Modeling and evaluation of knowledge discovery in wholesale and retail industry

By
Tokuyo Mizuhara

A thesis submitted to
Saint Mary's University, Halifax, Nova Scotia
in Partial Fulfillment of the Requirements for
the Degree of Master of Applied Science (in Computer Science).

September 2009, Halifax, Nova Scotia

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-55991-8
Our file Notre référence
ISBN: 978-0-494-55991-8

Certification

Modeling and Evaluation of Knowledge Discovery in Wholesale and Retail Industry

by

Tokuyo Mizuhara

A Thesis Submitted to Saint Mary's University, Halifax, Nova Scotia,
in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Applied Science

September 1, 2009, Halifax, Nova Scotia

Examining Committee:

Approved:	Dr. Jingtao Yao, External Examiner Department of Computer Science, University of Regina
Approved:	Dr. Pawan Lingras, Senior Supervisor Department of Mathematics and Computing Science
Approved:	Dr. Sageev Oore, Supervisory Committee Member Department of Mathematics and Computing Science
Approved:	Dr. Dawn Jutla, Supervisory Committee Member Department of Finance, Information Systems and Management Science
Approved:	Dr. Hai Wang, Program Representative
Approved:	Dr. Hugh Broders, Graduate Studies Representative

ACKNOWLEDGEMENTS

Sincere gratitude is given to all the people who made this thesis possible!

It is difficult to overstate my gratitude to my supervisor, Dr. Pawan Lingras. Without his knowledge, patience, inspiration and his great efforts to explain things clearly and simply, this thesis could not have been completed. He conducted me through this interesting topic and guided me on the way of doing research. His instruction during my thesis writing period has benefitted me in many ways, which includes summarizing contents from different perspectives and highlighting important points.

I gratefully acknowledge the examining committee members, Dr. Dawn Jutla, Dr. Sageev Oore, and the external examiner, Dr. JingTao Yao. Their involvement on the research helped me to open my mind to look wider and further. Furthermore, I convey my appreciation to my friends who did proof reading for my thesis. Their comments and supports assisted me with this work through these remarkable and pleasant years.

I would like to thank the wholesale and the retail company which provided data for my study. I also want to highlight my thanks to Department of Mathematics and Computing Science, Faculty of Graduate Studies and Research, and Faculty of Science for providing me the opportunity to study at Saint Mary's University and the financial support. I really appreciate the help from people in the Department of Mathematics and Computing Science.

Finally, I would like to thank my parents, friends and my fiancée Xiyu Xue for their unconditional support, patience and constant encouragements, which helped me to overcome every difficulty and complete this thesis.

Abstract

Modeling and evaluation of knowledge discovery in wholesale and retail industry

By Tokuyo Mizuhara

Abstract: This thesis demonstrates an enterprise-wide Knowledge Discovery in Databases (KDD) process CRISP for wholesale and retail industry, which can facilitate business decision-making processes and improve corporate profits. While part of the KDD process described here is well documented, the modeling and evaluations used in the commercial products is not reported in literature. Hence, the focus of this thesis is on the development and evaluation of models used in the knowledge discovery. Description of the underlying models will help the decision makers better understand the quality and limitations of the KDD process.

The usefulness of KDD process CRISP is illustrated for two companies, i.e. a multinational retailer and a small chain of specialty grocery stores. The detailed steps highlight business understanding, data exploration, data preparation, data modeling, results evaluation, and interpretation. The methodologies applied in this thesis include prediction, clustering and association to discover knowledge about products/suppliers, consumers, and business units.

Date: September 1st, 2009

Contents

List of Figures

Chapter 1

Introduction

1.1 The overview of data mining and its applications

We are in an "Information Age". Industry, government and organization are collecting and managing their data at a dramatic pace [1] [2] [3]. The growth in data storage and computing capability has put pressures on corporations to make use of information that had not previously been available. As a result, gathering data has become fairly easy and inexpensive, yet large-scale data analysis remains a formidable challenge. Manual analysis of massive datasets is not practical; therefore, data analysis techniques must be capable of extracting useful information that may be acted upon by decision makers [1]. This widely recognized requirement has lead to a rapid development of a new academic field – Knowledge Discovery in Databases (KDD), defined by Fayyad in 1996. KDD is based on other fields, i.e. machine learning, pattern recognition [4], database systems [5], statistics [6], artificial intelligence, data visualization, and high-performance computing [7]. Typically, KDD systems use algorithms and techniques from these fields to extract patterns and models from data.

