Chapter 2 Data Mining Tools

2.1 Introduction

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. In the pre-modelling stage, an organisation has to identify the business issue and translate it into a data mining application. Next, the organisation has to assess the data that are accessible vis-à-vis those that are needed and proceed to obtain and prepare the data for mining. In the modelling stage, the data are analysed. This is the core of data mining and there is a wide range of data mining tools available to do the data analysis, ranging from statistical methods to artificial intelligence models to machine learning techniques.

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). Some authors (e.g., Berry and Linoff, 1997 and 2000) have classified data mining tools into more detailed categories.

2.2 Description and Visualisation

Description and visualisation can contribute greatly towards understanding a data set (especially a large one) and detecting hidden patterns in the data (especially complicated data containing complex interactions and non-linear

relationships). They are frequently performed before modelling is attempted to understand and/or detect relationships among variables. Description and visualisation can also help greatly in the summarisation of data and in the presentation and reporting of results.

Description refers to the summarisation of data to facilitate understanding. An example of description is the profiling of data sets in order to understand their characteristics, similarities and differences. Standard description tools include summary statistics such as measures of central tendency (e.g., mean), measures of dispersion (e.g., standard deviation) and counts (e.g., cross-tabulation). For example, a market segment can be profiled as having a mean age of 35 years and primarily (80%) females.

Graphical approaches (e.g., distributions and plots) can also help to describe data and the relationships in data. For instance, histograms and pie charts can be used to describe data and plots/graphs to display relationships (e.g., a scatter plot of age on the x-axis and expenditure on the y-axis).

Visualisation can be considered an enhanced graphical approach that allows user input and interaction. An example is a rotating multidimensional plot that permits the user to define multiple dimensions (i.e., multiple variables) in the plot as well as the direction and angle of rotation to facilitate viewing complex relationships. Colours can also enhance visualisation tools.

Another very useful visualisation tool is the web graph, which shows graphically the existence and strength of relationships among variables and among levels of different variables. Conceptually, it is similar to link analysis, which examines how variables and their different levels are linked. To illustrate, “churn” may be a variable with two levels (e.g., no-churn and voluntary churn) and mobile phone plans may be a variable with three levels (e.g., Economy Plan, Standard Plan and Unlimited Plan). The web graph may show, for instance, a strong link between churn and mobile phone plans in general and between voluntary churn and Economy Plan in particular in that there are many cases of customers on the Economy Plan who leave the telephone company

(or telco) voluntarily. The web graph is especially useful if an organisation is interested in the links among many variables. In contrast, cross-tabulations that are too large and that have too many dimensions are very difficult to interpret.

2.2.1 Illustration of Description and Visualisation

In data mining, description and visualisation tools can be used to understand people, products and processes and to study the relationships among variables. Sometimes, the results from such analyses are an end in themselves (e.g., the profiling of clusters as discussed in section 2.4). However, the results are usually used as a means to construct data mining models (e.g., to predict certain target variables).

To illustrate description and visualisation, suppose that MailPurchase – a mail order company – has a database of 1400 customers (each identified by a customer number). Suppose further that the following data are captured in the database:

1) Whether the customer has purchased a promoted product in any of the quarterly marketing campaigns last year (= 1 for a purchaser and = 0 for a non-purchaser) [variable denoted as “status”];
2) Average monthly expenditure on the company’s products last year [variable denoted as “expend”];
3) Average number of purchases per quarter last year [variable denoted as “numpur”];
4) Age of customer as at 1 January last year [variable denoted as “age”];
5) Gender of customer (= 1 for male and = 2 for female) [variable denoted as “gender”];
6) Annual income of customer as at 1 January last year (in $’000) [variable denoted as “income”];
7) Race of customer (= 1 for Chinese, = 2 for Malay, = 3 for Indian and =4 for Others) [variable denoted as “race”];

8) Marital status of customer as at 1 January last year (= 1 for single, = 2 for married and = 3 for married with children) [variable denoted as "marital"]; and

9) Whether the customer is a member of the loyalty card programme last year (= 1 for member and = 0 for non-member) [variable denoted as "member"].

Assume that MailPurchase can extract the data as an SPSS (.sav) file. (SPSS [Statistical Package for Social Sciences] is a very user-friendly and powerful statistical software). An illustrative listing of the data is given in Figure 2.1.

To understand its customers better, MailPurchase is interested to generate the demographic and transactional profile of the customers. In particular, MailPurchase wants to compute the mean, minimum value, maximum value and standard deviation of the quantitative variables (namely expend, numpur, age and income) and the frequency distribution of the qualitative variables (namely, status, gender, race, marital and member). For this illustration, the SPSS product Clementine (a very user-friendly and powerful data mining software) is used. The results are summarised in Figure 2.2.

As shown, the mean average monthly expenditure of customers on MailPurchase's products last year is $207.65, with a minimum of $10.00, a maximum of $450.00 and a standard deviation (i.e., dispersion or variation) of $136.92. Also, 737 (i.e., 52.64%) of MailPurchase's customers have purchased a promoted product last year during the quarterly marketing campaigns. Conversely, 663 (i.e., 47.36%) of MailPurchase's customers have not purchased a promoted product last year. The other variables in Figure 2.2 can be similarly interpreted. It can be seen that description can help MailPurchase understand more about its customers' demographic and transactional profiles.

