Chapter 1 Overview of Data Mining

1.1 Background

The term “data mining” is not new in that it has been used for a long time to denote the idea of unscientific “fishing” or “dredging” of data in data analysis. That is, if an analyst is searching for a particular conclusion, then there is a good chance that this conclusion can be “found” by repeatedly analysing the data in various ways, including inappropriate ways. For a long time, the term “data mining” has had a negative connotation.

“Data mining” as used today, however, refers to an entirely different concept from that of unscientific data fishing or dredging. The new concept of data mining can be considered a recently developed methodology and technology, coming into prominence only in 1994 (Trybula, 1997). It uses techniques from the disciplines of statistics/mathematics, machine learning and artificial intelligence. It aims to identify valid, novel, potentially useful and understandable correlations and patterns in data (Chung and Gray, 1999) by combing through copious data sets to uncover patterns and relationships that are too subtle or complex for humans to detect (Kreuze, 2001).

Since the mid-1990’s, data mining has been used intensively and extensively by financial institutions (e.g., for credit scoring and fraud detection), marketers (e.g., for direct marketing and cross-/up-selling), retailers (e.g., for market segmentation and store layout) and manufacturers (e.g., for quality control and maintenance scheduling), among others.

The increasing popularity and application of data mining can be explained by a few important developments in the last fifteen years. Firstly, advances in both computer hardware and software have made many data mining applications more accessible and affordable to businesses now than ever before. These include cheaper and more powerful computers and more user-friendly, comprehensive and advanced data mining software.

Secondly, challenging business problems (such as the detection of fraud) and the increasingly competitive business environment have led organisations to search for more powerful analytical tools. In these areas, data mining is able to make and has made significant contributions. The regulation, liberalisation and competitiveness of the financial sector have often been cited as the driver of data mining applications among financial institutions (Berger, 1999). Some of the most sophisticated data mining applications have been implemented by banks (e.g., fraud detection and credit scoring).

Thirdly, with the data explosion experienced by many organisations collecting increasingly larger amounts of data (e.g., on transactions and customers), organisations have begun to realise that data are not useful for decision making unless they can be transformed into information – an important asset. Further, the masses of data generated by organisations are often too complex and voluminous to be processed and analysed by traditional methods. In this respect, data mining provides the means to analyse large databases to generate valuable business information.

Fourthly, success stories of data mining applications and aggressive marketing by data mining consultants and software vendors have resulted in increasing numbers of organisations jumping onto the data mining bandwagon. This has enabled increasingly more organisations to reap the benefits of data mining. As stated by Davis (1999), more companies are now using data mining as the foundation for strategies that help them outsmart competitors, identify new customers and lower costs. Data mining has a range of techniques to help solve and address a host of business (and non-business) problems and issues.

1.2 Definition

Despite the increasing popularity and application of data mining, there is surprisingly no common accepted definition of data mining. In fact, Trybula (1997) has commented that data mining terminology is inconsistent and poorly defined. Some authors, for example, equate data mining with modelling (see Banecko and Russo, 1999) while others define it as an extended form of on-line analytical processing or OLAP (see Bodnar, 1998). The slightly confused notion of data mining is made worse by some enthusiastic software vendors and business consultants who present data mining as a panacea for almost all business problems and/or as an unavoidable investment for almost guaranteed business success.

Among the more useful definitions of data mining are those of Kincade (1998) and Milley (2000). They have defined data mining, respectively, as "the process of finding previously unknown patterns and trends in databases and using that information to build predictive models" and "the process of data selection, exploration and building models using vast data stores to uncover previously unknown patterns". These definitions capture the essential aspects of data mining.

Firstly, data mining is a process. That is, it comprises several iterative steps and is not a one-off effort to apply an analytical technique to a data set. Data mining is a methodology in that it represents a particular approach to data analysis. It is also a technology in that it uses sophisticated algorithms from diverse disciplines such as statistics/mathematics, machine learning and artificial intelligence. Data mining uses considerable computing power too.

