Data Mining and Knowledge Discovery

Sally I. McClean
University of Ulster

GLOSSARY

Association rules link the values of a group of attributes, or variables, with the value of a particular attribute of interest which is not included in the group.

Data mining process takes place in four main stages: Data Pre-processing, Exploratory Data Analysis, Data Selection, and Knowledge Discovery.

Data mining tools are software products; a growing number of such products are becoming commercially available. They may use just one approach (single paradigm), or they may employ a variety of different methods (multi-paradigm).

Deviation detection is carried out in order to discover Interestingness in the data. Deviations may be detected either for categorical or numerical data.

Interestingness is central to Data Mining where we are looking for new knowledge which is nontrivial. It allows the separation of novel and useful patterns from the mass of dull and trivial ones.

Knowledge discovery in databases (KDD) is the main objective in Data Mining. The two terms are often used synonymously, although some authors define Knowledge Discovery as being carried out at a higher level than Data Mining.

DATA MINING is the process by which computer programs are used to repeatedly search huge amounts of data, usually stored in a Database, looking for useful new patterns. The main developments that have led to the emergence of Data Mining have been in the increased volume of data now being collected and stored electronically, and an accompanying maturing of Database Technology. Such developments have meant that traditional Statistical Methods and Machine Learning Technologies have had to be extended to incorporate increased demands for fast and scaleable algorithms.

In recent years, Database Technology has developed increasingly more efficient methods for data processing and

data access. Simultaneously there has been a convergence between Machine Learning Methods and Database Technology to create value-added databases with an increased capability for intelligence. There has also been a convergence between Statistics and Database Technology.

I. DATA MINING AND KNOWLEDGE DISCOVERY

A. Background

The main developments that have led to the emergence of Data Mining as a promising new area for the discovery of knowledge have been in the increased amount of data now available, with an accompanying maturing of Database Technology. In recent years Database Technology has developed efficient methods for data processing and data access such as parallel and distributed computing, improved middleware tools, and Open Database Connectivity (ODBC) to facilitate access to multi-databases.

Various Data Mining products have now been developed and a growing number of such products are becoming commercially available. Increasingly, Data Mining systems are coming onto the market. Such systems ideally should provide an integrated environment for carrying out the whole Data Mining process thereby facilitating end-user Mining, carried out automatically, with an interactive user interface.

B. The Disciplines

Data Mining brings together the three disciplines of Machine Learning, Statistics, and Database Technology. In the Machine Learning field, many complex problems are now being tackled by the development of intelligent systems. These systems may combine Neural Networks, Genetic Algorithms, Fuzzy Logic systems, Case-Based Reasoning, and Expert Systems. Statistical Techniques have become well established as the basis for the study of Uncertainty. Statistics embraces a vast array of methods used to gather, process, and interpret quantitative data. Statistical Techniques may be employed to identify the key features of the data in order to explain phenomena, and to identify subsets of the data that are interesting by virtue of being significantly different from the rest. Statistics can also assist with prediction, by building a model from which some attribute values can be reliably predicted from others in the presence of uncertainty. Probability Theory is concerned with measuring the likelihood of events under uncertainty, and underpins much of Statistics. It may also be applied in new areas such as Bayesian Belief Networks, Evidence Theory, Fuzzy Logic systems and Rough Sets.

Database manipulation and access techniques are essential to efficient Data Mining; these include Data Visualization and Slice and Dice facilities. It is often the case that it is necessary to carry out a very large number of data manipulations of various types. This involves the use of a structured query language (SQL) to perform basic operations such as selecting, updating, deleting, and inserting data items. Data selection frequently involves complex conditions containing Boolean operators and statistical functions, which thus require to be supported by SQL. Also the ability to join two or more databases is a powerful feature that can provide opportunities for Knowledge Discovery.

C. Data Mining Objectives and Outcomes

Data Mining is concerned with the search for new knowledge in data. Such knowledge is usually obtained in the form of rules which were previously unknown to the user and may well prove useful in the future. These rules might take the form of specific rules induced by means of a rule induction algorithm or may be more general statistical rules such as those found in predictive modeling. The derivation of such rules is specified in terms of Data Mining tasks where typical tasks might involve classifying or clustering the data.

A highly desirable feature of Data Mining is that there be some high-level user interface that allows the end-user to specify problems and obtain results in as friendly as matter as possible. Although it is possible, and in fact common, for Data Mining to be carried out by an expert and the results then explained to the user, it is also highly desirable that the user be empowered to carry out his own Data Mining and draw his own conclusions from the new knowledge. An appropriate user interface is therefore of great importance.

Another secondary objective is the use of efficient data access and data processing methods. Since Data Mining is increasingly being applied to large and complex databases, we are rapidly approaching the situation where efficient methods become a *sine qua non*. Such methods include Distributed and Parallel Processing, the employment of Data Warehousing and accompanying technologies, and the use of Open Database Connectivity (ODBC) to facilitate access to multi-databases.

D. The Data Mining Process

The Data Mining process may be regarded as taking place in four main stages (Fig. 1):

- Data Pre-processing
- Exploratory Data analysis
- Data Selection
- Knowledge Discovery

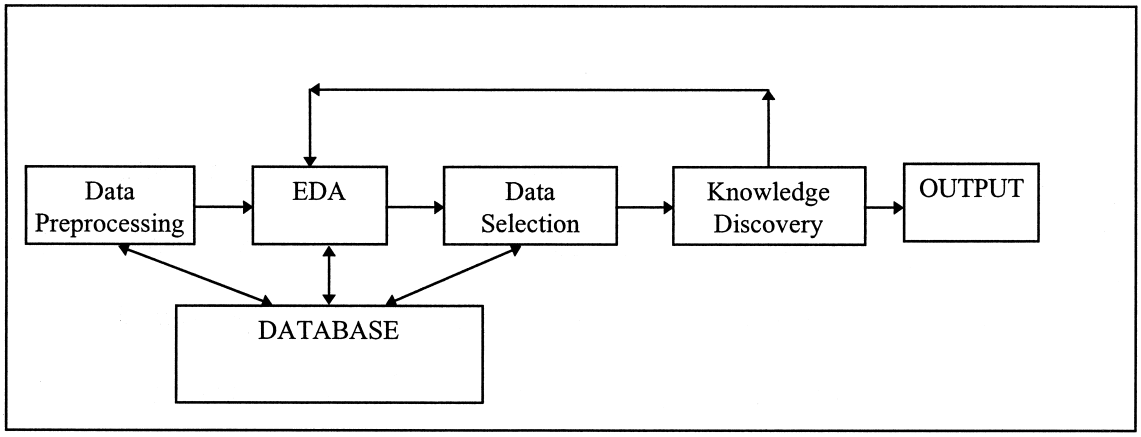

FIGURE 1 The Data Mining Process.

Data Pre-processing is concerned with data cleansing and reformatting, so that the data are now held in a form that is appropriate to the Mining algorithms and facilitates the use of efficient access methods. Reformatting typically involves employing missing value handling and presenting the data in multidimensional views suitable for the multidimensional servers used in Data Warehousing.

In Exploratory Data Analysis (EDA), the miner has a preliminary look at the data to determine which attributes and which technologies should be utilized. Typically, Summarization and Visualization Methods are used at this stage.

For Data Selection, we may choose to focus on certain attributes or groups of attributes since using all attributes at once is likely to be too complex and time consuming. Alternatively, for large amounts of data, we may choose to sample certain tuples, usually chosen at random. We may then carry out Knowledge Discovery using the sample, rather than the complete data, thus speeding up the process enormously. Variable reduction techniques or new variable definition are alternative methods for circumventing the problems caused by such large data sets.

Knowledge Discovery is the main objective in Data Mining and many different technologies have been employed in this context. In the Data Mining Process we frequently need to iterate round the EDA, Data Selection, Knowledge Discovery part of the process, as once we discover some new knowledge, we often then want to go back to the data and look for new or more detailed patterns.

Once new knowledge has been mined from the database, it is then reported to the user either in verbal, tabular or graphical format. Indeed the output from the Mining process might be an Expert System. Whatever form the output takes, it is frequently the case that such information is really the specification for a new system that will use the knowledge gained to best advantage for the user and domain in question. New knowledge may feed into the business process which in turn feeds back into the Data Mining process.

