Genetic Wrappers for Constructive Induction in High-Performance Data Mining

William H. Hsu[1,2] bhsu@cis.ksu.edu
Michael Welge[2] welge@ncsa.uiuc.edu
Thomas Redman[2] redman@ncsa.uiuc.edu
David Clutter[2] clutter@ncsa.uiuc.edu

[1]*Lab for Knowledge Discovery in Databases*
http://ringil.cis.ksu.edu/KDD
Kansas State University
Manhattan, KS 66506

[2]*Automated Learning Group*
http://www.ncsa.uiuc.edu/STI/ALG
National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana-Champaign
Champaign, IL 61820

Abstract

We present an application of genetic algorithm-based design to configuration of high-level optimization systems, or *wrappers*, for relevance determination and constructive induction. Our system combines genetic wrappers with elicited knowledge on attribute relevance and synthesis. We discuss decision support issues in a large-scale commercial data mining project (cost prediction for multiple automobile insurance markets), and report experiments using *D2K*, a Java-based visual programming system for data mining and information visualization, and several commercial and research tools. Our GA system, *Jenesis* [HWRC00], is deployed on several network-of-workstation systems (Beowulf clusters). It achieves a linear speedup, due to a high degree of task parallelism, and improved test set accuracy, compared to decision tree learning with only constructive induction and state-space search-based wrappers [KJ97].

1 GENETIC WRAPPERS AND KDD

Our commercial decision support project applies a large demographic and historical customer database from the Allstate *One Company* project to a predictive classification problem in automobile insurance underwriting. The data mining (DM) pipeline we have developed comprises descriptive statistics, interactive visualization, data aggregation, data clustering, relevance determination, and supervised inductive learning for classification [HWRC00]. The research presented here focuses on genetic optimization of the inductive learning steps, where relevance determination is the objective of our attribute subset selection and synthesis (*aka* feature selection, extraction, and construction) stage. For this purpose, we have developed a genetic algorithm that calibrates hyperparameters in a validation-based model selection wrapper [RPG+97, HWRC00] written in a *D2K*-based GA, *Jenesis* [HWRC00].

2 EXPERIMENTAL RESULTS

As Table 1 shows, the *Jenesis* wrapper produces a very significant improvement in classification accuracy and a large reduction in the number of attributes used. Genetic wrappers also escape local optima better than state-space search-based wrappers [KJ97]. 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 [HWRC00].

Data Set (*ALLVAR-2*, 5 bins)	**Results**	
	Cross-Validated Accuracy	**Attributes Selected**
Unwrapped (ID3)	21.37%	47/97
Jenesis	50.00%	23/97

Table 1. Results for One Company, MLDBR data sets

3 REFERENCES

[HWRC00] W. H. Hsu, M. Welge, T. Redman, and D. Clutter. Constructive Induction Wrappers in High-Performance Commercial Data Mining and Decision Support Systems. NCSA Technical Report, http://chili.ncsa.uiuc.edu/jenesis.html, 2000.

[KJ97] R. Kohavi and G. H. John. Wrappers for Feature Subset Selection. *Artificial Intelligence 97*(1-2):273-324, 1997.

[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.