# DATA MINING: CONCEPTS, APPLICATIONS AND TECHNIQUES

**Wang Guohua and Tay E H Francis**

Department of Mechanical and Production Engineering
National University of Singapore
10 Kent Ridge Crescent
Singapore 119260

## ABSTRACT

Data mining has been widely used in many heterogeneous areas. It has been used to extract valuable knowledge or rules that are hidden behind a vast amount of raw data. Nowadays there is urgent demand to introduce this important technology to business managers and executives. This technology will help them make decisions that will give them a competitive edge over others in the marketplace. Today's operational data represents the current state of a company. When it is combined with historical data, it can tell you where you are and where you are going. The aim of this paper is to introduce the concepts, applications and techniques that are used in data mining. Finally, a case study is provided to illustrate the process and the benefits of data mining.

## 1. OVERVIEW

Each year companies accumulate more and more information in their databases. However, it is not enough for these companies to store their data in a safe place that can be accessed easily. The point is that their databases contain huge treasures of information that had never been seen before in many of the companies' processes. This information can be used to improve the process, to allow the company to predict sales or demand trends of the market, and to react flexibly to them. Moreover, this information decreases in its value as time passes, and the cost of storing this data has already been incurred. Therefore, there is an urgent demand to extract useful knowledge from large amounts of raw data in an efficient and effective way. Unfortunately this information is hidden within the midst of mountains of data that has been stored over a long time. They cannot be retrieved using conventional database management systems. To solve this problem data mining technology comes into the picture.

Data mining is the data-driven extraction of information from large databases. It is the process of automated presentation of patterns, rules or functions to a knowledge user for review and examination[1]. According to Keim[2], data mining can help someone find the right information at the right moment so as to make correct decisions. On the other hand, data mining aims to compress this large amount of data. They will only unearth the so-called "nuggets" from it.

Although statistics can be used to achieve some performance out of data mining, new techniques should be introduced since there are inherent drawbacks in statistics. Hitherto, people have been using statistics to predict linear trends for years. This is a very important activity of data mining[3]. However, trying to forecast complex non-linear or chaotic time series is another matter. For example, if an ARMA method is used to model a predictable model, the choice of order and parameters for such a model is crucial and somewhat difficult to determine. Fortunately artificial neural networks have shown themselves to be excellent tools for modelling complex time series problems. This is the reason why artificial intelligence (AI) techniques are recommended for data mining.

Almost all the large companies have their own databases where much historical data is stored. However, the larger the databases are, the more difficult it is to manage them. According to the report of the International Data Corporation, Figure 1 shows the amount of information that is being stored on mainframe computer systems from 1990 to 1998 (predicted). A tremendous growth can be found from 1995 to 1998 (1998. predicted)[5].

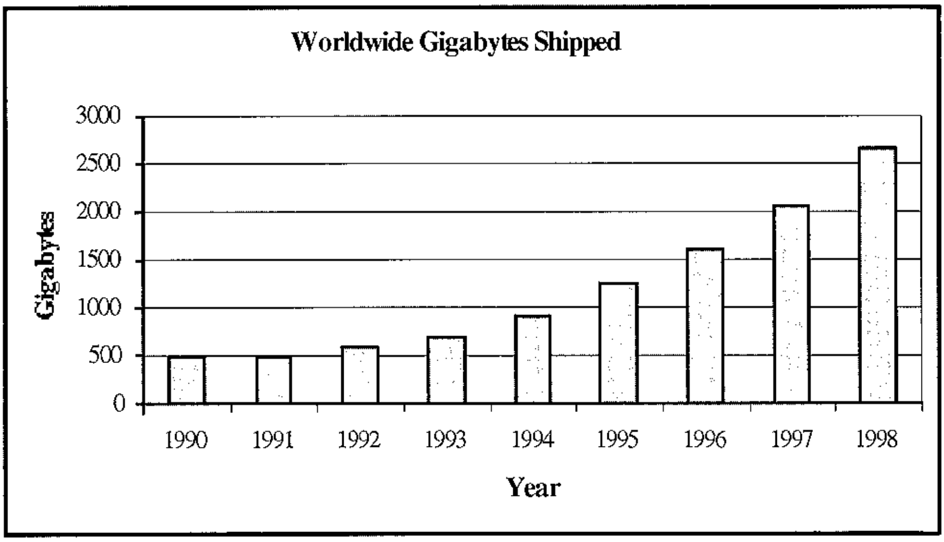

*Figure 1. Growth in mainframe data storage (Source: International Data Corp.)[5].*