E. Data Mining Tasks

1. Rule Induction

Rule induction uses a number of specific beliefs in the form of database tuples as evidence to support a general belief that is consistent with these specific beliefs. A collection of tuples in the database may form a relation that is defined by the values of particular attributes, and relations in the database form the basis of rules. Evidence from within the database in support of a rule is thus used to induce a rule which may be generally applied.

Rules tend to be based on sets of attribute values, partitioned into an antecedent and a consequent. A typical "if then" rule, of the form "if antecedent = true, then consequent = true," is given by "if a male employee is aged over 50 and is in a management position, then he will hold an additional pension plan." Support for such a rule is based on the proportion of tuples in the database that have the specified attribute values in both the antecedent and the consequent. The degree of confidence in a rule is the proportion of those tuples that have the specified attribute values in the antecedent, which also have the specified attribute values in the consequent.

Rule induction must then be combined with rule selection in terms of interestingness if it is to be of real value in Data Mining. Rule-finding and evaluation typically require only standard database functionality, and they may be carried out using embedded SQL. Often, if a database is very large, it is possible to induce a very large number of rules. Some may merely correspond to well-known domain knowledge, whilst others may simply be of little interest to the user. Data Mining tools must therefore support the selection of **interesting** rules.

2. Classification

A commonly occurring task in Data Mining is that of classifying cases from a dataset into one of a number of well-defined categories. The categories are defined by sets of attribute values, and cases are allocated to categories according to the attribute values that they possess. The selected combinations of attribute values that define the classes represent **features** within the particular context of the classification problem. In the simplest cases, classification could be on a single binary-valued attribute, and the dataset is partitioned into two groups, namely, those cases with a particular property, and those without it. In general it may only be possible to say which class the case is "closest to," or to say how likely it is that the case is in a particular category.

Classification is often carried out by **supervised Machine Learning**, in which a number of training examples (tuples whose classification is known) are presented to the system. The system "learns" from these how to classify other cases in the database which are not in the training set. Such classification may be probabilistic in the sense that it is possible to provide the probability that a case is any one of the predefined categories. **Neural Networks** are one of the main Machine Learning technologies used to carry out classification. A probabilistic approach to classification may be adopted by the use of **discriminant functions**.

3. Clustering

In the previous section, the classification problem was considered to be essentially that of learning how to make decisions about assigning cases to known classes. There are, however, different forms of classification problem, which may be tackled by **unsupervised learning**, or clustering. Unsupervised classification is appropriate when the definitions of the classes, and perhaps even the number of classes, are not known in advance, e.g., market segmentation of customers into similar groups who can then be targeted separately.

One approach to the task of defining the classes is to identify clusters of cases. In general terms, **clusters** are groups of cases which are in some way similar to each other according to some measure of **similarity**. Clustering algorithms are usually iterative in nature, with an initial classification being modified progressively in terms of the class definitions. In this way, some class definitions are discarded, whilst new ones are formed, and others are modified, all with the objective of achieving an overall goal of separating the database tuples into a set of cohesive categories. As these categories are not predetermined, it is clear that clustering has much to offer in the process of Data Mining in terms of discovering **concepts**, possibly within a concept hierarchy.

4. Summarization

Summarization aims to present concise measures of the data both to assist in user comprehension of the underlying structures in the data and to provide the necessary inputs to further analysis. Summarization may take the form of the production of graphical representations such as bar charts, histograms, and plots, all of which facilitate a visual overview of the data, from which sufficient insight might be derived to both inspire and focus appropriate Data Mining activity. As well as assisting the analyst to focus on those areas in a large database that are worthy of detailed analysis, such visualization can be used to help with the analysis itself. Visualization can provide a "drill-down" and "drill-up" capability for repeated transition between summary data levels and detailed data exploration.

5. Pattern Recognition

Pattern recognition aims to classify objects of interest into one of a number of categories or classes. The objects of interest are referred to as **patterns**, and may range from printed characters and shapes in images to electronic waveforms and digital signals, in accordance with the data under consideration. Pattern recognition algorithms are designed to provide automatic identification of patterns, without the need for human intervention. Pattern recognition may be **supervised**, or **unsupervised**.

The relationships between the observations that describe a pattern and the classification of the pattern are used to design **decision rules** to assist the recognition process. The observations are often combined to form **features**, with the aim that the features, which are smaller in number than the observations, will be more reliable than the observations in forming the decision rules. Such **feature extraction** processes may be application dependent, or they may be general and mathematically based.

6. Discovery of Interestingness

The idea of interestingness is central to Data Mining where we are looking for new knowledge that is non-trivial. Since, typically, we may be dealing with very large amounts of data, the potential is enormous but so too is the capacity to be swamped with so many patterns and rules that it is impossible to make any sense out of them. It is the concept of interestingness that provides a framework for separating out the novel and useful patterns from the myriad of dull and trivial ones.

Interestingness may be defined as deviations from the norm for either categorical or numerical data. However, the initial thinking in this area was concerned with categorical data where we are essentially comparing the deviation between the proportion of our target group with

a particular property and the proportion of the whole population with the property. Association rules then determine where particular characteristics are related.

An alternative way of computing interestingness for such data comes from statistical considerations, where we say that a pattern is interesting if there is a statistically significant association between variables. In this case the measure of interestingness in the relationship between two variables A and B is computed as:

$$\text{Probability of } (A \text{ and } B)\text{-Probability of } (A)^* \text{ Probability of } (B).$$

Interestingness for continuous attributes is determined in much the same way, by looking at the deviation between summaries.

7. Predictive Modeling

In Predictive Modeling, we are concerning with using some attributes or patterns in the database to predict other attributes or extract rules. Often our concern is with trying to predict behavior at a future time point. Thus, for business applications, for example, we may seek to predict future sales from past experience.

Predictive Modeling is carried out using a variety of technologies, principally Neural Networks, Case-Based Reasoning, Rule Induction, and Statistical Modeling, usually via Regression Analysis. The two main types of predictive modeling are **transparent** (explanatory) and **opaque** (black box). A transparent model can give information to the user about why a particular prediction is being made, while an opaque model cannot explain itself in terms of the relevant attributes. Thus, for example, if we are making predictions using Case-Based Reasoning, we can explain a particular prediction in terms of similar behavior commonly occurring in the past. Similarly, if we are using a statistical model to predict, the forecast is obtained as a combination of known values which have been previously found to be highly relevant to the attribute being predicted. A Neural Network, on the other hand, often produces an opaque prediction which gives an answer to the user but no explanation as to why this value should be an accurate forecast. However, a Neural Network can give extremely accurate predictions and, where it may lack in explanatory power, it more than makes up for this deficit in terms of predictive power.

8. Visualization

Visualization Methods aim to present large and complex data sets using pictorial and other graphical representations. State-of-the-art Visualization techniques can thus assist in achieving Data Mining objectives by simplifying the presentation of information. Such approaches are often concerned with summarizing data in such a way as to facilitate comprehension and interpretation. It is important to have the facility to handle the commonly occurring situation in which it is the case that too much information is available for presentation for any sense to be made of it—the "haystack" view. The information extracted from Visualization may be an end in itself or, as is often the case, may be a precursor to using some of the other technologies commonly forming part of the Data Mining process.

Visual Data Mining allows users to interactively explore data using graphs, charts, or a variety of other interfaces. Proximity charts are now often used for browsing and selecting material; in such a chart, similar topics or related items are displayed as objects close together, so that a user can traverse a topic landscape when browsing or searching for information. These interfaces use colors, filters, and animation, and they allow a user to view data at different levels of detail. The data representations, the levels of detail and the magnification, are controlled by using mouse-clicks and slider-bars.

Recent developments involve the use of "virtual reality," where, for example, statistical objects or cases within a database may be represented by graphical objects on the screen. These objects may be designed to represent people, or products in a store, etc., and by clicking on them the user can find further information relating to that object.

9. Dependency Detection

The idea of dependency is closely related to interestingness and a relationship between two attributes may be thought to be interesting if they can be regarded as dependent, in some sense. Such patterns may take the form of statistical dependency or may manifest themselves as **functional dependency** in the database. With functional dependency, all values of one variable may be determined from another variable. However, statistical dependency is all we can expect from data which is essentially random in nature.

