Knowledge Discovery in Databases: An Information Retrieval Perspective

Ong Cheng Soon
MIMOS Berhad
Technology Park Malaysia
57000 Kuala Lumpur
Malaysia
Tel: +603 89965000
Email: csong@mimos.my

ABSTRACT

The current trend of increasing capabilities in data generation and collection has resulted in an urgent need for data mining applications, also called knowledge discovery in databases. This paper identifies and examines the issues involved in extracting useful grains of knowledge from large amounts of data.

It describes a framework to categorise data mining systems. The author also gives an overview of the issues pertaining to data pre-processing, as well as various information gathering methodologies and techniques. It covers some popular tools such as classification, clustering, and generalisation. A summary of statistical and machine learning techniques used currently is also provided.

Keywords: Data Mining, Knowledge Discovery, Database, Statistical Techniques, Machine Learning.

1 INTRODUCTION

Current improvements in technology have enabled businesses and individuals to access massive amounts of data. The capabilities of both generating and collecting data have been increasing rapidly [1], [2], [3]. The resulting information overload has generated an urgent need for new techniques and tools that intelligently and automatically transforms the processed data into useful information and knowledge.

Data Mining, which is also referred to as knowledge discovery in databases, is the process of nontrivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules, constraints, and regularities) from data in databases [4]. In this context, it also implies the use of automated or semi-automated, computational intelligence driven systems to perform the task. Data mining derives its name from the similarities between searching for valuable business information in a large database and mining a mountain for a vein or even just a grain of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides.

As the name implies, knowledge discovery in databases involves the extraction of interesting knowledge, regularities or high level information from the relevant sets of data in databases. Databases can then be investigated from different angles and hence become rich sources for knowledge generation and verification. The discovered knowledge can be applied to information management, query processing, decision making, process control and many other applications. Emerging information services applications such as the World Wide Web and on-line services also require data mining to better understand user behaviour, improve services and increase business opportunities.

1.1 A Data Mining Architecture

Data Mining techniques must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. The raw data from the warehouse is preprocessed to remove noise and also to impart domain knowledge on the data. The knowledge discovery stage then extracts the knowledge which must then be post processed to facilitate human understanding. Post processing usually takes the form of representing the discovered knowledge in a user friendly display. Figure 1 illustrates an architecture for advanced analysis in a large data warehouse.

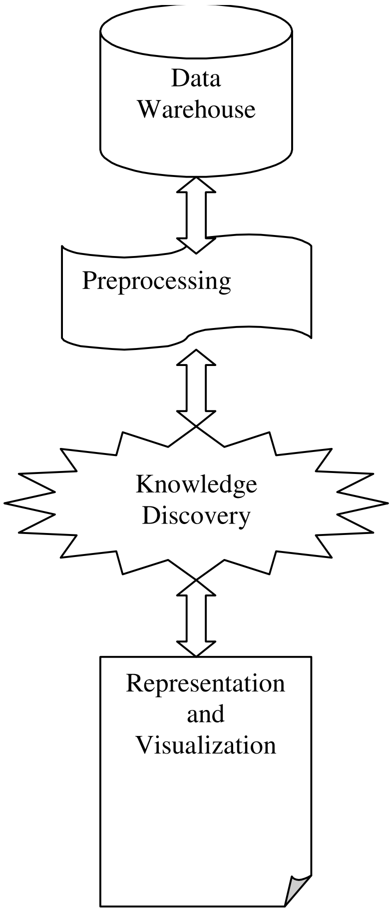

Fig.1: Integrated data mining architecture.

Knowledge discovery from databases is based on three main enabling technologies:

- Increased computing power
- Statistical and learning algorithms
- Improved data collection and management

1.2 Structure of this paper

Previous work [5], [6], [7] focuses on evaluations of specific tools and applications available in the marketplace. This paper identifies and examines the issues involved in data mining or knowledge discovery in databases, from the perspective of information retrieval. It focuses on the extraction of information from large amounts of data. Section 2 describes a method to classify and compare the numerous data mining tools that are available on the market today. The issues pertaining to data preparation are discussed in section 3. Section 4 gives a brief overview of the types of information that can be extracted, while section 5 investigates the various technologies that can be used to discover the information. In Section 6, the author summarises the areas covered in this paper and briefly mentions some other issues not discussed.

