Comparing a Genetic Algorithm with a Rule Induction Algorithm in the Data Mining Task of Dependence Modeling

Edgar Noda
CEFET-PR
CPGEI.
Av. 7 de Setembro, 3165.
Curitiba – PR.
80230-901. Brazil
edgar@dainf.cefetpr.br

Alex A. Freitas
PUC-PR
PPGIA-CCET.
R. Imaculada Conceicao, 1155
Curitiba – PR. 80.215-901. Brazil
alex@ppgia.pucpr.br
http://www.ppgia.pucpr.br/~alex

Heitor S. Lopes
CEFET-PR
CPGEI.
Av. 7 de Setembro, 3165.
Curitiba - PR.
80230-901. Brazil
hslopes@cpgei.cefetpr.br

1 OVERVIEW OF THE WORK

In this paper we compare two kinds of data mining algorithm: a genetic algorithm (GA) and a rule induction one. The data mining task for which the two algorithms were developed is dependence modeling. This task can be regarded as a generalization of the classification task [Noda et al. 1999]. In classification there is a single goal attribute to be predicted, while in dependence modeling there is more than one goal attribute. Hence, different rules can predict different goal attributes.

The GA used in our experiments, called GA-Nuggets-2.0 (version 2.0), has been introduced in [Noda et al. 1999]. This GA combines some characteristics of GA-Nuggets [Freitas 1999] with an information-theoretic, objective measure of rule interestingness [Freitas 1998], to favor the discovery of interesting rules. GA-Nuggets-2.0 uses binary tournament and uniform crossover. It uses new insert-condition and remove-condition operators that try to directly control the size of the rules, to favor the discovery of shorter rules. Once all operators have been applied to an individual and its corresponding rule antecedent is formed, the algorithm chooses the best consequent for each rule (individual) in such a way that maximizes its fitness.

The other algorithm used in our experiments is a greedy rule induction algorithm. It starts with an empty rule antecedent (the IF part of an IF-THEN rule), and then it iteratively adds one rule condition at a time to that antecedent, while there is an improvement in rule quality. This kind of greedy strategy is commonplace in rule induction algorithms, including decision-tree ones [Quinlan 1993]. In order for the rule induction algorithm to discover the same number of rules as GA-Nuggets-2.0, we run it once for every possible goal-attribute value to be predicted.

Both GA-Nuggets-2.0 and the greedy rule induction algorithm use the same fitness (evaluation) function, which consists of two parts. The first one measures the degree of interestingness of the rule, while the second one measures its predictive accuracy. The value of the fitness function is a weighted average of the interestingness and predictive accuracy of the rule.

2 EXPERIMENTS AND CONCLUSIONS

The experiments have used three datasets obtained from the well-known UCI repository of machine learning datasets. The datasets were Nursery, Zoo and Auto-mpg-hp. The results were obtained by performing a 5-fold cross-validation procedure. For GA-Nuggets-2.0 the population size was 50 for the Nursery dataset and 100 for the Zoo and Auto dataset. The number of generations was 100 for all three datasets. For each of the two algorithms (GA-Nuggets-2.0 and greedy rule induction) we measured the quality of each of the discovered rules. Hence, a comparison between the rules discovered by the two algorithms is fair, since both algorithms discover the same number and type of rules - i.e. one rule for each goal attribute-value to be predicted – and use the same fitness (evaluation) function.

We found that both algorithms discover highly interesting rules – according to the definition of rule interestingness used in the evaluation function. However, overall the predictive accuracy of the rules discovered by the GA turned out to be better than the predictive accuracy of the rules discovered by the rule induction algorithm.

REFERENCES

Freitas, A. A. (1998) On objective measures of rule surprisingness. *Proc. PKDD-98. Lecture Notes in Artificial Intelligence 1510*, 1-9. Springer-Verlag.

Freitas, A. A. (1999) A genetic algorithm for generalized rule induction. In: R. Roy et al. *Advances in Soft Computing – Eng. Design and Manufacturing,* 340-353. Springer-Verlag.

Noda, E.; Freitas, A.A. and Lopes, H.S. (1999) Discovering interesting prediction rules with a genetic algorithm. *Proc. CEC-99*, 1322-1329. Washington D.C., USA.

Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning,* Morgan Kaufmann.