Suri. Automated diagnosis of coronary artery disease affected patients using lda, pca, ica and discrete wavelet transform. *Knowledge-Based Systems*, 37:274–282, 2013. (Cited on pages 95, 96, 97, 98, and 99.)
 - [102] Dimitra Gkatzia. *Data-driven approaches to content selection for data-to-text generation*. PhD thesis, Heriot-Watt University, 2015. (Cited on page 32.)
 - [103] Dimitra Gkatzia. Content selection in data-to-text systems: A survey. *arXiv preprint arXiv:1610.08375*, 2016. (Cited on page 29.)
 - [104] Eli Goldberg, Norbert Driedger, and Richard I Kittredge. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53, 1994. (Cited on page 31.)
 - [105] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220, 2000. (Cited on pages 99 and 107.)
 - [106] Dina Goldin, Ricardo Mardales, and George Nagy. In search of meaning for time series subsequence clustering: matching algorithms based on a new distance measure. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 347–356. ACM, 2006. (Cited on page 116.)
 - [107] Machon Gregory and Ben Shneiderman. Shape identification in temporal data sets. In *Expanding the Frontiers of Visual Analytics and Visualization*, pages 305–321. Springer, 2012. (Cited on pages 140 and 141.)
 - [108] Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. Topics in semantic representation. *Psychological review*, 114(2):211, 2007. (Cited on pages 19, 20, and 33.)
 - [109] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003. (Cited on page 38.)
 - [110] Isabelle Guyon, Steve Gunn, Masoud Nikraves, and Lofti A Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2008. (Cited on page 38.)

- [111] James Hampton and Helen Moss. Concepts and meaning: Introduction to the special issue on conceptual representation. *Language and cognitive processes*, 18(5-6):505–512, 2003. (Cited on pages 21 and 33.)
- [112] James A Hampton. Typicality, graded membership, and vagueness. *Cognitive Science*, 31(3):355–384, 2007. (Cited on page 62.)
- [113] Jing He, Yanchun Zhang, Guangyan Huang, Yefei Xin, Xiaohui Liu, Hao Lan Zhang, Stanley Chiang, and Hailun Zhang. An association rule analysis framework for complex physiological and genetic data. In *International Conference on Health Information Science*, pages 131–142. Springer, 2012. (Cited on page 121.)
- [114] Jing He, Yanchun Zhang, Guangyan Huang, Yefei Xin, Xiaohui Liu, HaoLan Zhang, Stanley Chiang, and Hailun Zhang. An association rule analysis framework for complex physiological and genetic data. In Jing He, Xiaohui Liu, ElizabethA. Krupinski, and Guandong Xu, editors, *Health Information Science*, volume 7231 of *Lecture Notes in Computer Science*, pages 131–142. Springer Berlin Heidelberg, 2012. (Cited on page 98.)
- [115] Åke Hjalmarson. Heart rate: an independent risk factor in cardiovascular disease. *European Heart Journal Supplements*, 9(suppl_F):F3–F7, 2007. (Cited on page 95.)
- [116] Michael Holender, Rakesh Nagi, Moises Sudit, and J Terry Rickard. Information fusion using conceptual spaces: Mathematical programming models and methods. In *Information Fusion, 2007 10th International Conference on*, pages 1–8. IEEE, 2007. (Cited on pages 3, 21, 23, and 24.)
- [117] Fei Hu, Meng Jiang, Laura Celentano, and Yang Xiao. Robust medical ad hoc sensor networks (masn) with wavelet-based ecg data mining. *Ad Hoc Networks*, 6(7):986–1012, 2008. (Cited on pages 96, 97, and 98.)
- [118] Guangyan Huang, Yanchun Zhang, Jie Cao, Michael Steyn, and Kersi Taraporewalla. Online mining abnormal period patterns from multiple medical sensor data streams. *World Wide Web*, pages 1–19, 2013. (Cited on pages 93, 99, 100, 101, and 102.)
- [119] James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, and Cindy Sykes. Automatic generation of natural language nursing shift summaries in neonatal intensive care: Bt-nurse. *Artificial intelligence in medicine*, 56(3):157–172, 2012. (Cited on pages 30, 31, 103, and 132.)

