A Summary of the Papers Presented at the AAAI-99 & GECCO-99 Workshop on Data Mining with Evolutionary Algorithms: Research Directions

Alex A. Freitas

PUC-PR, PPGIA-CCET
Rua Imaculada Conceicao, 1155
Curitiba – PR, 80215-901. Brazil
alex@ppgia.pucpr.br
http://www.ppgia.pucpr.br/~alex

This extended abstract summarizes the papers presented at the AAAI-99 & GECCO-99 Workshop on Data Mining with Evolutionary Algorithms: Research Directions. The general goal of the workshop is to discuss promising and necessary research directions in the integration of two areas, namely data mining and evolutionary algorithms. The workshop consists of six paper presentations, as follows.

The paper by Hsu, Pottenger, Welge, Wu and Yang, titled Genetic Algorithms for Selection and Partitioning of Attributes in Large-Scale Data Mining Problems, focuses on constructive induction (attribute synthesis) with genetic algorithms. The paper first presents a brief introduction to constructive induction. Next, it summarizes existing search-based approaches to knowledge-based construction induction. Then it discusses some relevant issues related to GA-based constructive induction, particularly chromosome encoding. Finally, it addresses critical scalability issues for industrial applications, including the exploitation of parallelism in genetic algorithms for constructive induction.

The paper by Thomas & Sycara, titled The Importance of Simplicity and Validation in Genetic Programming for Data Mining in Financial Data, describes a genetic programming-based system for trading rule discovery and evaluates its performance over real world exchange rate data in the dollar/yen and dollar/dm markets. They focus on the issues of rule complexity and how to fight overfitting.

The paper by Weiss, Mining Predictive Patterns in Sequence of Events, addresses the problem of predicting rare events from a sequence of events. The author has developed a GA-based system, called Timeweaver, that, given a pre-specified “target” event, learns to identify patterns in the data that successfully predict the future occurrence of that event. In essence, Timeweaver uses a Michigan-style GA to evolve a set of prediction rules. It also uses a niching strategy to ensure that a diverse set of rules is achieved. Timeweaver has been applied to the task of predicting telecommunication failures from time-stamped alarm messages.

The paper by Guerra-Salcedo & Whitley, titled Feature Seelction Mechanisms for Ensemble Creation, combines the outputs of several runs of a genetic algorithm into one ensemble of classifiers. They have experimented with two classifiers, namely C4.5 and Euclidean Decision Tables, and with several ensemble-construction methods, including bagging and boosting. The main goal of this paper is to compare the accuracy of ensembles of classifiers constructed using three different feature selection methods. First, features selected by a genetic algorithm. Second, features selected by another ensemble-construction method proposed in the literature, based on a random selection of features. Third, ensembles constructed using the complete set of features available.

The paper by Thompson, titled Genetic Algorithms as Post Processors for Data Mining, is also related to ensemble of classifiers. However, instead of using a GA to construct an ensemble of classifiers, the author proposes to use a GA to prune an ensemble of classifiers. The GA works with a real-valued encoding, and it is used to optimize the weightings of the classifiers in the ensemble. The paper also discusses some research directions in the use of GAs for post processing of the discovered knowledge, including the supplementation of this knowledge – e.g. the discovery of new rules covering data that is not correctly classified by the current rules.

The paper by Monmarche, titled On Data Clustering with Artificial Ants, presents a hybrid method for clustering, combining the stochastic principles of the artificial ants paradigm with a well-known statistical method for clustering, the K-means algorithm. The results of this hybrid algorithm are compared against the results achieved by two other clustering algorithms, namely K-means and ISODATA.

The Proceedings of the Workshop will be published by AAAI Press as an AAAI Technical Report.