EXAMENSARBETE INOM TEKNIK,
GRUNDNIVÅ, 15 HP

STOCKHOLM, *SVERIGE 2016*

Football result prediction using simple classification algorithms, a comparison between k-Nearest Neighbor and Linear Regression

PIERRE RUDIN

KTH
SKOLAN FÖR DATAVETENSKAP OCH KOMMUNIKATION

Football result prediction using simple classification algorithms, a comparison between k-Nearest Neighbor and Linear Regression

Pierre Rudin

Degree Projec in Computer Science
DD143X
Supervision: Alex Kozlov
Examiner: Örjan Ekeberg

CSC KTH 2016-05-12

Abstract

Ever since humans started competing with each other, people have tried to accurately predict the outcome of such events. Football is no exception to this and is extra interesting as subject for a project like this with the ever growing amount of data gathered from matches these days. Previously predictors had to make there predictions using there own knowledge and small amounts of data.

This report will use this growing amount of data and find out if it is possible to accurately predict the outcome of a football match using the k-Nearest Neighbor algorithm and Linear regression. The algorithms are compared on how accurately they predict the winner of a match, how precise they predict how many goals each team will score and the accuracy of the predicted goal difference.

The results are graphed and presented in tables. A discussion analyzes the results and draw the conclusion that booth algorithms could be useful if used with a good model, and that Linear Regression out performs k-NN.

Sammanfattning

Ända sedan vi människor började tävla mot varandra, har folk försökt förutspå vinnaren i tävlingarna. Fotboll är inget undantag till detta och är extra intressant för den här studien då den tillgängliga mängden data från fotbollsmatcher ständigt ökar. Tidigare har egna kunskaper och små mängder data använts för att förutspå resultaten.

Den här rapporten kommer dra nytta av den växande mängden data för att ta reda på om det är möjligt att med hjälp av k-Nearest Neighbor algoritmen och Linjär regression förutspå resultat i fotbollsmatcher. Algoritmerna kommer jämföras utifrån hur exakt de förutspår vinnaren i matcher, hur många mål de båda lagen gör samt hur precist algoritmerna förutspår målskilnaden i matcherna.

Resultaten presenteras både i grafer och i tabeller. En diskusion förs för att analysera resultaten och kommer fram till att båda algoritmerna kan vara användbara om modelen är välkonstruerad, och att Linjär regression är bättre lämpad än k-NN.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

For as long as we humans have been competing with each other, we have also engaged in betting activities in attempt to increase our fortune. We know that gambling took place in ancient Rome were it was legal to bet on chariot races and at the circus. [1]

Historically the possibilities to make accurate predictions in order to place good bets, has been limited to small amounts of data. Such as outcomes of previous events, injuries and standings or other point based ranking systems. Other than that the predictor have often been left with his own observations and experience.

Today on the other hand we got companies and organizations who have build huge databases with information of different events occurring in football matches. Almost every step a footballer takes on the pitch is recorded and stored in a database. It is easy to find information about how many passes a player delivers per game or get a heat map of a players movement during a game. [2]

1.1 Problem Statement

Can classification algorithms be used to accurately predict the outcome of football matches? This report will investigate and compare Linear Regression and k-Nearest Neighbor algorithm and give answer to which of the two are best suited for this type of predictions.

1.2 Scope

This report aims to compare two different classification algorithms and to find out if either of them is suited for sports forecasting. This will be done by looking at how precise they predict the final score, and how well they predict if a match

will end with a home win, draw or away win.

Before the algorithms can be implemented the footballers and there teams has to be modeled based on their abilities. The algorithms will then be implemented using a existing library.

In the process off this project some limitations had to be set, the study is only based on league matches, all cup fixtures and friendlies are not considered, neither is international fixtures.

1.3 Disposition

- The **Background** (2) introduces the reader to the algorithms used in this project, and gives a historical overview of football forecasting.
- Under **Method** (3) a more detailed description of the process is given, how players and teams are modeled, how data is collected and how the results are evaluated.
- Graphs and tables displaying results from the predictions are presented under the **Results** (4) section. Some statistical data for comparison and error calculations are also presented here.
- Under **Discussion** (5) the results are analyzed and a conclusion are presented.

