Enhancing Genetic Programming for Predictive Modeling

av

Rikard König

Akademisk avhandling

Avhandling för teknologe doktorsexamen i datavetenskap, som kommer att föras offentligt tisdagen den 11 mars 2014 kl. 13.15, M404, Högskolan i Borås

Opponent: docent Niklas Lavesson
Blekinge Tekniska Högskola
Karlskrona

Örebro universitet
Institutionen för Naturvetenskap och Teknik

701 82 ÖREBRO
Abstract

Rikard König (2014): Enhancing Genetic Programming for Predictive Modeling
Örebro Studies in Computer Science 58.

Traditional predictive modeling techniques are most often optimized using a local
greedy strategy and are tied to specific representations and optimization criteria, leading
to potentially suboptimal solutions with regard to both predictive performance and com-
prehensibility. However, using genetic programming (GP), both the model representa-
tion of the solution and the optimization criteria can be designed specifically for the
problem at hand and the model is optimized globally. On the other hand, GP has disad-
vantages that traditional techniques do not have, i.e., it is inherently inconsistent, struggles
with large search spaces, and produces programs that often contain introns. Further-
more, the advantages are rarely fully exploited in the few predictive modeling frame-
works that incorporate GP.

Hence, this thesis aims to enhance GP for predictive modeling, by exploiting the ad-
vantages and countering the deficiencies. A set of criteria for frameworks of predictive
modeling based on GP are suggested and several novel techniques that improve the
accuracy or comprehensibility of predictive models produced using GP are presented.

With regard to comprehensibility, three novel techniques are proposed, of which the
first produces accurate and comprehensible models through rule extraction, the second
removes introns for arbitrary representation, and the third generates alternative solutions
and helps select an accurate solution that conforms to domain knowledge.

A point made in this thesis is that inconsistency should be considered an advantage
and not a disadvantage. In fact, the technique of generating and selecting among alterna-
tive models is, for example, only possible due to inconsistency. Furthermore, this thesis
also presents two additional novel techniques which enhance probability estimates of the
produced programs and create accurate kNN-ensembles by exploiting inconsistency.

A step towards handling large search spaces is taken by applying local search in two
novel techniques. The first injects induced mutated decision trees, when a reasonable
performance is not achieved for classification tasks, while the second uses the least
squares method to increase the predictive performance on regression problems.

Finally, a GP framework for predictive modeling, i.e., G-REX, has been implement-
ed. G-REX fulfills most of the suggested criteria, realizes many of the presented tech-
niques, and is available to the public on www.grex.se.

Keywords: Genetic Programming, Predictive Modeling, Data Mining, Machine
Learning, Rule Extraction, Classification, Regression, Ensembles,
Comprehensibility.

Rikard König, Department
Örebro University, SE-701 82 Örebro, Sweden, rikard@konigs.org
Enhancing Genetic Programming for Predictive Modeling

av

Rikard König

Akademisk avhandling

Avhandling för teknologi doktorsexamen i datavetenskap, som kommer att försvaras offentligt tisdagen den 11 mars 2014 kl. 13.15, M404, Högskolan i Borås

Opponent: docent Niklas Lavesson
Blekinge Tekniska Högskola
Karlskrona

Örebro universitet
Institutionen för Naturvetenskap och Teknik

701 82 ÖREBRO
Abstract


Traditional predictive modeling techniques are most often optimized using a local greedy strategy and are tied to specific representations and optimization criteria, leading to potentially suboptimal solutions with regard to both predictive performance and comprehensibility. However, using genetic programming (GP), both the model representation of the solution and the optimization criteria can be designed specifically for the problem at hand and the model is optimized globally. On the other hand, GP has disadvantages that traditional techniques do not have, i.e., it is inherently inconsistent, struggles with large search spaces, and produces programs that often contain introns. Furthermore, the advantages are rarely fully exploited in the few predictive modeling frameworks that incorporate GP.

Hence, this thesis aims to enhance GP for predictive modeling, by exploiting the advantages and counteracting the deficiencies. A set of criteria for frameworks of predictive modeling based on GP are suggested and several novel techniques that improve the accuracy or comprehensibility of predictive models produced using GP are presented.

With regard to comprehensibility, three novel techniques are proposed, of which the first produces accurate and comprehensible models through rule extraction, the second removes introns for arbitrary representation, and the third generates alternative solutions and helps select an accurate solution that conforms to domain knowledge.

A point made in this thesis is that inconsistency should be considered an advantage and not a disadvantage. In fact, the technique of generating and selecting among alternative models is, for example, only possible due to inconsistency. Furthermore, this thesis also presents two additional novel techniques which enhance probability estimates of the produced programs and create accurate kNN-ensembles by exploiting inconsistency.

A step towards handling large search spaces is taken by applying local search in two novel techniques. The first injects induced mutated decision trees, when a reasonable performance is not achieved for classification tasks, while the second uses the least squares method to increase the predictive performance on regression problems.

Finally, a GP framework for predictive modeling, i.e., G-REX, has been implemented. G-REX fulfills most of the suggested criteria, realizes many of the presented techniques, and is available to the public on www.grex.se.

Keywords: Genetic Programming, Predictive Modeling, Data Mining, Machine Learning, Rule Extraction, Classification, Regression, Ensembles, Comprehensibility.

Rikard König, Department
Örebro University, SE-701 82 Örebro, Sweden, rikard@konigs.org