Recently data mining has received more and more attention from many large companies, such as IBM, SYBASE, ORACLE. It is worth noting that IBM has developed a software named IBM Neural Network Utility (NNU)[4] that could be used to provide support for data preparation before data mining, and some other types of neural network data mining algorithms[5]. It is easy to perceive here that data mining will draw more and more attention from business managers, executives and decision-makers. This is because they want to obtain unexpected high revenues from their companies. After all, they need to be more competitive than others.

Although data mining can be widely used in many areas, different applications can have similar processes. The first process step is data preparation. You may have retrieved a very large amount of data from a database, however, you do not need all of these retrieved data. What you should do now is to select some of the information that you are most interested in. The next step is to extract hidden rules or knowledge from the selected data using some specific data mining algorithms, which will be introduced later on in this paper. The last step is to analyse the results. In this phrase the participation of domain experts seems to be very important. With the collaboration of these experts and programmers some useful knowledge comes into being, which can be used to enhance decision support. Figure 2 illustrates a typical process of data mining.

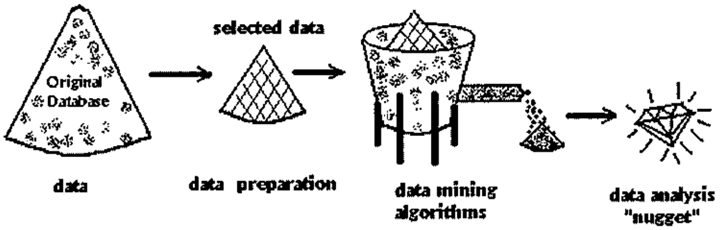

*Figure 2. Data mining process.*

## 2. APPLICATIONS AND PROCEDURES OF DATA MINING

Data mining has been widely used to solve many real-world problems. A few specific and successful applications from different areas are listed down below:

Marketing A number of marketers have the opinion that one of the most important aspects of business is to be able to understand their customers' individual needs. Data mining technology is being used to shed a light on the customers' preference and buying patterns. The goal of data mining is to realise the so-called "target marketing". This means that the results obtained from data mining for this area can be used to determine who the most "typical" customers for a certain set of products are. It is easy to understand that companies can achieve more profits if they are knowledgeable about what a typical customer looks like.

Health A doctor may want to give his diagnosis to a patient based on similar situations that he had come across before. This can be one area that data mining can be applied to. For some other reasons, GTE laboratories (the central research and development facility for GTE Corporation, one of the largest publicly held telecommunications companies in the world) have developed the Health-KEFIR (an advanced data mining system, KEFIR stands for KEy Findings Reporter). The Health-KEFIR can be used to find areas where costs (service fee and operation fee, among others) are most likely to increase in hospitals and among these select where specific actions probably will save the most money.

Science Data mining technology has started to help humans in their scientific research and discoveries. By traversing enormous sets of data, patterns have been found in molecular structures, genetic data, global climatic changes and more.

Finance Data mining has found widespread applications in the finance industry[6,10]. Some data mining techniques have been used to detect patterns in potential fraudulent transactions in consumer credit cards. They are also used to predict interest rate and exchange fluctuations in currency markets. Some others use data mining for credit risk assessments and for bankruptcy prediction in commercial lending and bond rating.

Manufacturing So far, data mining technology has successfully been used in manufacturing areas like production scheduling and planning, production quality control, optimisation of resource allocation, etc. Yong-zai Lu describes such applications of data mining in the control industry in his excellent book[7]. For instance, the goal of a production scheduling and planning system is to minimise the total manufacturing time for a given set of work pieces. At the same time, they are subjected to the constraints of limitation of machine tools and other working forces. If this goal is met, production efficiency will be improved[8].

