D

DATA MINING

Data mining is the process of automatic discovery of valid, novel, useful, and understandable patterns, associations, changes, anomalies, and statistically significant structures from large amounts of data. It is an interdisciplinary field merging ideas from statistics, machine learning, database systems and datawarehousing, and high-performance computing, as well as from visualization and human-computer interaction. It was engendered by the economic and scientific need to extract useful information from the data that has grown phenomenally in all spheres of human endeavor.

It is crucial that the patterns, rules, and models that are discovered be valid and generalizable not only in the data samples already examined, but also in future data samples. Only then can the rules and models obtained be considered meaningful. The discovered patterns should also be novel, that is, not already known to experts; otherwise, they would yield very little new understanding. Finally, the discoveries should be useful as well as understandable.

Typically data mining has two high-level goals: prediction and description. The former answers the question of *what* and the latter the question of *why*. For prediction, the key criterion is the accuracy of the model in making future predictions; how the prediction decision is arrived at may not be important. For description, the key criterion is the clarity and simplicity of the model describing the data in understandable terms. There is sometimes a dichotomy between these two aspects of data mining in that the most accurate prediction model for a problem may not be easily understandable, and the most easily understandable model may not be highly accurate in its predictions.

Steps in Data Mining

Data mining refers to the overall process of discovering new patterns or building models from a given dataset. There are many steps involved in the mining enterprise. These include data selection, data cleaning and preprocessing, data transformation and reduction, data mining task and algorithm selection, and finally, postprocessing and the interpretation of discovered knowledge. Here are the most important steps:

Understand the application domain: A proper understanding of the application domain is necessary to appreciate the data mining outcomes desired by the user. It is also important to assimilate and take advantage of available prior knowledge to maximize the chance of success.

Collect and create the target dataset: Data mining relies on the availability of suitable data that reflects the underlying diversity, order, and structure of the problem being analyzed. Therefore, it is crucial to collect a dataset that captures all the possible situations relevant to the problem being analyzed.

Clean and transform the target dataset: Raw data contain many errors and inconsistencies, such as noise, outliers, and missing values. An important element of this process is the unduplication of data records to produce a nonredundant dataset. Another important element of this process is the normalization of data records to deal with the kind of pollution caused by the lack of domain consistency.

Select features and reduce dimensions: Even after the data have been cleaned up in terms of eliminating duplicates, inconsistencies, missing values, and so on, there may still be noise that is irrelevant to the problem being analyzed. These noise attributes may confuse subsequent data mining steps, produce irrelevant rules and associations, and increase computational cost. It is therefore wise to perform a dimension-reduction or feature-selection step to separate those attributes that are pertinent from those that are irrelevant.

Apply data mining algorithms: After performing the preprocessing steps, apply appropriate data mining algorithms—association rule discovery, sequence mining, classification tree induction, clustering, and so on—to analyze the data.

Interpret, evaluate, and visualize patterns: After the algorithms have produced their output, it is still necessary to examine the output in order to interpret and evaluate the extracted patterns, rules, and models. It is only by this interpretation and evaluation process that new insights on the problem being analyzed can be derived.

Data Mining Tasks

In verification-driven data analysis the user postulates a hypothesis, and the system tries to validate it. Common verification-driven operations include querying and reporting, multidimensional analysis, and statistical analysis. Data mining, on the other hand, is discovery driven—that is, it automatically extracts new hypotheses from data. The typical data mining tasks include the following:

Association rules: Given a database of transactions, where each transaction consists of a set of items, association discovery finds all the item sets that frequently occur together, and also the rules among them. For example, 90 percent of people who buy cookies also buy milk (60 percent of grocery shoppers buy both).

Sequence mining: The sequence-mining task is to discover sequences of events that commonly occur together. For example, 70 percent of the people who buy Jane Austen's *Pride and Prejudice* also buy *Emma* within a month.

Similarity search: An example is the problem where a person is given a database of objects and a "query" object, and is then required to find those objects in the database that are similar to the query object. Another example is the problem where a person is given a database of objects, and is then required to find all pairs of objects in the databases that are within some distance of each other.

Deviation detection: Given a database of objects, find those objects that are the most different from the other objects in the database—that is, the outliers. These objects may be thrown away as noise, or they may be the "interesting" ones, depending on the specific application scenario.

Classification and regression: This is also called supervised learning. In the case of classification, someone is given a database of objects that are labeled with predefined categories or classes. They are required to develop from these objects a model that separates them into the predefined categories or classes. Then, given a new object, the learned model is applied to assign this new object to one of the classes. In the more general situation of regression, instead of predicting classes, real-valued fields have to be predicted.

Clustering: This is also called unsupervised learning. Here, given a database of objects that are usually without any predefined categories or classes, the individual is required to partition the objects into subsets or groups such that elements of a group share a common set of properties. Moreover, the partition should be such that the similarity between members of the same group is high and the similarity between members of different groups is low.

Challenges in Data Mining

Many existing data mining techniques are usually ad hoc; however, as the field matures, solutions are being proposed for crucial problems like the incorporation of prior knowledge, handling missing data, adding visualization, improving understandability, and other research challenges. These challenges include the following:

Scalability: How does a data mining algorithm perform if the dataset has increased in volume and in dimensions? This may call for some innovations based on efficient and sufficient sampling, or on a trade-off between in-memory and disk-based processing, or on an approach based on high-performance distributed or parallel computing.