Another type of dependency is that which results from some sort of causal mechanism. Such causality is often represented in Data Mining by using Bayesian Belief Networks which discover and describe. Such causal models allow us to predict consequences, even when circumstances change. If a rule just describes an association, then we cannot be sure how robust or generalizable it will be in the face of changing circumstances.

10. Uncertainty Handling

Since real-world data are often subject to uncertainty of various kinds, we need ways of handling this uncertainty. The most well-known and commonly used way of

handling uncertainty is to use classical, or Bayesian, probability. This allows us to establish the probabilities, or support, for different rules and to rank them accordingly. One well-known example of the use of Bayesian probability is provided by the Bayesian Classifier which uses Bayes' Theorem as the basis of a classification method. The various approaches to handling uncertainty have different strengths and weaknesses that may make them particularly appropriate for particular Mining tasks and particular data sets.

11. Sequence Processing

Sequences of data, which measure values of the same attribute at a sequence of different points, occur commonly. The best-known form of such data arises when we collect information on an attribute at a sequence of time points, e.g., daily, quarterly, annually. However, we may instead have data that are collected at a sequence of different points in space, or at different depths or heights. Statistical data that are collected at a sequence of different points in time are known as time series.

In general, we are concerned with finding ways of describing the important features of a time series, thus allowing Predictive Modeling to be carried out over future time periods. There has also been a substantial amount of work done on describing the relationship between one time series and another with a view to determining if two time series co-vary or if one has a causal effect on the other. Such patterns are common in economic time series, where such variables are referred to as leading indicators. The determination of such leading indicators can provide new knowledge and, as such, is a fertile area for Data Mining.

The methods used for Predictive Modeling for the purpose of sequence processing are similar to those used for any other kind of Predictive Modeling, typically Rule Induction and Statistical Regression. However, there may be particular features of sequences, such as seasonality, which must be incorporated into the model if prediction is to be accurate.

F. Data Mining Approaches

As has already been stated, Data Mining is a multidisciplinary subject with major input from the disciplines of Machine Learning, Database Technology and Statistics but also involving substantial contributions from many other areas, including Information Theory, Pattern Recognition, and Signal Processing. This has led to many different approaches and a myriad of terminology where different communities have developed substantially different terms for essentially the same concepts. Nonetheless, there is much to gain from such an interdisciplinary approach and the synergy that is emerging from recent developments in the subject is one of its major strengths.

G. Advantages of Data Mining

Wherever techniques based on data acquisition, processing, analysis and reporting are of use, there is potential for Data Mining. The collection of consumer data is becoming increasingly automatic at the point of transaction. Automatically collected retail data provide an ideal arena for Data Mining. Highly refined customer profiling becomes possible as an integral part of the retail system, eschewing the need for costly human intervention or supervision. This approach holds the potential for discovering interesting or unusual patterns and trends in consumer behavior, with obvious implications for marketing strategies such as product placement, customized advertising, and rewarding customer loyalty. The banking and insurance industries have also well-developed and specialized data analysis techniques for customer profiling for the purpose of assessing credit worthiness and other risks associated with loans and investments. These include using Data Mining methods to adopt an integrated approach to mining criminal and financial data for fraud detection.

Science, technology, and medicine are all fields that offer exciting possibilities for Data Mining. Increasingly it is the vast arrays of automatically recorded experimental data that provide the material from which may be formed new scientific knowledge and theory. Data Mining can facilitate otherwise impossible Knowledge Discovery, where the amount of data required to be assimilated for the observation of a single significant anomaly would be overwhelming for manual analysis.

Both modern medical diagnosis and industrial process control are based on data provided by automated monitoring systems. In each case, there are potential benefits for efficiency, costs, quality, and consistency. In the Health Care environment, these may lead to enhanced patient care, while application to industrial processes and project management can provide a vital competitive advantage.

Overall, however, the major application area for Data Mining is still Business. For example a recent survey of Data Mining software tools (Fig. 2) showed that over three-quarters (80%) are used in business applications, primarily in areas such as finance, insurance, marketing and market segmentation. Around half of the vendor tools surveyed were suited to Data Mining in medicine and industrial applications, whilst a significant number are most useful in scientific and engineering fields.

II. THE TECHNOLOGIES

A. Machine Learning Technologies

1. Inferencing Rules

Machine Learning, in which the development of Inferencing Rules plays a major part, can be readily applied

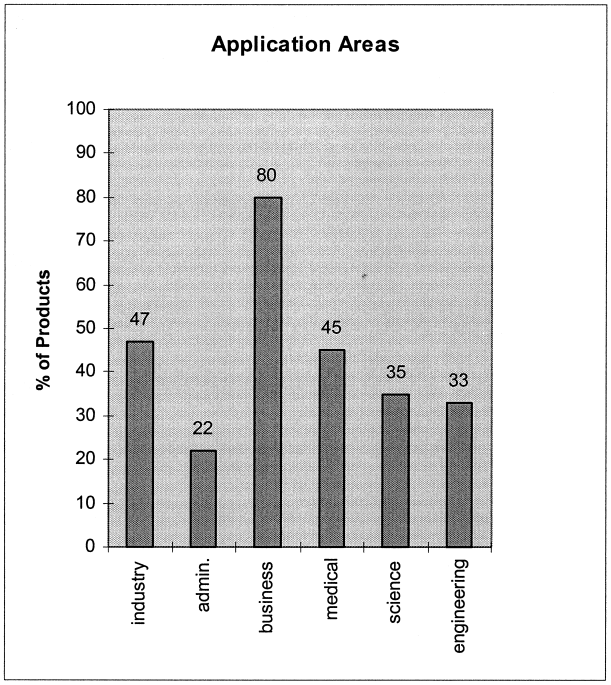

FIGURE 2 Application areas for Data Mining Tools.

to Knowledge Discovery in Databases. Records (tuples) in a database may be regarded as training instances that attribute-value learning systems may then use to discover patterns in a file of database records. Efficient techniques exist which can handle very large training sets, and include ways of dealing with incomplete and noisy data.

Logical reasoning may be automated by the use of a logic programming system, which contains a language for representing knowledge, and an **inference engine** for automated reasoning. The induction of logical definitions of relations has been named **Inductive Logic Programming (ILP)**, and may also be used to compress existing relations into their logical definitions. Inductive learning systems that use ILP construct logical definitions of target relations from examples and background knowledge. These are typically in the form of if–then rules, which are then transformed into clauses within the logic programming system. ILP systems have applications in a wide variety of domains, where Knowledge Discovery is achieved via the learning of relationships. Inference rules may be implemented as **demons**, which are processes running within a program while the program continues with its primary task.

2. Decision Trees

A Decision Tree provides a way of expressing knowledge used for classification. A Decision Tree is constructed by using a **training set** of cases that are described in terms of a collection of attributes. A sequence of tests is carried out on the attributes in order to partition the training set into ever-smaller subsets. The process terminates when each subset contains only cases belonging to a single class. Nodes in the Decision Tree correspond to these tests, and the leaves of the tree represent each subset. New cases (which are not in the training set) may be classified by tracing through the Decision Tree starting at its root and ending up at one of the leaves.

The choice of test at each stage in "growing" the tree is crucial to the tree's predictive capability. It is usual to use a selection criterion based on the gain in classification information and the information yielded by the test. In practice, when "growing" a Decision Tree, a small working set of cases is used initially to construct a tree. This tree is then used to classify the remaining cases in the training set: if all are correctly classified, then the tree is satisfactory. If there are misclassified cases, these are added to the working set, and a new tree constructed using this augmented working set. This process is used iteratively until a satisfactory Decision Tree is obtained. Overfitting of the data and an associated loss of predictive capability may be remedied by **pruning** the tree, a process that involves replacing sub-trees by leaves.

3. Neural Networks

Neural Networks are designed for pattern recognition, and they thus provide a useful class of Data Mining tools. They are primarily used for classification tasks. The Neural Network is first trained before it is used to attempt to identify classes in the data. Hence from the initial dataset a proportion of the data are partitioned into a **training set** which is kept separate from the remainder of the data. A further proportion of the dataset may also be separated off into a **validation set** that is used to test performance during training along with criteria for determining when the training should terminate, thus preventing **overtraining**.