2 A FRAMEWORK TO CATERGORISE DATA MINING APPLICATIONS

There are many different products and systems on the marketplace today. These systems are based on various new methods and technologies. The convergence of machine learning, statistics and database technologies have produced a multitude of different methods for extracting information from data. Data mining systems can be classified based on the kinds of knowledge to be discovered, and the kinds of techniques to be utilised, as shown below. Another issue is the kinds of database that the system encounters, with the accompanying noise and accuracy attributes.

2.1 Data Mining Models

Data mining can have two models or modes of operation, each of which has a different emphasis. The verification model takes a hypothesis from the user and tests the validity of it against the data. The discovery model differs in its focus in that the system automatically discovers important information hidden in the data.

2.1.1 Verification Model

The verification model allows the user to verify his or her hypothesis. The user is responsible for formulating the hypothesis and issuing the query on the data to affirm or negate the hypothesis. For example, consider a marketing division planning the launch of a new product. With a limited budget for a mailing campaign, this organisation will need to identify the section of the population most likely to buy the new product. The user formulates a hypothesis to identify potential customers and the characteristics they share. Historical data about customer purchase and demographic information can then be queried to reveal comparable purchases and the characteristics shared by those purchasers which in turn can be used to target a mailing campaign. The whole operation can be refined by `drilling down' so that the hypothesis reduces the `set' returned each time until the required limit is reached. This implies that the user iteratively selects more focused subsets of the population. This process is sometimes labelled as on-line analytical processing (OLAP).

The problem with this model is the fact that no new information is created in the retrieval process but rather the queries will always return records to verify or negate the hypothesis. The search process here is iterative in that the output is reviewed, a new set of questions or hypothesis formulated to refine the search and the whole process repeated. The user is discovering the facts about the data using a variety of techniques such as queries, multidimensional analysis and visualisation to guide the exploration of the data being inspected.

2.1.2 Discovery Model

In the discovery model, the data drives the search for nuggets of information. This process searches for frequently occurring patterns, trends and generalisations about the data without any intervention or guidance from the user. The discovery or data mining tools aim to reveal a large number of facts about the data in as short a time as possible. An example of such a model is a bank database that is mined to discover the many groups of customers to target for a mailing campaign. The data is searched with no hypothesis in mind other than for the data mining system to group the customers according to the common characteristics found.

2.2 Classifying data mining techniques

There have been many advances on researches and developments of data mining, and many data mining techniques and systems have recently been developed. Different classification schemes can be used to categorise data mining methods and systems [5]. It can be argued that data mining systems can be classified according to the type of database system. However, irrespective of the database system used, whether it is a relational database or a transactional database or even a simple text file, raw data can be conceptually represented as a matrix of values. For example, each entry in a relational database may correspond to a row in a matrix, with each column representing the attributes of the entry.

2.2.1 What kind of knowledge to be mined.

Several typical kinds of knowledge can be discovered by data mining systems, including rules, clustering, factor analysis, generalisation, classification and deviation analysis. Moreover, these systems can also be categorised according to various abstraction levels. This includes generalised knowledge, primitive level knowledge, and multiple level knowledge. A flexible data mining system may discover knowledge at multiple abstraction levels.

2.2.2 What kind of techniques to be utilised.

Data mining systems can also be categorised according to the underlying data mining techniques. For example, it can be categorised according to the driven method into autonomous knowledge miner, data driven miner, query driven miner, and interactive data miner. They can also be categorised according to its underlying data mining approach into generalisation based mining, pattern based mining, mining based on statistics or mathematical theories, and integrated approaches, etc.

2.3 Commercial Tools

There have been many successful commercial tools implementing the various algorithms and methods for data mining. Recent surveys [5], [6], [7] compare several popular desktop tools as well and high-end software for data mining. The surveys also discuss the strengths and weaknesses of the commercially available data mining tools as well as the costs associated with them.