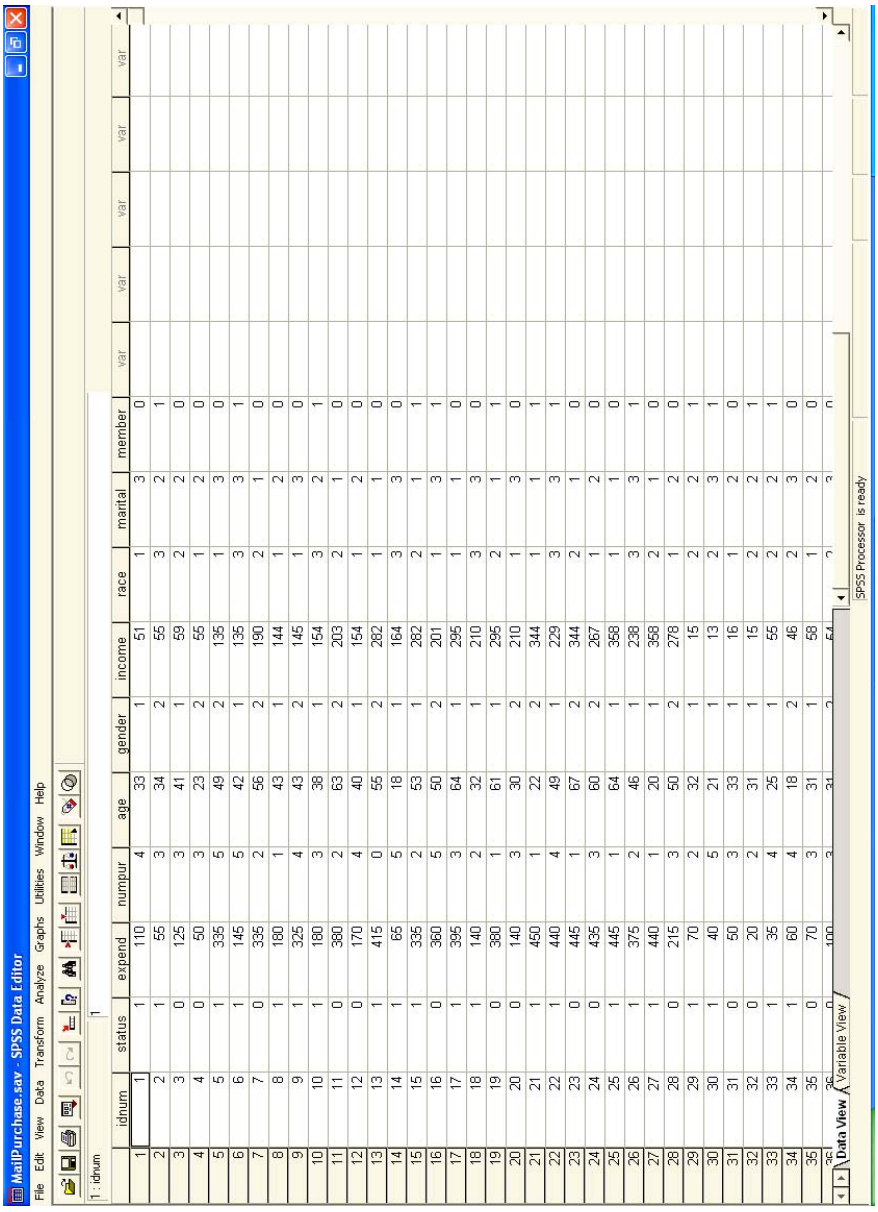

Figure 2.1 Illustrative Data of MailPurchase

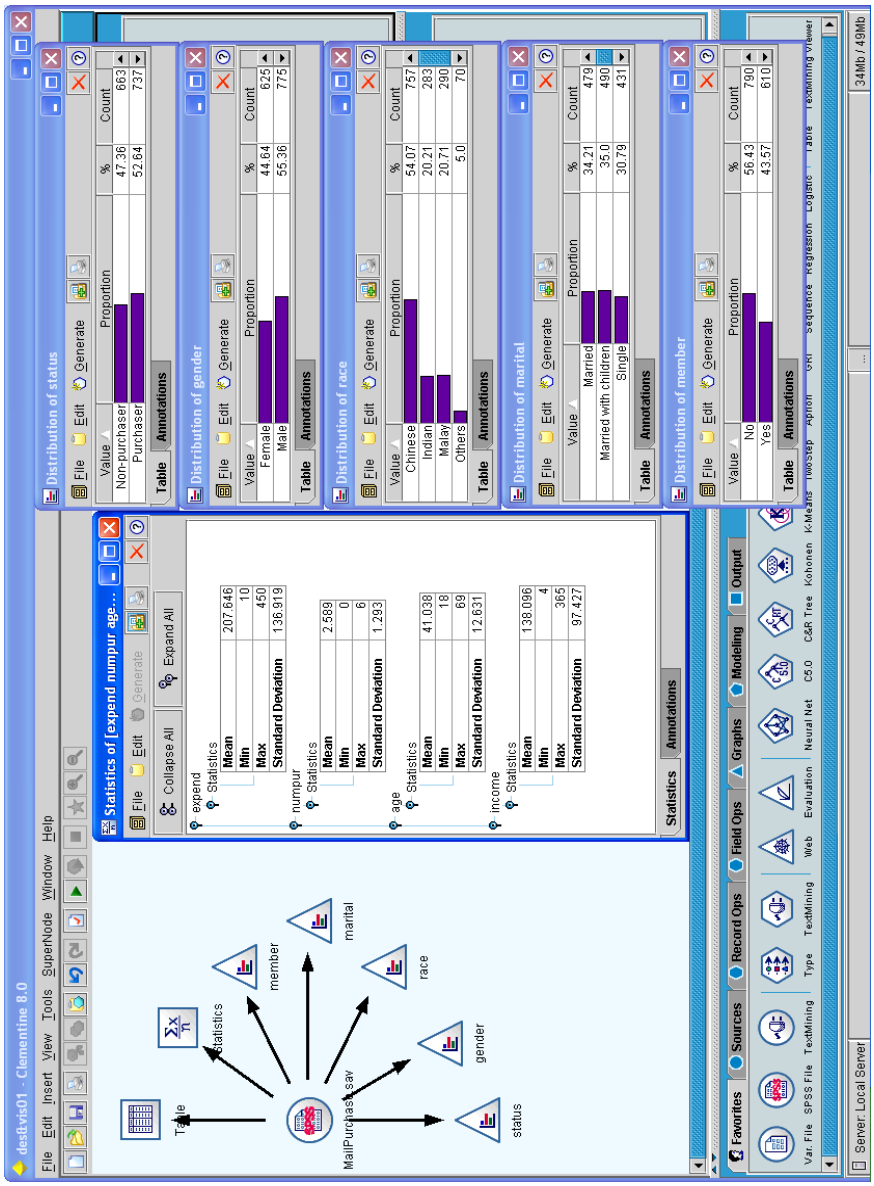

Figure 2.2 Descriptive Statistics for MailPurchase's Customers

Assume that MailPurchase is also interested to explore the relationship between average monthly expenditure on the company's products and the customers' age and income. It has decided to use a rotating plot for this purpose. An illustrative rotating plot using the SPSS statistics software is presented in Figure 2.3.

The figure shows that generally increasing expenditure is associated with increasing income. However, its association with age is non-linear in that mid-range ages seem to be associated with lower levels of expenditure. There appears to be an interaction effect of age and income on average monthly expenditure. In particular, different levels of expenditure are associated with particular combinations of age and income levels. For example, for customers with low income, expenditure increases with age. However, for customers with high income, a "U" shaped relationship exists between expenditure and age. As can be seen in Figure 2.3, the plot can be rotated by adjusting the horizontal and vertical dials on the left side of the plot. Rotation helps in exploring relationships among variables.

Finally, to illustrative the web graph, assume that MailPurchase is interested to explore the links among the variables status, gender, race, marital status and membership. In particular, it is interested to see how gender, race, marital status and membership are linked to the purchase/non-purchase of promoted products in past marketing campaigns. When a target variable of interest is identified (such as the status variable in this illustration), a directed web graph can be plotted. MailPurchase is not interested at this stage to explore the links among the input variables (i.e., among gender, race, martial status and membership). The directed web graph generated by the Clementine data mining software is shown in Figure 2.4.