- [120] Lidija Iordanskaja, Myunghee Kim, Richard Kittredge, Benoit Lavoie, and Alain Polguere. Generation of extended bilingual statistical reports. In *Proceedings of the 14th conference on Computational linguistics-Volume 3*, pages 1019–1023. Association for Computational Linguistics, 1992. (Cited on page 31.)
- [121] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010. (Cited on page 116.)
- [122] Tony Jebara. *Machine learning: discriminative and generative*, volume 755. Springer Science & Business Media, 2012. (Cited on page 83.)
- [123] Zhanpeng Jin, Yuwen Sun, and A. C. Cheng. Predicting cardiovascular disease from real-time electrocardiographic monitoring: An adaptive machine learning approach on a cell phone. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6889–6892, 2009. (Cited on page 100.)
- [124] Janusz Kacprzyk, Anna Wilbik, and S Zadrożny. Linguistic summarization of time series using a fuzzy quantifier driven aggregation. *Fuzzy Sets and Systems*, 159(12):1485–1499, 2008. (Cited on pages 111 and 139.)
- [125] Janusz Kacprzyk and Sławomir Zadrożny. Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. *Information Sciences*, 173(4):281–304, 2005. (Cited on page 28.)
- [126] Janusz Kacprzyk and Sławomir Zadrożny. Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries, and natural-language generation. *IEEE Transactions on Fuzzy Systems*, 18(3):461–472, 2010. (Cited on pages 27 and 28.)
- [127] Jayant Kalagnanam and Max Henrion. A comparison of decision analysis and expert rules for sequential diagnosis. In *Machine Intelligence and Pattern Recognition*, volume 9, pages 271–281. Elsevier, 1990. (Cited on page 98.)
- [128] Walter Karlen, Claudio Mattiussi, and Dario Floreano. Sleep and wake classification with ecg and respiratory effort signals. *IEEE Transactions on Biomedical Circuits and Systems*, 3(2):71–78, 2009. (Cited on pages 95 and 98.)
- [129] Eamonn Keogh. Instance-based learning. In *Encyclopedia of Machine Learning*, pages 549–550. Springer, 2011. (Cited on page 25.)
- [130] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. Segmenting time series: A survey and novel approach. *Data mining in time series databases*, 57:1–22, 2004. (Cited on page 110.)

- [131] Carsten Keßler. Conceptual spaces for data descriptions. In *The cognitive approach to modeling environments (CAME), workshop at GI-Science*, pages 29–35, 2006. (Cited on pages 4, 26, and 50.)
- [132] Jonghun Kim, Jaekwon Kim, Daesung Lee, and Kyung-Yong Chung. Ontology driven interactive healthcare with wearable sensors. *Multimedia Tools and Applications*, 71(2):827–841, 2014. (Cited on page 103.)
- [133] Rik Koncel-Kedziorski, Hannaneh Hajishirzi, and Ali Farhadi. Multi-resolution language grounding with weak supervision. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–396, 2014. (Cited on page 32.)
- [134] Sotiris Kotsiantis and Dimitris Kanellopoulos. Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1):71–82, 2006. (Cited on pages 113 and 121.)
- [135] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition*, pages 951–958, June 2009. (Cited on page 160.)
- [136] Martin Längkvist, Lars Karlsson, and Amy Loutfi. Sleep stage classification using unsupervised feature learning. *Advances in Artificial Neural Systems*, 2012. (Cited on page 102.)
- [137] Oscar D Lara and Miguel A Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, 15(3):1192–1209, 2013. (Cited on page 96.)
- [138] Kevin LeBlanc. Cooperative anchoring: sharing information about objects in multi-robot systems. 2010. (Cited on page 25.)
- [139] Ickjai Lee. Data mining coupled conceptual spaces for intelligent agents in data-rich environments. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 42–48. Springer, 2005. (Cited on page 26.)
- [140] Ickjai Lee and Bayani Portier. An empirical study of knowledge representation and learning within conceptual spaces for intelligent agents. In *6th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2007.*, pages 463–468. IEEE, 2007. (Cited on page 26.)
- [141] Kyong Ho Lee, Sun-Yuan Kung, and Naveen Verma. Low-energy formulations of support vector machine kernel functions for biomedical sensor applications. *Journal of Signal Processing Systems*, 69(3):339–349, 2012. (Cited on pages 93, 97, 98, 99, and 100.)

- [142] Martha Lewis and Jonathan Lawry. Hierarchical conceptual spaces for concept combination. *Artificial Intelligence*, 237:204–227, 2016. (Cited on page 66.)
- [143] Qiao Li and Gari D Clifford. Dynamic time warping and machine learning for signal quality assessment of pulsatile signals. *Physiological measurement*, 33(9):1491–1501, 2012. (Cited on page 98.)
- [144] Xiaokun Li and Fatih Porikli. Human state classification and predication for critical care monitoring by real-time bio-signal analysis. In *Proceedings of the 20th International Conference on Pattern Recognition*, pages 2460–2463, Washington, DC, USA, 2010. IEEE Computer Society. (Cited on page 101.)
- [145] T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005. (Cited on page 114.)
- [146] Moshe Lichman. UCI machine learning repository. Available online: <http://archive.ics.uci.edu/ml>, 2013. (Cited on page 70.)
- [147] Antonio Lieto, Antonio Chella, and Marcello Frixione. Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation. *Biologically Inspired Cognitive Architectures*, 19:1–9, 2017. (Cited on page 25.)
- [148] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11. ACM, 2003. (Cited on page 110.)
- [149] Catherine Loader. Smoothing: local regression techniques. In *Handbook of Computational Statistics*, pages 571–596. Springer, 2012. (Cited on page 110.)
- [150] Joan Albert López-Vallverdú, David Riaño, and John A Bohada. Improving medical decision trees by combining relevant health-care criteria. *Expert Systems with Applications*, 39(14):11782–11791, 2012. (Cited on page 98.)
- [151] George F Luger. *Artificial intelligence: structures and strategies for complex problem solving*. Pearson education, 2005. (Cited on page 25.)
- [152] Yi Mao, Wenlin Chen, Yixin Chen, Chenyang Lu, Marin Kollef, and Thomas Bailey. An integrated data mining approach to real-time clinical monitoring and deterioration warning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data*

