Chapter 1.10
Distributed Data Mining

Grigorios Tsoumakas
Aristotle University of Thessaloniki, Greece

Ioannis Vlahavas
Aristotle University of Thessaloniki, Greece

INTRODUCTION

The continuous developments in information and communication technology have recently led to the appearance of distributed computing environments, which comprise several, and different sources of large volumes of data and several computing units. The most prominent example of a distributed environment is the Internet, where increasingly more databases and data streams appear that deal with several areas, such as meteorology, oceanography, economy and others. In addition the Internet constitutes the communication medium for geographically distributed information systems, as for example the earth observing system of NASA (*eos.gsfc.nasa.gov*). Other examples of distributed environments that have been developed in the last few years are *sensor networks* for process monitoring and *grids* where a large number of computing and storage units are interconnected over a high-speed network.

The application of the classical knowledge discovery process in distributed environments requires the collection of distributed data in a data warehouse for central processing. However, this is usually either ineffective or infeasible for the following reasons:

(1) *Storage cost*. It is obvious that the requirements of a central storage system are enormous. A classical example concerns data from the astronomy science, and especially images from earth and space telescopes. The size of such databases is reaching the scale of exabytes (10^{18} bytes) and is increasing at a high pace. The central storage of the data of all telescopes of the planet would require a huge data warehouse of enormous cost.
(2) *Communication cost*. The transfer of huge data volumes over network might take extremely much time and also require an unbearable financial cost. Even a small volume of data might create problems in wireless network environments with limited

bandwidth. Note also that communication may be a continuous overhead, as distributed databases are not always constant and unchangeable. On the contrary, it is common to have databases that are frequently updated with new data or data streams that constantly record information (e.g remote sensing, sports statistics, etc.).

(3) *Computational cost*. The computational cost of mining a central data warehouse is much bigger than the sum of the cost of analyzing smaller parts of the data that could also be done in parallel. In a grid, for example, it is easier to gather the data at a central location. However, a distributed mining approach would make a better exploitation of the available resources.

(4) *Private and sensitive data*. There are many popular data mining applications that deal with sensitive data, such as people's medical and financial records. The central collection of such data is not desirable as it puts their privacy into risk. In certain cases (e.g. banking, telecommunication) the data might belong to different, perhaps competing, organizations that want to exchange knowledge without the exchange of raw private data.

This article is concerned with Distributed Data Mining algorithms, methods and systems that deal with the above issues in order to discover knowledge from distributed data in an effective and efficient way.

BACKGROUND

Distributed Data Mining (DDM) (Fu, 2001; Park & Kargupta, 2003) is concerned with the application of the classical Data Mining procedure in a distributed computing environment trying to make the best of the available resources (communication network, computing units and databases). Data Mining takes place both locally at each distributed site and at a global level where the local knowledge is fused in order to discover global knowledge.

A typical architecture of a DDM approach is depicted in Figure 1. The first phase normally involves the analysis of the local database at each distributed site. Then, the discovered knowledge is usually transmitted to a merger site, where the integration of the distributed local models is performed. The results are transmitted back to the distributed databases, so that all sites become updated with the global knowledge. In some approaches, instead of a merger site, the local models are broadcasted to all other sites, so that each site can in parallel compute the global model.

Distributed databases may have homogeneous or heterogeneous schemata. In the former case, the attributes describing the data are the same in each distributed database. This is often the case when the databases belong to the same organization (e.g. local stores of a chain). In the latter case the attributes differ among the distributed databases. In certain applications a key attribute might be present in the heterogeneous databases, which will allow the association between tuples. In other applications the target attribute for prediction might be common across all distributed databases.

MAIN FOCUS

Distributed Classification and Regression

Approaches for distributed classification and regression are mainly inspired from methods that appear in the area of ensemble methods, such as Stacking, Boosting, Voting and others. Some distributed approaches are straightforward adaptations of ensemble methods in a distributed computing environment, while others extend the existing approaches in order to minimize the communication and coordination costs that arise.