Secondly, data mining is usually applied to large data sets. In fact, it is the existence of large data sets that has motivated the development of alternative analytical methods that are different from traditional statistical methods. While there is no suggestion that data mining can be used only on large data sets, it can be argued that the power and benefits of data mining can best be brought out where large data sets are involved.

Thirdly, data mining focuses on the exploration and discovery of previously unknown patterns and trends. This aspect is not surprising because data mining has been and still is mainly developed and applied in the commercial world, where gaining a competitive edge is critical for good business performance, survival and growth. The aim is to generate information unknown to competitors so that the chance of organisational success can be enhanced.

Fourthly, data mining helps organisations and managers make better decisions. Data mining is only a means to an end, whether it is to solve a business problem, capitalise on a business opportunity or respond to a business threat. Data mining transforms raw data into competitive information and business intelligence to provide valuable inputs for better decision making.

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. This definition of data mining will be used in this book. This definition covers both commercial applications (e.g., credit scoring and customer relationship management) and non-commercial applications (e.g., public health and criminal investigation).

1.3 Methodology

As discussed above, data mining is a process comprising several iterative steps. The methodology of data mining refers to these steps, which together form the approach to data mining. The methodology of data mining can be broadly divided into three major stages: pre-modelling, modelling and post-modelling. The emphasis of this book is on modelling, which is the core of data mining. In this context, modelling refers to the exploratory or discovery activities in data analysis. The methodology of data mining is summarised in Figure 1.1. The double-headed arrows show that the three stages can be interactive and iterative, and not necessarily sequential.

1.3.1 Pre-modelling Stage

One of the most important reasons for failure in a data mining application is the overemphasis on data analysis at the expense of the business problem the application is intended to solve and the business issue it is meant to address. The business problem or issue should underlie the entire data mining process and its identification should signal the first step in the process. To facilitate discussion, from this point onwards the term “business problem” will be used to include also “business issue”.

To state formally, the first step in data mining is to identify the business problem. For example, the business problem may be the desire to improve the response rate of a direct mail marketing campaign or to reduce customer churn (i.e., to reduce the number of customers leaving an organisation to patronise other organisations).

Stating a business problem by itself does not automatically suggest a data mining application. While data mining can help solve many business problems, not all business problems can or should be tackled via data mining applications (e.g., some business problems are best solved by internal organisational restructuring or via management science methods). For business problems that are appropriate for data mining, it is necessary to translate the business problem into a data mining application. This is the second step in the pre-modelling stage.