A Neural Network is perhaps the simplest form of parallel computer, consisting of a (usually large) set of simple processing elements called **neurons**. The neurons are connected to one another in a chosen configuration to form a network. The types of connectivities or network architectures available can vary widely, depending on the application for which the Neural Network is to be used. The most straightforward arrangements consist of neurons set out in layers as in the **feedforward network**. Activity feeds from one layer of neurons to the next, starting at an initial input layer.

The **Universal Approximation Theorem** states that a single layer net, with a suitably large number of hidden nodes, can well approximate any suitably smooth function. Hence for a given input, the network output may be compared with the required output. The total mean square error function is then used to measure how close the actual

output is to the required output; this error is reduced by a technique called **back-error propagation**. This approach is a **supervised** method in which the network learns the connection weights as it is taught more examples. A different approach is that of **unsupervised learning**, where the network attempts to learn by finding statistical features of the input training data.

4. Case-Based Reasoning

Case-Based Reasoning (CBR) is used to solve problems by finding similar, past cases and adapting their solutions. By not requiring specialists to encapsulate their expertise in logical rules, CBR is well suited to domain experts who attempt to solve problems by recalling approaches which have been taken to similar situations in the past. This is most appropriate in domains which are not well understood, or where any rules that may have been devised have frequent exceptions. CBR thus offers a useful approach to building applications that support decisions based on past experience.

The quality of performance achieved by a case-based reasoner depends on a number of issues, including its experiences, and its capabilities to adapt, evaluate, and repair situations. First, partially matched cases must be retrieved to facilitate reasoning. The retrieval process consists of two steps: recalling previous cases, and selecting a best subset of them. The problem of retrieving applicable cases is referred to as the **indexing problem**. This comprises the matching or similarity-assessment problem, of recognizing that two cases are similar.

5. Genetic Algorithms

Genetic Algorithms (GA's) are loosely based on the biological principles of genetic variation and natural selection. They mimic the basic ideas of the evolution of life forms as they adapt to their local environments over many generations. Genetic Algorithms are a type of **evolutionary algorithm**, of which other types include Evolutionary Programming and Evolutionary Strategies.

After a new generation is produced, it may be combined with the population that spawned it to yield the new current population. The size of the new population may be curtailed by selection from this combination, or alternatively, the new generation may form the new population. The genetic operators used in the process of generating offspring may be examined by considering the contents of the population as a **gene pool**. Typically an individual may then be thought of in terms of **a binary string** of fixed length, often referred to as a **chromosome**. The genetic operators that define the offspring production process are usually a combination of **crossover** and **mutation** operators. Essentially these operators involve swapping part of the binary string of one parent with the corresponding part for the other parent, with variations depending on the particular part swapped and the position and order in which it is inserted into the remaining binary string of the other parent. Within each child, mutation then takes place.

6. Dynamic Time-Warping

Much of the data from which knowledge is discovered are of a temporal nature. Detecting patterns in sequences of time-dependent data, or time series, is an important aspect of Data Mining, and has applications in areas as diverse as financial analysis and astronomy. Dynamic Time-Warping (DTW) is a technique established in the recognition of patterns in speech, but may be more widely applied to other types of data.

DTW is based on a dynamic programming approach to aligning a selected template with a data time series so that a chosen measure of the distance between them, or error, is minimized. The measure of how well a template matches the time series may be obtained from the table of cumulative distances. A **warping path** is computed through the grid from one boundary point to another, tracing back in time through adjacent points with minimal cumulative distance. This warping path defines how well the template matches the time series, and a measure of the fit can be obtained from the cumulative distances along the path.

B. Statistical and other Uncertainty-Based Methods

1. Statistical Techniques

Statistics is a collection of methods of enquiry used to gather, process, or interpret quantitative data. The two main functions of Statistics are to describe and summarize data and to make inferences about a larger population of which the data are representative. These two areas are referred to as Descriptive and Inferential Statistics, respectively; both areas have an important part to play in Data Mining. Descriptive Statistics provides a toolkit of methods for data summarization while Inferential Statistics is more concerned with data analysis.

Much of Statistics is concerned with statistical analysis that is mainly founded on statistical inference or hypothesis testing. This involves having a Null Hypothesis (H_o): which is a statement of null effect, and an Alternative Hypothesis (H_1): which is a statement of effect. A test of significance allows us to decide which of the two hypotheses (H_o or H_1) we should accept. We say that a result is significant at the 5% level if the probability that the discrepancy between the actual data and what is expected assuming the null hypothesis is true has probability less that 0.05 of

occurring. The significance level therefore tells us where to **threshold** in order to decide if there is an effect or not.

Predictive Modeling is another Data Mining task that is addressed by Statistical methods. The most common type of predictive model used in Statistics is **linear regression**, where we describe one variable as a linear combination of other known variables. A number of other tasks that involve analysis of several variables for various purposes are categorized by statisticians under the umbrella term multivariate analysis.

Sequences of data, which measure values of the same attribute under a sequence of different circumstances, also occur commonly. The best-known form of such data arises when we collect information on an attribute at a sequence of time points, e.g., daily, quarterly, annually. For such time-series data the trend is modeled by fitting a regression line while fluctuations are described by mathematical functions. Irregular variations are difficult to model but worth trying to identify, as they may turn out to be of most interest.

Signal processing is used when there is a continuous measurement of some sort—the signal—usually distorted by noise. The more noise there is, the harder it is to extract the signal. However, by using methods such as filtering which remove all distortions, we may manage to recover much of the original data. Such filtering is often carried out by using **Fourier transforms** to modify the data accordingly. In practice, we may use **Fast Fourier Transforms** to achieve high-performance signal processing. An alternative method is provided by **Wavelets**.

All of Statistics is underpinned by classical or Bayesian probability. Bayesian Methods often form the basis of techniques for the automatic discovery of *classes* in data, known as **clustering** or **unsupervised learning**. In such situations Bayesian Methods may be used in computational techniques to determine the optimal set of classes from a given collection of unclassified instances. The aim is to find the most likely set of classes given the data and a set of prior expectations. A balance must be struck between data fitting and the potential for class membership prediction.

2. Bayesian Belief Networks

Bayesian Belief Networks are graphical models that communicate causal information and provide a framework for describing and evaluating probabilities when we have a network of interrelated variables. We can then use the graphical models to evaluate information about external interventions and hence predict the effect of such interventions. By exploiting the dependencies and interdependencies in the graph we can develop efficient algorithms that calculate probabilities of variables in graphs which are often very large and complex. Such a facility makes this technique suitable for Data Mining, where we are often trying to sift through large amounts of data looking for previously undiscovered relationships.

A key feature of Bayesian Belief Networks is that they discover and describe causality rather than merely identifying associations as is the case in standard Statistics and Database Technology. Such causal relationships are represented by means of **DAGs (Directed Acyclic Graphs)** that are also used to describe conditional independence assumptions. Such conditional independence occurs when two variables are independent, conditional on another variable.

3. Evidence Theory

Evidence Theory, of which Dempster–Shafer theory is a major constituent, is a generalization of traditional probability which allows us to better quantify uncertainty. The framework provides a means of representing data in the form of a **mass function** that quantifies our degree of belief in various propositions. One of the major advantages of Evidence Theory over conventional probability is that it provides a straightforward way of quantifying ignorance and is therefore a suitable framework for handling missing values.

We may use this Dempster–Shafer definition of mass functions to provide a lower and upper bound for the probability we assign to a particular proposition. These bounds are called the **belief** and **plausibility**, respectively. Such an interval representation of probability is thought to be a more intuitive and flexible way of expressing probability, since we may not be able to assign an exact value to it but instead give lower and upper bounds.

The Dempster–Shafer theory also allows us to transform the data by changing to a higher or lower granularity and reallocating the masses. If a rule can be generalized to a higher level of aggregation then it becomes a more powerful statement of how the domain behaves while, on the other hand, the rule may hold only at a lower level of granularity.