Chapter 2

Background

2.1 k-Nearest Neighbor

k-Nearest Neighbor builds on the idea that based on classified data points a new unclassified data point may be classified by looking at the K closest data points surrounding it. This means that k-NN uses instance-based learning, where training data set is used to classify the unknown data point. As shown in the picture below the colored squares are classified and the white square is the record we want to classify. In this case K is set to three and a majority vote would classify the white square as red. k-NN is a classification algorithm but is widely used to predict and to make estimations. [3]

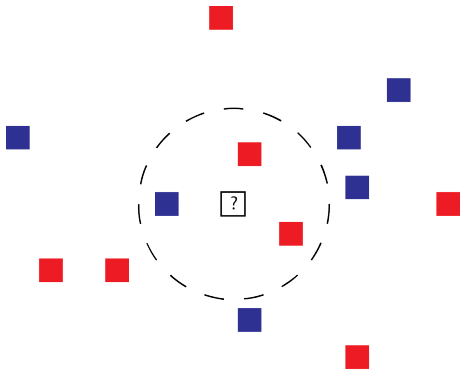

Figure 2.1: k-Nearest Neighbor, with K=3

To determine which points are closest to the point being classified the Euclidean distance is measured between all classified points and the unclassified point.

Large K values are in general more precise then small K values since it generally

will eliminate a lot of noise. This however is no guarantee and may especially when predicting, make the result flat i.e. almost all results will be the same or very similar. The best way to evaluate K probably is to run tests, and see which one is most accurate. Historically K values between 3 and 10 is considered optimal in most situations.[4]

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Figure 2.2: Euclidean distance

2.2 Linear Regression

Linear regression is a algorithm that tries to find a straight line that goes through a scattered plot of points as close to all points as possible i.e. find the best-fitted line. Such line is called a regression line.

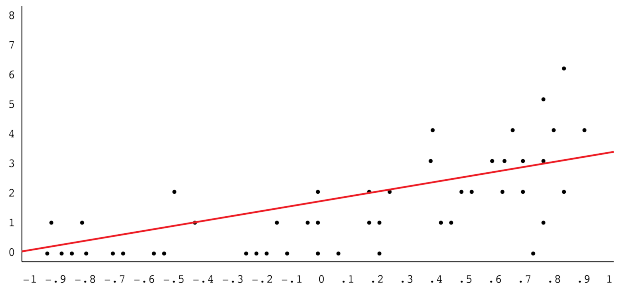

Figure 2.3: Linear Regression

The line is fitted using the least squares method, where the distance between the line and the dots are the value being squared. The line with the shortest sum of squared distances is the one considered best fitted.[5][6]

2.3 Data mining

New technologies have lead to an increase in the amount of statistic data that organizations gather about football. Today almost everything a footballer does on a football field is put on record and stored in databases. All this new data has started to [2][7]

Data mining are generally divided into three steps that covers the whole process

1. Data gathering - Early on in all projects of this type it is needed to collect data, this is a task that can be done in many different ways. Common for all methods are the importance of collecting good data. When the data is collected it needs to be stored somewhere, generally in a database.

2. Extraction and cleaning - It is rarely that the data comes in a form where it is usable. In this phase all unnecessary data is deleted and the useful is manipulated to be usable in analysis.

3. Analyzing and algorithms - Lastly a effective analyzing model is used to analyze the data. [8]

Chapter 3

Method

3.1 Data

Statistical data from 70000 fixtures were collected using a web-scrapping script written in Python, the script uses Selenium WebDriver and lxml libraries to read web pages and search them for information. The script scrapped Whoscored.com on statistics from there detailed tournaments i.e. the tournaments where they offer extended statistics. There detailed tournaments list contains 12 leagues from 11 different countries. Considering the possible quality difference between the domestic football played in different countries it was decided to only use domestic competition for this project i.e. Champions League and other international competitions were not used in this project.

Out off these 70000 fixtures 18000 had enough statistical data to use in this project. To get stability in the model two years of data was used to calculate the abilities of players and teams. Therefore the first two years of detailed statistics in each league were only used for calculations and no predictions were done on those fixtures. This left 10000 fixtures on which predictions were made.

3.2 Modeling

To quantify the abilities of players and teams a mathematical model is created, this model is based on statistical data from fixtures within 2 years prior to the game that's subject for prediction. There are three types of statistical data:

1. Successful attempt ratio,
2. Per game ratio i.e. how many times per game does something happen,
3. Successful attempts per game i.e. the product of (1) and (2).

It is impossible to definitely quantify players abilities, which makes it difficult to create a good model. To get a value for how good a player are their rated

abilities on the video game FIFA 16 were used as base values, and it also gives a good scale on which players are rated, 0-99.

$$rating = \sum_{i=1}^{n} factor_i \cdot stat_i$$

Figure 3.1: The model used for player ratings

Using the values from FIFA 16, statistical data and Linear Regression the factors were calculated. This procedure were used to quantify players passing, defending and shooting abilities. To quantify the teams abilities to pass, shoot and defend, all individual ratings for the starting players were summed.