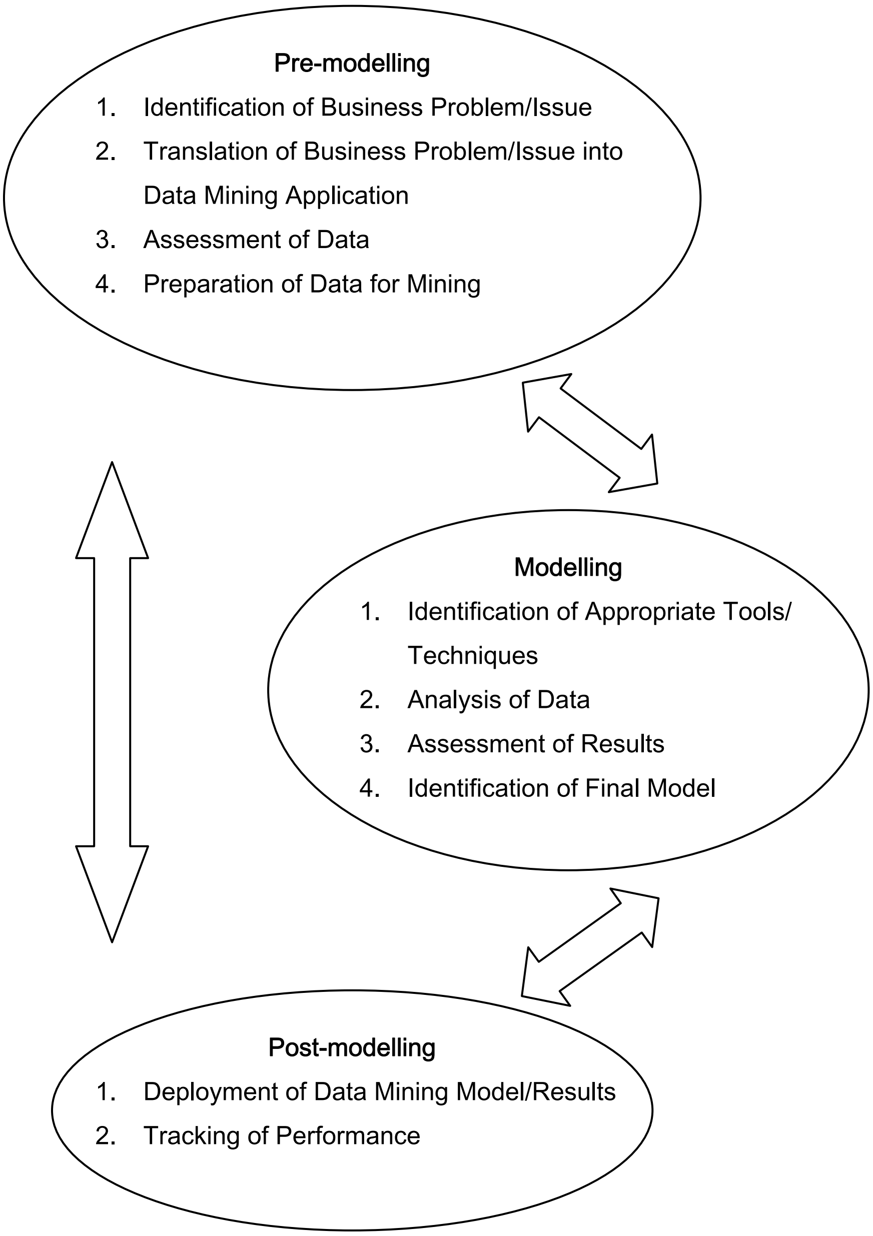
Pre-modelling
1. Identification of Business Problem/Issue
2. Translation of Business Problem/Issue into Data Mining Application
3. Assessment of Data
4. Preparation of Data for Mining
Modelling
1. Identification of Appropriate Tools/ Techniques
2. Analysis of Data
3. Assessment of Results
4. Identification of Final Model
Post-modelling
1. Deployment of Data Mining Model/Results
2. Tracking of Performance

Figure 1.1 Data Mining Methodology

For instance, to improve the response rate of a direct mail marketing campaign, an organisation may conclude that it is essential that the product brochures be sent only to customers with a reasonably good chance of purchase. This may mean that it is useful to know the relationship between customer demographic characteristics (e.g., age and gender) and past purchasing behaviour (e.g., nature and frequency of products purchased) and the probability of purchasing a particular product being promoted via direct mail marketing. Translated into a data mining application, this means that the objective of the data mining application is to model the relationship between customer demographic characteristics and past purchasing behaviour so as to predict the probability of purchase.

Data mining is not possible without data. Hence, the third step is to assess the data needed and available for the data mining application. Following from the above example on direct mail marketing, data on customer demographic characteristics, past purchasing behaviour and purchase/non-purchase of the particular promoted product are needed for modelling to proceed. Data can come from internal sources (e.g., existing customer or transaction records) or external sources (e.g., data compiled from marketing surveys). If the needed data are not available, then they will have to be obtained (e.g., by purchasing the data from data providers) or generated (e.g., by running a test marketing campaign). An alternative is to exclude from the model data that are not available. That is, if an organisation does not have any data on customer's income, then income cannot be included in the data mining model.