3 DATA PRE-PROCESSING

Before data from a warehouse can be used for knowledge discovery, it has to be pre-processed. This includes identifying, gathering, cleaning, and labelling the data. Some thought also has to go into specifying the questions to be asked of it, finding the right way to view it, and to discover useful patterns. Despite the central importance of actually modelling the data, that stage actually takes up only a small proportion of the project effort. Increased automation has not absolved researchers of the need to think in statistical terms, including keeping a lookout for the unexpected. However, modern statistical modelling tools do make it possible for an analyst to think about the problem at a higher level of abstraction. It also allows the researcher to try numerous approaches, and to estimate the uncertainty of conclusions arising out of complex processes. The analyst can iterate through several stages of a solution design before settling on a representation scheme (or even a blend of them). When one is comparing data mining techniques, or attempting to extract the most out of a database, it makes sense to try some of these accessible modern statistical and machine learning algorithms.

Data mining systems rely on databases to supply the raw data for input. This raises problems in that databases tend to be dynamic, incomplete, noisy, and large. Other problems arise as a result of the adequacy and relevance of the information stored.

3.1 Limited Information

A database is often designed for purposes different from data mining. Sometimes the properties or attributes that would simplify the learning task are not present nor can they be requested from the real world. Incomplete data cause problems because if some attributes essential to knowledge about the application domain are not present in the data it may be impossible to discover significant knowledge about a given domain. For example, one cannot diagnose malaria from a patient database if that database does not contain the patients' red blood cell count, since the malaria depletes red blood cells.

3.2 Noise and missing values

Databases are usually contaminated by errors so it cannot be assumed that the data they contain is entirely correct. Attributes that rely on subjective or measurement judgements can give rise to errors such that some examples may end up misclassified. Error in either the values of attributes or class information are known as noise. Obviously, it is desirable to eliminate noise from the classification information as this affects the overall accuracy of the generated rules. Missing data can be treated by discovery systems in a number of ways such as; simply disregard missing values, omit the corresponding records, infer missing values from known values, treat missing data as a special value to be included additionally in the attribute domain, or average over the missing values using Bayesian techniques. Noisy data in the sense of being imprecise is characteristic of all data collection and typically fit a regular statistical distribution such as Gaussian while wrong values are data entry errors. Statistical methods can treat problems of noisy data, and separate different types of noise.

3.3 Uncertainty

Uncertainty refers to the severity of the error and the degree of noise in the data. Data precision is an important consideration in a discovery system. Approaches to the modelling and management of uncertainty and inaccuracy in database systems can be categorised into two broad categories, quantitative and qualitative. Quantitative techniques use numerical factors for uncertainty, and manipulate these factors to obtain numerical measures for the uncertainty of derived data. Qualitative techniques are often based on partitioning the data into ``definite'' and ``indefinite'' components. A data mining system may have to take these factors into account when dealing with noisy data.

3.4 Size, updates, and irrelevant fields

Databases tend to be large and dynamic. Their contents are ever changing as information is added, modified or removed. From the data mining perspective, the issue is ensuring that the rules are up-to-date and consistent with the most current information. Also the learning system has to be time-sensitive as some data values vary over time and the discovery system is affected by the `timeliness' of the data. Another issue is the relevance or irrelevance of the fields in the database to the current focus of discovery. For example, postcodes are fundamental to any studies trying to establish a geographical connection to an item of interest, however the number of adults in the household does not provide any useful information for the same study.

4 KNOWLEDGE TO BE MINED

The knowledge extracted from the data can come in many forms. Some of them are human friendly, others are not. For example, propositional if/then rules are easy to interpret whereas a neural network classifier does not lend itself to analysis of the classification. There is also value in producing meta data, or information that describes the database. This generalisation will allow the human user to perform the final knowledge extraction step. This section will examine the different facets of knowledge that can be extracted from data.

4.1 Classification

Classification is the process in which items are categorised according to some predefined taxonomy. For example, using the dimensions of the sepal and petal, the species of iris can be determined. To perform classification, data mining tools have to infer a model from the database, and in the case of supervised learning this requires the user to define one or more classes. The database contains one or more attributes that denote the class of a tuple and these are known as predicted attributes whereas the remaining attributes are called predicting attributes. A combination of values for the predicted attributes defines a class. When learning classification rules, the system has to find the rules that predict the class from the predicting attributes. First, the user has to define conditions for each class, then the data mining system then constructs descriptions for the classes. Basically the system should, given a case or tuple with certain known attribute values, be able to predict which class this case belongs to. Once classes are defined the system can then infer rules that govern the classification. Therefore, the system should be able to find the description of each class, and use it to classify the data into different categories.