Another important advantage of Evidence Theory is that the Dempster–Shafer law of combination (the orthogonal sum) allows us to combine data from different independent sources. Thus, if we have the same frame of discernment for two mass functions which have been derived independently from different data, we may obtain a unified mass assignment.

4. Fuzzy Logic

Fuzzy logic maintains that all things are a matter of degree and challenges traditional two-valued logic which holds

that a proposition is either true or it is not. Fuzzy Logic is defined via a **membership function** that measures the degree to which a particular element is a member of a set. The membership function can take any value between 0 and 1 inclusive.

In common with a number of other Artificial Intelligence methods, fuzzy methods aim to simulate human decision making in uncertain and imprecise environments. We may thus use Fuzzy Logic to express expert opinions that are best described in such an imprecise manner. Fuzzy systems may therefore be specified using natural language which allows the expert to use vague and imprecise terminology. Fuzzy Logic has also seen a wide application to control theory in the last two decades.

An important use of fuzzy methods for Data Mining is for classification. Associations between inputs and outputs are known in fuzzy systems as **fuzzy associative memories** or **FAM**s. A FAM system encodes a collection of compound rules that associate multiple input statements with multiple output statements We combine such multiple statements using logical operators such as conjunction, disjunction and negation.

5. Rough Sets

Rough Sets were introduced by Pawlak in 1982 as a means of investigating structural relationships in data. The technique, which, unlike classical statistical methods, does not make probability assumptions, can provide new insights into data and is particularly suited to situations where we want to reason from qualitative or imprecise information. Rough Sets allow the development of **similarity measures** that take account of semantic as well as syntactic distance. Rough Set theory allows us to eliminate redundant or irrelevant attributes. The theory of Rough Sets has been successfully applied to knowledge acquisition, process control, medical diagnosis, expert systems and Data Mining. The first step in applying the method is to generalize the attributes using domain knowledge to identify the concept hierarchy. After generalization, the next step is to use reduction to generate a minimal subset of all the generalized attributes, called a **reduct**. A set of general rules may then be generated from the reduct that includes all the important patterns in the data. When more than one reduct is obtained, we may select the best according to some criteria. For example, we may choose the reduct that contains the smallest number of attributes.

6. Information Theory

The most important concept in Information Theory is **Shannon's Entropy**, which measures the amount of information held in data. Entropy quantifies to what extent the data are spread out over its possible values. Thus high entropy means that the data are spread out as much as possible while low entropy means that the data are nearly all concentrated on one value. If the entropy is low, therefore, we have high information content and are most likely to come up with a strong rule.

Information Theory has also been used as a measure of interestingness which allows us to take into account how often a rule occurs and how successful it is. This is carried out by using the **J-measure**, which measures the amount of information in a rule using Shannon Entropy and multiplies this by the probability of the rule coming into play. We may therefore rank the rules and only present the most interesting to the user.

C. Database Methods

1. Association Rules

An Association Rule associates the values of a given set of attributes with the value of another attribute from outside that set. In addition, the rule may contain information about the frequency with which the attribute values are associated with each other. For example, such a rule might say that "75% of men, between 50 and 55 years old, in management positions, take out additional pension plans."

Along with the Association Rule we have a **confidence threshold** and a **support threshold**. Confidence measures the ratio of the number of entities in the database with the designated values of the attributes in both A and B to the number with the designated values of the attributes in A. The support for the Association Rule is simply the proportion of entities within the whole database that take the designated values of the attributes in A and B.

Finding Association Rules can be computationally intensive, and essentially involves finding all of the covering attribute sets, A, and then testing whether the rule "A implies B," for some attribute set B separate from A, holds with sufficient confidence. Efficiency gains can be made by a combinatorial analysis of information gained from previous passes to eliminate unnecessary rules from the list of candidate rules. Another highly successful approach is to use **sampling** of the database to estimate whether or not an attribute set is covering. In a large data set it may be necessary to consider which rules are interesting to the user. An approach to this is to use **templates**, to describe the form of interesting rules.

2. Data Manipulation Techniques

For Data Mining purposes it is often necessary to use a large number of data manipulations of various types. When searching for Association Rules, for example, tuples

with certain attribute values are grouped together, a task that may require a sequence of conditional data selection operations. This task is followed by counting operations to determine the cardinality of the selected groups. The nature of the rules themselves may require further data manipulation operations such as summing or averaging of data values if the rules involve comparison of numerical attributes. Frequently knowledge is discovered by combining data from more than one source—knowledge which was not available from any one of the sources alone.

3. Slice and Dice

Slice and Dice refers to techniques specifically designed for examining cross sections of the data. Perhaps the most important aspect of Slice and Dice techniques is the facility to view cross sections of data that are not physically visible. Data may be sliced and diced to provide views in orthogonal planes, or at arbitrarily chosen viewing angles. Such techniques can be vital in facilitating the discovery of knowledge for medical diagnosis without the requirement for invasive surgery.

4. Access Methods

For Data Mining purposes it is often necessary to retrieve very large amounts of data from their stores. It is therefore important that access to data can be achieved rapidly and efficiently, which effectively means with a minimum number of input/output operations (I/Os) involving physical storage devices. Databases are stored on direct access media, referred to generally as **disks**. As disk access times are very much slower than main storage access times, acceptable database performance is achieved by adopting techniques whose objective is to arrange data on the disk in ways which permit stored records to be located in as few I/Os as possible.

It is valuable to identify tuples that are logically related, as these are likely to be frequently requested together. By locating two logically related tuples on the same page, they may both be accessed by a single physical I/O. Locating logically related tuples physically close together is referred to as clustering. Intra-file clustering may be appropriate if sequential access is frequently required to a set of tuples within a file; inter-file clustering may be used if sets of tuples from more than one file are frequently requested together.

D. Enabling Technologies

1. Data Cleansing Techniques

Before commencing Data Mining proper, we must first consider all data that is erroneous, irrelevant or atypical, which Statisticians term **outliers**.

Different types of outliers need to be treated in different ways. Outliers that have occurred as a result of human error may be detected by consistency checks (or integrity constraints). If outliers are a result of human ignorance, this may be handled by including information on changing definitions as **metadata**, which should be consulted when outlier tests are being carried out. Outliers of distribution are usually detected by outlier tests which are based on the deviation between the candidate observation and the average of all the data values.

2. Missing Value Handling

When we carry out Data Mining, we are often working with large, possibly heterogeneous data. It therefore frequently happens that some of the data values are missing because data were not recorded in that case or perhaps was represented in a way that is not compatible with the remainder of the data. Nonetheless, we need to be able to carry out the Data Mining process as best we can. A number of techniques have been developed which can be used in such circumstances, as follows:

- All tuples containing missing data are eliminated from the analysis.
- All missing values are eliminated from the analysis.
- A typical data value is selected at random and imputed to replace the missing value.

3. Advanced Database Technology

The latest breed of databases combines high performance with multidimensional data views and fast, optimized query execution. Traditional databases may be adapted to provide query optimization by utilizing Parallel Processing capabilities. Such **parallel databases** may be implemented on parallel hardware to produce a system that is both powerful and scaleable. Such postrelational Database Management Systems represent data through nested multidimensional tables that allow a more general view of the data of which the relational model is a special case. **Distributed Databases** allow the contributing heterogeneous databases to maintain local autonomy while being managed by a global data manager that presents a single data view to the user. **Multidimensional servers** support the multidimensional data view that represents multidimensional data through nested data. In the three-dimensional case, the data are stored in the form of a **data cube**, or in the case of many dimensions we use the general term **data hypercube**.

The Data Warehousing process involves assembling data from heterogeneous sources systematically by using **middleware** to provide connectivity. The data are

then cleansed to remove inaccuracies and inconsistencies and transformed to give a consistent view. **Metadata** that maintain information concerning the source data are also stored in the warehouse. Data within the warehouse is generally stored in a distributed manner so as to increase efficiency and, in fact, parts of the warehouse may be replicated at local sites, in **data marts**, to provide a facility for departmental decision-making.

4. Visualization Methods

Visualization Methods aim to present complex and voluminous data sets in pictorial and other graphical representations that facilitate understanding and provide insight into the underlying structures in the data. The subject is essentially interdisciplinary, encompassing statistical graphics, computer graphics, image processing, computer vision, interface design and cognitive psychology.