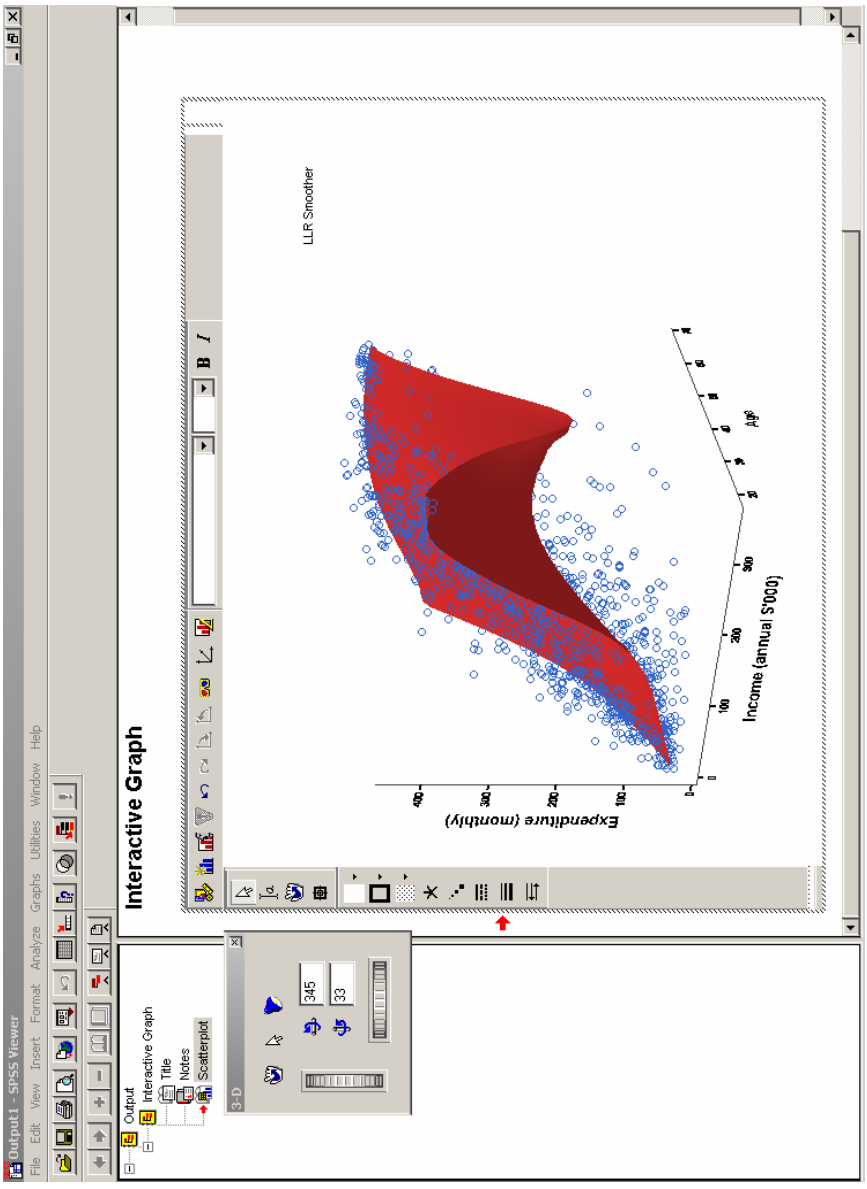

Figure 2.3 Rotating Plot for Average Monthly Expenditure, Age and Income

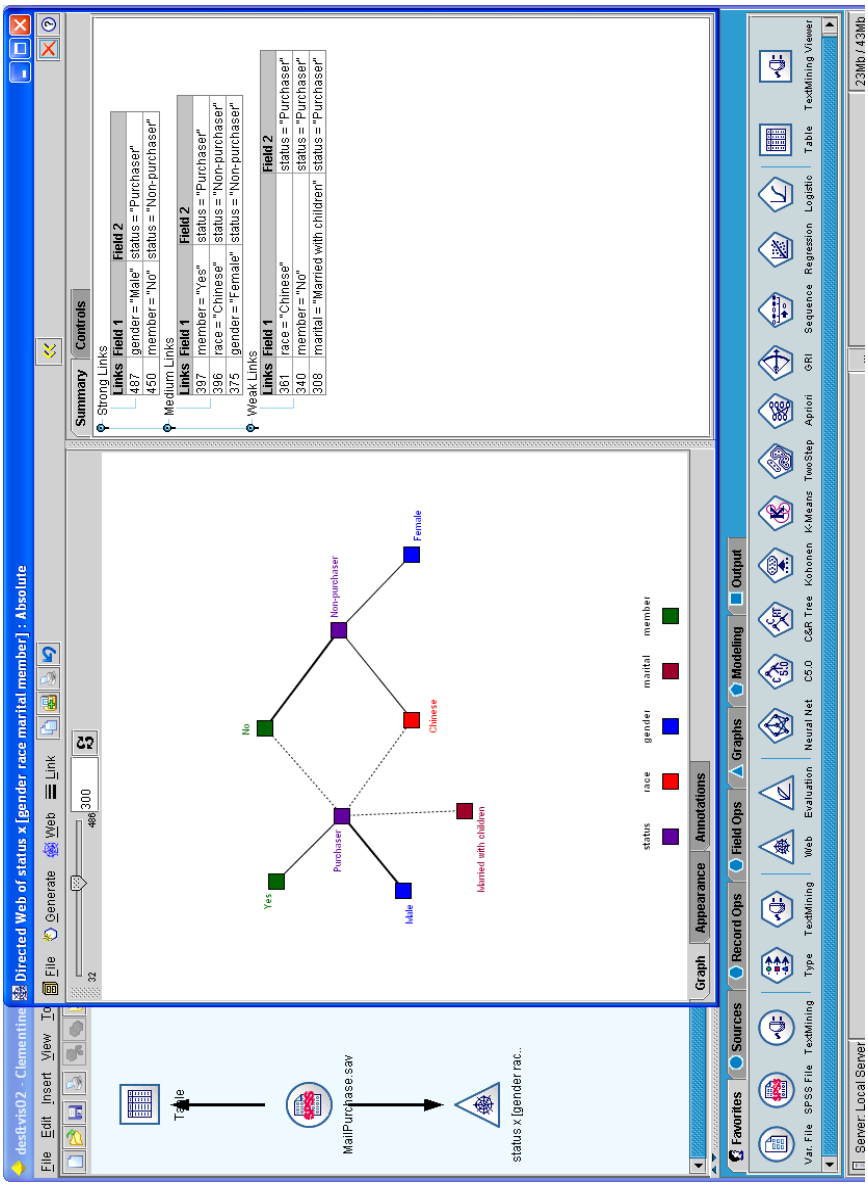

Figure 2.4 Directed Web Graph for Selected Variables of Interest

In generating the directed web graph, certain parameters can be specified by MailPurchase. In this illustration, a strong link is specified as 450 or more links (or joint occurrences of purchase/non-purchase and another input variable [e.g., male or female]). Further, a moderate link is specified as 375 to 449 links and a weak link as 301 to 374 links. Only links above 300 will be shown in the directed web graph. A stronger link between two variables (i.e., a darker line in the directed web graph) can be interpreted as a stronger association between them.

As shown in Figure 2.4, a "purchaser" status is strongly associated with male customers (487 links), moderately associated with members (397 links), and weakly associated with Chinese (361 links), non-members (340 links) and customers who are married with children (308 links). On the other hand, a "non-purchaser" status is strongly associated with non-members (450 links) and moderately associated with Chinese (396 links) and female customers (375 links). Among other things, such information can be useful in planning for the next marketing campaign to increase customer response by targeting the appropriate customers.

To summarise, it can be concluded that description and visualisation are important and useful tools for data analysis.

2.3 Association Analysis

In association analysis, the objective is to determine which variables/items go together. It is a tool that looks for groupings or patterns among a set of items. For example, market basket analysis refers to a technique that generates probabilistic statements such as: if customers purchase coffee, there is a 0.35 probability that they also purchase bread. Such statements (also called rules in data mining) are intuitive and easy to understand. Also, good association is expected to have predictive value. However, many applications of association

analysis are exploratory in nature, with a view to better understand groupings and patterns in the data set.

Association rules can be useful for store layout, items bundling, discount and promotion decisions, cross-selling … etc. For instance, if a supermarket finds that eggs, bread and butter are usually purchased together, then it can put these items in close proximity to each other in the store for customer convenience and to facilitate purchase. Alternatively, it can bundle the items together as a promotion package. Market basket analysis can be applied not only to items purchased concurrently but also to items purchased sequentially. In this case, an association (sequence) rule may be: if customers purchase a flower pot, there is a 0.30 probability that they will purchase potting soil in the next purchase/visit.

A large set of association rules can be generated for a large data set. When the number of items for association analysis increases, the number of combinations increases exponentially. An excessive number of association rules can be dysfunctional in that decision makers may be overloaded with rules that are not useful. Hence, it is necessary to either focus only on the items are of most interest to an organisation or reduce the number of items by grouping minor items into major categories (e.g., 1.0 cm, 1.5 cm, 2.0 cm and 2.5 cm nails may be grouped as just one major category “nails”). In data mining, the reduction of groups is referred to as binning.

Association analysis is based on a co-occurrence table. To illustrate, suppose that an organisation is interested in the four items A to D, whose frequency of co-occurrence is summarised in Table 2.1. For simplicity, assume that 100 items of A, 150 of B, 200 of C and 250 of D are purchased and that only one- and two-item purchases (i.e., transactions) are of interest. Higher dimension co-occurrences tables (e.g., three-item and four-item purchases) can also be constructed.