4.2 Clustering

Clustering is the technique of grouping together pattern vectors that in some sense belong together because of similar characteristics. The clusters formed with this family of techniques should be internally homogenous (members are similar to one another) and externally heterogeneous (members are not like members of other clusters) [8], [9].

Cluster analysis is an exploratory data analysis tool for solving classification problems. Its objective is to sort cases (people, things, events, etc) into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Each cluster thus describes, in terms of the data collected, the class to which its members belong; and this description may be abstracted through use from the particular to the general class or type. Cluster analysis is thus a tool of discovery. It may reveal associations and structure in data, which though not previously evident, nevertheless are sensible and useful, once found. The results of cluster analysis may contribute to the definition of a formal classification scheme, such as a taxonomy for related animals, insects or plants; or suggest statistical models with which to describe populations; or indicate rules for assigning new cases to classes for identification and diagnostic purposes; or provide measures of definition, size and change in what previously were only broad concepts; or find exemplars to represent classes.

4.3 Generalisation, Estimation & Forecasting

Generalisation requires prior knowledge of the problem domain [10], [11]. For any practical application, selection of relevant indicators and learning method are important to a successful generalisation. Another issue is the selection of training cases that represent the problem domain adequately.

Forecasting is one of the most common uses of data mining. In forecasting, a user takes past data on a given variable, such as sales, and projects that variable into the future. Forecasting can help an organisation plan appropriate strategies for long-term growth. Many of these techniques can pick up not only long-term linear trends, but also short-term cyclical fluctuation. Forecasting in the traditional sense is no longer enough. Forecasts must be usable by those who make decisions that will set the stage for future success. The forecasting process does not conclude with an accurate forecast. Business planners must also understand several critical factors. For example, identifying the major factors influencing success, improving the results instead of just anticipating them, and involving the users with the deepest understanding of the data in the forecasting process.

4.4 Rules

Rules are one of the most easily interpreted information nuggets. It basically is a list of instructions or procedures that the human user can utilize to identify interesting portions of the data. There are two main types of rules: propositional rules and association rules. Each has various different versions, including probabilistic rules and datalog rules.

Propositional rules is a set of "If / Then" rules, which are inferred from the data to classify the different cases. They imply a correlation between the factors in question. However, they should not be misinterpreted to imply causality. Typically, information regarding accuracy and coverage for each case provides clues as to how significant each rule is.

Association rules are models that examine the extent to which values of one field depend on, or are predicted by, values of another field. Association discovery finds rules about items that appear together in an event such as a purchase transaction. The rules have user-stipulated support, confidence, and length. The rules find objects or events that "go together", which is different than "predicted by". These models are often referred to as Market Basket Analysis when they are applied to retail industries to study the buying patterns of their customers. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X ➔Y , where X and Y are sets of items. [12], [13]. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. An example of an association rule is: ``30% of transactions that contain beer also contain diapers; 2% of all transactions contain both of these items". Here 30% is called the confidence of the rule, and 2% the support of the rule.

4.5 Trend and Deviation Analysis

Trend analysis is the discovery of a specific sub-sequence of history that satisfies the users' criterion [14]. Trends may include temporal factors or it may be just a sequence of events. Trend analysis uncovers time-related patterns that deal with variation of quantities and measures. Such a situation is typical of a direct mail application. For example, a catalogue merchant has the information of the sets of products that the customer buys in every purchase order. A sequential pattern function will analyse such collections of related records and will detect frequently occurring patterns of products bought over time. A sequential pattern operator could also be used to discover the set of purchases that frequently precedes the purchase of a microwave oven. Sequential pattern mining functions are quite powerful and can be used to detect the set of customers associated with some frequent buying patterns.

Deviation analysis investigates how patterns and data characteristics vary among comparable data sets. In some sense, deviation analysis is the opposite of rule induction or clustering. While in rule discovery we look for similar characteristics that lead to a conclusion, e.g. customer demographic clusters that lead to profitability, in comparative analysis we look for differences among data sets, e.g. what makes profitable customers different from unprofitable ones. The use of these functions for example, on a set of medical insurance claims can lead to the identification of frequently occurring sequences of medical procedures applied to patients which can help identify good medical practices as well as to detect possible medical insurance fraud.