Fayyad defines KDD as "*the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data*" [8]. KDD is also known as data mining. These terms represent the overall process of seeking reproducible patterns and unsuspected relationships called “knowledge nuggets” from datasets (databases). The nuggets represent some aspect of data that is both understandable and useful to end-users.

1.2 Data mining in business

The business value of KDD techniques stems from its role in decision making. Companies increase their competitive advantage by addressing business issues, or uncovering useful information that is otherwise undetectable to decision makers. Data mining processes provide the framework to achieving these goals with efficiency and accuracy while presenting data in novel ways that help further the aims of the company. Such a broad usefulness emphasizes why KDD is being applied to almost any industry imaginable whether it be banking, finance, telecommunications, manufacturing, medicine, internet or other fields [7] [9].

The widespread adoption of KDD has resulted in the development of specific commercial applications that each respond to unique business requirements. Examples are intelligent miner from IBM [10]; SAS Enterprise Miner from SAS (http://www.sas.com/); Clementine from SPSS (http://www.spss.com/); SQL Server Business Intelligence Development Studio from Microsoft [11]. However, even though data mining applications are very powerful tools, they are not self-sufficient applications [12]. Data mining process requires both domain knowledge as well as analytical skills in order for the user to define business problems, structure the analysis, and interpret the results. This can be a challenge.

1.3 Conventional data mining process

CRISP-DM (CRoss Industry Standard Process) is a well-known KDD process (http://www.crisp-dm.org/). It was proposed by SPSS in late 1996 and aims to be an industry tool and an application-neutral model [13]. It does not depend on a particular data mining technique, but describes the process of data mining project's life cycle.

CRISP-DM divides data mining process into 6 phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [13] [14] [15]. These phases help domain experts understand the KDD process and provide a road map to follow while planning and carrying out a KDD project. The process is shown in Figure 1-1.

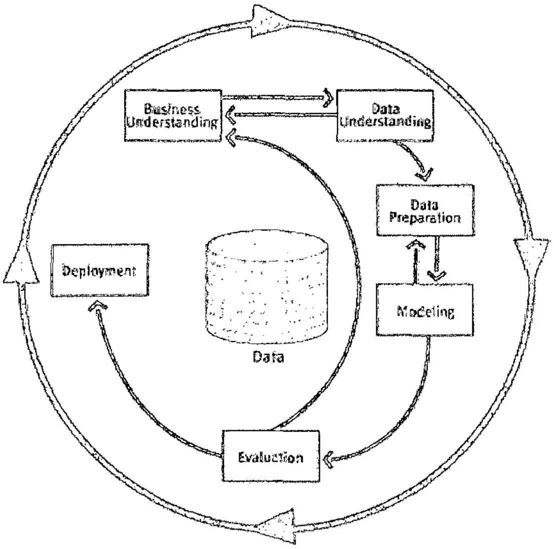


Figure 1-1: Phases of the CRISP-DM process model [13]

Business understanding

Before application, data mining experts should begin by understanding the aims for which data mining is intended to achieve. Knowing the domain's characteristics and purpose is a first step in the application of the KDD process. KDD experts need to learn things about the domain from a user's points of view, since any knowledge that does not line up with the business goal is useless.

Data understanding

In order to discover all available information and to process different activities, knowing and understanding the data is the only way that KDD and domain experts are able to constitute the KDD project vision and generate KDD plans. The quality of the chosen data determines the outcome of the process.

Data preparation

Data preparation includes data extraction, data reduction, data cleaning, data integration and data transformation.