3.3 Algorithms

In order to in a simple way implement the algorithms correctly the scikit-learn library were used. Scikit-learn includes both Linear Regression and k-Nearest Neighbor and they booth take the same data as input which makes it easy to prepare and run tests on booth algorithms.

H def.	H pass.	H shot.	A def.	A pass.	A shot.	H goal	A goal
543	412	600	512	477	502	2	1
488	563	492	492	672	519	0	0
612	494	503	552	549	608	1	3
844	712	695	672	651	712	3	0
592	555	497	695	562	601	1	0
712	607	682	543	812	647	**?**	**?**

Table 3.1: Algorithm input, the first five rows represents training data and the last row the data needed to make a prediction, H - home team, A - away team.

The algorithms are compared on their ability to predict the winner of a match, the amount of goals that are scored and which team who scores them.

3.3.1 k-Nearest Neighbor

To begin with it had to be decided which K value were to be used, this was done by implementing and comparing performance for different K values. The chosen K value was the one who's predictions had the best hit ratio. The best K value found was 7. k-NN outputs a integer goal prediction i.e. it will predict exactly how many goals each team will score. Which makes it easy to conclude the predicted winner of the fixture or if it will end in a draw. If the home team is predicted to score more goals then the away team it is considered that they

are the predicted winner, it is considered a predicted draw if booth teams are predicted to score equally many goals and a predicted away win if the away team are predicted to score the most goals.

3.3.2 Linear Regression

Since Linear Regression doesn't output integer goal predictions, limits for home wins, draws and away wins are handled slightly different then for k-NN.

$$Homewin = goal_{Home} > goal_{Away} + .5$$

$$draw = goal_{Away} \leq goal_{Home} \leq goal_{Away} + .5$$

$$Awaywin = goal_{Home} < goal_{Away}$$

Figure 3.2: Linear Regression outcome limits

These limits were selected since they produced the best result of the tested limits. When comparing the predictions with the number of goals that are scored the prediction is used as it is i.e. not rounded to integers.

3.4 Validation

3.4.1 Root mean square error

To quantify how big the general error differing the prediction from the actual value is RMSE is used.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}}$$

Figure 3.3: Root Mean Square Error, $X_{obs,i}$ is the observed value, $X_{model,i}$ is the predicted value and n are the number of predictions made.

3.4.2 Mean Percentage Error

MPE is a statistical error measurement giving a mean percentage value off the difference between forecasts and observations.

Since this measurement uses actual values of forecast errors, positive and negative errors even out the calculated error i.e. a error of -30 % and one of 30 % results in a mean error of 0 %. This can be useful when determining if the algorithm generally gives to high or to low predictions.

$$MPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{a_t - f_t}{a_t}$$

Figure 3.4: Mean percentage error, a_t is the actual value, f_t is the forecast, and n is the quantity of forecasts.

3.4.3 Random predictions

A random set of predictions are generated using the built-in function RAND() in MySQL. These predictions are used to test if the tested algorithms perform better then selecting winner in a fixture random. The algorithms may be consider to no use if they don't perform better then the random predictions.

Chapter 4

Results

In the following chapter all results will be presented. All data is presented in tables and the outcome/goal difference is also graphed. The "accuracy of predictions" should be read as the time when a outcome is predicted and also occurs, and not as the ratio between the number of times the prediction was correct and the number of predictions.

4.1 Benchmark Data

4.1.1 Real world data

These tables presents how often outcomes and final score happens in the investigated leagues.

1	X	2
.4667	.2655	.2678

Table 4.1: Accuracy of predictions using MySQL built-in RAND() function

	0	1	2	3	4	5	5+
0	.082	.1084	.0819	.0436	.0172	.0055	.0028
1	.0714	.1208	.0903	.0445	.0181	.006	.0025
2	.0438	.0617	.0517	.0251	.0103	.0031	.0011
3	.0176	.0257	.0173	.0098	.0042	.0012	.0003
4	.0067	.008	.0054	.0022	.0011	.0003	.0002
5	.0017	.002	.0013	.0006	.0002	.0001	0
5+	.0007	.0006	.0005	.0002	.0001	0	0

Table 4.2: How many goals home and away team scores and ratio of how common each score are. Home team on top, away team to left.

4.1.2 Random predictions

Random predictions are made to control how the algorithmic predictions performs.