The fourth and last step in the pre-modelling stage is also the most tedious step in data mining: the preparation of the data for mining. In many instances, the needed data are available but not in the same database or the same standard format. Efforts have to be made to extract and combine data from different databases or sources and make the data consistent. More importantly, existing data may be incomplete or may contain errors. These data problems have to be dealt with too (say, by filling in missing data from other sources and correcting erroneous data).

Data preparation for data mining also includes data transformation and data derivation. Data transformation refers to the application of mathematical operations to the original data values. For example, to induce a linear relationship with a target variable (e.g., customer spending), an organisation may use the square-root of income instead of income as a predictor variable in the data mining model. In this instance, the variable income is transformed to the square-root of income. Data derivation refers to the generation of new variables from existing variables. To illustrate, instead of using current assets and current liabilities separately and independently as variables to predict the financial status of corporate customers, an organisation may find that the ratio of current assets to current liabilities (also known as current ratio in accounting) may be a better predictor variable. In this instance, current ratio is derived from current assets and current liabilities. Other common derivations include percentages and differences.

1.3.2 Modelling Stage

The modelling stage can often be deemed to be the core of data mining. This is the stage where data analysis is performed. Generally, for any data mining application, several data mining tools or techniques can be used (e.g., logistic regression, neural networks or decision trees can be used for churn modelling to determine customer turnover). Conversely, a particular data mining tool or technique can be used to achieve different data mining objectives (e.g., clustering can be used for market segmentation or fraud detection). Given the above, as the first step in the modelling stage, it is essential to identify the appropriate tools or techniques to use in a particular data mining application. What tool or technique is appropriate depends on the nature of the data mining application (e.g., clustering or predictive modelling) as well as the nature of the data (e.g., whether the data are qualitative [e.g., gender] or quantitative [e.g., age]). More on data mining tools/techniques will be discussed in later chapters.

Once the appropriate tools/techniques are identified, the next step is to perform the actual data analysis (or modelling). After this is done, it is necessary to assess the results (the third step of the modelling stage). This relates to the objective of the data mining application. For example, if the objective is to perform market segmentation, then it is necessary to assess if the clustering results lead to interpretable, useful and actionable market segments. Similarly, if the objective is to predict customer churn, then it is necessary to assess if the predictive modelling results give sufficiently accurate predictions. Accuracy rates can be computed by applying the data mining model to a sample of data not used in the construction of the model. Assessment of results can be statistical in nature too (such as evaluating the statistical significance of the model or its parameters/coefficients). Data mining is an iterative process. Hence, the assessment results may lead the organisation to re-look at the data, re-select the variables for inclusion, re-sample the observations or re-perform the analysis … etc.

As mentioned earlier, for any data mining application, there may be more than one data mining tool/technique that can meet its objective. Predictive modelling, for example, can be done by using logistic regression, neural networks or decision trees. If two or more models give acceptable results, then there is a need to identify the final (i.e., champion or best) model. In particular, the different acceptable models can be compared with respect to their accuracy rates and the one that is the most accurate can be selected as the final model. Identification of the final model can be aided by data mining statistics such as lift value/chart or profit value/chart. A more elaborate discussion of these will be given in Chapter 4. The comparison of different model results to identify the final model is the last step of the modelling stage.

1.3.3 Post-modelling Stage

The post-modelling stage relates to the actions to be taken after the data analysis is completed. The first step is the deployment of the data mining model or results. What this involves depends on the objective of the data mining application. For example, if the objective is market segmentation, then the clustering results will be used as inputs for decision making (e.g., designing new products for different market segments, reaching different market segments through different advertisement channels, formulating different strategies for different market segments … etc.). If the objective is predictive modelling, then the data mining model may have to be embedded into the operational or other systems of the organisation to provide inputs into decision support systems or to generate information for decision making.

