On the Accuracy of Meta-learning for Scalable Data Mining

PHILIP K. CHAN  
pkc@cs.fit.edu

Computer Science, Florida Institute of Technology, Melbourne, FL 32901

SALVATORE J. STOLFO  
sal@cs.columbia.edu

Department of Computer Science, Columbia University, New York, NY 10027

Abstract. In this paper, we describe a general approach to scaling data mining applications that we have come to call meta-learning. Meta-Learning refers to a general strategy that seeks to learn how to combine a number of separate learning processes in an intelligent fashion. We desire a meta-learning architecture that exhibits two key behaviors. First, the meta-learning strategy must produce an accurate final classification system. This means that a meta-learning architecture must produce a final outcome that is at least as accurate as a conventional learning algorithm applied to all available data. Second, it must be fast, relative to an individual sequential learning algorithm when applied to massive databases of examples, and operate in a reasonable amount of time. This paper focussed primarily on issues related to the accuracy and efficacy of meta-learning as a general strategy. A number of empirical results are presented demonstrating that meta-learning is technically feasible in wide-area, network computing environments.

Keywords: machine learning, meta-learning, scalability, data mining, classifiers.

1. Introduction

Many believe that we are poised once again for a radical shift in the way we learn and work, and in the amount of new knowledge we will acquire. The coming age of high performance network computing, and widely available “data highways” will transform the “information age” into the “knowledge age” by providing new opportunities in defense, commerce, education and science for sharing and utilizing information. However, with this new technological capability comes along a number of hard technical problems, many centered on the issue of scale. It is perhaps obvious that having massive amounts of data and information available anywhere and anytime enables many new opportunities to acquire new knowledge. The field of data mining studies how precisely this will be achieved in an efficient and transparent fashion.

One means of acquiring new knowledge from databases is to apply various machine learning algorithms that compute descriptive representations of the data as well as patterns that may be exhibited in the data. The field of machine learning has made substantial progress over the years and a number of algorithms have been popularized and applied to a host of applications in diverse fields. Thus, we may simply apply the current generation of learning algorithms to very large databases and wait for a response! However, the question is how long might we wait? Indeed, do the current generation of machine learning algorithms scale from tasks common today that include thousands of data items to new learning tasks encompassing as much as two orders of magnitude or more of data that is
physically distributed? Furthermore, many existing learning algorithms require all the data to be resident in main memory, which is clearly untenable in many realistic databases. In certain cases, data is inherently distributed and cannot be localized on any one machine for a variety of practical reasons. In such situations it is infeasible to inspect all of the data at one processing site to compute one primary “global” classifier. We call the problem of learning useful new knowledge from large inherently distributed databases the scaling problem for machine learning.

Our approach to solve the scaling problem is to execute a number of learning processes (each implemented as a distinct serial program) on a number of data subsets (a data reduction technique) in parallel (eg. over a network of separate processing sites) and then to integrate the collective results through a process we call meta-learning (Chan and Stolfo, 1993). Without any integration, as we discuss later, individual results generated from the data subsets are far from desired. Here, meta-learning serves as the means of “gluing” multiple knowledge sources together.

We note with interest that this general meta-learning approach is independent of the underlying learning algorithms that may be employed. Furthermore, it is independent of the computing platform used. Thus, our meta-learning approach is intended to be scalable as well as portable and extensible. However, we may not be able to guarantee the accuracy of the final result to be as good as an individual learning algorithm applied to the entire data set since a considerable amount of information may not be accessible to each of the separate learning processes. It is this primary issue we study in this paper.

2. Related Work

In a relational database context, a typical data mining task is to explain and predict the value of some attribute of the data given a collection of tuples with known attribute values. An existing relation with attribute values drawn from some domain is thus treated as training data for a learning algorithm that computes a logical expression, a concept description or a classifier, that is later used to predict a value of the desired attribute for some “test datum” whose desired attribute value is unknown.

Before the details of our approach are discussed, we first summarize closely related work by others in improving the accuracy of learning algorithms applied to large amounts of data. Machine learning researchers clearly desire more accurate learning algorithms. One recent approach has focussed on integrating by some means multiple strategies or multiple algorithms. Some research has concentrated on methods to improve an existing algorithm by using the algorithm itself to generate purposely biased distributions of training data. The most notable work in this area is due to Schapire (Schapire, 1990). Schapire proves, under the theoretical PAC (Probabilistic Approximately Correct) learning model (Valiant, 1984), that his boosting technique can improve a “weak” learner to achieve arbitrary high accuracy.

Other researchers have proposed implementing learning systems by integrating in some fashion a number of different algorithms to boost overall accuracy. The basic notion behind this integration is to complement the different underlying learning strategies embodied by different learning algorithms by effectively reducing the space of incorrect classifications of a learned concept.