1	X	2	Total
.2174	.0688	.075	.3612
.2102	.068	.0773	.3555
.2102	.071	.0775	.3587
.2157	.0651	.0747	.3555
.2104	.0666	.0794	.3564
.2158	.0678	.0735	.3571
.2118	.0724	.0722	.3564
.2094	.0671	.077	.3535
.2127	.0731	.0794	.3652
.2064	.0699	.0768	.3531
.2120	.069	.0763	.3573

Table 4.3: Accuracy of predictions using MySQL built-in RAND() function, the last line contain mean values of the predictions above

4.2 k-Nearest Neighbor

1	X	2	Totalt
.2364	.0936	.078	.408

Table 4.4: Accuracy of predictions using k-NN

Home goals	Away goals	Total goals	Goal difference
1.5655	1.3958	1.4831	1.951

Table 4.5: RMSE of score prediction using k-Nearest Neighbor

Home goals	Away goals	Total goals	Goal difference
32.438 %	52.440 %	41.685 %	84.356 %

Table 4.6: MPE of score prediction using k-Nearest Neighbor

+/-	1	X	2	Count
-5	0	0	1	5
-4	0	.5	.5	10
-3	.3425	.2329	.4247	73
-2	.3537	.2585	.3878	410
-1	.3836	.2607	.3558	1619
0	.4241	.2835	.2924	3287
1	.4833	.2465	.2702	2957
2	.5634	.2309	.2057	1230
3	.6156	.2347	.1497	294
4	.7213	.1639	.1148	61
5	.7143	.2857	0	7
6	1	0	0	1

Table 4.7: Predicted goal difference using k-NN

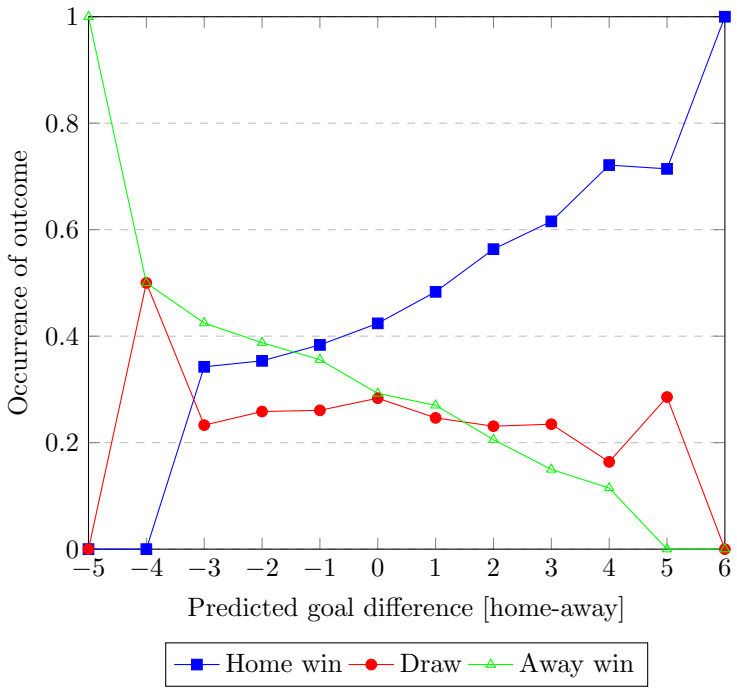

Figure 4.1: Graph of predicted goal difference using k-NN

4.3 Linear Regression

1	X	2	Totalt
.2613	.0915	.114	.4668

Table 4.8: Accuracy of predictions using Linear Regression

Home goals	Away goals	Total goals	Goal difference
1.2334	1.0985	1.1682	1.6319

Table 4.9: RMSE of score prediction using Linear Regression

Home goals	Away goals	Total goals	Goal difference
-1.281 %	17.014 %	7.177 %	81.796 %

Table 4.10: MPE of score prediction using Linear Regression

+/-	1	X	2	Count
-3	0	0	1	4
-2	.1135	.2162	.6703	185
-1	.2806	.2658	.4536	2220
0	.4134	.2912	.2954	3128
1	.5695	.2451	.1854	4002
2	.7657	.136	.0982	397
3	1	0	0	18