For exploratory Data Mining purposes, we require flexible and interactive visualization tools which allow us to look at the data in different ways and investigate different subsets of the data. We can highlight key features of the display by using color coding to represent particular data values. Charts that show relationships between individuals or objects within the dataset may be color-coded, and thus reveal interesting information about the structure and volume of the relationships. Animation may provide a useful way of exploring sequential data or time series by drawing attention to the changes between time points. **Linked windows**, which present the data in various ways and allow us to trace particular parts of the data from one window to another, may be particularly useful in tracking down interesting or unusual data.

5. Intelligent Agents

The potential of Intelligent Agents is increasingly having an impact on the marketplace. Such agents have the capability to form their own goals, to initiate action without instructions from the user and to offer assistance to the user without being asked. Such software has been likened to an intelligent personal assistant who works out what is needed by the boss and then does it. Intelligent Agents are essentially software tools that interoperate with other software to exchange information and services. They act as an intelligent layer between the user and the data, and facilitate tasks that serve to promote the user's overall goals. Communication with other software is achieved by exchanging messages in an **agent communication language**. Agents may be organized into a federation or agency where a number of agents interact to carry out different specialized tasks.

6. OLAP

The term OLAP (On-line Analytical Processing) originated in 1993 when Dr. E. F. Codd and colleagues developed the idea as a way of extending the relational database paradigm to support business modeling. This development took the form of a number of rules that were designed to facilitate fast and easy access to the relevant data for purposes of management information and decision support. An OLAP Database generally takes the form of a multidimensional server database that makes management information available interactively to the user. Such multidimensional views of the data are ideally suited to an analysis engine since they give maximum flexibility for such database operations as Slice and Dice or drill down which are essential for analytical processing.

7. Parallel Processing

High-performance parallel database systems are displacing traditional systems in very large databases that have complex and time-consuming querying and processing requirements. Relational queries are ideally suited to parallel execution since they often require processing of a number of different relations. In addition to parallelizing the data retrieval required for Data Mining, we may also parallelize the data processing that must be carried out to implement the various algorithms used to achieve the Mining tasks. Such processors may be designed to (1) share memory, (2) share disks, or (3) share nothing. Parallel Processing may be carried out using shared address space, which provides hardware support for efficient communication. The most scaleable paradigm, however, is to share nothing, since this reduces the overheads. In Data Mining, the implicitly parallel nature of most of the Mining tasks allows us to utilize processors which need only interact occasionally, with resulting efficiency in both speed-up and scalability.

8. Distributed Processing

Distributed databases allow local users to manage and access the data in the local databases while providing some sort of global data management which provides global users with a global view of the data. Such global views allow us to combine data from the different sources which may not previously have been integrated, thus providing the potential for new knowledge to be discovered. The constituent local databases may either be homogeneous and form part of a design which seeks to distribute data storage and processing to achieve greater efficiency, or they may be heterogeneous and form part of a legacy system where

the original databases might have been developed using different data models.

E. Relating the Technologies to the Tasks

Data Mining embraces a wealth of methods that are used in parts of the overall process of Knowledge Discovery in Databases. The particular Data Mining methods employed need to be matched to the user's requirements for the overall KDD process.

The tools for the efficient storage of and access to large datasets are provided by the Database Technologies. Recent advances in technologies for data storage have resulted in the availability of inexpensive high-capacity storage devices with very fast access. Other developments have yielded improved database management systems and Data Warehousing technologies. To facilitate all of the Data Mining Tasks, fast access methods can be combined with sophisticated data manipulation and Slice and Dice techniques for analysis of Data Warehouses through OLAP to achieve the intelligent extraction and management of information.

The general tasks of Data Mining are those of description and prediction. Descriptions of the data often require Summarization to provide concise accounts of some parts of the dataset that are of interest. Prediction involves using values of some attributes in the database to predict unknown values of other attributes of interest. Classification, Clustering, and Pattern Recognition are all Data Mining Tasks that can be carried out for the purpose of description, and together with Predictive Modeling and Sequence Processing can be used for the purpose of prediction. All of these descriptive and predictive tasks can be addressed by both Machine Learning Technologies such as Inferencing Rules, Neural Networks, Case-Based Reasoning, and Genetic Algorithms, or by a variety of Uncertainty Methods.

Data Mining methods are used to extract both patterns and models from the data. This involves modeling dependencies in the data. The model must specify both the structure of the dependencies (i.e., which attributes are inter-dependent) and their strengths. The tasks of Dependency Detection and modeling may involve the discovery of empirical laws and the inference of causal models from the data, as well as the use of Database Technologies such as Association Rules. These tasks can be addressed by Machine Learning Technologies such as Inferencing Rules, Neural Networks and Genetic Algorithms, or by Uncertainty Methods, including Statistical Techniques, Bayesian Belief Networks, Evidence Theory, Fuzzy Logic and Rough Sets.

The tasks of Visualization and Summarization play a central role in the successful discovery and analysis of patterns in the data. Both of these are essentially based on Statistical Techniques associated with Exploratory Data Analysis. The overall KDD process also encompasses the task of Uncertainty Handling. Real-world data are often subject to uncertainty of various kinds, including missing values, and a whole range of Uncertainty Methods may be used in different approaches to reasoning under uncertainty.

III. DATA MINING FOR DIFFERENT DATA TYPES

A. Web Mining and Personalization

Developments in Web Mining have been inexorably linked to developments in e-commerce. Such developments have accelerated as the Internet has become more efficient and more widely used. Mining of click streams and session log analysis allows a web server owner to extract new knowledge about users of the service, thus, in the case of e-commerce, facilitating more targeted marketing. Similarly, personalization of web pages as a result of Data Mining can lead to the provision of a more efficient service.

B. Distributed Data Mining

Recent developments have produced a convergence between computation and communication. Organizations that are geographically distributed need a decentralized approach to data storage and decision support. Thus the issues concerning modern organizations are not just the size of the database to be mined, but also its distributed nature. Such developments hold an obvious promise not only for what have become traditional Data Mining areas such as Database Marketing but also for newer areas such as e-Commerce and e-Business.

Trends in DDM are inevitably led by trends in Network Technology. The next generation Internet will connect sites at speeds of the order of 100 times faster than current connectivity. Such powerful connectivity to some extent accommodates the use of current algorithms and techniques. However, in addition, new algorithms, and languages are being developed to facilitate distributed data mining using current and next generation networks.

Rapidly evolving network technology, in conjunction with burgeoning services and information availability on the Internet is rapidly progressing to a situation where a large number of people will have fast, pervasive access to a huge amount of information that is widely accessible. Trends in Network Technology such as bandwidth developments, mobile devices, mobile users, intranets, information overload, and personalization leads to the conclusion that mobile code, and mobile agents, will, in the near future, be a critical part of Internet services. Such developments must be incorporated into Data Mining technology.

C. Text Mining

Text may be considered as sequential data, similar in this respect to data collected by observation systems. It is therefore appropriate for Data Mining techniques that have been developed for use specifically with sequential data to be also applied to the task of Text Mining. Traditionally, text has been analyzed using a variety of information retrieval methods, including natural language processing. Large collections of electronically stored text documents are becoming increasingly available to a variety of end-users, particularly via the World Wide Web. There is great diversity in the requirements of users: some need an overall view of a document collection to see what types of documents are present, what topics the documents are concerned with, and how the documents are related to one another. Other users require specific information or may be interested in the linguistic structures contained in the documents. In very many applications users are initially unsure of exactly what they are seeking, and may engage in browsing and searching activities.

General Data Mining methods are applicable to the tasks required for text analysis. Starting with textual data, the Knowledge Discovery Process provides information on commonly occurring phenomena in the text. For example, we may discover combinations of words or phrases that commonly appear together. Information of this type is presented using **episodes**, which contain such things as the base form of a word, grammatical features, and the position of a word in a sequence. We may measure, for example, the support for an episode by counting the number of occurrences of the episode within a given text sequence.

For Text Mining, a significant amount of pre-processing of the textual data may be required, dependent on the domain and the user's requirements. Some natural language analysis may be used to augment or replace some words by their parts of speech or by their base forms. Post-processing of the results of Text Mining is usually also necessary.

