Chapter 8 Conclusion

8.1 Summary

Data mining is a very powerful and useful methodology and technology for generating information for decision making. Future developments are expected to make data mining even more powerful and useful. Despite this, data mining is not without limitations. Before highlighting the limitations of data mining and discussing some future directions, a summary of the content of the earlier chapters of the book is first presented below.

8.1.1 Overview of Data Mining

Data mining can be considered a relatively recent methodology and technology, coming into prominence only in 1994 (Trybula, 1997). It uses techniques from the disciplines of statistics and mathematics, machine learning and artificial intelligence.

Data mining can be described by the following characteristics: (1) it is a process; (2) it is usually applied to large data sets; (3) it focuses on the exploration and discovery of previously unknown patterns, trends and relationships; and (4) it helps organisations and managers make better decisions. Combining all the above, data mining can be defined as the process of analysing mostly large data sets to explore and discover previously unknown patterns, trends and relationships to generate information for better decision making.

Conclusion

The increasing popularity and application of data mining can be explained by: (1) advances in both computer hardware and software that have made many data mining applications more accessible and affordable to businesses now than ever before; (2) challenging business problems such as the detection of fraud and the increasingly competitive business environment that have led organisations to the search for more powerful analytical tools.; (3) the data explosion experienced by many organisations collecting increasingly larger amounts of data that has led organisations to realise that data are not useful for decision making unless they can be transformed into information; and (4) the success stories of data mining applications and aggressive marketing by data mining consultants and software vendors that have resulted in increasing numbers of organisations wanting to explore or use data mining.

As a process, the data mining methodology comprises three major stages. In the pre-modelling stage, the steps involved are: (1) identification of the business problem; (2) translation of the business problem into a data mining application; (3) assessment of the data needed and available for the data mining application; and (4) preparation of the data for mining.

The modelling stage can be deemed to be the core of data mining. This is the stage where data analysis is performed. The following steps are involved in the modelling stage: (1) identification of the appropriate tools or techniques to use in a particular data mining application; (2) performance of the actual data analysis (or modelling); (3) assessment of the results; and (4) identification of the final model for deployment.

Finally, the post-modelling stage relates to the actions to be taken after the data analysis is completed. It comprises the following two steps: (1) deployment of the data mining model or results; and (2) tracking of the performance of the data mining application.

The data mining methodology is interactive and iterative. It is often necessary to go back and forth between the different stages as well as steps.

8.1.2 Data Mining Tools and Issues

Data mining tools can be broadly classified based on what they can do, namely: (1) description and visualisation; (2) association and clustering; and (3) classification and estimation (i.e., predictive modelling). A summary is given in Table 8.1.

Description and visualisation contribute towards understanding a data set and detecting hidden patterns, trends and relationships in the data. As such, they are frequently performed before modelling. Description (e.g., descriptive statistics and graphs) and visualisation (e.g., rotating plots and web graphs) also help greatly in the summarisation of data and in the presentation and reporting of results.

Association analysis (e.g., market basket analysis) looks for groupings or patterns among a set of items. The resulting association rules are intuitive and easy to understand and can have predictive value. Hence, association rules are frequently used for making decisions in areas such as store layout, items bundling, and discount and promotion campaigns. Association analysis can be extended to include a target variable, customer characteristics, items that are not purchased (instead of purchased) and time.

Association analysis is most useful in the retail industry. However, it can also be used in industries such as banking and finance as well as healthcare. Association analysis can be performed using the Apriori and generalised rule induction (GRI) algorithms.

Clustering aims to group similar (homogeneous) objects into the same cluster and dissimilar (heterogeneous) objects into different clusters. It is usually used for market segmentation and to identify the cluster profile of the different segments. In some instances, cluster membership is used as an input variable to predict a target variable of interest (e.g., which cluster of customers is more likely to purchase a particular product).

