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 datadriven, 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. Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

on task-decomposable machine learning and on two large-scale commercial datamining 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 26 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: <u>www.igi-</u> <u>global.com/chapter/control-inductive-bias-supervised-</u> learning/7595

Related Content

Finding the Semantic Relationship Between Wikipedia Articles Based on a Useful Entry Relationship

Lin-Chih Chen (2017). International Journal of Data Warehousing and Mining (pp. 33-52).

www.irma-international.org/article/finding-the-semantic-relationship-between-wikipedia-articlesbased-on-a-useful-entry-relationship/188489

A Survey on Database Performance in Virtualized Cloud Environments

Todor Ivanov, Ilia Petrovand Alejandro Buchmann (2012). *International Journal of Data Warehousing and Mining (pp. 1-26).*

www.irma-international.org/article/survey-database-performance-virtualized-cloud/67571

Pattern Discovery in Biosequences: From Simple to Complex

Simona Este Romboand Luigi Palopoli (2008). Data Mining Patterns: New Methods and Applications (pp. 85-105).

www.irma-international.org/chapter/pattern-discovery-biosequences/7561

Mining Flow Patterns in Spatio-Temporal Data

Wynne Hsu, Mong Li Leeand Junmei Wang (2008). *Temporal and Spatio-Temporal Data Mining (pp. 157-188).* www.irma-international.org/chapter/mining-flow-patterns-spatio-temporal/30266

New Trends in Fuzzy Clustering

Zekâi Sen (2013). Data Mining in Dynamic Social Networks and Fuzzy Systems (pp. 248-288).

www.irma-international.org/chapter/new-trends-fuzzy-clustering/77531