The basic definition of sequential patterns can be generalised by incorporating further features. These include introduction of time constraints between adjacent elements of a sequence, a more flexible definition of a transaction, and inclusion of data taxonomies [15].

5 TECHNIQUES TO BE USED

There is currently no single technique that satisfies all the requirements of data mining. Hence, researchers and developers tend to use a plethora of tools, each of which has a specific purpose. Viewing the information gained from different perspectives may also provide further insight to the human user. It is to be noted that the techniques used for data mining are closely tied to the knowledge that the user intends to gain from the process. Hence, the technique used is frequently determined by the end result required of the knowledge discovery system. This section gives an overview of the most popular technologies used in data mining systems.

5.1 Decision Trees

Decision trees are one of the most commonly used methods in data mining. The term decision tree is different from what can be found in decision science, in which case decision trees are used to support decisions with uncertainty. In data mining, decision tree is used as a classifying tool to solve a large number of classification problems. Decision tree uses a tree structure to represent knowledge. The construction of the tree is called tree induction [16]. This process is supervised; i.e. it employs a set of pre-classified examples to develop a decision. Once such a tree is inducted, it is used in a predictive mode to classify new records into these same predefined classes.

Decision trees are excellent tools for making decisions where a lot of complex information needs to be taken into account. They provide an effective structure in which alternative decisions and the implications of taking those decisions can be laid down and evaluated. They also help in forming an accurate, balanced picture of the risks and rewards that can result from a particular choice.

5.2 Fuzzy Logic

Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth -- truth values between "completely true" and "completely false". Lotfi Zadeh introduced it as a means to model the uncertainty of natural language [17]. Rather than regarding fuzzy theory as a single theory, we should regard the process of ``fuzzification" as a methodology to generalize any specific theory from a crisp (discrete) to a continuous (fuzzy) form.

Fuzzy systems use fuzzy logic, which comprises of fuzzy sets and fuzzy rules that combines numerical and linguistic data. Fuzzy logic uses natural language terms such as: cold, warm, hot, small, average, large, etc. Such terms are not precise and cannot be represented in normal set theory. Fuzzy sets allow members to be partial members as well as the normal multi-set membership. This partial membership enables the system to process data that has characteristics that are difficult to quantify.

5.3 Evolutionary algorithms

The idea of applying the biological principle of natural evolution to artificial systems, introduced more than three decades ago, has seen impressive growth in the past few years. Usually grouped under the term evolutionary algorithms or evolutionary computation, we find the domains of genetic algorithms, evolution strategies, evolutionary programming, and genetic programming [18]. Evolutionary algorithms are ubiquitous nowadays, having been successfully applied to numerous problems from different domains, including optimization, automatic programming, machine learning, economics, operations research, ecology, population genetics, studies of evolution and learning, and social systems [19].

A evolutionary algorithm is an iterative procedure that consists of a population of individuals, each one represented by a finite string of symbols, known as the genome, encoding a possible solution in a given problem space. This space, referred to as the search space, comprises all possible solutions to the problem at hand. More precisely, evolutionary algorithms maintain a population of structures, which evolve according to rules of selection and other operators, which are referred to as genetic operators, such as recombination and mutation. Each individual in the population receives a measure of its fitness in the environment. Reproduction focuses attention on high fitness individuals, thus exploiting the available fitness information. Recombination and mutation perturb those individuals, providing general heuristics for exploration. The fittest individual after a certain number of generations is then considered to be the optimal solution to the problem. Although simplistic from a biologist's viewpoint, these algorithms are sufficiently complex to provide robust and powerful adaptive search mechanisms.

5.4 Neural Networks

A neural network is a software (or hardware) simulation of a biological brain. The purpose of a neural network is to learn to recognise patterns in data. Neural networks self-adapt to learn from information, providing powerful models representing knowledge about a specific problem [20]. Much of the inspiration for the field of neural networks comes from the desire to produce artificial systems capable of sophisticated, perhaps "intelligent", computations similar to those that the human brain routinely performs. The human brain consists of billions of neurons with even more connections between them. A neural network tries to emulate this marvel of nature by having models of neurons and their corresponding synapses.