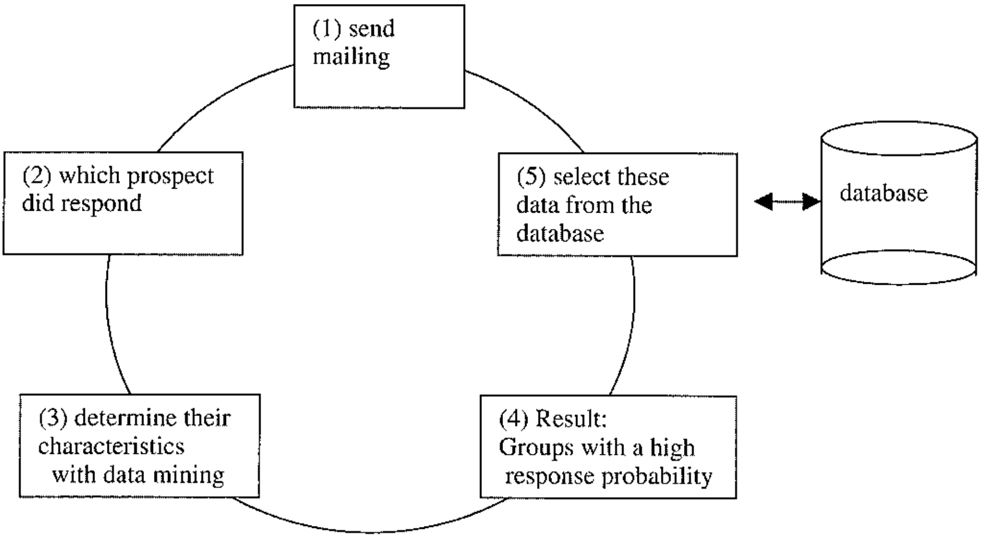
(1) send mailing
(2) which prospect did respond
(3) determine their characteristics with data mining
(4) Result: Groups with a high response probability
(5) select these data from the database
database

*Figure 3. Selecting prospects using data mining.*

To understand the basic concepts and procedures of data mining technology, here is a simple example to illustrate these concepts. Suppose a direct mailing company X want to send mails (1) to a number of prospects (Figure 3). However the response percentage is low, say 2% (2). The response is analysed using data mining techniques (3). Next, try to find out the differences between the customers who response and those who did not response. The result consists of database subgroups that have a significantly higher response probability (4), e.g., for all young couples with double incomes, 24% replied to the last mailing. The groups with the highest response probability are selected as the target for the next mailing (5). Data mining thus increases the response probability.

It is found that data mining can be used to find out the relationship between one's personal characteristics, e.g. age, gender, hometown, and the probability that one will respond to a mailing. Such relationships can be used to select from the mailing database customers with the highest probability of response to a mailing, which can help the company to mail its prospects selectively. Hence they can avoid low response probabilities caused by random mailing. Once this can be avoided, their response probability can be maximised.

## 3. DATA MINING ACTIVITIES

Data mining technology contains several activities, this section tries to introduce some basic and important activities as listed below. But this list is not exhaustive.

- Association
- Classification
- Clustering
- Sequential Patterns
- Similar Time Sequence

Association Some others would like to call association as relationships. Suppose a database of transactions is given, where each transaction consists of a set of items. One can discover all associations where the presence of one set of items in a transaction implies the presence of another set of items. Below is a well-known and interesting association discovered with data mining technology.

"30% of customers who buy diapers also buy beer."

Classification This is a very important activity of data mining. Classification refers to the determination of whether or not an object will fit the predefined profile of a group. After classification has been done, discovering the profile of each group in terms of the attributes of an object is needed. This will help to predict the group of a new object.

A profile of a group may look like this:

"Buyers of expensive sports cars are typically young urban professionals whereas luxury sedans are often bought by elderly wealthy persons."

