Genetic Algorithms for Attribute Synthesis in Large-Scale Data Mining

William H. Hsu **William M. Pottenger** **Michael Welge** **Jie Wu** **Ting-Hao Yang**

Automated Learning Group, National Center for Supercomputing Applications
605 East Springfield Avenue, Champaign IL 61820
{bhsu | billp | welge | jiewu | tingy}@ncsa.uiuc.edu http://www.ncsa.uiuc.edu/STI/ALG

Abstract

The goal of this research is to apply genetic implementations of algorithms for *selection*, *partitioning*, and *synthesis* of attributes in large-scale concept learning problems. Domain knowledge about these operators has been shown to reduce the number of fitness evaluations for candidate attributes. Our research examines the genetic encoding of attribute selection, partitioning, and synthesis specifications, and the encoding of domain knowledge about operators in a fitness function. The purpose of this approach is to improve upon existing search-based algorithms (or *wrappers*) in terms of training sample efficiency. Several GA implementations of alternative attribute synthesis algorithms are applied to concept learning problems in industrial KDD applications.

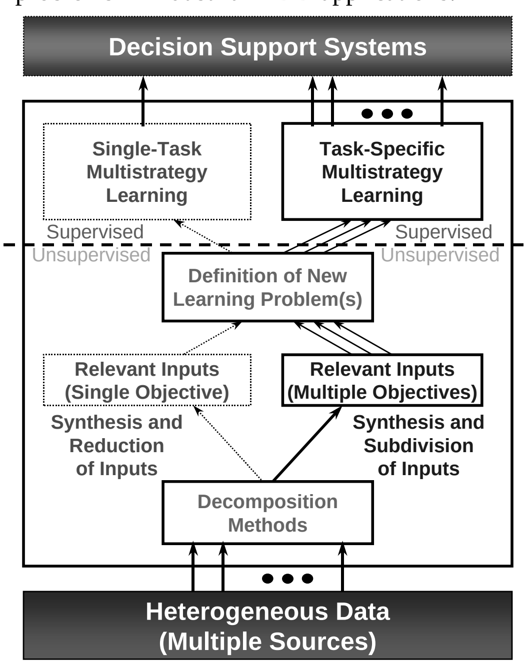

Figure 1. Attribute-based transformations in KDD

This research addresses the problems of reducing and decomposing large-scale concept learning problems in knowledge discovery in databases (KDD). The approach described here adapts the methodology of *wrappers* for performance enhancement and attribute subset selection [KJ97] to a genetic optimization problem. 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 functions are defined in terms of classification accuracy on cross-validation data (or continuations of time series data), given a particular supervised learning technique (or *inducer*) [KS96].

The synthesis of a new group of attributes (also known as the *feature construction* problem) in inductive concept learning is an optimization problem. Its control parameters include the attributes selected as relevant [KJ97, Hs98], how they are grouped (with respect to multiple targets), and how new attributes are defined in terms of *ground* (original) attributes. This synthesis and selection problem is a key initial step in *constructive induction* [DR95] – the reformulation of a learning problem in terms of its inputs (attributes) and outputs (concept class descriptors). Figure 1 illustrates the role of attribute selection (reduction of inputs) and partitioning (subdivision of inputs) in constructive induction (the "unsupervised" component of this generic KDD process). In this framework, the input consists of *heterogeneous* data (that originating from multiple sources). Supervised learning is to acquire the performance element (time series classification [Hs98] and other forms of pattern recognition that are important for decision support). Our applications include insurance risk valuation, precision agriculture, and record and document clustering for information retrieval.

References

[DR95] S. K. Donoho and L. A. Rendell. Rerepresenting and restructuring domain theories: A constructive induction approach. *Journal of Artificial Intelligence Research*, 2:411-446, 1995.

[Hs98] W. H. Hsu. *Time Series Learning With Probabilistic Network Composites*. Ph.D. thesis, UIUC. URL: http://www.ncsa.uiuc.edu/People/bhsu/thesis.html.

[KS96] R. Kohavi and D. Sommerfield. *MLC++ v2.0*. URL: http://www.sgi.com/Technology/mlc.

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