To illustrate, suppose that an organisation is keen to promote selected products to its existing customers. To facilitate this, the organisation has constructed a data mining model to predict the probability of purchase based on the data available for existing customers. This model can be embedded into the operational system at the call centre so that for every customer who contacts the call centre, the operators know the probability of the customer's propensity to purchase the selected products. This information can then be used by the operator to promote the products.

The last step in the data mining methodology shown in Figure 1.1 is the tracking of performance. This is necessary because of changes in the environment in which an organisation operates. Such changes may lead to a deterioration of the performance of data mining models, which may be outdated. For example, the variables and relationships that help predict a target variable (e.g., fraudulent behaviour) may change over time and a data mining model constructed in the past may no longer be useful at present. Hence, tracking is important as deteriorating performance may signal the need to look at the data mining model again and to build an updated model if necessary. The need to re-look at data mining applications is not limited to predictive modelling. Even data mining results such as market segments or association rules may be outdated.

Finally, the double-headed arrows connecting the three stages of the data mining methodology in Figure 1.1 indicate that data mining is an interactive and iterative process. Very often, it is necessary to move back and forth among the stages or steps when developing a data mining application. For instance, poor modelling results may mean looking at the data again. Hence, data mining is not a strictly sequential process.

1.4 Organisation of the Book

The remainder of this book is organised as follows. Chapter Two introduces the wide range of data mining tools covering description and visualisation as well as association and clustering while Chapter Three covers predictive modelling and its tools. In data mining, the terms "tools" and "techniques" have been used interchangeable and sometimes with some slight differentiation. For ease of discussion, the term "tools" will be used from this point onwards to refer to statistical, artificial intelligence and machine learning methods and techniques that can be used in data mining.

Chapter Four discusses selected data mining issues that are of special significance in predictive modelling. These include model validation, optimal cut-off points and evaluation charts. Chapters Five to Seven present comprehensive case studies illustrating the potential applications of data mining for a retailer, a service provider and a manufacturer, respectively. The primary purpose of these case studies is to help organisations get started in data mining. SPSS software (e.g., Clementine – a very user-friendly and competitively priced data mining software) will be used in the illustrations.

Finally, the concluding chapter presents a summary, highlights some limitations of data mining and suggests some future directions. An appendix and an index are also given after Chapter Eight. The appendix summarises the software used in the illustrations to facilitate organisations that are interested to get started in data mining. The index provides page references to key topics discussed in the book.

This book aims to be as practical as possible in introducing readers to data mining and in guiding them to develop data mining applications. It aims also to be as non-technical as possible so that readers without a background in statistics or mathematics can benefit from it. Further, it aims to be concise so that readers can get to the gist of the content without being unnecessarily distracted by details. However, references are provided where appropriate for readers who are interested to know the content in greater depth.

Chapter References

Banecko, J. J. and Russo, A. W. (1999), “Taking the mystery out of mining and modelling”, *Direct Marketing*, Vol. 62 No. 4, pp. 42-43.

Berger, C. (1999), “Data mining to reduce churn”, *Target Marketing*, Vol. 22 No. 8, pp. 26-28.

Bodnar, G. H. (1998), “Data warehouses and data mining”, *Internal Auditing*, Vol. 13 No. 3, pp. 37-42.

Chung, H. M. and Gray, P. (1999), “Data mining”, *Journal of Management Information Systems*, Vol. 16 No. 1, pp. 11-13.

Davis, B. (1999), “Data mining transformed”, *InformationWeek*, No. 751, pp. 86-88.

Kincade, K. (1998), “Data mining: digging for healthcare gold”, *Insurance & Technology*, Vol. 23 No. 2, pp. IM2-IM7.

Kreuze, D. (2001), "Debugging hospitals", *Technology Review*, Vol. 104 No. 2, p. 32.

Milley, A. (2000), "Healthcare and data mining", *Health Management Technology*, Vol. 21 No. 8, pp. 44-47.

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