Rule Acquisition with a Genetic Algorithm

Robert Cattral
Intelligent Systems Research Unit
School of Computer Science
Carleton University
Ottawa, On K1S 5B6

Franz Oppacher
Intelligent Systems Research Unit
School of Computer Science
Carleton University
Ottawa, On K1S 5B6

Dwight Deugo
Intelligent Systems Research Unit
School of Computer Science
Carleton University
Ottawa, On K1S 5B6

Abstract

This paper describes the implementation and the functioning of RAGA (Rule Acquisition with a Genetic Algorithm), a genetic-algorithm-based data mining system suitable for both supervised and certain types of unsupervised knowledge extraction from large and possibly noisy databases. RAGA differs from a standard Genetic Algorithm in several crucial respects, including the following: (i) its 'chromosomes' are variable-length symbolic structures, i.e. association rules that may contain n-place predicates ($n \geq 0$), (ii) besides typed crossover and mutation operators, it uses macromutations as generalization and specialization operators to efficiently explore the space of rules, and (iii) it evolves a default hierarchy of rules. Several data mining experiments with the system are described.

Data mining, also known as KDD, or *Knowledge Discovery in Databases*, refers to the attempt to extract previously unknown and potentially useful relations and other information from databases and to present the acquired knowledge in a form that is easily comprehensible to humans.

RAGA meets the comprehensibility requirement by working with a population of variable-length, symbolic rule structures that can accommodate not just feature-value pairs but arbitrary n-place predicates ($n \geq 0$), while exploiting the proven ability of the Genetic Algorithm (Holland 75, Mitchell 96) to efficiently search large spaces.

IF-THEN Rules

An important type of knowledge acquired by many data mining systems takes the form of *if-then rules*. Such rules state that the presence of one or more items implies or predicts the presence of other items. A typical rule has the form

$$If\ X_1 \wedge X_2 \wedge \ldots \wedge X_n,\ then\ Y.$$

In RAGA, different rules will often have a different number of conjuncts in the antecedent and in the consequent, and a user-supplied parameter limits the maximum number of such conjuncts. Each part of the antecedent as well as the expression in the consequent can contain n-place predicates.

To assess the quality of such a rule, data mining systems detmine its confidence and support. The *confidence* for a given rule is a measure of how often the consequent is true, given that the antecedent is true. The *support* indicates how often the rule holds in a set of data.

RAGA operates such that the confidence and support values for potential rules are maximized in accordance with user specification. The target values vary depending on the application, and the type of data being examined.

Experiments

RAGA has been tested as a classification system using several data sets, the first of which contains 8124 sample descriptions of 23 species of gilled mushrooms in the Agaricus and Lepiota Family (drawn from [Lincoff 81] and presented in [Schlimmer 87]). The data set uses 22 attributes, and classifies each mushroom as either edible (51.8%) or poisonous.

Each of 9 test runs produced between 14 and 25 rules. Each rule set yields 100% accuracy for the entire set. This compares favorably with STAGGER (Schlimmer 87) and HILLARY (Iba et al. 88) which approach 95% classification accuracy after training on 1000 instances.

References

[Holland 75] J. Holland (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan press.

[Iba et al. 88] W. Iba, J. Wogulis, P. Langley. *Trading off Simplicity and Coverage in Incremental Concept Learning*. Proceedings of 5th Int. Conference on Machine Learning. Morgan Kaufmann, Ann Arbor, Michigan.

[Lincoff 81] G. H. Lincoff (1981). *The Audubon Society Field Guide to North American Mushrooms*. Alfred A. Knopf, New York.

[Mitchell 96] Melanie Mitchell. *An Introduction to Genetic Algorithms*. MIT Press.

[Schlimmer 87] J. S. Schlimmer (1987). *Concept Acquisition Through Representational Adjustment*. Doctoral dissertation, University of California, Irvine.

Proceedings of the Genetic and Evolutionary Computation Conference

Edited by

Wolfgang Banzhaf

Jason Daida

Agoston E. Eiben

Max H. Garzon

Vasant Honavar

Mark Jakiela

Robert E. Smith

July 13–17, 1999
Orlando, Florida

Volume 1