Table 8.1 Summary of Data Mining Tools

Functions	Objectives	Illustrative Tools
Description and Visualisation	1. To understand data. 2. To detect patterns.	1. Descriptive statistics (e.g., measures of central tendency and pie charts). 2. Visualisation (e.g., rotating plots and web graphs).
Association and Clustering	1. To detect groupings among items. 2. To group similar objects in the same cluster and different objects into different clusters.	1. Apriori algorithm. 2. Generalised rule induction. 3. Sequence detection algorithm. 4. Statistical clustering methods (e.g., K-means and TwoStep clustering). 5. Other clustering methods (e.g., Kohonen nets).
Predictive Modelling	1. To classify objects into one of several categories. 2. To estimate a metric variable.	1. Statistical models (e.g., multiple and logistic regression). 2. Artificial intelligence models (e.g., neural networks). 3. Machine learning techniques (e.g., decision trees).
Illustrative Applications: 1) Visualisation: data exploration. 2) Association: market basket analysis and web mining. 3) Clustering: market segmentation and fraud (i.e., outlier) detection. 4) Predictive modelling: churn modelling and customer value estimation.		

Besides market segmentation, clustering can also be used in other contexts (e.g., the identification of abnormal observations and the use of cluster membership to identify the incidence of an event of interest such as fraud). Clustering can be performed using statistical (e.g., K-means clustering) or other methods (e.g., Kohonen nets).

Predictive modelling is probably the most common and important analysis in data mining applications. It includes both classification and estimation, depending on whether the prediction is for a non-metric or metric target variable, respectively. Classification is commonly used for applications such as cross-/up-selling (to predict if a prospect is likely to purchase a product/service), churn modelling (to predict if an existing customer is likely to turnover), credit scoring (to predict if a prospect or existing customer is likely to default on payments) and fraud detection/control (to predict if a transaction is likely to be fraudulent).

Besides giving a predicted classification, predictive modelling can also generate a probability or confidence index that is associated with the classification. For example, in addition to predicting a particular loan applicant as a low-risk borrower, the model may predict the probability of payment default as, say, 0.10. This information may have implications on the amount of loan to be given, the interest rate to be charged, the repayment period to be applied, the collaterals to be secured … etc.

Estimation can be used in data mining applications to predict any metric target variable, including customer lifetime value, amount of expenditure on a particular product/service, duration of insurance policies and so on. The tools most commonly used for predictive modelling are regression (a traditional statistical method), neural networks (an artificial intelligence model) and decision trees (a machine learning technique). Generally, predictive modelling attempts to predict a target variable on the basis of one or more input variables.

In the context of predictive modelling, there are a few important data mining issues that should be considered. The first is the assessment (or validation) of the adequacy of prediction models. In model validation, it is important that a validation sample is used instead of the construction sample so that bias in the validation results is minimised. The best approach is to have separate construction and validation data sets. If this is not feasible, the n-fold validation method is a good alternative. Besides assessing the adequacy of prediction models in use, model validation can also serve as a means to reduce the potential over-fitting of a model to the data in neural networks and decision trees.

For the prediction of a non-metric target variable, another important issue is the determination of the optimal cut-off point of the model. There are two important factors that affect the optimal cut-off point, namely, the prior probabilities of the event and non-event occurring and the relative misclassification costs. The latter refers to the cost of misclassifying a particular category of the target variable as another category. Both prior probabilities and misclassification costs should be considered in determining and adjusting the optimal cut-off point in order to reduce total misclassifications or the total cost of applying the prediction model.

Another important data mining issue is the assessment and comparison of different acceptable prediction models. For this purpose, evaluation charts (besides accuracy rates) can be used. Non-financial evaluation charts include response charts, gains charts and lift charts and financial evaluation charts include profit charts and ROI charts. These charts are based on the concept of hit rates (which is different from the concept of accuracy rates) and can be cumulative or non-cumulative.

Besides comparing the evaluation charts of different prediction models to assess relative effectiveness, these charts can also be compared to benchmarks such as the baseline model and the exact model to assess the acceptability of the models. For financial assessments, break-even points can also be computed to assess the risks of business actions and sensitivity analysis can be performed to assess the impact of changes in selected variables of interest on the financial outcome. Sensitivity analysis results can also indicate risk.

8.1.3 Lessons from the Case Studies

Chapters 5 to 7 present case studies illustrating the potential applications of data mining for a retailer, a service provider and a manufacturer, respectively. These case studies can serve as simple templates to help organisations get started in data mining. They are generally driven by the need to respond to increasingly intense competition, the desire to transform ever-growing data into information, and the hope to gain a competitive edge by harnessing the benefit and usefulness of data mining.