Neural networks are non-linear by design, but they do not require explicit specification of a functional form (such as quadratic or cubic) as does non-linear regression. The advantage is that there is no need to have any specific model in mind when running the analysis. Moreover, neural networks can find interaction effects (such as effects from the combination of age and gender) which must be explicitly specified in regression. The disadvantage is that it is harder to interpret the resultant model with its layers of weights and arcane transformations. Neural networks are therefore useful in predicting a target variable when the data is highly non-linear with interactions, but they are not very useful when these relationships in the data need to be explained. They have been reported to be suitable tools for such applications as forecasting, credit scoring, response model scoring, and risk analysis.

5.5 Rule Induction

Rule induction is one of the most common forms of knowledge discovery. It is a technique for discovering propositional and association rules from data to classify the different cases. Because it looks for all possible interesting patterns in a data set, the technique is powerful. But it can be overwhelming with the large number of rules that can be generated. Because the rules are independent, many may contradict each other and they may not cover all possible situations. They also should not be misinterpreted to imply causality. Typically, information regarding accuracy and coverage for each case provides clues as to how significant each rule is.

Association models examine the extent to which values of one field depend on, or are predicted by, values of another field. Association discovery finds rules about items that appear together in an event such as a purchase transaction. There are various algorithms for mining association rules from a set of transactions [12], [13]. The challenge is to find all rules that satisfy user specified minimum support and minimum confidence constraints. Applications include discovering affinities for market basket analysis and cross marketing, catalogue design, loss leader analysis, store layout, customer segmentation based on buying patterns, etc.

5.6 Statistics

The classical techniques using mathematical modeling are still among the fastest and most accurate methods of data mining [5], [6]. Most data mining suites include basic statistical techniques that produce comparable results to so called modern machine learning techniques. A brief description of some popular techniques is provided below.

5.6.1 Correlation

Correlation is a measure of relation between two variables. For example, a high correlation between purchases of certain products such as cheese and crackers indicates that these products are likely to be purchased together. Correlations may be either positive or negative. A positive correlation indicates that a high level of one variable will be accompanied by a high value of the correlated variable. A negative correlation indicates that a high level of one variable will be accompanied by a low value of the correlated variable. For example, for a retail outlet, positive correlations are useful for finding products that tend to be purchased together. Negative correlations can be useful for diversifying across markets in a company's strategic portfolio.

5.6.2 ANOVA

ANOVA, which stands for Analysis of Variance, is a statistical technique which tests differences in mean values of a dependent variable between two or more categories of independent variables. The ANOVA statistical test answers questions arising from the examination and analysis of the study's results. For example, "...did the outcome differ across experimental groups?..." or "...Were there differences on some variable between populations or groups?...", etc. Repeated measures of ANOVA can help answer further questions regarding the before and after differences between the control group and the experimental group.

5.6.3 t-Tests

If you want to test whether there is a difference between two particular regions, you need to use t-tests. In contrast, an ANOVA will only tell you whether there are any differences in multiple regions. T-tests are general tests to see whether two variables have equal means, variances, or entire distributions (when population variance is unknown). They can also be used to test whether the means or variances are equal to a specific number. They are often used in regression when testing the significance of a variable. In this case the test is whether a coefficient in the regression is equal to zero. A "significant" coefficient means that the coefficient is significantly different from zero, which allows the rejection of the hypothesis that the variable has no effect on the result. Hence, a significant coefficient implies that the variable in question affects the outcome of the situation.

5.6.4 Linear Regression

Linear regression is a method that fits a straight line through data. If the line is upward sloping it means that an independent variable such as the size of a sales force has a positive effect on a dependent variable such as revenue. If the line is downward sloping there is a negative effect. The steeper the slope, the more effect the independent variable has on the dependent variable. Although many business models are not linear, many models such as this one between revenue and the size of a sales force can be linearised by a log transformation. Models of any functional form can be estimated by using non-linear regression.

5.6.5 Logistic Regression

Logistic regression estimates the probability of a certain event occurring. It uses observed factors coupled with occurrences or non-occurrences of the event to model the probability of occurrence under different factor conditions. Logistic regression describes the relationship between a dichotomous response variable and a set of explanatory variables. The explanatory variables may be continuous or (with dummy variables) discrete.

