DATA MINING

Elena Marchiori
Free University Amsterdam, Faculty of Science, Department of Mathematics and Computer Science, Amsterdam, The Netherlands

Keywords: Data mining, knowledge discovery in databases, association rules, data generalization, machine learning, information management.

Contents

Summary

In the recent past, many organizations have produced a large amount of machine-readable data in the form of files and databases. The explosive growth in data has generated an urgent need for techniques and tools for extracting useful hidden information from the data, in terms, e.g., of patterns or rules. Information from the data is used to analyze the user behaviour, to improve the services provided, and to increase the business opportunities. Data mining refers to the discovery process of hidden information from the data. This article presents a short overview of data mining goals, techniques, and applications.

1. Introduction

In the 1980s, all major organizations built infra-structural databases, containing gigabytes of data about their clients, competitors and products. This explosion of data in the modern society has produced a new source of hidden knowledge, consisting of regularities and patterns in the data.

In international organizations, this kind of knowledge provides an essential method for understanding the behaviour of the users in order to improve the services provided by

the organization.

This justifies the sudden expansion and popularity of a new field called Knowledge Discovery in Databases (KDD). KDD is a multi-disciplinary field of research which exploits and integrates techniques from different areas, like machine learning, statistics, database technology, expert systems, and data visualization. These techniques are integrated for the non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data in databases.
The main stages of the KDD process can be summarized as follows:

- *data cleaning.* Noisy, erroneous, missing, or irrelevant data are handled.
- *data integration.* Multiple heterogeneous data sources may be integrated into one.
- *data selection.* Data relevant to the analysis task **are** retrieved **from** the database.
- *data coding.* Data are transformed into forms appropriate for extracting useful knowledge.
- *data mining.* Advanced techniques are applied in order to extract useful knowledge.
- *reporting.* Visualization and knowledge representation techniques are used to present the mined knowledge to the user.

A methodology for knowledge discovery in databases consists of the sequential application of the stages above enumerated; at every stage it is possible to jump back one or more stages. For instance, polluted data can be , discovered during the application of the data coding or data mining stage, hence the process has to jump back to the data cleaning stage in order to eliminate the discovered pollution.

Observe that the term Data Mining (DM) refers to the discovery stage of the KDD process (stage 5 in the above description). However, many authors consider Data Mining and KDD as the same process.

The aim of this article is to introduce the reader to the field of data mining by describing its goals, techniques, and applications. For each of these topics a list of relevant issues is presented and each issue is briefly described. The paper is mainly based on material listed in the references. The references can be used by the reader who wishes to deepen his/her knowledge on knowledge discovery in databases.

2. Goals

The goals of data mining depend on the form of the data to be analyzed and on the type of information one is interested in. One can distinguish the following main data mining tasks.

- *Class Description.* Data mining for class description finds a concise characterization of a collection of data and distinguishes it from others. For instance, class description can be used to compare European versus Asian sales of a company, identify the important factors which discriminate the two classes,

and present a summarized overview.

- *Prediction.* The aim of predictive data mining is to find a description of how certain attributes within the data will behave in the future. For example, in business applications, the analysis of buying transactions to predict what consumers will buy under certain discounts, or how much sales volume a store would generate in a given period.
- *Identification.* Data patterns can be used to identify the existence of an event or an activity. For example, in biological applications, the existence of a gene may be identified by certain sequences of nucleotide symbols in the DNA sequence.
- *Association.* The discovery of association relationships or correlations among attributes of the data is a central task in data mining. For instance, data mining techniques may find rules of the form '98% of customers that purchase tires and automobile accessories also have automotive services carried out'. Association analysis is widely used in transaction data analysis for marketing, catalog design, and other business decision making processes.
- *Sequential pattern analysis. A* sequence of actions or events is sought. For example, if a patient underwent cardiac bypass surgery for blocked arteries and an aneurysm and later developed high blood urea within a year of surgery, then the patient is likely to suffer from kidney failure within the next 18 months. Detection of sequential patterns is equivalent to detecting associations among events with certain temporal relationships.
- *Classification.* Data mining can partition the data so that different classes or categories can be identified based on combinations of attributes. For example, customers in a supermarket can be categorized into discount-seeking shoppers, shoppers in a rush, loyal regular shoppers, and infrequent shoppers.
- *Clustering.* A cluster is a collection of objects that are near each other with respect to a given similarity measure. Data mining for clustering identifies clusters embedded in the data. For instance, one may cluster the houses in an area according to their house category, floor area, and geographical locations.
- *Time-series analysis.* A large set of time-series data is analyzed to find certain regularities and interesting characteristics, including search for similar sequences, and mining sequential patterns, periodicities, trends and deviations. For example, a pattern in solar magnetic wind may be used to predict changes in earth atmospheric conditions.

3. Techniques

Data mining techniques are a 'anything goes' affair which uses approaches from multiple disciplines, like statistics, machine learning, information retrieval, and high performance computing. Different methods, like neural networks, evolutionary computation, pattern recognition, spatial data analysis, signal processing, probabilistic graph theory, and inductive logic programming, can be adapted and integrated into hybrid systems for data mining.