Table 2.1 Co-occurrence Table for Items A to D

Item	A	B	C	D	Total
A	40	10	10	40	100
B	10	60	80	0	150
C	10	80	110	0	200
D	40	0	0	210	250
Total	100	150	200	250	700

From Table 2.1, it can be seen that if C is purchased, then there is a 0.05 probability (i.e., 10/200) that A is also purchased and a 0.40 probability (i.e., 80/200) that B is also purchased. Association rules can be expressed in the following format:

Conclusion/Consequent ← Condition/Antecedent i + Condition/Antecedent j + …

Three measures are usually given with association rules, namely instances, support (also called coverage) and confidence (also called accuracy). Instances refer to the number of occurrences or co-occurrences of the antecedent(s) and support refers to this number as a percentage of the total number of transactions. Hence, support indicates how general an association rule is. Confidence refers to the number of transactions with the conclusion/consequent as a percentage of the number of transactions with the condition(s)/antecedent(s). Therefore, it measures how effective (or accurate) an association rule is.

Focusing on items B (as a consequent) and C (as an antecedent), the instances value is 200 for C since there are 200 transactions where C is purchased. Accordingly, the support is 0.2857 or 28.57% (i.e., 200 instances out of a total of 700 transactions). Also, the confidence is 0.4000 or 40.00%

(i.e., 80 instances out of 200 transactions where C is purchased, B is also purchased). These can be expressed as follows:

$$B \leftarrow C \ (200, 28.57\%, 40.00\%)$$

where the figures in brackets are the instances, support and confidence, respectively.

In its most simple form, association analysis begins by generating simple association rules involving two items. These rules are then validated against the data set to identify those that satisfy the requirements related to support and/or confidence as specified by the user/organisation. These "interesting" rules are stored and then used to generate the next set of higher-dimension association rules. That is, the process as described above is used to generate three-item interesting association rules, four-item interesting association rules … etc., up to n-item interesting association rules (where n is defined by the user/organisation). As expected, the number of co-occurrences (i.e., instances) decreases at each successive step.

Association analysis can be extended to include more sophisticated applications. For example, a particular item can be identified as a target variable and association rules that have that item as a conclusion/consequent can be generated. This enables decisions to be made with respect to that item (e.g., promotion or discount). Also, in the case of market basket analysis, in addition to items being included in the analysis to understand which ones are usually purchased together, customer characteristics (e.g., age and gender) can also be included. This enables the organisation to study, for instance, the relationship between purchasing patterns and demographic patterns. Further, instead of examining items that are purchased, it is also possible to include items that are not purchased. In this case, an associate rule may be: if peanuts and potato chips are purchased and cola is not purchased, there is a 65% chance that beer is also purchased. Another example that involves a different context is: if high blood pressure is present and daily exercise is absent, there is a 70% chance that high cholesterol is also present.

Finally, time can be incorporated into association analysis. Instead of examining which items are usually purchased together, it is possible to examine the sequence in which the items are purchased. A sequence association rule generated may be: if paint and brushes are purchased at time t (e.g., current visit), then there is a 75% chance that drills will be purchased at time t+1 (e.g., next visit), where t indicates the time period. Time-incorporated association analysis is especially useful in applications such as the study of repair sequences or web navigation sequences. For example, to increase on-line purchases, an on-line bookstore may try to isolate the sequences of web navigation that are likely to lead to on-line purchases. Actions (e.g., web re-design) can then be taken to promote sales-generating sequences/behaviour.

2.3.1 Illustration of Association Analysis

Continuing with the mail order company MailPurchase, suppose that its database of 1400 customers also contains the items purchased by each customer in the last major promotion campaign. To help develop future major promotion campaigns, MailPurchase is keen to examine what items have been purchased by each customer. In particular, it wishes to focus on a list of high margin items. MailPurchase has decided to perform association analysis on the following seventeen items:

1) Item A: Flower arrangements;
2) Item B: Puzzles and games;
3) Item C: Card games;
4) Item D: Table clocks;
5) Item E: Desktop decorations;
6) Item F: Blouses;
7) Item G: Tee-shirts;
8) Item H: Sports shorts;
9) Item I: Stationery gift sets;

10) Item J: Gardening tools;
11) Item K: Garden decorations;
12) Item L: Magazine subscriptions;
13) Item M: Cards for special occasions;
14) Item N: Sports shoes;
15) Item O: Walking shoes;
16) Item P: Hats and hair accessories; and
17) Item Q: Watch gift sets.

For this illustration, the Apriori node in SPSS Clementine is used. The data are contained in the SPSS file MailPurchase_Assoc.sav. Assume that MailPurchase is interested only in useful association rules, which it defines as rules with support of at least 20% and confidence of at least 60%. In other words, the association rules generated should be applicable to at least 20% of the customers in the database and when the antecedents (of an association rule) hold, there should be at least a 60% chance that the consequent will also hold. The results are shown in Figure 2.5.

As shown, there are five association rules generated. The first rule indicates that there are 381 instances/customers (which is equivalent to a support of 27.2% of the 1400 customers) who have purchased cards for special occasions (which is the antecedent). Further, when cards for special occasions are purchased, there is a 68.8% chance (i.e., confidence) that stationery gift sets are also purchased.

The lift value is 1.426. It indicates how much more the probability of purchase of stationery gift sets for customers who have also purchased cards for special occasions (also called conditional probability) is compared to that for all customers (also called prior probability). Figure 2.5 (lower right panel) shows that 675 (or 48.21%) of all customers purchase stationery gift sets (see “T” column and “Total %” row).