Many of the data mining applications illustrated in the case studies revolve around customer analytics and focus on "getting to know your customers". In particular, data mining can be very useful for customer relationship management (e.g., customer valuation, cross-/up-selling and churn modelling) and market segmentation (e.g., customer profiling and response-based segmentation). Such data mining applications are well documented in the literature (see, for example, Peterson, 2003; and Bloemer et al., 2002).

For the retailer in Chapter 5, three data mining applications are developed. Firstly, for a baby products promotion, web graphs and association analysis are applied on transactional data to bundle items together to increase the amount spent by customers. Secondly, to uncover buying patterns of existing customers so that tailored services can be provided to selected customer segments, customer segmentation using clustering is performed using transaction and demographic variables. In this application, data transformation is done to improve the clustering solution.

Cluster profile and profitability are also derived and their relationship examined. Thirdly, churn modelling is attempted based on demographic variables and prior transaction summary measures. The objective is to predict (as well as profile) customers who are likely to churn so that pre-emptive and pro-active actions can be taken to reinforce their loyalty before they are lured away by competitors. The distinction between model construction and model deployment is made and the need for latency month(s) explained. The need for data partitioning and model validation and derived variables is also emphasised. This illustration covers the major issues in predictive modelling, including the assessment and evaluation of competing models.

For the service provider in Chapter 6, four data mining applications are developed. In the first application, description and visualisation are used to identify and profile dissatisfied customers to facilitate improving products and services to meet customers' expectations. In the second application, data mining is used to help the service provider (in this case, a travel agency) to develop new products (i.e., optional tour packages) using association analysis. In the third application, the most profitable customer segments are identified and profiled with a view to reduce churn and improve profitability. Demographic as well as transaction variables are used as clustering criteria. Statistical analyses are also performed in this application to further analyse segment profitability. Finally, in the fourth application, data mining is used to do target mailing by predicting the probability of response/non-response to a mailing campaign. The model results are expected to increase the response rate and at the same time reduce the cost of mailing campaigns. With the neural network model as the champion model to be deployed, a decision tree is constructed on the model predictions to reveal the relationships captured by the neural network model. The customers who are likely to respond to mailing campaigns are also profiled.

In contrast to Chapters 5 and 6, Chapter 7 focuses on a manufacturing environment and looks at potential data mining applications for a manufacturer. Data explosion has also occurred in manufacturing because of the automation and computerisation of many industrial and manufacturing processes. These data can be a great asset to an organisation if they can be transformed into useful information. Data mining applications in manufacturing are well documented in the literature (see, for example, Gertosio and Dussauchoy, 2004; and Hou et al., 2003).

Three data mining applications are illustrated in Chapter 7. The first application involves profiling "bad-fit" and "good-fit" workers to facilitate employing the right kind of workers for the right factory jobs. Visualisation and association analysis are used for this purpose. Such applications are not limited to just the manufacturing industry. The objective of the second data mining application is to uncover the relationships between the strain rate of rubber (one of the raw materials used in the manufacturing process) and storage conditions. The uncovered relationships can help determine the optimal conditions for storage. A two-stage analysis is performed. In the first stage, segmentation is done to group the observations into several segments based on the storage conditions. In the second stage, visualisation is used to examine the relationships between the strain rate and the storage conditions (i.e., cluster profile) within each cluster. Finally, in the last data mining application, the objective is to understand the relationship between manufacturing parameters and output quality. The findings can help detect faults as well as facilitate the formulation of operating guidelines for machine operators on the specific settings of process parameters to correct detected quality problems. Decision trees are constructed in this application to generate operating guidelines. This application also incorporates differential relative misclassification costs and creates derived variables.

The potential data mining applications illustrated in Chapters 5 to 7 are by no means exhaustive. Instead, they aim to link the earlier discussion on data mining to specific applications. There is no doubt that data mining is a very powerful methodology and technology that can be applied in many different contexts and functions (e.g., fraud detection, risk management and customer relationship management) and in many different organisations (including non-profit and non-commercial organisations). With some imagination and creativity (in addition to the pre-requisites of data mining to be discussed in the next section), data mining can go a long way towards enhancing the competitive advantage of organisations.

8.2 Limitations of Data Mining

In developing data mining applications, it is important to be mindful of the limitations of data mining. Some of the common ones are discussed below.