A large set of data analysis methods have been developed in statistics over many years of studies. Machine learning has also contributed substantially to classification and induction problems. Neural networks have shown their effectiveness in classification, prediction, and cluster analysis tasks. However, with increasing large amounts of data

stored in databases for data mining, these methods face challenges on efficiency and scalability. Efficient data structures, indexing and data accessing techniques developed in database research contribute to high performance data mining.

This section reviews some of the most important machine-learning and pattern recognition algorithms that are used in data mining: query tools, visualization techniques, case-based learning (k-nearest neighbor), decision trees, association rules, neural networks, genetic algorithms, and inductive logic programming.

3.1 Query Tools

The first step in mining a data set should always be a rough analysis of the data using traditional query tools.

For instance, by applying simple structured query languages, like SQL, one can obtain useful insight on basic aspects and structure of the data.

Observe that SQL is a structured language that assumes the user is aware of the database schema. It allows to view the same information along multiple dimensions, by means of operations of relational algebra that allow a user to select from tables (rows and columns of data) or to join related information from tables based on common fields.

As a consequence, with SQL one can uncover only shallow data, hence to retrieve information that is easily accessible from the data set. Thus SQL does not really belong to the data mining techniques. Nevertheless, most of the interesting information (around 80%) can be retrieved from the database using SQL. More sophisticated techniques are needed for mining the remaining interesting information (around 20%), which consists of hidden knowledge that can be of strategic importance for large organizations.

3.2 Visualization Techniques

Visualization techniques provide another tool that can be used at the beginning of the analysis of a data set in order to get a rough idea of the structure and distribution of the data.

For instance, an elementary technique that can be of great help for a preliminary data analysis is the so-called scatter diagram. In this technique, information on two attributes is displayed in a Cartesian space.

Scatter diagrams can be used to identify interesting sub-sets of the data set which can be further mined using more advanced techniques for extracting useful hidden information. The search for interesting projections of data sets constitutes a whole field of research, known as projection pursuit.

3.3 K-Nearest Neighbor

The representation of the records in the dataset as points in a multi-dimensional space is very useful for the analysis of the data. Using this representation, the concept of

neighborhood can be defined, where records that are close to each other in the space are considered to belong to each other neighborhood. This notion is used in a simple yet powerful learning technique, called k-nearest neighbor, where k denotes the number of neighbors that are used.

The basic idea of the k-nearest neighbor learning algorithm is 'do as your neighbors do'. For instance, in order to predict the behaviour of a certain individual, the best k neighbors of that individual are considered, and the average of the behaviour of the neighbors provides the prediction for the behaviour of that individual.
The k-neighbor technique is an elegant and simple search method. However, it has a number of drawbacks which limit its general applicability.

For instance, the k-neighbor algorithm has a quadratic computational complexity (in the number of records of the data set) which prevents its application to very large datasets.

Another problem is related to the number of attributes of a record. A record consisting of many independent attributes is represented by a point in a high-dimensional search space. In high dimensional spaces, every two points have almost the same distance, thus the k-neighbor technique does not provide any useful information, since all pairs of points are neighbors.

Finally, the k-neighbor technique does not provide a theory to understand the structure of the data. This latter drawback can be overcome by the technique described in the next subsection.

-
-
-

Bibliography

P. Adriaans and D. Zantinge. *Data Mining.* Addison-Wesley, 1996. [Nicely readable introduction to data mining]

R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Basese*, pages 478-499, 1993. [Introduces the Apriori algorithm]

M.S. Chen, J. Han, and P.S. Yu. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):866-883, 1996.

R. Elmasri and S.B. Navathe. *Fundamentals of Database Systems.* Addison-Wesley, 2000.

U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.). *Advances in Knowledge Discovery*

and Data Mining. AAAI/MIT Press, 1996. [Gives a good overview]

J. Han. Data mining. *Encyclopedia of Distibuted Computing,* 1999.

M. Holsheimer and A.P.J.M. Siebes. Data mining: the search for knowledge in databases. Technical report, CWI, CS-R9406 1994.

N. Lavrac and S. Dzeroski. *Inductive Logic Programming: Techniques and Applications.* Ellis Horwood, 1994.

R.S. Michalski, I. Bratko, and M. Kubat (eds.). *Machine Learning and Data Mining: Methods and Applications.* Wiley, England, 1998.

J. R. Quinlan. Learning logical definitions from relations. *Machine Learning,* 5(3):239-266, 1990.

S. Russell and P. Norvig. *Artificial Intelligence: a Modern Approach.* Prentice-Hall International, 1995.[General introduction to Artificial Intelligence from an agent-based perspective]

I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, 1999.

Biographical Sketch

Elena Marchiori received the Laurea degree in Mathematics from the University of Padova, Italy, in 1988 and the PhD degree from the same university, in 1992. She has been a researcher at the University of Leiden, The Netherlands, at CWI Amsterdam, and at the University Ca' Foscari of Venice, Italy. From November 1999 she is a researcher at the Free University of Amsterdam.

Her research interests include optimization, data mining and machine learning.In particular, learning methods based on computational intelligence and their use in computational biology (bio-informatics).