- mining*, pages 1140–1148, New York, NY, USA, 2012. ACM. (Cited on pages 96, 97, 99, and 100.)
- [153] Benjamin M. Marlin, David C. Kale, Robinder G. Khemani, and Randall C. Wetzel. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 389–398, New York, NY, USA, 2012. ACM. (Cited on pages 94, 99, 100, and 101.)
 - [154] Benjamin M Marlin, David C Kale, Robinder G Khemani, and Randall C Wetzel. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 389–398. ACM, 2012. (Cited on page 108.)
 - [155] Marvin L Minsky. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI magazine*, 12(2):34, 1991. (Cited on page 3.)
 - [156] George B Moody and Roger G Mark. A database to support development and evaluation of intelligent intensive care monitoring. In *Computers in Cardiology*, 1996, pages 657–660. IEEE, 1996. (Cited on page 107.)
 - [157] Lailil Muflikhah, Yeni Wahyuningsih, et al. Fuzzy rule generation for diagnosis of coronary heart disease risk using subtractive clustering method. 2013. (Cited on page 122.)
 - [158] Arijit Mukherjee, Arpan Pal, and Prateep Misra. Data analytics in ubiquitous sensor-based health information systems. In *Proceedings of the 2012 Sixth International Conference on Next Generation Mobile Applications, Services and Technologies*, pages 193–198, Washington, DC, USA, 2012. IEEE Computer Society. (Cited on page 93.)
 - [159] Vishal Nangalia, David R Prytherch, and Gary B Smith. Health technology assessment review: Remote monitoring of vital signs-current status and future challenges. *Critical Care*, 14(5):233, 2010. (Cited on page 92.)
 - [160] Alex Nanopoulos, Rob Alcock, and Yannis Manolopoulos. Feature-based classification of time-series data. *International Journal of Computer Research*, 10(3):49–61, 2001. (Cited on page 141.)
 - [161] Kali Vara Prasad Narahariseti and Manan Bawa. Comparison of different signal processing methods for reducing artifacts from photoplethysmograph signal. In *Proceedings of the IEEE International Conference on*

- Electro/Information Technology (EIT)*, pages 1–8. IEEE, 2011. (Cited on page 99.)
- [162] Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002. (Cited on page 83.)
 - [163] Adam Niewiadomski and Izabela Superson. Multi-subject type-2 linguistic summaries of relational databases. In *Frontiers of Higher Order Fuzzy Sets*, pages 167–181. Springer, 2015. (Cited on page 28.)
 - [164] Nils J Nilsson. *Artificial intelligence: a new synthesis*. Elsevier, 1998. (Cited on page 20.)
 - [165] Vilém Novák. Linguistic characterization of time series. *Fuzzy Sets and Systems*, 285:52–72, 2016. (Cited on pages 63, 139, and 140.)
 - [166] Miho Ohsaki, Yoshinori Sato, Hideto Yokoi, and Takahira Yamaguchi. A rule discovery support system for sequential medical data, in the case study of a chronic hepatitis dataset. In *Workshop Notes of the International Workshop on Active Mining, at IEEE International Conference on Data Mining*, 2002. (Cited on page 121.)
 - [167] Patricia Ordóñez, Tom Armstrong, Tim Oates, and Jim Fackler. Classification of patients using novel multivariate time series representations of physiological data. In *10th International Conference on Machine Learning and Applications and Workshops (ICMLA)*, volume 2, pages 172–179. IEEE, 2011. (Cited on page 98.)
 - [168] Mukta Paliwal and Usha A Kumar. Neural networks and statistical techniques: A review of applications. *Expert systems with applications*, 36(1):2–17, 2009. (Cited on page 97.)
 - [169] Alexandros Pantelopoulos and Nikolaos G Bourbakis. Prognosis - a wearable health-monitoring system for people at risk: Methodology and modeling. *IEEE Transactions on Information Technology in Biomedicine*, 14(3):613–621, 2010. (Cited on page 95.)
 - [170] Abhilasha M. Patel, Pankaj K. Gakare, and A. N. Cheeran. Real time ecg feature extraction and arrhythmia detection on a mobile platform. *International Journal of Computer Applications*, 44(23):40–45, April 2012. (Cited on pages 99, 100, and 101.)
 - [171] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005. (Cited on page 40.)

