Genetic Algorithms for Reformulation of Large-Scale KDD Problems with Many Irrelevant Attributes

William H. Hsu
bhsu@cis.ksu.edu

Yuhong Cheng
ych7945@cis.ksu.edu

Haipeng Guo
hpguo@cis.ksu.edu

Steven M. Gustafson
steveg@cis.ksu.edu

Laboratory for Knowledge Discovery in Databases
http://ringil.cis.ksu.edu/KDD
Computing and Information Sciences (CIS) Department, Kansas State University
Manhattan, KS 66506

Abstract

The goal of this research is to apply genetic implementations of algorithms for *selection*, *partitioning*, and *synthesis* of attributes in large-scale data mining problems. Domain knowledge about these operators has been shown to reduce the number of fitness evaluations for candidate attributes. We report results on genetic optimization of attribute selection problems and current work on attribute partitioning, synthesis specifications, and the encoding of domain knowledge about operators in a fitness function. The purpose of this approach is to reduce overfitting in inductive learning and produce more general genetic versions of existing search-based algorithms (or *wrappers*) for KDD performance tuning [KS98, HG00]. Several GA implementations of alternative attribute synthesis algorithms are applied to concept learning problems in military and commercial KDD applications. One of these, *Jenesis*, is deployed on several network-of-workstation clusters. It is shown to achieve strongly improved test set accuracy, compared to unwrapped decision tree learning and search-based wrappers [KS98].

1 GA WRAPPERS AND RELEVANCE

Our research addresses reduction, decomposition, and **reformulation** of large-scale concept learning problems in knowledge discovery in databases (KDD) by means of GA-based optimization. The approach described here adapts the methodology of *wrappers* for performance enhancement and attribute subset selection [KS98] to a more general problem for selection, partitioning, and synthesis of attributes (feature subset selection, partitioning, construction, and extraction) [RPG+97]. Systems of this type apply *relevance determination* criteria to attributes from those specified for the original data set. The selected and synthesized attributes are used to define new data clusters that are used as intermediate training targets. The purpose of this *change of representation* step is to improve the accuracy of supervised learning using the reformulated data. Fitness is defined in terms of classification accuracy on cross-validation data (or continuations of time series data), given a particular supervised learning technique (or *inducer*) [KS98]. Our decision support projects use very large databases (from vehicular sensors in the military application, and from historical and demographic customer records in the commercial application) for predictive classification [HG00]. The research presented here focuses on genetic optimization of the inductive learning steps (accuracy and model minimization). For this purpose, we are developing a "model-tuning" GA that abstracts the feature selection and construction wrappers.

2 PRELIMINARY RESULTS

The *Jenesis* wrapper is shown to achieve significant improvement in classification accuracy and reduction in the number of attributes used, both in simple test corpora containing irrelevant attributes and our KDD test beds [HG00]. Genetic wrappers also escape local optima better than state-space search-based wrappers [KS98]. Results for data sets from the Irvine database repository that are known to contain irrelevant attributes are also positive. For implementation and experimental details, we refer the interested reader to [HG00].

3 REFERENCES

[HG00] W. H. Hsu and S. M. Gustafson. Constructive Induction Wrappers in High-Performance Data Mining and Decision Support. KDD Technical Report, http://ringil.cis.ksu.edu/KDD, to appear, 2000.

[KS98] R. Kohavi and D. Sommerfield. *MLC++: Machine Learning Library in C++, Utilities v2.0*. URL: http://www.sgi.com/Technology/mlc. 1998.

[RPG+97] M. Raymer, W. Punch, E. Goodman, P. Sanschagrin, and L. Kuhn, Simultaneous Feature Extraction and Selection using a Masking Genetic Algorithm, In *Proceedings of ICGA-97*, pp. 561-567, San Francisco, CA, July, 1997.