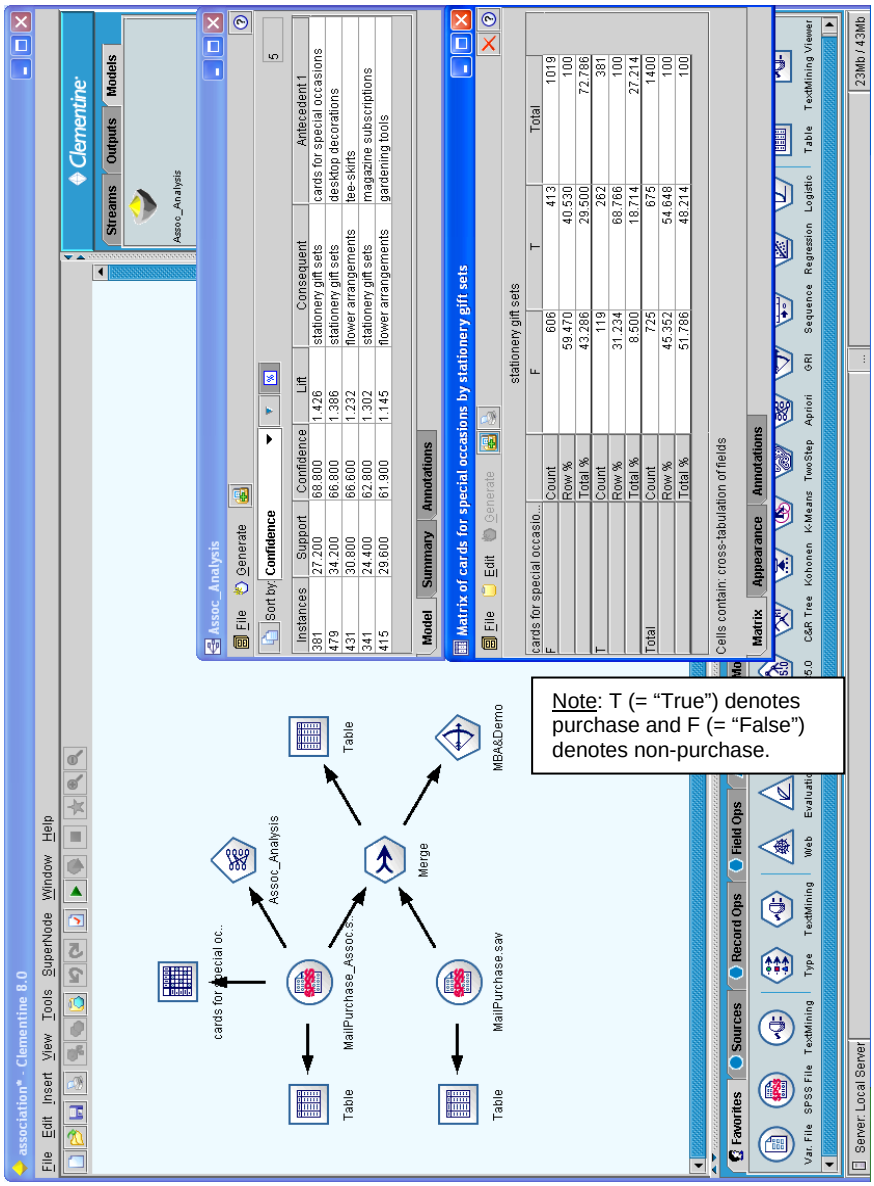

Figure 2.5 Association Analysis of High Margin Items

For customers who have purchased cards for special occasions (see "T" row), 262 (or 68.77%) also purchase stationery gift sets. This is also the confidence of the first association rule. Therefore, the lift value is 68.77/48.21 or 1.426.

The other association rules can be interpreted in a similar manner. Such association rules can help MailPurchase in future promotion campaigns (e.g., deciding on which items to bundle together or which items to give discounts or promote).

Suppose that to help plan future marketing campaigns, MailPurchase is also keen to investigate if demographic characteristics may be associated with purchasing patterns. For this, the GRI (Generalized Rule Induction) node in SPSS Clementine is used, with minimum support of 20% and minimum confidence of 75%. Also, data from the SPSS files MailPurchase.sav (comprising demographic characteristics) and MailPurchase_Assoc.sav (comprising purchasing patterns) are merged and used to generate the rules.

As shown in Figure 2.6, four association rules are generated. The interpretation of these rules is similar to that discussed earlier for Figure 2.5. For example, the second rule indicates that there are 287 Chinese who are singles – this is about 20.49% of the 1400 customers in the database. Of these customers, there is a 83.00% chance that they purchase flower arrangements. Similarly, the third rule indicates that there are 309 male customers who are married with children, which is about 22.06% of the 1400 customers in the database. Of these customers, there is a 79.00% chance that they purchase blouses. As before, the lift value indicates the ratio of the conditional probability of purchase of a particular item (conditional on the purchase of other item[s]) to the prior probability of the purchase of that particular item in the database. It indicates the effectiveness of the association rule (i.e., the higher the lift, the more effective [or accurate] the rule is).

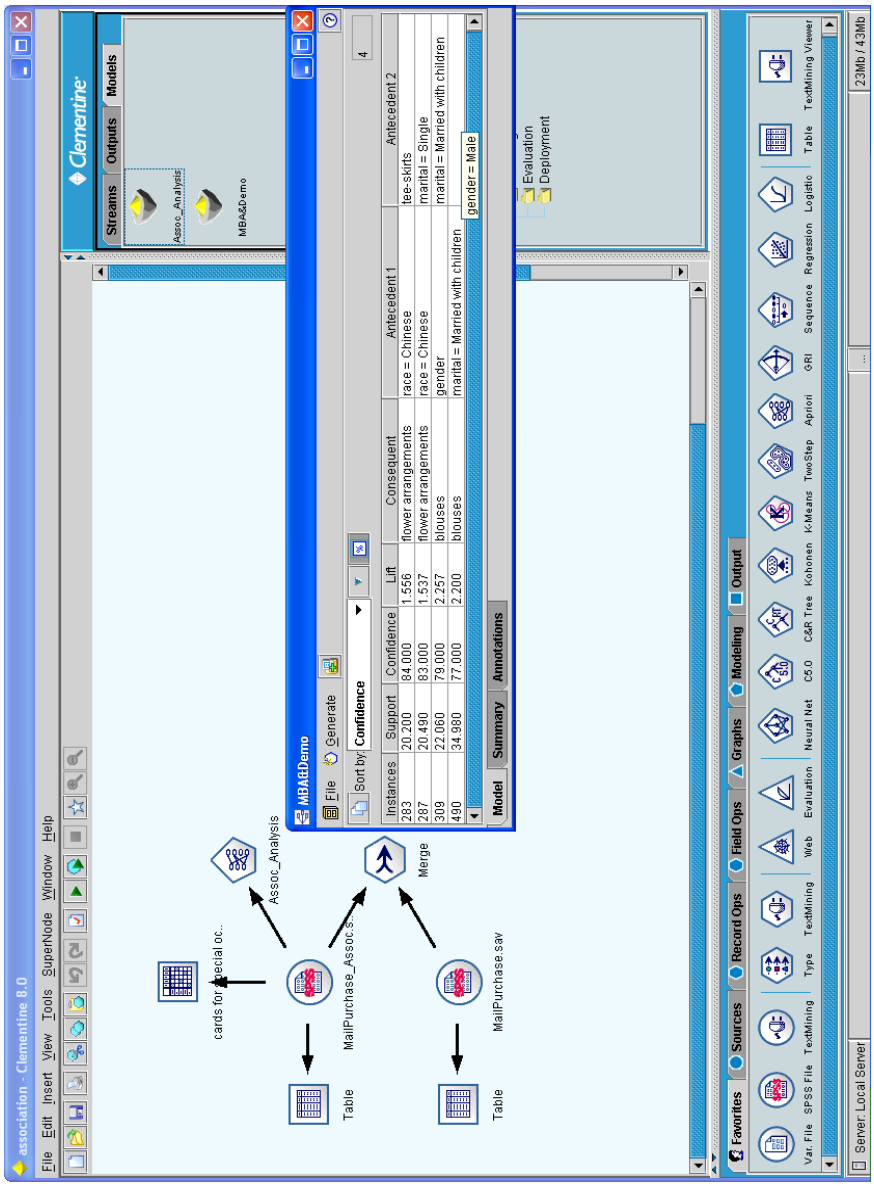

Figure 2.6 Association Analysis Incorporating Demographic Variables

To summarise, association analysis is useful when organisations want to group items into small sets (e.g., sets of items that are purchased together). This is most useful in the retail industry, where association analysis is commonly referred to as market basket analysis.

The applications of association analysis, however, are not limited to the retail industry. In the banking and finance industry, for example, association analysis can be used for customer relationship management. One possible application is to study the types of transactions that are frequently done together at an ATM or a service counter. The results can help a bank better serve its customers, say, by re-configuring the ATM to meet customer needs better or by re-designing the queuing system for the service counters to facilitate transactions.

As another example, the healthcare industry can also benefit from performing association analysis. For instance, association analyses can look at different treatment regimes, patient characteristics, illness symptoms and different outcomes to examine if certain combinations of treatment regimes, patient characteristics and illness symptoms are associated with certain outcomes. These results can then be used to help identify best practices or to better understand the factors affecting outcomes.

From the illustrations above, it can be seen that association results are easy to understand and association rules are easy to use. However, domain business knowledge is essential in interpreting and using the rules.

2.4 Clustering

Clustering is an exploratory technique that attempts to discover natural groupings in data. The objective is to group similar (homogeneous) objects into the same cluster and dissimilar (heterogeneous) objects into different clusters on the basis of distances among these objects. A graphical illustration of this is given in Figure 2.7, which consists of several points plotted on the X_1 and X_2

axes. As shown, points that are close together (i.e., similar on X_1 and X_2) can be grouped into the same cluster. Further, points that are far apart (i.e., dissimilar on X_1 and X_2) can be grouped into different clusters. The variables X_1 and X_2 are referred to as the clustering criteria as they form the basis for clustering the points. In a clustering application, it is common to use between five to fifteen variables as clustering criteria. Figure 2.7 is only a simple illustration – in a real-life clustering application, the grouping of observations is unlikely to be as obvious and neat.