Often times, the data mining algorithms cannot be applied directly to the databases. The databases are normally accessed by multiple users. The day-to-day work should not be interrupted by the application of the KDD. For safety and security reasons, target datasets need to be created in order to apply the process.

As soon as target datasets are generated, the experts need to spend time exploring rough data to inspect the KDD subject oriented data. For the attributes which are not related to this study, they need to be deducted for saving physical space and extracting the KDD subject oriented data. The deduction of unnecessary attributes also emphasizes the focus.

However, data selection needs to be reviewed several times by domain experts. Data aggregation, data dimension reduction and data compression are included in this area.

The next step is data cleaning. Data from real world usually have incomplete, noisy, unknown values or inconsistent data. The improper data might be caused by operation error or improper system implementation flows. Applying appropriate treatment to such data can ensure the quantity of data mining. Some treatments will be presented in chapter 2.

Data integration involves combining data from multiple resources into a coherent data store. The sources may include multiple databases or flat files [16]. The data form transformation, which is able to transfer data into necessary format, is also included in this phase.

Modeling

After data preparation, the data experts should select suitable mining techniques and algorithms. Based on its goals, the experts are able to pick up one or more techniques from association, classification, prediction and clustering. How to build the model not only depends on the goals, but also depends on the available data. It is a comprehensive and vital point of data mining. It directly decides how close data mining outputs to real data.

Evaluation

At this stage, the KDD experts, together with the domain experts interpret the discovered information. If the results are satisfactory, one can move to the next step. If the results are not satisfactory, the experts will be required to go back to the first stage, review the business goal, and reconstruct the model to ensure that it achieves its business objectives. An important objective is to determine if there is an important business issue that has not been fully considered. At the end of this phase, a decision on the use of the KDD outcome needs to be reached.

Deployment

The goal of the KDD process is not simply to discover information. The knowledge gained should be structured and presented in a practical way.

Even though CRISP-DM described the road map of KDD, some challenges still exist and obstruct the spread of KDD in the business world.

1.4 Objectives of the thesis

As described in session 1.2, some powerful data mining tools already exist i.e. SAS or SPSS. However, due to the complexity of data mining process, Fayyad pointed out that only a few E-commerce companies such as Amazon and Google have been successful in practical use of KDD as of year 2003. The reason for limited use of KDD is because it needs significant investment of company resources [2]. Large amounts of data are still not utilized. There are some emerging IT consulting organizations that provide data mining services to wholesale and retail companies. However, these companies do not elaborate on the modeling and evaluation of the data mining processes. Only data mining results are available for end-users. Once a wholesale and retail company has signed up with the consultant, then the association will necessarily have to continue for a long period of time, because data mining is an evolving process and knowledge changes as time goes by. Once the latest data is available, data mining has to be repeated and results have to be updated. The cost of such data mining consultancy can be prohibitive for small to medium wholesale and retail companies.

This thesis describes an elaborate generic enterprise-wide data mining process based on the CRISP model for wholesale and retail businesses. For completeness, this thesis will demonstrate the complete data mining process for a wholesale and retail organization that includes data preparation, modeling, and evaluation. However, the focus of this thesis is on the modeling and evaluation phases. It is hoped that knowledge engineers will be able to employ various modeling possibilities described in this thesis for improving their sales operations. This thesis demonstrates the feasibility of the proposed modeling and evaluation phases for two businesses in the wholesale and retail sector. The process described in this thesis covers customer relationship management, inventory management and supplier analysis. Clustering, prediction and association data mining techniques are applied in this study.

1.5 Organization of the thesis

This study focuses on how KDD will be able to help wholesale and retail operations. In chapter 2, the general data mining methods and techniques are presented. Data preparation, clustering, classification, prediction, association and results evaluation techniques are included. Applications of data mining are also presented through two case studies in this chapter.

In chapter 3, the elaboration on CRISP data mining process for wholesale and retail operations is described with a particular emphasis on modeling and evaluation phases. The overview of wholesale and retail industry and the outline of data mining applied in this industry will also be described. Furthermore, this chapter also introduces two wholesale and retail companies which provided data for this study.

