IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB7997

Chapter V

Optimization of the Knowledge Discovery Process in Very Large Databases

M. Mehdi Owrang
American University, USA

Current database technology involves processing a large volume of data in order to discover new knowledge. The high volume of data makes the discovery process computationally expensive. In addition, real-world databases tend to be incomplete, redundant and inconsistent which could lead to discovery of redundant and inconsistent knowledge. We propose use of domain knowledge to reduce the size of the database being considered for discovery and to optimize the hypothesis (representing the pattern to be discovered) by eliminating implied, unnecessary and redundant conditions from the hypothesis. The benefits can be greater efficiency and the discovery of more meaningful, non-redundant, non-trivial and consistent rules. Experimental results are provided and analyzed.

INTRODUCTION

Modern database technology involves processing a large volume of data in databases in order to discover new knowledge. Knowledge discovery is defined as the nontrivial extraction of implicit, previously unknown and potentially useful information from data (Adriaans et al., 1996; Fayyad, 1996; Fayyad et al., 1996; Ganti et al., 1999; Groth, 1998). While promising, the available discovery schemes

and tools are limited in many ways. Some databases are so large that they make the discovery process computationally expensive. Databases containing on the order of $N=10^9$ records are becoming increasingly common, for example, in the astronomical sciences. Similarly, the number of fields can easily be on the order of 10^2 or 10^3, for example, in medical diagnostic applications (Fayyad, 1996).

If we apply discovery algorithms to discover all the correlations between concepts in a real database, we will generally observe the production of a set of results whose size is just too large to be handled in a useful manner. Another major concern in knowledge discovery, in addition to the large size of the databases, is data redundancy in the databases. Databases include data redundancies that could lead to discovering redundant knowledge. A common form of redundancy is a functional dependency in which a field is defined as a function of other fields, for example, profit = sales - expenses. The discovered knowledge may contain redundancy when two pieces of knowledge are exactly the same (rules having the same premises and conclusions) or semantically equivalent. In addition, the discovered knowledge may indeed be a previously known fact (i.e., a domain knowledge) rather than a new discovery. In addition to data redundancy, data inconsistency in databases is another issue in knowledge discovery. Databases are normally incomplete; thus, discovered knowledge may be inconsistent and inaccurate. It may be impossible to discover significant knowledge about a given domain (i.e., medicine) if some attributes essential to knowledge about the application domain are not present in the data. For example, we cannot diagnose Malaria from a patient database if the data does not contain the patient's red blood cell counts.

A major challenge in knowledge discovery is computational efficiency. The vastness of the data and the data redundancy that exists in databases force us to use techniques for optimizing the discovery of consistent and accurate patterns. A pattern represents the useful knowledge to be discovered from the database.

There are several approaches to knowledge discovery for handling the large volume of data and minimizing search efforts. These techniques include: parallel processor architecture, providing some measure of "interestingness of patterns," elimination of irrelevant attributes, data sampling, data segmentation, and data summarization. These techniques are used to reduce the size of the databases for discovery and to define a bias in searching for interesting patterns. These techniques are described in greater detail later in the chapter.

The human user almost always has some previous concepts or knowledge about the domain represented by the database. This information, known as domain or background knowledge, can be defined as any information that is not explicitly presented in the data (Adriaans et al., 1996; Fayyad, 1996; Owrang, 1997), including the relationship (or lack of it) that exists among attributes, constraints imposed on data and redundant data definition. The domain knowledge reduces search time by optimizing the hypothesis associated with knowledge discovery. A hypothesis represents the pattern to be discovered.

Once the concept of domain knowledge is defined, it can be incorporated into a Knowledge Discovery System (KDS). A KDS is a system that finds knowledge