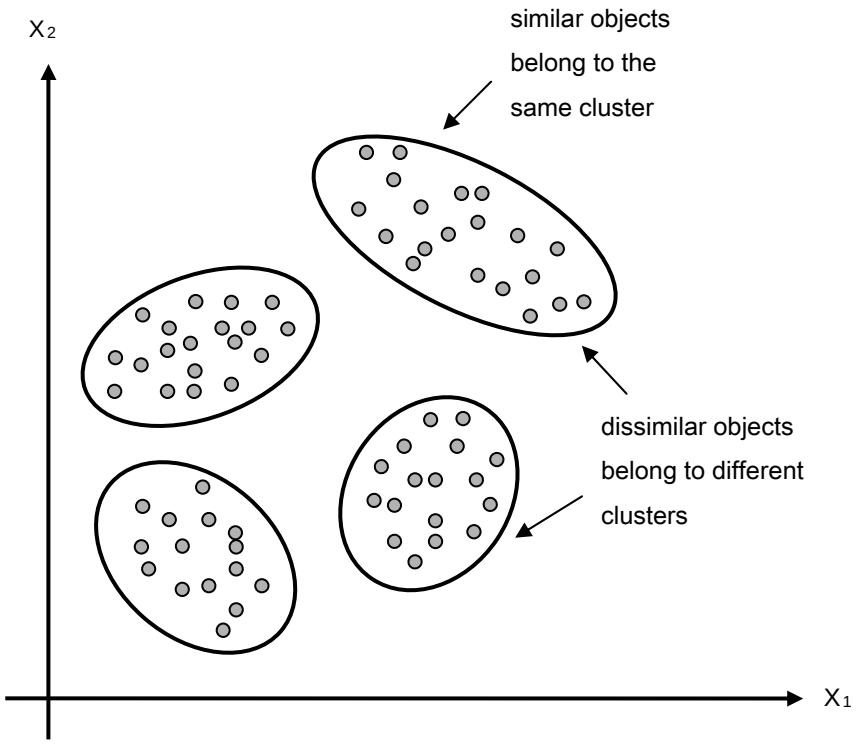

Figure 2.7 Graphical Illustration of Clustering Concepts

Clustering is usually used in data mining applications to do market segmentation and to identify the cluster profile of the different segments. Knowing how the market is segmented and the characteristics of the different segments helps in decisions such as an organisation's competitive positioning, the products to market to particular market segments, and the avenues and communications that can be used to reach the targeted segments.

In some instances, cluster membership is used as an input variable to predict a target variable of interest (e.g., which cluster is more likely to purchase a particular product). Here, clustering serves, among other things, as a data reduction device in that instead of describing an observation with a list of variables (e.g., the clustering criteria), the observation can now be described by its cluster membership.

One of the most commonly used data mining tools for clustering is K-means clustering. This technique is very efficient in clustering large data sets. In K-means clustering, "K" refers to the number of clusters and "means" the cluster centroids (i.e., the centre or average of a group of observations in a cluster). Usually, different K values (i.e., different number of clusters) are explored and the clustering results are evaluated to identify the most useful clustering solution. The K-means clustering algorithm works as follows.

Firstly, K well-spaced observations are selected as initial cluster centroids. Secondly, the distance of each observation to each centroid is computed. The most commonly used distance measure is the Euclidean distance. Given two variables X_1 and X_2 and two points (X_{11}, X_{21}) and (X_{12}, X_{22}), the Euclidean distance is:

$$\sqrt{[(X_{11} - X_{12})^2 + (X_{21} - X_{22})^2]}.$$

For p variables, the Euclidean distance between points i and j is:

$$\sqrt{\sum(X_{pi} - X_{pj})^2} \text{ summed over all the p variables.}$$

Thirdly, based on the distance computed, each observation in the data set is assigned to the nearest centroid. Observations assigned to the same centroid then form a cluster. There will be K number of clusters. Fourthly,

based on the observations assigned to each cluster, new cluster centroids are computed. Then, the iteration starts again from the second step.

The iterative process stops when the clustering solution is stable (e.g., when the cluster centroids do not change significantly) or when a specified number of iterations has been performed. The clustering results show cluster membership (i.e., which observation belongs to which cluster). After obtaining the cluster membership, the next step is to describe the characteristics of the clusters. For an organisation to use the clustering results, it is important that it understands the profile of the different clusters.

Another technique for clustering which is very different from statistical clustering methods such as K-means is the Kohonen network (also called self-organising map or SOM). This technique is based on neural network concepts (which will be discussed in the next chapter). Essentially, a Kohonen network can be deemed to be a system of nodes where each node gathers observations that are similar. That is, each node forms a cluster. An interesting feature of a Kohonen network is that the clustering results can be represented on a map (or grid) of clusters, where clusters that have more similar profiles are closer together. More discussion on the Kohonen network is presented in Chapter 3 under the topic of neural networks.

As clustering is an exploratory technique, it is useful to have some guidelines. Some rules-of-thumb are: (1) the number of clusters in the clustering solution should not be excessive [e.g., not more than ten]; (2) the number of observations (e.g., customers) per cluster should be at least 5-10% of the data set [to avoid having very small clusters]; and (3) the number of clustering criteria should not be excessive (e.g., not more than fifteen). These rules-of-thumb help ensure that the clustering solution is reasonably easy to interpret and hence useful. As in the case of association analysis, domain business knowledge is essential in the interpretation and use of clustering results.

Despite the rules-of-thumb, the utility of the clustering results ultimately depends on whether the cluster profiles are meaningful and useful, or actionable. Further, to validate the clustering results, it is advisable to check if similar clustering results can be obtained from different samples and different clustering methods. More confidence can be placed on clustering results that are more stable.

Both K-means and Kohonen network require that the number of clusters be specified. Generally, different numbers of clusters can be explored and the one that gives the most useful clustering solution selected. A more objective way to determine an appropriate number of clusters is to use the TwoStep clustering algorithm. This is a proprietary algorithm developed by SPSS which uses a statistical criterion (a likelihood cum penalty function) to determine the appropriate number of clusters. It works well on large data sets and is available in SPSS Clementine.

Finally, clustering results may be affected by outliers (i.e., observations that are very different from "typical" observations) and dominated by clustering criteria (i.e., variables used to cluster the observations) that have large values. Hence, before performing clustering, it is advisable to remove outliers and to standardise the clustering criteria (e.g., via data transformation so that each clustering criterion has a mean of 0 and a standard deviation of 1).

2.4.1 Illustration of Clustering

In the illustration described in section 2.2.1, MailPurchase is a mail order company with a database of 1400 customers. For each customer, the following data are captured:

1) Status: whether the customer has purchased a promoted product in any of the quarterly marketing campaigns last year;
2) Expend: average monthly expenditure on the company's products last year;

3) Numpur: average number of purchases per quarter last year;
4) Age: age of customer as at 1 January last year;
5) Gender: gender of customer;
6) Income: annual income of customer as at 1 January last year (in $'000);
7) Race: race of customer;
8) Marital: marital status of customer as at 1 January last year; and
9) Member: whether the customer is a member of the loyalty card programme last year.

(More details are given in section 2.2.1).

Suppose that to understand its customers better, MailPurchase is interested to segment its customers in the database by using clustering. In particular, MailPurchase wants to know if the customers can be segmented meaningfully based on purchasing patterns (namely, expend and numpur) and demographic characteristics (namely, age, gender, income, race, marital and member). MailPurchase has decided that for the clustering results to be useful, each cluster should have at least 200 customers. Also, there should not be too many customer segments – four clusters are probably desirable.

For this application, MailPurchase has decided to use SPSS Clementine (in particular, the K-means clustering node) to do the clustering. The clustering results are summarised in Figures 2.8 and 2.9. As can be seen, four clusters are produced with cluster sizes of 381, 391, 206 and 422, respectively. The cluster profiles in tabular form are also given in Figure 2.8.