5.7 Statistical Classifiers

5.7.1 Nearest neighbour method

Nearest neighbour techniques classify each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k is greater than or equal to 1). This is also sometimes called the k-nearest neighbour technique. The k-nearest neighbour classifier labels an unknown object O with the label of the majority of the k nearest neighbours. A neighbour is deemed nearest if it has the smallest distance, in the Euclidean sense, in feature space. For k = 1, this is the label of its closest neighbour in the learning set. The k-nearest neighbour method is intuitively a very attractive method, since it corresponds with the human notion of similarity. A disadvantage of this method is its large computing power requirement, since for classifying an object its distance to all the objects in the learning set has to be calculated.

5.7.2 Bayesian Classifiers

Classical inferential models do not permit the introduction of prior knowledge into the calculations. For the rigors of the scientific method, this is an appropriate response to prevent the introduction of extraneous data that might skew the experimental results. However, there are times when the use of a priori knowledge would be a useful contribution to the evaluation process.

The Bayes classifier is a mechanism that minimizes the classification error. In order to do this, it needs the underlying probability density functions of the classes. The minimal error this classifier makes, the Bayes error, is a theoretical minimum to the error any classifier can make [9]. In practical problems however, an infinite number of learning objects is never available. For that reason, the true probability density functions and a priori probability information is not available. In this case the error will probably be higher than the Bayes error. Classifier construction is then based on the assumption that the learning objects available represent the true probability density function, i.e. that they are realizations of stochastic variables having the true distribution function. Therefore, classifiers will be built that minimize the error on the learning set in the hope that this will also minimize the error made on new objects.

Bayesian networks have he ability to offer assistance in a wide range of endeavors. They support the use of probabilistic inference to update and revise belief values. Bayesian networks readily permit qualitative inferences without the computational inefficiencies of traditional joint probability determinations. In doing so, they support complex inference modeling including rational decision making systems, value of information and sensitivity analysis. As such, they are useful for causality analysis and through statistical induction they support a form of automated learning. This learning can involve parametric discovery, network discovery, and causal relationship discovery

5.8 Visualisation

Data visualisation makes it possible for the analyst to gain a deeper, more intuitive understanding of the data and as such can work well alongside data mining. Data mining allows the analyst to focus on certain patterns and trends and explore in-depth using visualisation. On its own data visualisation can be overwhelmed by the volume of data in a database, but in conjunction with data mining can help with the exploration [21], [22], [23].

Visualisation tools take advantage of human perception as an aid to analysis. What numbers cannot show you, corresponding pictures often can. For example, a linear trend in data might not be evident from a table of data. However, a scatterplot that shows a series of points lined up on a straight line provides immediate insight into the linear relationship between the variables. With high power computer graphics, visualisation tools can also be effective presentation tools. Once a discovery is made, the analyst must convey that discovery using an easily accessible language such as pictures.

An inherent danger of visualisation techniques is the fact that it is dependent on the interpretation of the human user. Hence, there is a subjective factor in the quantitative result that is gained from the data mining system. Another limitation is the fact that a computer monitor is a two dimensional display, and human perception is limited to the three dimensional world that we live in. Therefore, visualisation tools have to compress multidimensional data. It is possible that vital information may be lost in the transformation from multiple dimensions to two or three dimensions.

6 SUMMARY

Since data mining poses many challenging research issues, direct applications of methods and techniques developed in related studies in machine learning, statistics, and database systems cannot, on its own, fully solve all of such problems. It is necessary to perform dedicated studies to invent new data mining methods or develop integrated techniques for efficient and effective data mining. In this sense, data mining itself has formed an independent new field.

This paper gives an introduction to the field, and goes on to describe a high level architecture for data mining. The author then groups the models of operation of a data mining system into two categories, namely the verification and discovery models. As data pre-processing is a highly important step in the process, it is examined in closer detail. The structure of the knowledge to be mined, such as association rules and classifications, is one of the major considerations when deciding on a data mining system. The other major consideration is the tools or techniques used to achieve the knowledge.

Some of the issues not covered include the human perspective, or the target user group. From the perspective of the business user, an intuitive interface and descriptive reporting are important. On the other hand, the technical user would be more interested in algorithmic options and a simple model design cycle.