- [172] François Petitjean, Alain Ketterlin, and Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693, 2011. (Cited on page 116.)
- [173] Vili Podgorelec, Peter Kokol, Bruno Stiglic, and Ivan Rozman. Decision trees: an overview and their use in medicine. *Journal of medical systems*, 26(5):445–463, 2002. (Cited on page 98.)
- [174] François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816, 2009. (Cited on pages 31 and 32.)
- [175] Gaurav N. Pradhan, Rita Chattopadhyay, and S. Panchanathan. Processing body sensor data streams for continuous physiological monitoring. In *Proceedings of the international conference on Multimedia information retrieval*, pages 479–486, New York, NY, USA, 2010. ACM. (Cited on page 97.)
- [176] Willard Van Orman Quine. *Ontological relativity and other essays*. Number 1. Columbia University Press, 1969. (Cited on pages 24 and 36.)
- [177] Alejandro Ramos-Soto, Alberto Jose Bugarin, and Senén Barro. Fuzzy sets across the natural language generation pipeline. *Progress in artificial intelligence*, 5(4):261–276, 2016. (Cited on pages 56 and 83.)
- [178] Alejandro Ramos-Soto, Alberto Jose Bugarin, and Senén Barro. On the role of linguistic descriptions of data in the building of natural language generation systems. *Fuzzy Sets and Systems*, 285:31–51, 2016. (Cited on pages 27 and 28.)
- [179] Alejandro Ramos-Soto, Alberto Jose Bugarin, Senén Barro, and Felix Diaz-Hermida. Automatic linguistic descriptions of meteorological data a soft computing approach for converting open data to open information. In *8th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE, 2013. (Cited on page 28.)
- [180] Alejandro Ramos-Soto, Alberto Jose Bugarin, Senén Barro, and Juan Taboada. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. *IEEE Transactions on Fuzzy Systems*, 23(1):44–57, 2015. (Cited on page 83.)
- [181] Martin Raubal. Formalizing conceptual spaces. In *Formal ontology in information systems, proceedings of the third international conference (FOIS 2004)*, volume 114, pages 153–164, 2004. (Cited on pages 24 and 35.)

- [182] Ehud Reiter. An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 97–104, 2007. (Cited on pages 30, 31, 32, 57, and 65.)
- [183] Ehud Reiter and Anja Belz. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558, 2009. (Cited on page 80.)
- [184] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(01):57–87, 1997. (Cited on pages 29 and 69.)
- [185] Ehud Reiter, Robert Dale, and Zhiwei Feng. *Building natural language generation systems*, volume 33. MIT Press, 2000. (Cited on pages 29, 56, 68, 69, 70, and 132.)
- [186] Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1):137–169, 2005. (Cited on pages 31 and 32.)
- [187] Ehud Reiter, Somayajulu G Sripada, and Roma Robertson. Acquiring correct knowledge for natural language generation. *Journal of Artificial Intelligence Research*, pages 491–516, 2003. (Cited on pages 29 and 31.)
- [188] John T Rickard. A concept geometry for conceptual spaces. *Fuzzy optimization and decision making*, 5(4):311–329, 2006. (Cited on pages 4, 24, 26, 62, and 66.)
- [189] John T Rickard, Janet Aisbett, and Greg Gibbon. Reformulation of the theory of conceptual spaces. *Information Sciences*, 177(21):4539–4565, 2007. (Cited on pages 4, 24, 35, and 56.)
- [190] Joanne K Rowling. *Harry Potter And The Prisoner Of Azkaban*. New York: Arthur A. Levine Books, 1999. (Cited on page 2.)
- [191] Jalal al-Din Rumi and Arthur John Arberry. *Tales from the Masnavi*. Curzon paperbacks. Curzon Press, 1993. (Cited on page 1.)
- [192] Lucia Sacchi, Riccardo Bellazzi, Cristiana Larizza, Riccardo Porreca, and Paolo Magni. Learning rules with complex temporal patterns in biomedical domains. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 23–32. Springer, 2005. (Cited on page 121.)
- [193] Lucia Sacchi, Cristiana Larizza, Carlo Combi, and Riccardo Bellazzi. Data mining with temporal abstractions: learning rules from time series. *Data Mining and Knowledge Discovery*, 15(2):217–247, 2007. (Cited on page 121.)

- [194] Ehssan Sakhaee. Elephant in the dark. Available online: <https://www.youtube.com/watch?v=iQFjmTEYxs8>. (Accessed March 3, 2018). (Cited on pages xi and 163.)
- [195] Osman Salem, Yaning Liu, and Ahmed Mehaoua. A lightweight anomaly detection framework for medical wireless sensor networks. In *Proceedings of the IEEE Wireless Communications and Networking Conference*, pages 4358–4363. IEEE, 2013. (Cited on pages 98 and 99.)
- [196] Erik Schaffernicht, Robert Kaltenhaeuser, Saurabh Shekhar Verma, and Horst-Michael Gross. On estimating mutual information for feature selection. In *International Conference on Artificial Neural Networks*, pages 362–367. Springer, 2010. (Cited on page 40.)
- [197] Tim Schlüter and Stefan Conrad. About the analysis of time series with temporal association rule mining. In *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 325–332. IEEE, 2011. (Cited on page 120.)
- [198] Steven Schockaert and Henri Prade. Interpolative and extrapolative reasoning in propositional theories using qualitative knowledge about conceptual spaces. *Artificial Intelligence*, 202:86–131, 2013. (Cited on page 25.)
- [199] Angela Schwering and Martin Raubal. Spatial relations for semantic similarity measurement. In *International Conference on Conceptual Modeling*, pages 259–269. Springer, 2005. (Cited on page 25.)
- [200] Christopher G Scully, Jinseok Lee, Joseph Meyer, Alexander M Gorbach, Domhnall Granquist-Fraser, Yitzhak Mendelson, and Ki H Chon. Physiological parameter monitoring from optical recordings with a mobile phone. *IEEE Transactions on Biomedical Engineering*, 59(2):303–306, 2012. (Cited on pages 99 and 101.)
- [201] Yuval Shoham. A framework for knowledge-based temporal abstraction. *Artificial intelligence*, 90(1-2):79–133, 1997. (Cited on page 122.)
- [202] Roger N Shepard et al. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987. (Cited on page 62.)
- [203] Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples*. Springer Science & Business Media, 2010. (Cited on page 141.)
- [204] Carina Silberer, Vittorio Ferrari, and Mirella Lapata. Models of semantic representation with visual attributes. In *Proceedings of the 51st Annual*

- Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 572–582, 2013. (Cited on page 160.)
- [205] Fábio Silva, Teresa Olivares, Fernando Royo, MA Vergara, and Cesar Analide. Experimental study of the stress level at the workplace using an smart testbed of wireless sensor networks and ambient intelligence techniques. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 200–209. Springer, 2013. (Cited on pages 94 and 99.)
 - [206] Pedro FB Silva, André RS Marçal, and Rubim M Almeida da Silva. Evaluation of features for leaf discrimination. In *Image Analysis and Recognition*, pages 197–204. 2013. (Cited on pages 36, 73, 74, and 75.)
 - [207] Craig Silverstein, Sergey Brin, and Rajeev Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data mining and knowledge discovery*, 2(1):39–68, 1998. (Cited on page 133.)
 - [208] Rajiv Ranjan Singh, Sailesh Conjeti, and Rahul Banerjee. An approach for real-time stress-trend detection using physiological signals in wearable computing systems for automotive drivers. In *14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1477–1482. IEEE, 2011. (Cited on pages 96, 97, 99, and 100.)
 - [209] Linda B Smith. From global similarities to kinds of similarities: the construction of dimensions in development. In *Similarity and analogical reasoning*, pages 146–178. Cambridge University Press, 1989. (Cited on page 25.)
 - [210] Sweta Sneha and Upkar Varshney. Enabling ubiquitous patient monitoring: Model, decision protocols, opportunities and challenges. *Decision Support Systems*, 46(3):606–619, 2009. (Cited on page 95.)
 - [211] Mohammad S Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 2010. (Cited on page 83.)
 - [212] Daby Sow, Deepak S. Turaga, and Michael Schmidt. Mining of sensor data in healthcare: A survey. In Charu C. Aggarwal, editor, *Managing and Mining Sensor Data*, pages 459–504. Springer US, 2013. (Cited on pages 92, 93, 96, and 97.)
 - [213] Daby Sow, Deepak S Turaga, and Michael Schmidt. Mining of sensor data in healthcare: a survey. In *Managing and mining sensor data*, pages 459–504. Springer, 2013. (Cited on page 105.)
 - [214] Thomas L Spalding and Brian H Ross. Concept learning and feature interpretation. *Memory & cognition*, 28(3):439–451, 2000. (Cited on page 50.)

- [215] Somayajulu Sripada, Ehud Reiter, and Ian Davy. Sumtime-mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10, 2003. (Cited on page 31.)
- [216] Michael Stacey and Carolyn McGregor. Temporal abstraction in intelligent clinical data analysis: A survey. *Artificial intelligence in medicine*, 39(1):1–24, 2007. (Cited on page 93.)
- [217] Yu Su, Moray Allan, and Frédéric Jurie. Improving object classification using semantic attributes. In *Proceedings of the British Machine Vision Conference*, pages 1–10, 2010. (Cited on page 159.)
- [218] Feng-Tso Sun, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss. Activity-aware mental stress detection using physiological sensors. In Martin Gris and Guang Yang, editors, *Mobile Computing, Applications, and Services*, volume 76, pages 211–230. Springer Berlin Heidelberg, 2012. (Cited on pages 94, 97, 98, and 101.)
- [219] Ron Sun. Artificial intelligence: Connectionist and symbolic approaches. *International Encyclopedia of the Social and Behavioral Sciences*, Oxford: Pergamon/Elsevier, pages 783–789, 2001. (Cited on pages 3 and 26.)
- [220] Pang-Ning Tan et al. *Introduction to data mining*. Pearson Education India, 2006. (Cited on page 120.)
- [221] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004. (Cited on page 126.)
- [222] Bhaskar Thakker and Anoop Lal Vyas. Support vector machine for abnormal pulse classification. *International Journal of Computer Applications*, 22(7):13–19, 2011. (Cited on pages 93, 97, 98, 100, and 101.)
- [223] Kari Torkkola. Information-theoretic methods. In *Feature Extraction*, pages 167–185. Springer, 2008. (Cited on page 40.)
- [224] Gracian Trivino and Michio Sugeno. Towards linguistic descriptions of phenomena. *International Journal of Approximate Reasoning*, 54(1):22–34, 2013. (Cited on page 28.)
- [225] Gabriella Vigliocco, Lotte Meteyard, Mark Andrews, and Stavroula Kousta. Toward a theory of semantic representation. *Language and Cognition*, 1(2):219–247, 2009. (Cited on page 20.)

