Knowledge Discovery in Databases

Aakash Sarmandal

Research Scholar, Jodhpur National University, Rajasthan, INDIA

Abstract: At an abstract level, the knowledge discovery in databases (KDD) field is concerned with the development of methods and techniques for making sense of data. The basic problem addressed by the KDD process is one of mapping low-level data into other forms that might be more compact

Keywords: Mapping of data, KDD process, misconception about data mining, Comford's Views, A&Z viewpoints.

---♦--

1. INTRODUCTION

Data mining and Knowledge Discovery in Databases have become commercially important techniques and active areas of research in recent years. Business applications of data mining software are commonplace and are commodities in many cases. In this work, a literature survey of data mining is given

The basic problem addressed by the KDD process is one of mapping low-level data (which are typically too voluminous to understand and digest easily) into other forms that might be more compact

(for example, a short report), more abstract (for example, a descriptive approximation or model of the process that generated the data), or more useful (for example, a predictive model for estimating the value of future cases). At the core of the process is the application of specific data-mining methods for pattern discovery and extraction. The traditional method of turning data into knowledge relies on manual analysis and interpretation. For example, in the health-care industry, it is common for specialists to periodically analyze current trends and changes in health-care data, say, on a quarterly basis.

The specialists then provide a report detailing the analysis to the sponsoring health-care organization; this report becomes the basis for future decision making and planning for health-care management. In a totally different type of application, planetary geologists sift through remotely sensed images of planets and asteroids, carefully locating and cataloging such geologic objects of interest as impact craters. Be it science, marketing, finance, health care, retail, or any other field, the classical approach to data analysis relies fundamentally on one or more analysts becoming innovaintimately familiar with the data and serving as an interface between the data and the users and products.

For these (and many other) applications, this form of manual probing of a data set is slow, expensive, and highly subjective. In fact, as data volumes grow dramatically, this type of manual data analysis is becoming completely impractical in many domains.

Databases are increasing in size in two ways: (1) the number *N* of records or objects in the database and (2) the number *d* of fields or attributes to an object. Databases containing on the order of *N* = 109 objects are becoming increasingly common, for example, in the astronomical sciences. Similarly, the number of fields *d* can easily be on the order of 102 or even 103, for example, in medical diagnostic applications. Who could be expected to digest millions of records, each having tens or hundreds of fields? We believe that this job is certainly not one for humans; hence, analysis work needs to be automated, at least partially.

The need to scale up human analysis capabilities to handling the large number of bytes that we can collect is both economic and scientific. Businesses use data to gain competitive advantage, increase efficiency, and provide more valuable services to customers. Data we capture about our environment are the basic evidence we use to build theories and models of the universe we live in. Because computers have enabled humans to gather more data than we can digest, it is only natural to turn to computational techniques to help us unearth meaningful patterns and structures from the massive volumes of data. Hence, KDD is an attempt to address a problem that the digital information era made a fact of life for all of us: data overload shows data mining as one step in the overall KDD process:

1. Identify and develop an understanding of the application domain.

2. Select the data set to be studied.

3. Select complimentary data sets. Integrate the data sets.

4. Code the data. Clean the data of duplicates and errors. Transform the data.

5. Develop models and build hypotheses.

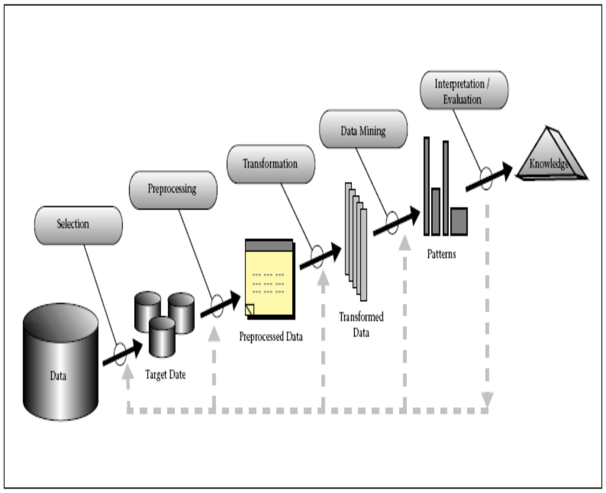
6. Select appropriate data mining algorithms.

7. Interpret results. View results using appropriate visualization tools.

8. Test results in terms of simple proportions and complex predictions.

9. Manage the discovered knowledge.

Although data mining is only a part of the KDD process, data mining techniques provide the algorithms that fuel the KDD process. The KDD process shown above is a never-ending process. Data mining is the essence of the KDD process. If data mining is being discussed, it is understood that the process of KDD is being used. In this work, we will focus on data mining algorithms.



Adriaans and Zantinge (A&Z) (Adriaans and Zantinge 1996, 5) emphasize that the KDD community reserves the term data mining for the discovery stage of the KDD process. Their definition of KDD is as follows: "... the non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data." Similarly, Berzal *et al.* define data mining as "a generic term which covers research results, techniques and tools used to extract useful information from large databases." Also, A&Z point out that KDD draws on techniques from the fields of expert systems, machine learning, statistics, visualization, and database technology.

Comaford addresses some misconceptions about data mining (Comaford 1997). In Comaford's view, data mining is not the same thing as data warehousing or data analysis. Data mining is a dynamic process that enables a more intelligent use of a data warehouse than data analysis. Data mining builds models that can be used to make predictions without additional SQL queries. Data mining techniques apply to both small and very large data sets. Instead of considering just the size of the data set, one must include appropriate width, depth, and volume as three important requirements. Effective data mining requires many attributes for the database records (width), a large number of records that are instances of the database entities (depth) and many entities determined by the database design (volume). Data mining is most appropriate for customer-oriented applications instead of for general business applications. Data mining does not necessarily require artificial intelligence (AI). If a data mining algorithm uses AI, it should be invisible to the user. That is, Comaford does not see data mining as a general business tool except for customer-oriented applications. For commercial data mining applications, this assessment of data mining may be true. This assessment underscores the need for data mining applications for technical data.

A&Z take a different viewpoint than Comaford in regard to width, depth, and volume. According to Comaford, join operations eliminate the need for a volume definition by collapsing a database's attributes of interest into a set of related records. A&Z, on the other hand, consider data mining as an exploration of a multidimensional space of data. Consider a database with one entity and with a million records. If the database has one attribute, it has only one dimension. Suppose this dimension is scaled from 0 to 100 with a resolution of one part per hundred. For one million records there are on average 10,000 records per unit of space or per unit length in the one-dimensional case. For two attributes and two dimensions, there are on average 100 records per unit area. For three attributes, there is on average only one record per unit volume. To put this number in perspective, consider that the vacuum of space contains about one to two atoms per cubic inch (Elert 1987). Thus, the data mining space of a three attribute database with one million records is an extremely low density space. Furthermore, if the database has ten attributes, then the density of records is 10-14 records per unit hypervolume. The point of this analogy is that hyperspace becomes relatively empty as the number of attributes increase above three even for very large databases. The density of records in hyperspace is thus a consideration in choosing a data mining technique.

REFERENCES

1. Adriaans and Zantinge 1996, 5,

2. Elert 1987,

3. Comaford 1999

4. Agrawal, R., Shim, K. 1996. Developing tightly-coupled applications on IBM DB2/CS relational database system: methodology and experience. In Proceedings of the Second International Conference on Knowledge Discovery in Databases and Data Mining Held in Portland, Oregon August 1996, 287-290.

5. Smith, K. A and Gupta, J. N. D. 2000. Neural Networks in Business: Techniques and Applications for the Operations Researcher. Computers & Operations Research 27: 1023-1044.