IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB8921

Chapter II

Control of Inductive Bias in Supervised Learning Using Evolutionary Computation: A Wrapper-Based Approach

William H. Hsu
Kansas State University, USA

ABSTRACT

In this chapter, I discuss the problem of feature subset selection *for supervised inductive learning approaches to knowledge discovery in databases (KDD), and examine this and related problems in the context of controlling* inductive bias. *I survey several combinatorial search and optimization approaches to this problem, focusing on data-driven, validation-based techniques. In particular, I present a* wrapper *approach that uses genetic algorithms for the search component, using a validation criterion based upon model accuracy and problem complexity, as the fitness measure. Next, I focus on design and configuration of high-level optimization systems (wrappers) for relevance determination and constructive induction, and on integrating these wrappers with elicited knowledge on attribute relevance and synthesis. I then discuss the relationship between this model selection criterion and those from the minimum description length (MDL) family of learning criteria. I then present results on several synthetic problems*

This chapter appears in the book, *Data Mining: Opportunities and Challenges,* edited by John Wang.

on task-decomposable machine learning and on two large-scale commercial data-mining and decision-support projects: crop condition monitoring, and loss prediction for insurance pricing. Finally, I report experiments using the Machine Learning in Java (MLJ) *and* Data to Knowledge (D2K) *Java-based visual programming systems for data mining and information visualization, and several commercial and research tools. Test set accuracy using a genetic wrapper is significantly higher than that of decision tree inducers alone and is comparable to that of the best extant search-space based wrappers.*

INTRODUCTION

This chapter introduces the problems for change of representation (Benjamin, 1990) in supervised inductive learning. I address the focal problem of inductive learning in data mining and present a multi-strategy framework for automatically improving the representation of learning problems. This framework incorporates methodological aspects of *feature subset selection and feature (attribute) partitioning, automated problem decomposition*, and *model selection*. The focus is on *wrapper-based* methods as studied in recent and continuing research.

As an example, I present a new metric-based model selection approach (composite learning) for decomposable learning tasks. The type of data for which this approach is best suited is heterogeneous, time series data – that arising from multiple sources of data (as in sensor fusion or multimodal human-computer interaction tasks, for example). The rationale for applying multi-strategy learning to such data is that, by systematic analysis and transformation of learning tasks, both the efficiency and accuracy of classifier learning may be improved for certain time series problems. Such problems are referred to in this chapter as *decomposable*; the methods addressed are: task decomposition and subproblem definition, quantitative model selection, and construction of hierarchical mixture models for data fusion. This chapter presents an integrated, multi-strategy system for decomposition of time series classifier learning tasks.

A typical application for such a system is learning to predict and classify hazardous and potentially catastrophic conditions. This prediction task is also known as *crisis monitoring*, a form of pattern recognition that is useful in decision support or *recommender* systems (Resnick & Varian, 1997) for many time-critical applications. Examples of crisis monitoring problems in the industrial, military, agricultural, and environmental sciences are numerous. They include: crisis control automation (Hsu *et al.*, 1998), online medical diagnosis (Hayes-Roth *et al.*, 1996), simulation-based training and critiquing for crisis management (Gaba *et al.*, 1992; Grois, Hsu, Voloshin, & Wilkins *et al.*, 1998), and intelligent data visualization for *real-time decision-making* (Horvitz & Barry, 1995).

Motivation: Control of Inductive Bias

The broad objectives of the approach I present here are to increase the robustness of inductive machine learning algorithms and develop learning systems that can be automatically tuned to meet the requirements of a knowledge discovery (KD) performance element. When developers of learning systems can map a KD application to a set