New data formats: To date, most data mining research has focused on structured data, because it is the simplest and most amenable to mining. However, support for other data types is crucial. Examples include unstructured or semistructured (hyper)text, temporal, spatial, and multimedia databases. Mining these is fraught with challenges, but it is necessary because multimedia content and digital libraries proliferate at astounding rates.

Handling data streams: In many domains the data changes over time and/or arrives in a constant stream. Extracted knowledge thus needs to be constantly updated.

Database integration: The various steps of the mining process, along with the core data mining methods, need to be integrated with a database system to provide common representation, storage, and retrieval. Moreover, enormous gains are possible when these are combined with parallel database servers.

Privacy and security issues in mining: Privacy-preserving data mining techniques are invaluable in cases where one may not look at the detailed data, but one is allowed to infer high-level information. This also has relevance for the use of mining for national security applications.

Human interaction: While a data mining algorithm and its output may be readily handled by a computer scientist, it is important to realize that the ultimate user is often not the developer. In order for a data mining tool to be directly usable by the ultimate user, issues of automation—especially in the sense of ease of use—must be addressed. Even for computer scientists, the use and incorporation of prior knowledge into a data mining algorithm is often a challenge; they too would appreciate data mining algorithms that can be modularized in a way that facilitates the exploitation of prior knowledge.

Data mining is ultimately motivated by the need to analyze data from a variety of practical applications— from business domains such as finance, marketing, telecommunications, and manufacturing, or from scientific fields such as biology, geology, astronomy, and medicine. Identifying new application domains that can benefit from data mining will lead to the refinement of existing techniques, and also to the development of new methods where current tools are inadequate.

Mohammed J. Zaki

FURTHER READING

Association for Computing Machinery's special interest group on knowledge discovery and data mining. Retrieved August 21, 2003, from http://www.acm.org/sigkdd.

Dunham, M. H. (2002). Data mining: Introductory and advanced topics. Upper Saddle River, NJ: Prentice Hall.

Han, J., & Kamber, M. (2000). *Data mining: Concepts and techniques.* San Francisco: Morgan Kaufman.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge, MA: MIT Press.

Kantardzic, M. (2002). *Data mining: Concepts, models, methods, and algorithms.* Somerset, NJ: Wiley-IEEE Press.

Witten, I. H., & Frank, E. (1999). *Data mining: Practical machine learning tools and techniques with Java implementations.* San Francisco: Morgan Kaufmann.

DATA VISUALIZATION

Data visualization is a new discipline that uses computers to make pictures that elucidate a concept, phenomenon, relationship, or trend hidden in a large quantity of data. By using interactive three-dimensional (3D) graphics, data visualization goes beyond making static illustrations or graphs and emphasizes interactive exploration.

The pervasive use of computers in all fields of science, engineering, medicine, and commerce has resulted in an explosive growth of data, presenting people with unprecedented challenges in understanding data. Data visualization transforms raw data into pictures that exploit the superior visual processing capability of the human brain to detect patterns and draw inferences, revealing insights hidden in the data. For example, data visualization allows us to capture trends, structures, and anomalies in the behavior of a physical process being modeled or in vast amounts of Internet data. Furthermore, it provides us with a visual and remote means to communicate our findings to others.

Since publication of a report on visualization in scientific computing by the U.S. National Science Foundation in 1987, both government and industry have invested tremendous research and development in data-visualization technology, resulting in advances in visualization and interactive techniques that have helped lead to many scientific discoveries, better engineering designs, and more timely and accurate medical diagnoses.

Visualization Process

A typical data-visualization process involves multiple steps, including data generation, filtering, mapping, rendering, and viewing. The data-generation step can be a numerical simulation, a laboratory experiment, a collection of sensors, an image scanner, or a recording of Web-based business transactions. Filtering removes noise, extracts and enhances features, or rescales data. Mapping derives appropriate representations of data for the rendering step. The representations can be composed of geometric primitives such as point, lines, polygons, and surfaces, supplemented with properties such as colors, transparency, textures. Whereas the visualization of a computerized tomography (CT) scan of a fractured bone should result in an image of a bone, plenty of room for creativity exists when making a visual depiction of the trend of a stock market or the chemical reaction in a furnace.

Rendering generates two-dimensional or three-dimensional images based on the mapping results and other rendering parameters, such as the lighting model, viewing position, and so forth. Finally, the resulting images are displayed for viewing. Both photorealistic and nonphotorealistic rendering techniques exist for different purposes of visual communication. Nonphotorealistic rendering, which mimics how artists use brushes, strokes, texture, color, layout, and so forth, is usually used to increase the clarity of the spatial relationship between objects, improve the perception of an object's shape and size, or give a particular type of media presentation.

Note that the filtering and mapping steps are largely application dependent and often require domain knowledge to perform. For example, the filtering and mapping steps for the visualization of website structure or browsing patterns would be quite different from those of brain tumors or bone fractures.

A data-visualization process is inherently iterative. That is, after visualization is made, the user should be able to go back to any previous steps, including the data-generation step, which consists of a numerical or physical model, to make changes such that more information can be obtained from the revised visualization. The changes may be made in a systematic way or by trial and error. The goal is to improve the model and understanding of the corresponding problem via this visual feedback process.