Firstly, the quality of data mining results and applications depends on the availability and quality of data (Chopoorian et al., 2001). For example, to construct a credit scoring model, sufficient numbers of “good” (i.e., low risk) and “bad” (i.e., high risk) cases have to be available. For rare occurrences (e.g., fraud), data availability can be a problem. Also, the data needed for data mining often exist in different settings and systems (e.g., in the administration database and transactions database). Hence, they have to be not only collected but also integrated before data mining can be done. In addition, for the available data, problems such as missing data, corrupted data, inconsistent data … etc. have to be resolved. It has been estimated that data preparation comprises about 75% of the resources needed for a data mining project.

Secondly, a sufficiently exhaustive mining of data may throw up patterns of some kind that are a product of random fluctuations (Hand, 1998). This is especially so for large data sets with many variables. Hence, many interesting and/or significant patterns, trends and relationships found in data mining may not be useful. Further, from a statistical perspective, while data mining is well developed for modelling (i.e., prediction), it is not as well developed for effect assessment. Murray (1997) and Hand (1998) have warned against using data mining for data dredging or fishing (i.e., randomly trawling through data in the hope of identifying patterns) because of the statistical problems involved.

Thirdly, successful application of data mining requires the user to be knowledgeable in the domain area of application as well as in data mining methodology and tools. Domain area knowledge is important as it is critical to identifying the appropriate business issues for which to develop data mining applications. Also, it is essential for specifying the appropriate models and correctly interpreting the results. In addition, without a sufficient knowledge of data mining methodology and tools, an organisation or user may not be aware of or be able to avoid the pitfalls of data mining (see, for example, McQueen and Thorley, 1999). To do the actual data mining, knowledge of data mining software is necessary too.

IT (i.e., information technology) expertise is also necessary for tasks such as the extraction and preparation of data for mining and the deployment of data mining models (e.g., to embed the models into operating systems). Further, statistical and research expertise is essential because a data mining project can be considered a large-scale research project that involves a lot of statistical and research issues. Therefore, collectively, the data mining team should possess the following knowledge and skills: domain knowledge, data mining knowledge and skills, IT expertise and statistical and research expertise.

Finally, organisations developing data mining applications need to make a substantial investment of their resources (e.g., time, effort and money) in data mining, which can be a complex and time-consuming task. Data mining projects can fail for a variety of reasons, including the lack of management support and organisational commitment, unrealistic user expectations, poor project management, inadequate data mining expertise … etc. Data mining requires intensive planning and technological preparation work. In addition, top management has to be convinced of the usefulness of data mining and be willing to change work processes, if necessary. Further, all parties involved in the data mining effort have to collaborate and cooperate (Gillespie, 2002). In particular, given the collective knowledge and skills needed, different departments in an organisation have to work closely together. Any reluctance to cooperate among departments can be a severe obstacle to data mining success (Luxton, 2002).

8.3 Future Directions

Given the usefulness of data mining, there is no doubt that it is going to be increasingly more important to organisations, both commercial and not-commercial ones. It may be useful to anticipate some of the future directions that data mining and its related areas may take. These can be classified under the major categories of applications, data, software and hardware.

8.3.1 Applications

At the infancy stage of the data mining era in the 1990s, applications are mainly generic and focussed primarily on helping organisations gain a competitive advantage over their competitors (e.g., churn modelling). Since the turn of the century, however, specialised data mining applications have increasingly been developed for industries such as biomedicine, finance and telecommunications. Also, specialised data mining applications have been developed for major functions such as customer relationship management, e-commerce and DNA assay. The trend of developing specialised data mining packages is expected to continue more aggressively in the future. In this respect, Fayyad and Uthurusamy (2002) have commented that future trends in data mining will emphasise highly specialised applications and solutions more than new data mining tools.

Overall, data mining applications are expected to increase in popularity and usage in the future. This expectation is consistent with the results of several surveys of data mining and of the most important technologies and skills for the future (see, for example, Baker and Baker, 1998; Anonymous, 1999; Deck, 1999; Sheier, 2001; Anonymous, 2002; West, 2002; and Anonymous, 2004).