- [226] Thi Hong Nhan Vu, Namkyu Park, Yang Koo Lee, Yongmi Lee, Jong Yun Lee, and Keun Ho Ryu. Online discovery of heart rate variability patterns in mobile healthcare services. *Journal of Systems and Software*, 83(10):1930–1940, 2010. (Cited on pages 95 and 99.)
- [227] Wei Wang, Honggang Wang, Michael Hempel, Dongming Peng, Hamid Sharif, and Hsiao-Hwa Chen. Secure stochastic ecg signals based on gaussian mixture model for e-healthcare systems. *IEEE Systems Journal*, 5(4):564–573, 2011. (Cited on pages 95, 98, and 99.)
- [228] Andrew R Webb. *Statistical pattern recognition*. John Wiley & Sons, 2003. (Cited on page 40.)
- [229] Eric W. Weisstein. Totally ordered set. From MathWorld—A Wolfram Web Resource, 2016. (Cited on page 66.)
- [230] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005. (Cited on pages 25 and 39.)
- [231] Stephen Gang Wu, Forrest Sheng Bao, Eric You Xu, Yu-Xuan Wang, Yi-Fan Chang, and Qiao-Liang Xiang. A leaf recognition algorithm for plant classification using probabilistic neural network. In *IEEE International Symposium on Signal Processing and Information Technology*, pages 11–16. IEEE, 2007. (Cited on page 74.)
- [232] Baile Xie and Hlaing Minn. Real-time sleep apnea detection by classifier combination. *IEEE Transactions on Information Technology in Biomedicine*, 16(3):469–477, 2012. (Cited on pages 95 and 97.)
- [233] Ronald R Yager. A new approach to the summarization of data. *Information Sciences*, 28(1):69–86, 1982. (Cited on page 27.)
- [234] Jinn-Yi Yeh, Tai-Hsi Wu, and Chuan-Wei Tsao. Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems*, 50(2):439–448, 2011. (Cited on pages 94 and 98.)
- [235] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, and Lei Hua. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, 36(4):2431–2448, 2012. (Cited on pages 92, 94, and 105.)
- [236] Jin Yu, Jim Hunter, Ehud Reiter, and Somayajulu Sripada. Recognising visual patterns to communicate gas turbine time-series data. In *Applications and Innovations in Intelligent Systems*, pages 105–118. Springer, 2003. (Cited on page 140.)

- [237] Jin Yu, Ehud Reiter, Jim Hunter, and Chris Mellish. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(01):25–49, 2007. (Cited on pages 29, 32, 68, 139, and 140.)
- [238] Lotfi Zadeh. From computing with numbers to computing with words—from manipulation of measurements to manipulation of perceptions. *International Journal of Applied Mathematics and Computer Science*, 12(3):307–324, 2002. (Cited on page 27.)
- [239] Lotfi A Zadeh. Fuzzy logic = computing with words. *IEEE transactions on fuzzy systems*, 4(2):103–111, 1996. (Cited on pages 27 and 28.)
- [240] Lotfi A Zadeh. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy sets and systems*, 90(2):111–127, 1997. (Cited on page 28.)
- [241] Lotfi A Zadeh. A new direction in ai: Toward a computational theory of perceptions. *AI magazine*, 22(1):73, 2001. (Cited on page 27.)
- [242] Frank Zenker and Peter Gärdenfors. Applications of conceptual spaces. (Cited on pages 25 and 42.)
- [243] Ying Zhu. Automatic detection of anomalies in blood glucose using a machine learning approach. *Journal of Communications and Networks*, 13(2):125–131, 2011. (Cited on pages 93, 94, 98, 99, 100, and 101.)

PUBLICATIONS *in the series*
ÖREBRO STUDIES IN TECHNOLOGY

1. Bergsten, Pontus (2001) *Observers and Controllers for Takagi – Sugeno Fuzzy Systems*. Doctoral Dissertation.
2. Iliev, Boyko (2002) *Minimum-time Sliding Mode Control of Robot Manipulators*. Licentiate Thesis.
3. Spännar, Jan (2002) *Grey box modelling for temperature estimation*. Licentiate Thesis.
4. Persson, Martin (2002) *A simulation environment for visual servoing*. Licentiate Thesis.
5. Boustedt, Katarina (2002) *Flip Chip for High Volume and Low Cost – Materials and Production Technology*. Licentiate Thesis.
6. Biel, Lena (2002) *Modeling of Perceptual Systems – A Sensor Fusion Model with Active Perception*. Licentiate Thesis.
7. Otterskog, Magnus (2002) *Produktionstest av mobiltelefonantennner i mod-växlande kammare*. Licentiate Thesis.
8. Tolt, Gustav (2003) *Fuzzy-Similarity-Based Low-level Image Processing*. Licentiate Thesis.
9. Loutfi, Amy (2003) *Communicating Perceptions: Grounding Symbols to Artificial Olfactory Signals*. Licentiate Thesis.
10. Iliev, Boyko (2004) *Minimum-time Sliding Mode Control of Robot Manipulators*. Doctoral Dissertation.
11. Pettersson, Ola (2004) *Model-Free Execution Monitoring in Behavior-Based Mobile Robotics*. Doctoral Dissertation.
12. Överstam, Henrik (2004) *The Interdependence of Plastic Behaviour and Final Properties of Steel Wire, Analysed by the Finite Element Metod*. Doctoral Dissertation.
13. Jennergren, Lars (2004) *Flexible Assembly of Ready-to-eat Meals*. Licentiate Thesis.
14. Jun, Li (2004) *Towards Online Learning of Reactive Behaviors in Mobile Robotics*. Licentiate Thesis.
15. Lindquist, Malin (2004) *Electronic Tongue for Water Quality Assessment*. Licentiate Thesis.
16. Wasik, Zbigniew (2005) *A Behavior-Based Control System for Mobile Manipulation*. Doctoral Dissertation.

