Chapter LXIII
Learning Classifiers from Distributed Data Sources

Doina Caragea
Kansas State University, USA

Vasant Honavar
Iowa State University, USA

INTRODUCTION

Recent development of high throughput data acquisition technologies in a number of domains (e.g., biological sciences, atmospheric sciences, space sciences, commerce) together with advances in digital storage, computing, and communications technologies have resulted in the proliferation of a multitude of physically distributed data repositories created and maintained by autonomous entities (e.g., scientists, organizations). The resulting increasingly data-rich domains offer unprecedented opportunities in computer assisted data-driven knowledge acquisition in a number of applications, including, in particular, data-driven scientific discovery, data-driven decision-making in business and commerce, monitoring and control of complex systems, and security informatics.

Machine learning (Duda, Hart & Stork, 2000; Mitchell, 1997) offers one of the most cost-effective approaches to analyzing, exploring, and extracting knowledge (i.e., features, correlations, and other complex relationships and hypotheses that describe potentially interesting regularities) from data. However, the applicability of current machine learning approaches in emerging data-rich applications is severely limited by a number of factors:

a. Data repositories are large in size, dynamic, and physically distributed. Consequently, it is neither desirable nor feasible to gather all of the data in a centralized location for analysis. Hence, there is a need for efficient

algorithms for analyzing and exploring multiple distributed data sources without transmitting large amounts of data.

b. Autonomously developed and operated data sources often differ in their structures and organizations (e.g., relational databases, flat files, etc.) and the operations that can be performed on the data sources (e.g., types of queries—relational queries, statistical queries, keyword matches). Hence, there is a need for theoretically well-founded strategies for efficiently obtaining the information needed for analysis within the operational constraints imposed by the data sources.

The purpose of this entry is to precisely define the problem of learning classifiers from distributed data and summarize recent advances that have led to a solution to this problem (Caragea, Silvescu & Honavar, 2004; Caragea, Zhang, Bao, Pathak & Honavar, 2005).

BACKGROUND: PROBLEM SPECIFICATION

Given a dataset D, a hypothesis class H, and a performance criterion P, an algorithm L for learning (from centralized data D) outputs a hypothesis $h \in H$ that optimizes P. In pattern classification applications, h is a classifier (e.g., a decision tree, a support vector machine, etc.) (see Figure 1). Data D typically consist of a set of training examples. Each training example is an ordered tuple of attribute values where one of the attributes corresponds to a class label and the remaining attributes represent inputs to the classifier. The goal of learning is to produce a hypothesis that optimizes the performance criterion (e.g., minimizes classification error on the training data) and the complexity of the hypothesis.

In a distributed setting, a dataset D is distributed among the sites 1,...,n containing the dataset fragments D_1,...,D_n. Two common types of data fragmentation are *horizontal fragmentation* and *vertical fragmentation*. In the case of horizontal fragmentation, each site contains a subset of data tuples that make up D, *for example,*

$$(D = \bigcup_{i=1}^{n} D_i).$$

In the case of vertical fragmentation, each site stores the subtuples of data tuples (corresponding to a subset of the attributes used to define data tuples in D). In this case, D can be constructed by taking the *join* of the individual datasets D_1,...,D_n (assuming a unique identifier for each data tuple is stored with the corresponding subtuples). More generally, the data may be fragmented into a set of relations, as in the case of tables of a relational database, but distributed across multiple sites (i.e., $D = \otimes_{i=1}^{n} D_i$), where $\otimes$ denotes the *join* operation (Friedman, Getoor, Koller & Pfeffer, 1999; Ozsu & Valduriez, 1999). If a dataset D is distributed among the sites 1,…,n containing dataset fragments D_1,..., D_n, we assume that the individual datasets D_1,...,D_n collectively contain (in principle) all the information needed to construct dataset D.

The distributed setting typically imposes a set of constraints Z on the learner (absent in the centralized setting). For example, the constraints Z may prohibit the transfer of raw data from each of the sites to a central location, while allowing the learner to obtain certain types of statistics from the individual sites (e.g., counts of instances that have specified values for some subset of attributes). In some applications of data mining (e.g., knowledge discovery from clinical records), Z might include constraints designed to preserve privacy.

Figure 1. Learning from centralized data