D. Temporal Data Mining

Temporal Data Mining often involves processing time series, typically sequences of data, which measure values of the same attribute at a sequence of different time points. Pattern matching using such data, where we are searching for particular patterns of interest, has attracted considerable interest in recent years. In addition to traditional statistical methods for time series analysis, more recent work on sequence processing has used association rules developed by the database community. In addition Temporal Data Mining may involve exploitation of efficient methods of data storage, fast processing and fast retrieval methods that have been developed for temporal databases.

E. Spatial Data Mining

Spatial Data Mining is inexorably linked to developments in Geographical Information Systems. Such systems store spatially referenced data. They allow the user to extract information on contiguous regions and investigate spatial patterns. Data Mining of such data must take account of spatial variables such as distance and direction. Although methods have been developed for Spatial Statistics, the area of Spatial Data Mining per se is still in its infancy. There is an urgent need for new methods that take spatial dependencies into account and exploit the vast spatial data sources that are accumulating. An example of such data is provided by remotely sensed data of images of the earth collected by satellites.

F. Multimedia Data Mining

Multimedia Data Mining involves processing of data from a variety of sources, principally text, images, sound, and video. Much effort has been devoted to the problems of indexing and retrieving data from such sources, since typically they are voluminous. A major activity in extracting knowledge from time-indexed multimedia data, e.g., sound and video, is the identification of episodes that represent particular types of activity; these may be identified in advance by the domain expert. Likewise domain knowledge in the form of metadata may be used to identify and extract relevant knowledge. Since multimedia contains data of different types, e.g., images along with sound, ways of combining such data must be developed. Such problems of Data Mining from multimedia data are, generally speaking, very difficult and, although some progress has been made, the area is still in its infancy.

G. Security and Privacy Aspects of Data Mining

As we have seen, Data Mining offers much as a means of providing a wealth of new knowledge for a huge range of applications. The knowledge thus obtained from databases may be far in excess of the use to which the data owners originally envisaged for the database. However, such data may include sensitive information about individuals or might involve company confidential information. Care must therefore be taken to ensure that only authorized personnel are permitted to access such databases. However, it may be possible to get around this problem of preserving the security of individual level data by using anonymization techniques and possibly only providing a sample of

the data for Mining purposes. In addition, it is often the case that, for purposes of Data Mining, we do not need to use individual level data but instead can utilize aggregates.

For Database Technology, intrusion detection models must be developed which protect the database against security breaches for the purpose of Data Mining. Such methods look for evidence of users running huge numbers of queries against the database, large volumes of data being downloaded by the user, or users running their own imported software on portions of the database.

H. Metadata Aspects of Data Mining

Currently, most data mining algorithms require bringing all together data to be mined in a single, centralized data warehouse. A fundamental challenge is to develop distributed versions of data mining algorithms so that data mining can be done while leaving some of the data in place. In addition, appropriate protocols, languages, and network services are required for mining distributed data to handle the mappings required for mining distributed data. Such functionality is typically provided via metadata.

XML (eXtensible Markup Language) is fast emerging as a standard for representing data on the World Wide Web. Traditional Database Engines may be used to process semistructured XML documents conforming to Data Type Definitions (DTDs). The XML files may be used to store metadata in a representation to facilitate the mining of multiple heterogeneous databases. **PMML** (predictive Mark-up Language) has been developed by the Data Mining community for the exchange of models between different data sites; typically these will be distributed over the Internet. Such tools support interoperability between heterogeneous databases thus facilitating Distributed Data Mining.

IV. KEY APPLICATION AREAS

A. Industry

Industrial users of databases are increasingly beginning to focus on the potential for embedded artificial intelligence within their development and manufacturing processes. Most industrial processes are now subject to technological control and monitoring, during which vast quantities of manufacturing data are generated. Data Mining techniques are also used extensively in process analysis in order to discover improvements which may be made to the process in terms of time scale and costs.

Classification techniques and rule induction methods are used directly for quality control in manufacturing. Parameter settings for the machinery may be monitored and evaluated so that decisions for automatic correction or intervention can be taken if necessary. Machine Learning technologies also provide the facility for failure diagnosis in the maintenance of industrial machinery.

Industrial safety applications are another area benefiting from the adoption of Data Mining technology. Materials and processes may need to be classified in terms of their industrial and environmental safety. This approach, as opposed to experimentation, is designed to reduce the cost and time scale of safe product development.

B. Administration

There is undoubtedly much scope for using Data Mining to find new knowledge in administrative systems that often contain large amounts of data. However, perhaps because the primary function of administrative systems is routine reporting, there has been less uptake of Data Mining to provide support and new knowledge for administrative purposes that in some other application areas.

Administrative systems that have received attention tend to be those in which new knowledge can be directly translated into saving money. An application of this type is provided by the Inland Revenue that collects vast amounts of data and may potentially save a lot of money by devising ways of discovering tax dodges, similarly for welfare frauds. Another successful application of Data Mining has been to the health care system where again new discoveries about expensive health care options can lead to huge savings. Data Mining is also likely to become an extremely useful tool in criminal investigations, searching for possible links with particular crimes or criminals.

C. Business

As we might expect, the major application area of Data Mining, so far, has been Business, particularly the areas of Marketing, Risk Assessment, and Fraud Detection.

In Marketing, perhaps the best known use of Data Mining is for customer profiling, both in terms of discovering what types of goods customers tend to purchase in the same transaction and groups of customers who all behave in a similar way and may be targeted as a group. Where customers tend to buy (unexpected) items together then goods may be placed on nearby shelves in the supermarket or beside each other in a catalogue. Where customers may be classified into groups, then they may be singled out for customized advertising, mail shots, etc. This is known as micro marketing. There are also cases where customers of one type of supplier unexpectedly turn out to be also customers of another type of supplier and advantage may be gained by pooling resources in some sense. This is known as cross marketing.

Another use of Data Mining for Business has been for Risk Assessment. Such assessment of credit worthiness of potential customers is an important aspect of this use which has found particular application to banking institutions where lending money to potentially risky customers is an important part of the day-to-day business. A related application has been to litigation assessment where a firm may wish to assess how likely and to what extent a bad debtor will pay up and if it is worth their while getting involved in unnecessary legal fees.

Case Study I. A supermarket chain with a large number of stores holds data on the shopping transactions and demographic profile of each customer's transactions in each store. Corporate management wants to use the customer databases to look for global and local shopping patterns.

D. Database Marketing

Database Marketing refers to the use of Data Mining techniques for the purposes of gaining business advantage. These include improving a company's knowledge of its customers in terms of their characteristics and purchasing habits and using this information to classify customers; predicting which products may be most usefully offered to a particular group of customers at a particular time; identifying which customers are most likely to respond to a mail shot about a particular product; identifying customer loyalty and disloyalty and thus improving the effectiveness of intervention to avoid customers moving to a competitor; identifying the product specifications that customers really want in order to improve the match between this and the products actually offered; identifying which products from different domains tend to be bought together in order to improve cross-marketing strategies; and detecting fraudulent activity by customers.

One of the major tasks of Data Mining in a commercial arena is that of market segmentation. Clustering techniques are used in order to partition a customer database into homogeneous segments characterized by customer needs, preferences, and expenditure. Once market segments have been established, classification techniques are used to assign customers and potential customers to particular classes. Based on these, prediction methods may be employed to forecast buying patterns for new customers.

E. Medicine

Potential applications of Data Mining to Medicine provide one of the most exciting developments and hold much promise for the future. The principal medical areas which have been subjected to a Data Mining approach, so far, may be categorized as: diagnosis, treatment, monitoring, and research.

The first step in treating a medical complaint is diagnosis, which usually involves carrying out various tests and observing signs and symptoms that relate to the possible diseases that the patient may be suffering from. This may involve clinical data, data concerning biochemical indicators, radiological data, sociodemographic data including family medical history, and so on. In addition, some of these data may be measured at a sequence of time-points, e.g., temperature, lipid levels. The basic problem of diagnosis may be regarded as one of classification of the patient into one, or more, possible disease classes.