To illustrate cluster profiling, it can be noted that Cluster 3 consists of male and single (i.e., non-married) customers with an average age of about 52.03 years, average annual income of about $192,155 and average monthly expenditure on MailPurchase's products of about $279.49. Further, the 206 customers in Cluster 3 have made an average quarterly number of purchases of 1.48 last year and 60.68% of them are not members of the loyalty card programme. From the clustering results, the profile of the other three clusters can also be described.

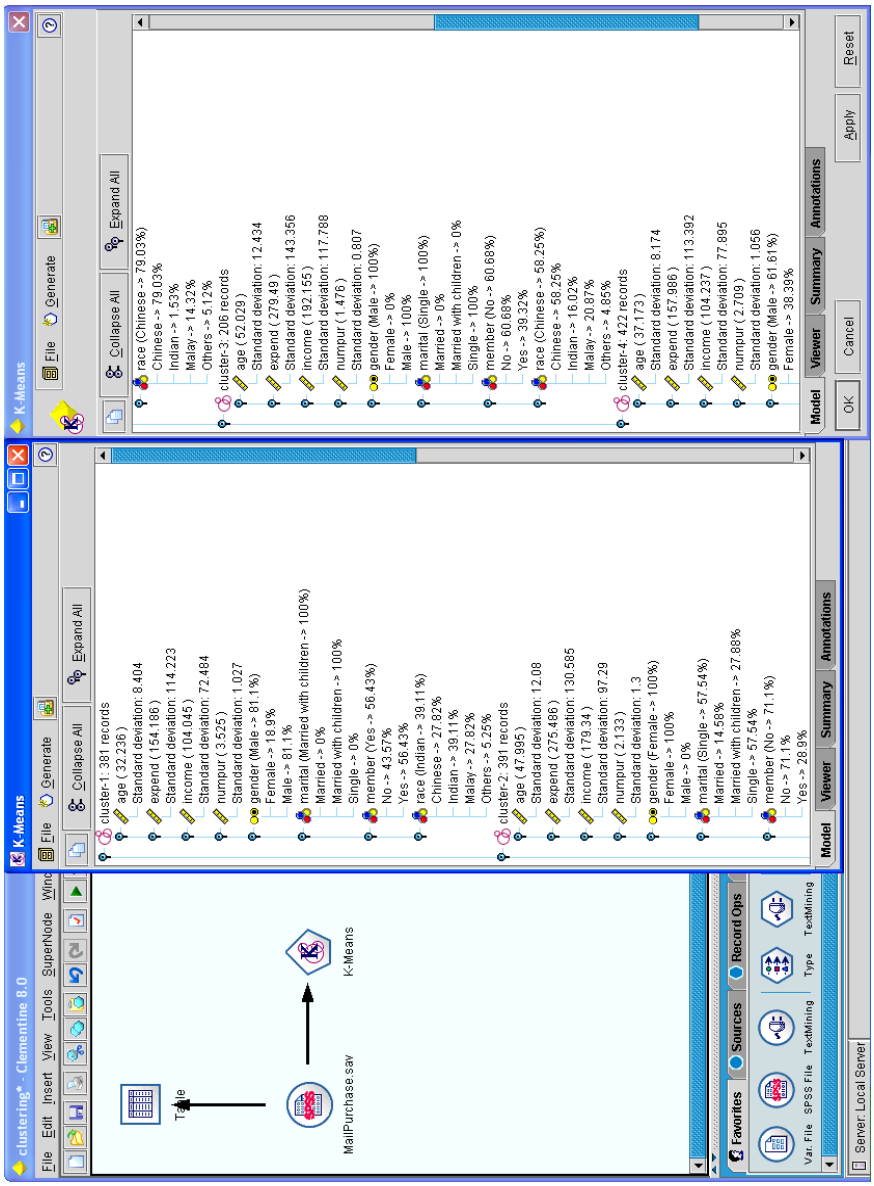

Figure 2.8 Clustering Results for MailPurchase – Tabular Format

Figure 2.9 summarises the various cluster profiles in a graphical format, with larger clusters on the left. The graphical view facilitates gaining a quick overview of the similarities and differences among clusters. For example, it can be seen that customers in Cluster 1 are younger than their counterparts in Cluster 3, and Cluster 4 comprises similar proportions of female and male customers whereas Cluster 2 comprises only female customers.

With the clustering results, MailPurchase can, among other things, understand its customer purchasing patterns and demographic characteristics better. This, in turn, can help MailPurchase to better evaluate customer needs and offer the appropriate products by market segment. Also, by using the clustering results, MailPurchase's marketing efforts can be more focused.

As an extension of the clustering application, suppose that MailPurchase is interested to see if the incidence of purchase/non-purchase in the quarterly marketing campaigns last year (i.e., the variable "status") differs across the four clusters. A frequency distribution with "status" as overlay and a cross-tabulation showing cluster number and purchase/non-purchase status is given in Figure 2.10. As shown in the cross-tabulation, Cluster 1 has the largest proportion of customers who have made purchases during the quarterly marketing campaigns (70.87%) and Cluster 2 has the smallest proportion of such customers (37.34%). With this information, MailPurchase will be able to target the right customers better.

2.5 Summary

Data mining tools can be broadly classified based on what they can do. Description and visualisation tools facilitate the understanding of a data set and the detection of patterns, trends and relationships. As such, they are frequently used to aid modelling.