Chapter 4 demonstrates the data mining process described in chapter 3 for a multinational wholesale and retail business. A comprehensive analysis of results is also given at the end of this chapter.

Chapter 5 confirms the generic applicability of the proposed data mining process for a smaller specialty retailer. The application of the proposed elaboration of CRISP data mining process for two widely different types of sales organization lends credibility to the process.

Chapter 6 discusses the results of our data mining studies. A comprehensive overview of the knowledge discovery possibilities for sales organizations will be discussed in this chapter. Summary, conclusions and directions for further research are also provided.

#

Chapter 2

Review of data mining methods and techniques

The key to data mining success involves discovering knowledge from large target datasets to identify proper techniques. Clustering, classification, prediction and association are the most popular and powerful methods in the data mining field. To deal with vague or imprecise concepts; rough sets and fuzzy sets are also applied in this field [1] [7] [18] [19] [20] [21]. In order to analyze datasets effectively and interpret analysis results graphically, data visualization techniques can be used to handle various complex datasets [36]. These techniques have been studied by statistics, machine learning, pattern recognition, and database researchers [22]. This chapter will present and explain several major data mining techniques, along with data preparation and results evaluation. The details of rough sets and fuzzy sets can be found in [1] [7] [18] [19] [20] [21].

2.1 Data preparation

Real world data tends to be incomplete (lacking attribute values or certain attribute of interest), noisy (containing errors or outliers) and inconsistent (e.g. containing discrepancies in the item codes used for different departments) [16]. Therefore, in order to prepare real world data for data mining purposes, the collected and extracted data has to undergo data reduction, data cleaning, data integration, and data transformation.

2.1.1 Data reduction

Data mining can discover useful information from large historical datasets. To accelerate the data mining process, the data size has to be reduced without changing the data mining results. This process consists of data aggregation, data dimension reduction, data compression and numerosity reduction [16] [25].

Data aggregation summarizes raw data. For example, the daily sales data may be aggregated to monthly or annual total sales amounts. Depending on the purpose of data mining, aggregation can represent the same information with smaller data size.

Dimension reduction allows data mining experts to remove subject-irrelevant attributes and redundant attributes. For example, the sales revenues before taxes and sales revenues after taxes can be considered as redundant attributes in data mining process as long as tax rates are not changed. Furthermore, data compression can be applied to reduce data by approximately reconstructing data. It allows reasonable information loss in the reconstruction process. Depending on the data mining needs, experts would use numbers that are larger or smaller than user specified threshold. The remaining numbers could be replaced with 0. Wavelet transforms are popular in the area of data compression. Details can be found in [16].

Histogram and sampling are two well-known statistical techniques used to achieve numerosity reduction [16]. Histogram counts the frequency of observations between a specified bin and previous one. It also displays the cumulative frequency of observations in all of the bins up to the specified bins. Histogram describes the distribution of the error rate efficiently and is very powerful at evaluating results. Sampling can be used as a data reduction technique as it allows large datasets to be represented by smaller random samples of subsets [45] [46].

2.1.2 Data cleaning

Data cleaning routines work to clean the data by filling in missing values, removing noisy records while identifying outliers and correcting inconsistencies [23]. Some of the possible data cleaning treatments include [7] [24]:

- Omit the record: If a record misses a particular value, then omit that record from the datasets
- Fill with a default value: For every attribute, decide a default value in advance. The default value is always used whenever missing data occurs. However, the missing data must not make up a large portion of the entire data. Otherwise, it will confuse the data mining process and lead to incorrect results. There is no easy way for a data mining technique to distinguish between the impure data and the original data.
- Fill with average data: The missing value is calculated by using the average value of an attribute or the average value of the previous value and next one.
- Fill with the most possible value: Predict the missing value using regression analysis. This treatment also biases the original data. The percentage of filled data should not be making up the majority of the entire dataset.

Noise data, which also indicates outliers, can be detected by a clustering technique. Similar data points are grouped into same clusters. Values that do not belong to main clusters can be considered as outliers [16].