This brief overview of current methodologies and techniques provides a framework to assist in categorising the multitude of systems available.

7 REFERENCES

[1] M.F. Schwartz, A. Emtage, B. Kahle, B.C. Neuman, "A comparison of Internet resource discovery approaches," *Computer Systems*, vol. 5, no. 4, 1992.

[2] S. Lawrence and C.L. Giles, "Searching the world wide web," *Science*, vol. 280, no. 4,:pp. 98-100, 1998.

[3] S. Lawrence, K. Bollacker, C.L. Giles, "Indexing and Retrieval of Scientific Literature," *Eight International Conference on Information and Knowledge Management*, pp 139-146, 1999.

[4] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus., "Knowledge discovery in databases: An overview," *In G. Piatetsky-Shapiro and W. J. Frawley, editors, Knowledge Discovery in Databases, AAAI/MIT Press,* pp. 1-27, 1991.

[5] M.A. King, J.F. Elder IV, B. Gomolka, E. Schmidt, M. Summers, K. Toop, "Evaluation of Fourteen Desktop Data Mining Tools," *IEEE International Conference on Systems, Man, and Cybernetics*, 1998.

[6] J.F. Elder IV, D.W. Abbot, "A comparison of Leading Data Mining Tools," *Fourth International Conference on Knowledge Discovery and Data Mining.* 1998

[7] M.S. Chen, J. Han, P.S. Yu, "Data Mining: An Overview from a Database Perspective", *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 866-883, 1996.

[8] L.K. Maisuria, C.S. Ong, W.K. Lai, "A comparison of artificial neural networks and cluster analysis for typing biometrics authentication," *Proceedings of the International Joint Conference on Neural Networks,* Paper #2072, 1999.

[9] R.O. Duda, P.E. Hart, "Pattern Classification and Scene Analysis", John Wiley & Sons, Inc, 1973.

[10] D.H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Computation*, vol. 8, pp. 1341-1390, 1996.

[11] D.H. Wolpert, "The existence of a priori distinctions between learning algorithms," *Neural Computation*, vol. 8, pp. 1391-1420. 1996.

[12] R. Agrawal, A. Arning, T. Bollinger, M. Mehta, J. Shafer, R. Srikant: "The Quest Data Mining System", *Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining*, 1996

[13] R. Agrawal, T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases," *In Proc. of the ACM SIGMOD Conference on Management of Data,* pp. 207-216, 1993.

[14] B. Lent, R. Agrawal, R. Srikant, "Discovering Trends in Text Databases", Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, 1997.

[15] R. Srikant, R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvements", *Proc. of the Fifth Int'l Conference on Extending Database Technology (EDBT),* 1996.

[16] J.R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81--106, 1986

[17] L. Zadeh, "Fuzzy Sets," Information and Control, vol. 8, pp. 338-353, Academic Press, New York, 1965.

[18] M. Tomassini. A survey of genetic algorithms. In D. Stauffer, editor, Annual Reviews of Computational Physics, volume III, pages 87-118. World Scientific, 1995.

[19] M. Mitchell. An Introduction to Genetic Algorithms. MIT Press, Cambridge, MA, 1996

[20] S. Haykin, "Neural Networks: A Comprehensive Foundation," 2nd edition, Prentice Hall, 1999.

[21] M. Derthick, J. Kolojejchick, and S.F. Roth, "An interactive visualization environment for data exploration," *Proceedings of Knowledge Discovery in Databases*, pp. 2-9, 1997.

[22] D.A. Keim, J.P. Lee, B. Thuraisinghaman, C. Wittenbrink, "Database Issues for Data Visualization: Supporting Interactive Database Exploration"

[23] D. A. Keim, H.P. Kriegel , "Visualization Techniques for Mining Large Databases: A Comparison," 1996

BIOGRAPHY

Ong Cheng Soon obtained his Bachelor of Science (Computer Science and Mathematics) from the University of Sydney in 1998. He obtained his Bachelor of Engineering (Information Systems) from the University of Sydney in 2000. Currently, he is a research associate in the Information Retrieval and Computational Intelligence Research Group, Software Lab at MIMOS Berhad. His research interests include machine learning, data mining and information retrieval.