Data Mining has tremendous potential as a tool for assessing various treatment regimes in an environment where there are a large number of attributes which measure the state of health of the patient, allied to many attributes and time sequences of attributes, representing particular treatment regimes. These are so complex and interrelated, e.g., the interactions between various drugs, that it is difficult for an individual to assess the various components particularly when the patient may be presenting with a variety of complaints (multi-pathology) and the treatment for one complaint might mitigate against another.

Perhaps the most exciting possibility for the application of Data Mining to medicine is in the area of medical research. Epidemiological studies often involve large numbers of subjects which have been followed-up over a considerable period of time. The relationship between variables is of considerable interest as a means of investigating possible causes of diseases and general health inequalities in the population.

Case Study II. A drug manufacturing company is studying the risk factors for heart disease. It has data on the results of blood analyses, socioeconomic data, and dietary patterns. The company wants to find out the relationship between the heart disease markers in the blood and the other relevant attributes.

F. Science

In many areas of science, automatic sensing and recording devices are responsible for gathering vast quantities of data. In the case of data collected by remote sensing from satellites in disciplines such as astronomy and geology the amount of data are so great that Data Mining techniques offer the only viable way forward for scientific analysis.

One of the principal application areas of Data Mining is that of space exploration and research. Satellites provide immense quantities of data on a continuous basis via remote sensing devices, for which intelligent, trainable image-analysis tools are being developed. In previous large-scale studies of the sky, only relatively small amount of the data collected have actually been used in manual attempts to classify objects and produce galaxy catalogs.

Not only has the sheer amount of data been overwhelming for human consideration, but also the amount of data required to be assimilated for the observation of a single significant anomaly is a major barrier to purely manual analysis. Thus Machine Learning techniques are essential for the classification of features from satellite pictures, and they have already been used in studies for the discovery of quasars. Other applications include the classification of landscape features, such as the identification of volcanoes on the surface of Venus from radar images. Pattern recognition and rule discovery also have important applications in the chemical and biomedical sciences. Finding patterns in molecular structures can facilitate the development of new compounds, and help to predict their chemical properties. There are currently major projects engaged in collecting data on the human gene pool, and rule-learning has many applications in the biomedical sciences. These include finding rules relating drug structure to activity for diseases such as Alzheimer's disease, learning rules for predicting protein structures, and discovering rules for the use of enzymes in cancer research.

Case Study III. An astronomy cataloguer wants to process telescope images, identify stellar objects of interest and place their descriptions into a database for future use.

G. Engineering

Machine Learning has an increasing role in a number of areas of engineering, ranging from engineering design to project planning. The modern engineering design process is heavily dependent on computer-aided methodologies. Engineering structures are extensively tested during the development stage using computational models to provide information on stress fields, displacement, load-bearing capacity, etc. One of the principal analysis techniques employed by a variety of engineers is the finite element method, and Machine Learning can play an important role in learning rules for finite element mesh design for enhancing both the efficiency and quality of the computed solutions.

Other engineering design applications of Machine Learning occur in the development of systems, such as traffic density forecasting in traffic and highway engineering. Data Mining technologies also have a range of other engineering applications, including fault diagnosis (for example, in aircraft engines or in on-board electronics in intelligent military vehicles), object classification (in oil exploration), and machine or sensor calibration. Classification may, indeed, form part of the mechanism for fault diagnosis.

As well as in the design field, Machine Learning methodologies such as Neural Networks and Case-Based Reasoning are increasingly being used for engineering project management in an arena in which large scale international projects require vast amounts of planning to stay within time scale and budget.

H. Fraud Detection and Compliance

Techniques which are designed to register abnormal transactions or data usage patterns in databases can provide an early alert, and thus protect database owners from fraudulent activity by both a company's own employees and by outside agencies. An approach that promises much for the future is the development of adaptive techniques that can identify particular fraud types, but also be adaptive to variations of the fraud. With the ever-increasing complexity of networks and the proliferation of services available over them, software agent technology may be employed in the future to support interagent communication and message passing for carrying out surveillance on distributed networks.

Both the telecommunications industry and the retail businesses have been quick to realize the advantages of Data Mining for both fraud detection and discovering failures in compliance with company procedures. The illegal use of telephone networks through the abuse of special services and tariffs is a highly organized area of international crime. Data Mining tools, particularly featuring Classification, Clustering, and Visualization techniques have been successfully used to identify patterns in fraudulent behavior among particular groups of telephone service users.

V. FUTURE DEVELOPMENTS

Data Mining, as currently practiced, has emerged as a subarea of Computer Science. This means that initial developments were strongly influenced by ideas from the Machine Learning community with a sound underpinning from Database Technology. However, the statistical community, particularly Bayesians, was quick to realize that they had a lot to contribute to such developments. Data Mining has therefore rapidly grown into the interdisciplinary subject that it is today.

Research in Data Mining has been led by the KDD (Knowledge Discovery in Databases) annual conferences, several of which have led to books on the subject (e.g., Fayyad *et al.*, 1996). These conferences have grown in 10 years from being a small workshop to a large independent conference with, in Boston in 2000, nearly 1000 participants. The proceedings of these conferences are still the major outlet for new developments in Data Mining.

Major research trends in recent years have been:

- The development of scalable algorithms that can operate efficiently using data stored outside main memory

- The development of algorithms that look for local patterns in the data—data partitioning methods have proved to be a promising approach
- The development of Data Mining methods for different types of data such as multimedia and text data
- The developments of methods for different application areas

Much has been achieved in the last 10 years. However, there is still huge potential for Data Mining to develop as computer technology improves in capability and new applications become available.

SEE ALSO THE FOLLOWING ARTICLES

ARTIFICAL NEURAL NETWORKS • COMPUTER ALGORITHMS • DATABASES • FOURIER SERIES • FUNCTIONAL ANALYSIS • FUZZY SETS, FUZZY LOGIC, AND FUZZY SYSTEMS • STATISTICS, BAYESIAN • WAVELETS

BIBLIOGRAPHY

Adriaans, P., and Zantinge, D. (1996). "Data Mining," Addison-Wesley, MA.

Berry, M., and Linoff, G. (1997). "Data Mining Techniques for Marketing, Sales and Customer Support," Wiley, New York.

Berson, A., and Smith, S. J. (1997). "Data Warehousing, Data Mining, and Olap," McGraw-Hill, New York.

Bigus, J. (1996). "Data Mining With Neural Networks," McGraw-Hill, New York.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). "Classification and Regression Trees," Wadsworth, Belmont.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A. (1997). "Discovering Data Mining from Concept to Implementation," Prentice-Hall, Upper Saddle River, NJ.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). "Advances in Knowledge Discovery and Data Mining," AAAI Press/The MIT Press, Menlo Park, CA.

Freitas, A. A., and Lavington, S. H. (1998). "Mining Very Large Databases With Parallel Processing," Kluwer, New York.

Groth, R. (1997). "Data Mining: A Hands on Approach to Information Discovery," Prentice-Hall, Englewood Cliffs, NJ.

Inmon, W. (1996). "Using the Data Warehouse," Wiley, New York.

Kennedy, R. L., Lee, Y., Van Roy, B., and Reed, C. D. (1997). "Solving Data Mining Problems Through Pattern Recognition," Prentice-Hall, Upper Saddle River, NJ.

Lavrac, N., Keravnou, E. T., and Zupan, B. (eds.). (1997). "Intelligent Data Analysis in Medicine and Pharmacology," Kluwer, Boston.

Mattison, R. M. (1997). "Data Warehousing and Data Mining for Telecommunications," Artech House, MA.

Mitchell, T. (1997). "Machine Learning," McGraw-Hill, New York.

Ripley, B. (1995). "Pattern Recognition and Neural Networks," Cambridge University Press, Cambridge.

Stolorz, P., and Musick, R. (eds.). (1997). "Scalable High Performance Computing for Knowledge Discovery and Data Mining," Kluwer, New York.

Weiss, S. M., and Indurkhya, N. (1997). "Predictive Data Mining: A Practical Guide" (with Software), Morgan Kaufmann, San Francisco, CA.

Wu, X. (1995). "Knowledge Acquisition from Databases," Ablex, Greenwich, CT.