Clustering Clustering refers to the partition of *n* objects into *m* categories based on the intrinsic common features they possess. This is an effective way of compressing large amounts of raw data. Clustering can also extract valuable information or knowledge that are needed from these raw data. For example, customers for a set of products can be clustered into several categories according to their personal characteristics, the products they bought, the amount of money they spent and some other similar information. By analysing the different features among groups, companies can discover the buying characteristics of each group and hence achieve the goal of "target marketing". They can then increase their revenues.

Here is a rule obtained from clustering:

"Customers in category one are elderly and spend twice on appliances compared with customers in category two who are younger."

Sequential Patterns Sequential patterns refers to the inter-transaction patterns in a database of transactions over a period of time, which can show that the presence of one set of items can be followed by another set of items. For example:

"10% of people with diabetes develop a treatable loss of eyesight."

Similar Time Sequence This refers to the search for sequences similar to a given one, or for all occurrences of similar sequences in a given a database of time sequences. However, it is dependent on criteria given in advance, which is then used to measure the similarities between two sequences. For example, a retailer wants to optimise purchasing and store-keeping. By applying the Similar Time Sequence technique, the retailer can find groups of products that have similar forecast seasonal sales for the next year. He then uses this information for combining sales and inventory replenishment.

## 4. PRINCIPAL TECHNIQUES FOR DATA MINING

So far, many techniques have been used to perform the common data mining activities such as associations, clustering, classification, modelling, sequential patterns and similar time sequence analysis. These techniques range from statistics to rough sets, and to neural networks. It can be said that data mining is an elegant combination of statistics, neural networks, fuzzy logic, evolutionary computation and machine learning techniques. These cover almost all artificial intelligence algorithms.

Table 1.1 lists the main techniques used in data mining and their applications for some specific activities of data mining technology[5].

Table 1.1. Data mining techniques and their applications

| Data Mining Function | Algorithms | Application Examples |
|---|---|---|
| Association<br>Classification | Statistics, Set theory<br>Decision trees, Neural networks, Genetic algorithms | Market basket analysis<br>Target marketing, Quality control, Risk assessment |
| Clustering | Neural networks, Statistics, Genetic algorithms | Market segmentation, Design reuse |
| Modelling | Linear and nonlinear regression, Curve fitting, Neural networks | Ranking/Scoring customers, Pricing models, Process control |
| Time-series Forecasting | ARMA (auto-regressive moving average) models, Neural networks | Sales forecasting, Interest rate prediction, Inventory control |
| Sequential Patterns | Statistics, Set theory | Market basket, Analysis over time |

## 5. A CASE STUDY

This section describes the detailed process of a data mining application. Thus one can understand the concepts of data mining clearly and also find out what data mining can do for us. This problem is the Market Segmentation (Target marketing) mentioned above.

### 5.1 Problem Description

An important challenge in business is to understand its customers. For example, for a given set of products, the businessman must find out who the customers most interested in them are, and what a "typical" customer might look like for a certain product. This is a crucial issue for any company.

In order to find some purchasing features for a certain category, all the customers need to be clustered into several categories based on some information of these customers. By analysing the makeup of each category, the purchasing features for each category could be found. Table 1.2 lists the available and selected information for a customer.

Table 1.2. Selected data for customer clustering

| Attribute | Logical Data Type | Values Range | Representation |
|---|---|---|---|
| Age | Continuous Numeric | 18-74 | Scaled: 0.0~1.0 |
| Sex | Categorical | Male, Female, Unknown | 1.0, 0.0, 0.5 |
| Marital Status | Categorical | Single, Married, Unknown | 1.0, 0.0, 0.5 |
| Homeowner | Categorical | Yes, No, Unknown | 1.0, 0.0, 0.5 |
| Sporting Goods | Continuous Numeric(CN)($) | 0~1500 | Scaled: 0.0~1.0 |
| Exercise Equipment | (CN)($) | 0~2500 | Scaled: 0.0~1.0 |
| Home Appliance | (CN)($) | 0~5000 | Scaled: 0.0~1.0 |
| Entertainment | (CN)($) | 0~2500 | Scaled: 0.0~1.0 |
| Furniture | (CN)($) | 0~5000 | Scaled: 0.0~1.0 |
| Total | (CN)($) | 0~15000 | Scaled: 0.0~1.0 |