8.3.2 Data

This book focuses on data mining, where data is taken to be numerical data (e.g., customers' demographic characteristics and transaction data). However, the data explosion in recent years has yielded not only numerical data (e.g., banking transactions) but also unstructured text data within organizations (e.g., documents on intranet, e-mail messages and customer comments). External sources (such as news and press articles, journals, government reports and regulatory filings) also provide ready sources of text data. Cody et al. (2002) have suggested that a critical component for organisational success is the ability to take advantage of all available data, including text data. Hence, one important future direction of data mining is the incorporation of text mining. The latter is capable of automatically identifying and indexing unstructured data concepts and providing graphical interface that supports a high-level view of the data as well as the capability to drill down to a very detailed level (Biggs, 2000).

Text mining has gained increasing popularity and usage in the last few years. Important advances have been made in this area. To move further ahead, Cody et al. (2002) have suggested the following areas for future research and development: (1) extracting quantitative facts from documents [e.g., the financial terms of a contract]; (2) deducing relationships between entities in documents [e.g., new product A competes with product B]; (3) measuring the level of subjective values such as severity or sentiment in documents [e.g., in customers' comments or complaints]; (4) integrating semantics to improve taxonomies and reasoning about data and text; and (5) integrating data mining and text mining. The need to develop new tools to incorporate ad hoc data as seamlessly and effectively as possible has also been emphasised by Kanzler (2002).

Finally, it is noted that data can also go beyond numerical and text data. One important type of data is images. The mining of images is quite established in some fields such as medical science (e.g., in the detection and interpretation of patterns in MRI scans) and biometrics/security (e.g., in the identification of faces as well as objects posing threats). Another important type of data is web/internet data. Web mining is fast becoming a standard and established form of data mining (Han and Kamber, 2001).

The power of data mining will be greatly enhanced if the mining of all types of data (numerical, text, web/internet and image) can be integrated.

8.3.3 Software

The future will see rapid and advanced developments in data mining software. Firstly, more as well as increasingly sophisticated data mining algorithms will be developed. These algorithms will, in turn, enable more advanced and new analysis of data. Such on-going efforts are well documented in the technical data mining literature (see, for example, Sorensen and Janssens, 2003; and Lingras and West, 2004). They will greatly enhance the power of data mining in discovering hidden patterns, trends and relationships.

Secondly, data mining software is expected to become more powerful, more user-friendly and more interactive in the future. Data mining is also expected to be more integrated with an organisation's information technology environment. These developments will make data mining more accessible to organisations and end-users. Also, the power of data mining will increasingly be in the hands of end-users (e.g., marketing personnel).

However, the benefits of this new and increased power will accrue to end-users and organisations only if they are in a position to use data mining appropriately and effectively.

Thirdly, data mining will be more multidimensional in terms of the type of data that it can analyse (see the previous section) as well as the channels used to aid knowledge discovery. Currently, a great part of knowledge discovery is visual. That is, visualisation is used frequently in data mining to discover data patterns, trends and relationships (e.g., multidimensional rotating plots or web graphs) and to summarise data mining results (e.g., graphical presentations of cluster profiles or decision trees). There have been on-going efforts to use audio signals to indicate data patterns or features of data mining results (Han and Kamber, 2001).

8.3.4 Hardware

Data are a critical component of data mining as the latter is not possible without the former. The advent of data warehousing has contributed greatly to data mining by storing, preparing and supplying data that can be mined. The future is likely to see future developments in hardware (and software) for data storage, extraction, manipulation, loading … etc. Also, data hardware of the future needs to be able to handle all types of data (e.g., numerical, text and image) as well as integrate with different information systems in an organisation.

Further, as data mining involves very intensive computations, it will benefit greatly from more powerful computers. This will enable more powerful data mining algorithms, approaches and software to be developed and used and faster turnaround time in getting data mining results.

8.4 Some Final Comments

As mentioned in the Preface, this book is written with a view to provide: (1) decision makers with an intuitive and practical introduction to data mining and its applications; (2) small and medium enterprises (or SMEs) with some fundamentals and case studies so that they can harness the benefits of data mining; and (3) students who will benefit from learning data mining (e.g., students in marketing or IT) with a book or guide.

The objectives listed above would have been achieved if readers can gain a good understanding of data mining and be able to develop data mining applications after reading the book. There is no doubt that data mining can be a very powerful technology and methodology for generating information from raw data to address business and other problems. This usefulness, however, will not be realised unless knowledge of data mining is put to good use.