Table 4.11: Predicted goal difference using Linear Regression

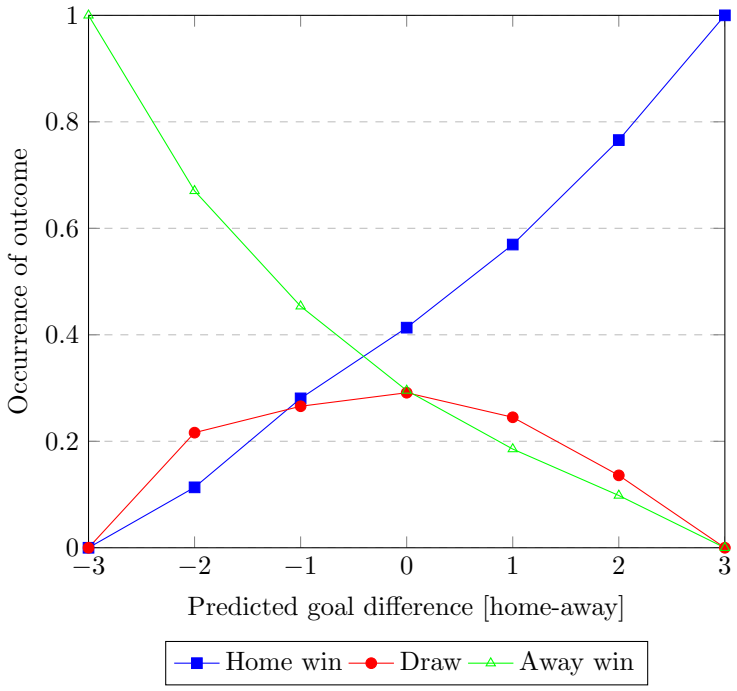

Figure 4.2: Graph of predicted goal difference using Linear Regression

Chapter 5

Discussion

5.1 Model

Probably the most decisive part of this project was to model players and teams abilities. This is mostly because putting a number on such abilities is arbitrary. And even though relatively large data is used to approximate the abilities of players, the variance in quantity between different statistic is big. So even if the model is good the variance in quantity can create big fluctuations in a players abilities between matches e.g. attacking players shoot much more then defending players, even though it is considered how often they shoot it a defensive player scoring on his only shot might be rated as superior to a attacking player

Another possible problem that's not investigated in this report is whether the model bias one of the algorithms. This however is quite likely to be true since the model may be constructed in infinitely many different ways and at least one of them should result in k-NN out performing Linear Regression.

5.2 k-Nearest Neighbor

In general when looking at the graph plotting predicted goal difference and outcomes, k-NN performs as expected, home wins increases as goal difference increases and the same pattern is present for away wins, as they decrease as the goal difference increase. It is not equally clear that the drawn games happen as desired, but they got a peak when the goal difference is 0, and decline in booth directions, until the numbers get statistical uncertain.

Compared to the random predictions k-NN performs better, but not convincingly better. Especially the away wins predictions can not be guaranteed to be any better then the random prediction. If it is the model biasing Linear regression or something else cannot be determined using data from this project.

5.3 Linear Regression

The prediction/outcome graph of Linear Regression behaves as expected, the number of matches won by the home team increases as the predicted goal difference increase, the number of away wins decrease in the same time, and the number of draws have a peak at the point where the goal difference is 0.

It is also clear that the Linear Regression performs significantly better then random predictions.

5.4 Conclusion

The problem statement consists of two questions. Is it possible to use k-NN or Linear Regression to predict football matches? Which of the two algorithms is most accurate in its predictions? To answer these questions both algorithms were implemented and the result were graphed and displayed in tables. To test if they are useful at all, they were compared to random predictions. It is considered that if they perform worse or equally to the random predictions that they are to no use.

Therefore the first question can be answered by comparing the random precision to the precision off the algorithmic predictions. Booth algorithms performs better then random event though its a thin margin in some cases for k-NN. Linear Regression on the other hand out performs booth k-NN and the random predictions. We can conclude that Linear Regression with current model works in this regard, and that k-NN probably do as well even though it's not as clear.

To answer the second question various results from the two algorithms are compared. And It is clear that Linear Regression performs better then k-NN. It is possible that this is because of that the model used in this project biases Linear Regression, to clearly state which of the two is best, it is needed to do further testing with different models. But whitin this project Linear Regression out performed k-NN.

It is however very interesting that this simple model produces such good results, and one cannot help but wondering, how well it could perform with a more complex model and/or more complex algorithms.

Bibliography

[1] Allen Moody. *Sports Betting Basics*. CreateSpace Independent Publishing Platform, 2013.

[2] Bernard Marr. How big data and analytics are changing football.

[3] Daniel T. Larose and Chantal D. Larose. *Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition*. John Wiley & Sons, Inc., 2014.

[4] Dr. Saed Sayad. K nearest neighbors - classification.

[5] David M. Lane. Introduction to linear regression.

[6] Thomas P. Ryan. *Statistical Methods for Quality Improvement, Third Edition*. John Wiley & Sons, Inc., 2011.

[7] John Goddard and andIoannis Asimakopoulos. Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23(1):51–66, 2004.

[8] Charu C. Aggarwal. *Data Mining*. Springer International Publishing, 2015.