17. Berntsson, Tomas (2005) *Replacement of Lead Baths with Environment Friendly Alternative Heat Treatment Processes in Steel Wire Production*. Licentiate Thesis.
18. Tolt, Gustav (2005) *Fuzzy Similarity-based Image Processing*. Doctoral Dissertation.
19. Munkevik, Per (2005) "Artificial sensory evaluation – appearance-based analysis of ready meals". Licentiate Thesis.
20. Buschka, Pär (2005) *An Investigation of Hybrid Maps for Mobile Robots*. Doctoral Dissertation.
21. Loutfi, Amy (2006) *Odour Recognition using Electronic Noses in Robotic and Intelligent Systems*. Doctoral Dissertation.
22. Gillström, Peter (2006) *Alternatives to Pickling; Preparation of Carbon and Low Alloyed Steel Wire Rod*. Doctoral Dissertation.
23. Li, Jun (2006) *Learning Reactive Behaviors with Constructive Neural Networks in Mobile Robotics*. Doctoral Dissertation.
24. Otterskog, Magnus (2006) *Propagation Environment Modeling Using Scattered Field Chamber*. Doctoral Dissertation.
25. Lindquist, Malin (2007) *Electronic Tongue for Water Quality Assessment*. Doctoral Dissertation.
26. Cielniak, Grzegorz (2007) *People Tracking by Mobile Robots using Thermal and Colour Vision*. Doctoral Dissertation.
27. Boustedt, Katarina (2007) *Flip Chip for High Frequency Applications – Materials Aspects*. Doctoral Dissertation.
28. Soron, Mikael (2007) *Robot System for Flexible 3D Friction Stir Welding*. Doctoral Dissertation.
29. Larsson, Sören (2008) *An industrial robot as carrier of a laser profile scanner. – Motion control, data capturing and path planning*. Doctoral Dissertation.
30. Persson, Martin (2008) *Semantic Mapping Using Virtual Sensors and Fusion of Aerial Images with Sensor Data from a Ground Vehicle*. Doctoral Dissertation.
31. Andreasson, Henrik (2008) *Local Visual Feature based Localisation and Mapping by Mobile Robots*. Doctoral Dissertation.
32. Bouguerra, Abdelbaki (2008) *Robust Execution of Robot Task-Plans: A Knowledge-based Approach*. Doctoral Dissertation.

33. Lundh, Robert (2009) *Robots that Help Each Other: Self-Configuration of Distributed Robot Systems*. Doctoral Dissertation.
34. Skoglund, Alexander (2009) *Programming by Demonstration of Robot Manipulators*. Doctoral Dissertation.
35. Ranjbar, Parivash (2009) *Sensing the Environment: Development of Monitoring Aids for Persons with Profound Deafness or Deafblindness*. Doctoral Dissertation.
36. Magnusson, Martin (2009) *The Three-Dimensional Normal-Distributions Transform – an Efficient Representation for Registration, Surface Analysis, and Loop Detection*. Doctoral Dissertation.
37. Rahayem, Mohamed (2010) *Segmentation and fitting for Geometric Reverse Engineering. Processing data captured by a laser profile scanner mounted on an industrial robot*. Doctoral Dissertation.
38. Karlsson, Alexander (2010) *Evaluating Credal Set Theory as a Belief Framework in High-Level Information Fusion for Automated Decision-Making*. Doctoral Dissertation.
39. LeBlanc, Kevin (2010) *Cooperative Anchoring – Sharing Information About Objects in Multi-Robot Systems*. Doctoral Dissertation.
40. Johansson, Fredrik (2010) *Evaluating the Performance of TEWA Systems*. Doctoral Dissertation.
41. Trincavelli, Marco (2010) *Gas Discrimination for Mobile Robots*. Doctoral Dissertation.
42. Cirillo, Marcello (2010) *Planning in Inhabited Environments: Human-Aware Task Planning and Activity Recognition*. Doctoral Dissertation.
43. Nilsson, Maria (2010) *Capturing Semi-Automated Decision Making: The Methodology of CASADEMA*. Doctoral Dissertation.
44. Dahlbom, Anders (2011) *Petri nets for Situation Recognition*. Doctoral Dissertation.
45. Ahmed, Muhammad Rehan (2011) *Compliance Control of Robot Manipulator for Safe Physical Human Robot Interaction*. Doctoral Dissertation.
46. Riveiro, Maria (2011) *Visual Analytics for Maritime Anomaly Detection*. Doctoral Dissertation.