Analysis of the makeup of the overall customer set is helpful for understanding the results of segmentation. For example, on average, the customer is 42 years old and has a yearly income of $35 000. On the whole, these customers spend $500 on sporting goods, $1 000 on exercise equipment and on other similar goods with similar prices. While these averages are interesting, they are a conglomeration of people, and do not represent any single customer. More valuable knowledge from the above information (Table 1.2) is wanted. One will find that customer segmentation with analysis can do it just as well.

Now go to the results of segmentation. In this case all the customers are clustered into 4 categories. Figure 4 shows the proportion of each category and their average ages. Figure 5 is a typical report for market segmentation, from which one can find the purchasing properties of each category. For example, for the largest category, the average age is 43. This age group makes up to 42.8% of the total number customers. This has a very similar makeup to the customer's set average. However, one big difference is that this category spends almost twice as much as on average on home appliances. The next largest category, 24.9%, is older (52). This category spends almost half the average on sports and exercise equipment, but spends about $500 or more than the average on appliances and entertainment. The knowledge obtained from clustering may look like these mentioned above. There is no doubt that it will give invaluable instruction to the sales managers.

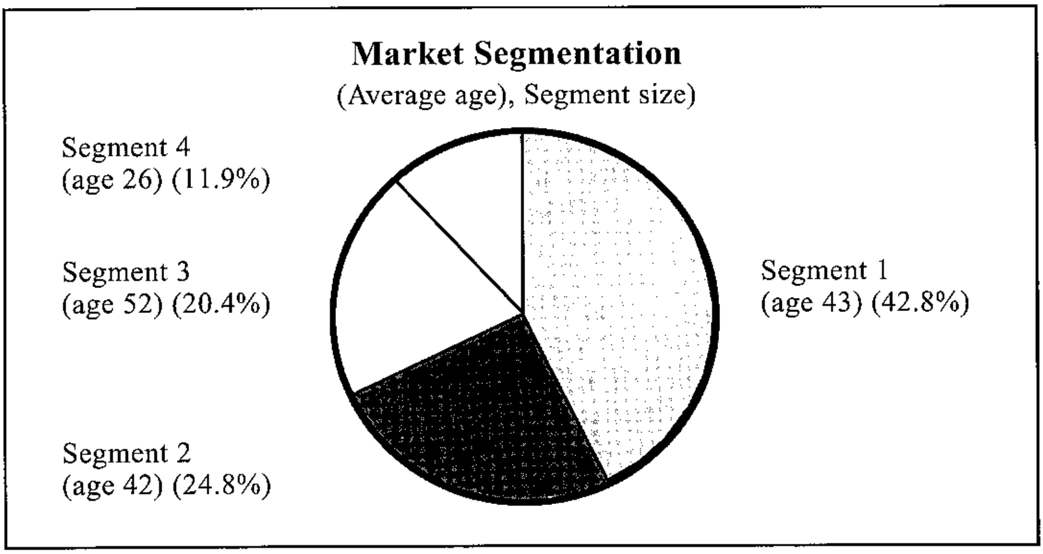

*Figure 4. Result of market segmentation.*

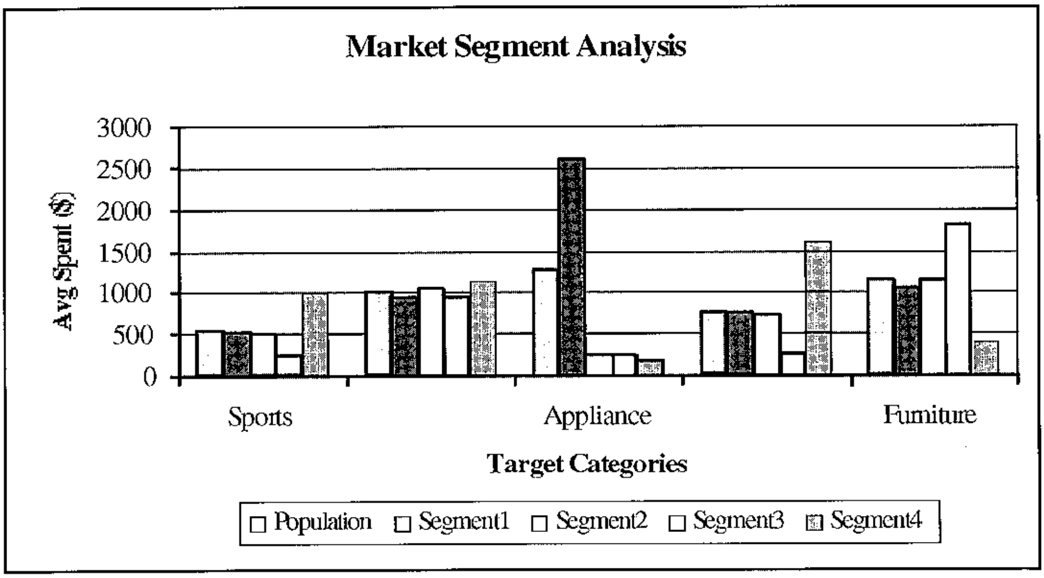

*Figure 5. Report of market segment analysis.*

## 5.2 Problem Solution

The basic steps to solve such a target market problem are given as follows. The first step is to extract the needed data from the customer database, that is only the 10 items out of those listed in Table 1.2 is necessary. The next step is pre-processing selected data to meet the demand of the selected data mining technique. For this problem a suitable algorithm is an artificial neural network. This is why the selected data are scaled in the range 0.0~1.0, which is the prerequisite for the training data set of an artificial neural network. (A very suitable algorithm for this clustering problem is Kohonen's self-organisation learning model)[9]. Following this will be the data mining process to cluster customers into several categories based on their commonly implicit information. Finally the makeup of each category is analysed and some useful profiles are extracted from it. These profiles can then be used further to help make right decisions for market makers.