Chapter References

Anonymous. (1999). "Data mining", *Chain Store Age*, October, p. 42.

Anonymous. (2002). "Data mining digs deep to improve on quality", *Professional Engineering*, Vol. 15 No. 11, p. 45.

Anonymous. (2004). "'Top 10 technologies' confirms interest in information security, spam control", *The CPA Journal*, Vol. 74 No. 4, p. 15.

Baker, S. and Baker, K. (1998). "Mine over matter", *The Journal of Business Strategy* , Vol. 19 No. 4, pp. 22-26.

Biggs, M. (2000). "Resurgent text-mining technology can greatly increase your firm's 'intelligence' factor", *InfoWorld*, Vol. 22 No. 2, p. 52.

Bloemer, J., Brijs, T., Swinnen, G. and Vanhoof, K. (2002), "Identifying latently dissatisfied customers and measures for dissatisfaction management", *The International Journal of Bank Marketing*, Vol. 20 No. 1, pp. 27-27.

Chopoorian, J. A., Witherell, R., Khalil, O. E. M. and Ahmed, M. (2001). "Mind your own business by mining your data", *SAM Advanced Management Journal*, Vol. 66 No. 2, pp. 45-51.

Cody, W. F., Kreulen, J. T., Krishna, V. and Spangler, W. S. (2002). "The integration of business intelligence and knowledge management", *IBM Systems Journal*, Vol. 41 No. 4, pp. 697-713.

Deck, S. (1999). "Data mining", *Computerworld*, Vol. 33 No. 13, p. 76.

Fayyad, U. and Uthurusamy, R. (2002). "Evolving data mining into solutions for insight", *Association for Computing Machinery*, Vol. 45 No. 8, pp. 28-31.

Gertosio, C. and Dussauchoy, A. (2004), "Knowledge discovery from industrial databases", *Journal of Intelligent Manufacturing*, Vol. 15 No. 1, pp. 29-37.

Gillespie, G., (2000). "There's gold in them thar' databases", *Health Data Management*, Vol. 8 No. 11, pp. 40-52.

Han, J. and Kamber, M. (2001), *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, Academic Press, California.

Hand, D. J. (1998). "Data mining: statistics and more?". *The American Statistician*, Vol. 52 No. 2, pp. 112-118.

Hou, T. H., Liu, W. L. and Lin, L. (2003), "Intelligent remote monitoring and diagnosis of manufacturing processes using an integrated approach of neural networks and rough sets", *Journal of Intelligent Manufacturing*, Vol. 14 No. 2, pp. 239-253.

Kanzler, J. (2002). "Business intelligence integration", *Computer Technology Review*, Vol. 22 No. 12, p. 22.

Lingras, P. and West, C. (2004), "Interval set clustering of web users with rough k-means", *Journal of Intelligent Information Systems*, Vol. 23, No. 1, pp. 5-16.

Luxton, R. (2002), “Marketing campaign systems – the secret to life-long customer loyalty”, *Journal of Database Marketing*, Vol. 9, No. 3, pp. 248-258.

McQueen, G. and Thorley, S. (1999). “Mining fool’s gold”. *Financial Analysts Journal*, Vol. 55 No. 2, pp. 61-72.

Murray, L. R. (1997). “Lies, damned lies and more statistics: the neglected issue of multiplicity in accounting research”. *Accounting and Business Research*, Vol. 27 No. 3, pp. 243-258.

Peterson, K. (2003), “Mining the data at hand”, *Chain Store Age*, Vol. 79 No. 6, p. 36.

Scheier, R. L. (2001), “Finding pearls in an ocean of data”, *Computerworld*, Vol. 35, No. 30, pp. 48-49.

Sorensen, K. and Janssens, G. K. (2003), “Data mining with genetic algorithms on binary trees”, *European Journal of Operational Research*, Vol. 151, No. 2, pp. 253-264.

Trybula, W. J. (1997), “Data mining and knowledge discovery”, *Annual Review of Information Science and Technology*, Vol. 32, pp. 197-229.

West, L. (2002), “Good news for CRM and outsourcing vendors: despite the economic slump, banks in Europe are still prepared to spend on certain areas of IT”, *Bank Technology News*, Vol. 15 No. 2, pp. 22-23.