47. Rashid, Md. Jayedur (2011) *Extending a Networked Robot System to Include Humans, Tiny Devices, and Everyday Objects*. Doctoral Dissertation.
48. Zain-ul-Abdin (2011) *Programming of Coarse-Grained Reconfigurable Architectures*. Doctoral Dissertation.
49. Wang, Yan (2011) *A Domain-Specific Language for Protocol Stack Implementation in Embedded Systems*. Doctoral Dissertation.
50. Brax, Christoffer (2011) *Anomaly Detection in the Surveillance Domain*. Doctoral Dissertation.
51. Larsson, Johan (2011) *Unmanned Operation of Load-Haul-Dump Vehicles in Mining Environments*. Doctoral Dissertation.
52. Lidström, Kristoffer (2012) *Situation-Aware Vehicles: Supporting the Next Generation of Cooperative Traffic Systems*. Doctoral Dissertation.
53. Johansson, Daniel (2012) *Convergence in Mixed Reality-Virtuality Environments. Facilitating Natural User Behavior*. Doctoral Dissertation.
54. Stoyanov, Todor Dimitrov (2012) *Reliable Autonomous Navigation in Semi-Structured Environments using the Three-Dimensional Normal Distributions Transform (3D-NDT)*. Doctoral Dissertation.
55. Daoutis, Marios (2013) *Knowledge Based Perceptual Anchoring: Grounding percepts to concepts in cognitive robots*. Doctoral Dissertation.
56. Kristoffersson, Annica (2013) *Measuring the Quality of Interaction in Mobile Robotic Telepresence Systems using Presence, Spatial Formations and Sociometry*. Doctoral Dissertation.
57. Memedi, Mevludin (2014) *Mobile systems for monitoring Parkinson's disease*. Doctoral Dissertation.
58. König, Rikard (2014) *Enhancing Genetic Programming for Predictive Modeling*. Doctoral Dissertation.
59. Erlandsson, Tina (2014) *A Combat Survivability Model for Evaluating Air Mission Routes in Future Decision Support Systems*. Doctoral Dissertation.
60. Helldin, Tove (2014) *Transparency for Future Semi-Automated Systems. Effects of transparency on operator performance, workload and trust*. Doctoral Dissertation.

61. Krug, Robert (2014) *Optimization-based Robot Grasp Synthesis and Motion Control*. Doctoral Dissertation.
62. Reggente, Matteo (2014) *Statistical Gas Distribution Modelling for Mobile Robot Applications*. Doctoral Dissertation.
63. Långkvist, Martin (2014) *Modeling Time-Series with Deep Networks*. Doctoral Dissertation.
64. Hernández Bennetts, Víctor Manuel (2015) *Mobile Robots with In-Situ and Remote Sensors for Real World Gas Distribution Modelling*. Doctoral Dissertation.
65. Alirezaie, Marjan (2015) *Bridging the Semantic Gap between Sensor Data and Ontological Knowledge*. Doctoral Dissertation.
66. Pashami, Sepideh (2015) *Change Detection in Metal Oxide Gas Sensor Signals for Open Sampling Systems*. Doctoral Dissertation.
67. Lagriffoul, Fabien (2016) *Combining Task and Motion Planning*. Doctoral Dissertation.
68. Mosberger, Rafael (2016) *Vision-based Human Detection from Mobile Machinery in Industrial Environments*. Doctoral Dissertation.
69. Mansouri, Masoumeh (2016) *A Constraint-Based Approach for Hybrid Reasoning in Robotics*. Doctoral Dissertation.
70. Albitar, Houssam (2016) *Enabling a Robot for Underwater Surface Cleaning*. Doctoral Dissertation.
71. Mojtahedzadeh, Rasoul (2016) *Safe Robotic Manipulation to Extract Objects from Piles: From 3D Perception to Object Selection*. Doctoral Dissertation.
72. Köckemann, Uwe (2016) *Constraint-based Methods for Human-aware Planning*. Doctoral Dissertation.
73. Jansson, Anton (2016) *Only a Shadow. Industrial Computed Tomography Investigation, and Method Development, Concerning Complex Material Systems*. Licentiate Thesis.
74. Sebastian Hällgren (2017) *Some aspects on designing for metal Powder Bed Fusion*. Licentiate Thesis.
75. Junges, Robert (2017) *A Learning-driven Approach for Behavior Modeling in Agent-based Simulation*. Doctoral Dissertation.
76. Ricão Canelhas, Daniel (2017) *Truncated Signed Distance Fields Applied To Robotics*. Doctoral Dissertation.

77. Asadi, Sahar (2017) *Towards Dense Air Quality Monitoring: Time-Dependent Statistical Gas Distribution Modelling and Sensor Planning*. Doctoral Dissertation.
78. Banaee, Hadi (2018) *From Numerical Sensor Data to Semantic Representations: A Data-driven Approach for Generating Linguistic Descriptions*. Doctoral Dissertation.