## 6. CONCLUSION

Data mining is hot around the world. Many big companies and researchers have recognised its importance. Furthermore, the development of hardware and software has paved the way for data mining technology to handle large databases. One can perceive that nowadays data mining will be used to tackle more heterogeneous problems in many different domains. The following problems in business industries and manufacturing enterprises deserve our particular attention; market segmentation, time sequence analysis, sales or inventory supplement prediction, production quality control, and job shop scheduling. One can find from this paper that successful application of data mining technology will lead to efficient management of databases and allow companies to achieve higher revenues.

## 7. REFERENCES

1. Samson Tai, *Data Mining: The software that find pattern never seen before*, Proceedings of the 8th international database workshop (Industrial Volume), Hongkong, 1997, pp103-109.
2. Daniel Keim, *Visual Support for Query Specification and Data Mining*, Verlag Shaker, Aachen, 1995, pp1-20.
3. McClave, J.T and P.G. Benson, *Statistics for Business and Economics*, second edition, San Francisco: Dellen, 1982.
4. IBM corporation, 1994a, *Neural Network Utility*: User Guide, SC 41-0223.
5. Joseph P. Bigus, *Data Mining with Neural Networks*, Macgraw-Hill, USA, 1996.
6. Disney, D. R., *Comment: For the real gold in customer data, dig deep*, The American Banker, May 10 1995.
7. Yong-zai, Lu, *Industrial Intelligent Control: Fundamentals and applications*, John Wiley and Sons, USA, 1996.
8. S. C., Lin, Goodman, E. D., etc. *A Genetic Algorithm Approach to Dynamic Job Shop Scheduling Problems*, Proc. Intl. Conf. On Genetic Algorithms and Their Applications, Morgan Kaufmann Publishers, San Francisco, July 1997, pp 481-488.
9. Kohonen T., *Self-Organisation and Association Memory*, Third Edition, Springer-Verlag, 1989.
10. Michael J.A. Berry & Gordon Linoff, Data Mining Techniques for Marketing, Sales and Customer Support, JohnWiley & Sons, USA, 1997.