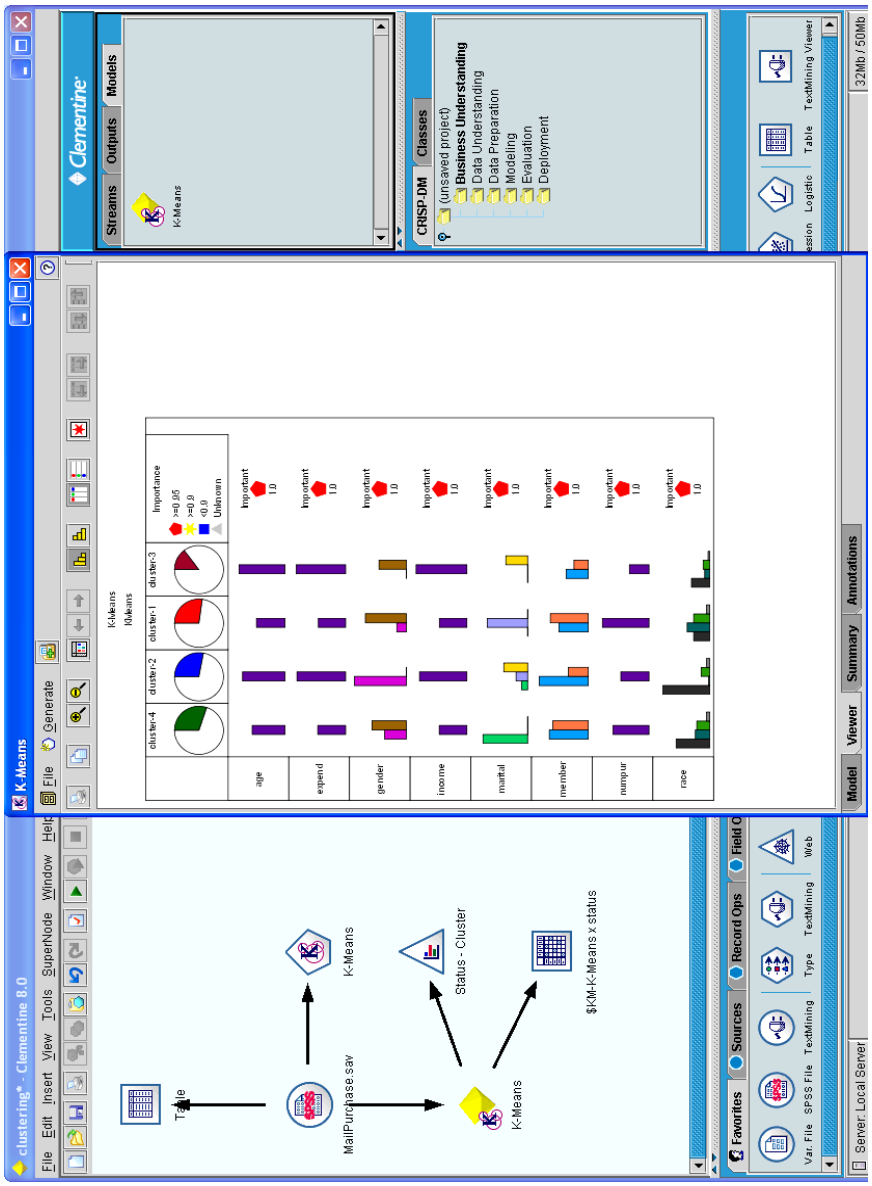

Figure 2.9 Clustering Results for MailPurchase – Graphical Format

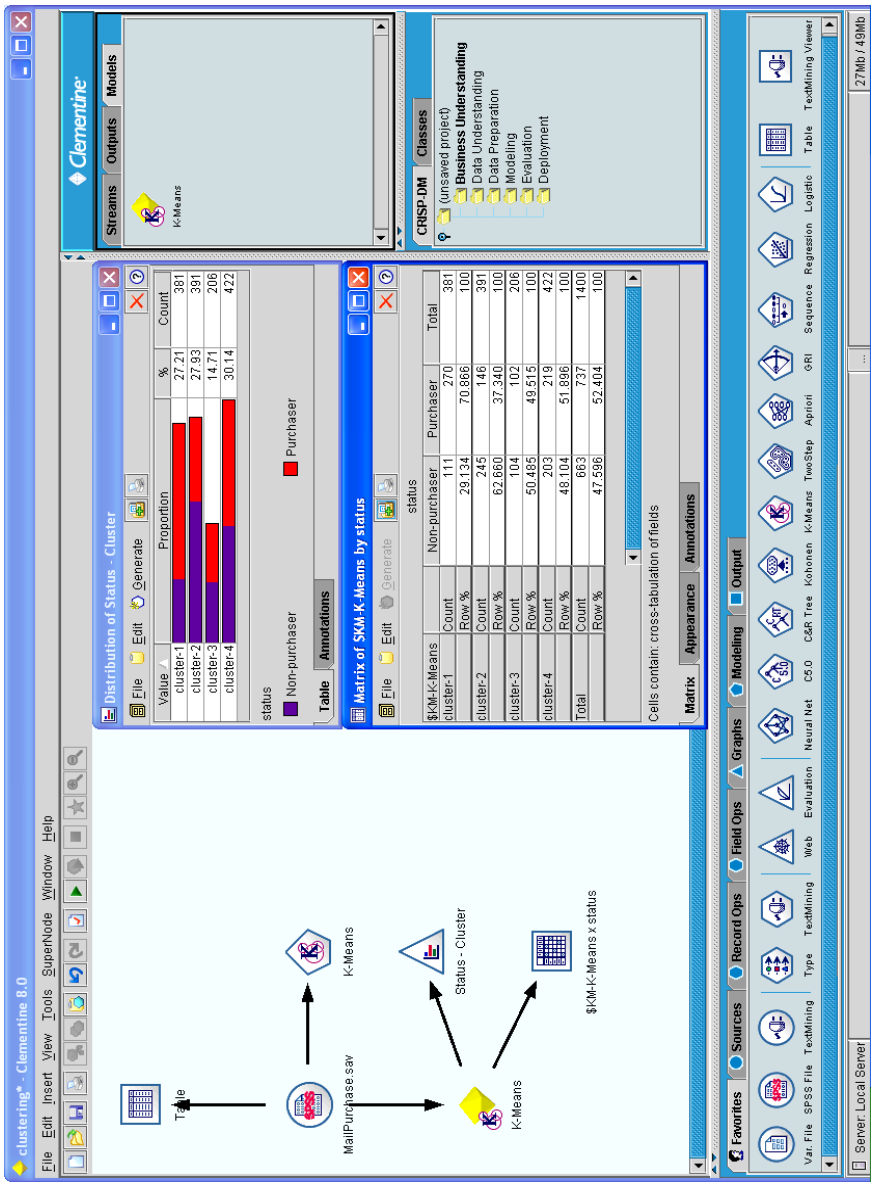

Figure 2.10 Distribution of Purchase/Non-Purchase Status across Clusters

Association analysis tools determine what variables/items go together. They are frequently used in market basket analysis in the retail sector to identify the items that are purchased together. Association analysis can also incorporate demographic characteristics (to answer "who buy what?" questions) and the time element (to answer "when buy what?" questions). Applications of association analysis are not restricted to the retail sector. As mentioned earlier, the banking and healthcare sectors can also benefit from association analysis. In fact, association analysis can be used in any setting where items (whether they be products, processes or people) are to be grouped together based on co-occurrences.

Clustering tools attempt to discover natural groupings in data so that similar items belong to the same cluster and dissimilar items to different clusters. Clustering is frequently used in market segmentation. As such, clustering results are useful for marketing decisions such as competitive positioning and target marketing, among others. On market segmentation, there may be situations where response-based market segmentation is more appropriate. Response-based market segmentation generally refers to the association of cluster profiles with different levels of a target variable. For example, an organisation may like to know what customer profiles are associated with the non-use, light use, moderate use and heavy use of a particular product. Such analyses can be done via predictive modelling instead of clustering.

Clustering can be used in other contexts besides segmentation. One example is the identification of abnormal observations, which may be indicated by very small clusters or clusters that are very dissimilar to other clusters. Another example is the use of cluster membership to identify the incidence of an event of interest. Suppose that a clustering exercise has resulted in ten clusters and it is observed that Cluster 3 and Cluster 7 have a high incidence of fraudulent transactions. This may suggest that observations belonging to these two clusters should be more carefully

monitored. Also, new observations can be assigned to the ten clusters based on the distance of the observations from the ten cluster centroids. An observation is then assigned to the cluster with the smallest distance.

Finally, readers who are interested to know more about data mining tools, association analysis and clustering may like to take a look at Chung and Gray (1999), Kantardzic (2003), Koh and Leong (2001), Lin et al. (2003), Peacock (1998), and Smith and Ng (2003).

Chapter References

Berry, M. J. A. and Linoff, G. S. (1997), *Data Mining Techniques: For Marketing, Sales, and Customer Support*. John Wiley & Sons, Inc., New York.

Berry, M. J. A. and Linoff, G. S. (2000), *Mastering Data Mining: The Art and Science of Customer Relationship Management*. John Wiley & Sons, Inc., New York.

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

Kantardzic, M. (2003), *Data Mining: Concepts, Models, Methods, and Algorithms*. Wiley Inter-Science, New Jersey.

Koh, H. C. and Leong, S. K., (2001), "Data mining applications in the context of casemix", *Annals, Academy of Medicine*, Vol. 30 No. 4 (Supplement), pp. 41-49.

Lin, Q. Y., Chen, Y. L., Chen, J. S. and Chen, Y. C. (2003), "Mining inter-organizational retailing knowledge for an alliance formed by competitive firms", *Information & Management*, Vol. 40 No. 5, pp. 431-442.

Peacock, P. R. (1998), "Data mining in marketing: Part 1", *Marketing Management* Vol. 6 No. 4, pp. 9-18.

Smith, K. A. and Ng, A. (2003), "Web page clustering using a self-organizing map of user navigation patterns", *Decision Support Systems*, Vol. 35 No. 2, pp. 245-256.