The inconsistent data is caused by operational mistakes and faulty system flow design. Therefore, the correction of inconsistent data needs to follow a consistent system flow. Several iterations might be necessary to complete this step.

2.1.3 Data integration

Data mining often requires data integration, which combines data from multiple data sources. Data integration includes schema integration, detecting and resolving data value conflicts, removing duplicates and redundant data [14] [16].

Identifying entities in several sources that refer to the same entity by data mining experts is a great concern in data mining process. For instance, data mining experts may have to conclude whether *cust_id* in one database and *customer_number* in another refer to the same entity. The data schema and relationship table can be used in this step for identification. Moreover, the data integration plan should be reviewed and approved by domain experts before actual implementation.

2.1.4 Data transformation

During data transformation, data are transformed into proper forms for the data mining application to use. The transformation process includes normalization, attribute construction, aggregation and generalization [16] [25].

Data normalization scales the data down to a small specified range, such as -1.0 to 1.0 or 0.0 to 1.0. The min-max normalization (refer to equation 2-1) is a common way to achieve this goal.

$$normalized\ x = \frac{x \text{ - min}}{\text{max - min}} \qquad \text{EQUATION 2-1}$$

Data attribute construction can add and construct new attributes from the given datasets to help the data mining process. For example, profits attribute can be calculated based on sales amount and costs in a given dataset.

Data aggregation (refer to section 2.1.1) can also be applied during the data transformation. Data generalization brings low-level data or raw data to higher-level concepts through the use of concept hierarchies. For example, *age* might be mapped to higher level concepts, like *young, middle-aged,* and *senior.*

2.2 Clustering

Clustering is a convenient technique for discovering data distribution and patterns in raw data. It partitions data into groups based on data similarity. This is called unsupervised learning; which does not require any predefined data records [22].

The clustering technique will be explained using Figure 2-1. Let us assume that nine given points need to be divided into certain groups depending on their characteristics. Based on their patterns, they can be divided into three groups; based on their shapes, they can be divided into another three groups. Both ways are correct. This technique can be applied to divide data into different groups based on different data attributes.

Clustering is widely used in wholesale and retail companies for marketing. In order for the companies to develop their marketing strategies for further business developments, the managers of companies must be very conscious of the characteristics of their customers, products and suppliers [26].

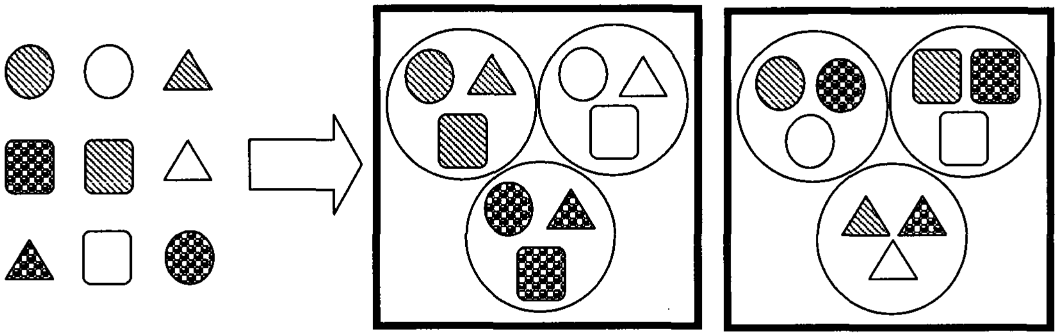


Figure 2-1: Clustering technique

2.2.1 K-means algorithm

K-Means algorithm is a commonly used clustering algorithm in practice [23]. The algorithm proceeds as follows.

Step 1: Ask the user how many clusters, *k*, the datasets should be partitioned into.

Step2: Randomly assign *k* data points to be the initial cluster center locations.

Step3: For each data point, discover the nearest cluster center. Thus, in a sense, each cluster center has a subset of records, thereby representing a partition of the datasets. Therefore, I have *k* clusters, C_1, $C_{2,}$ ..., C_k.

The distance calculated using the equation 2-2 estimates the nearest cluster center:

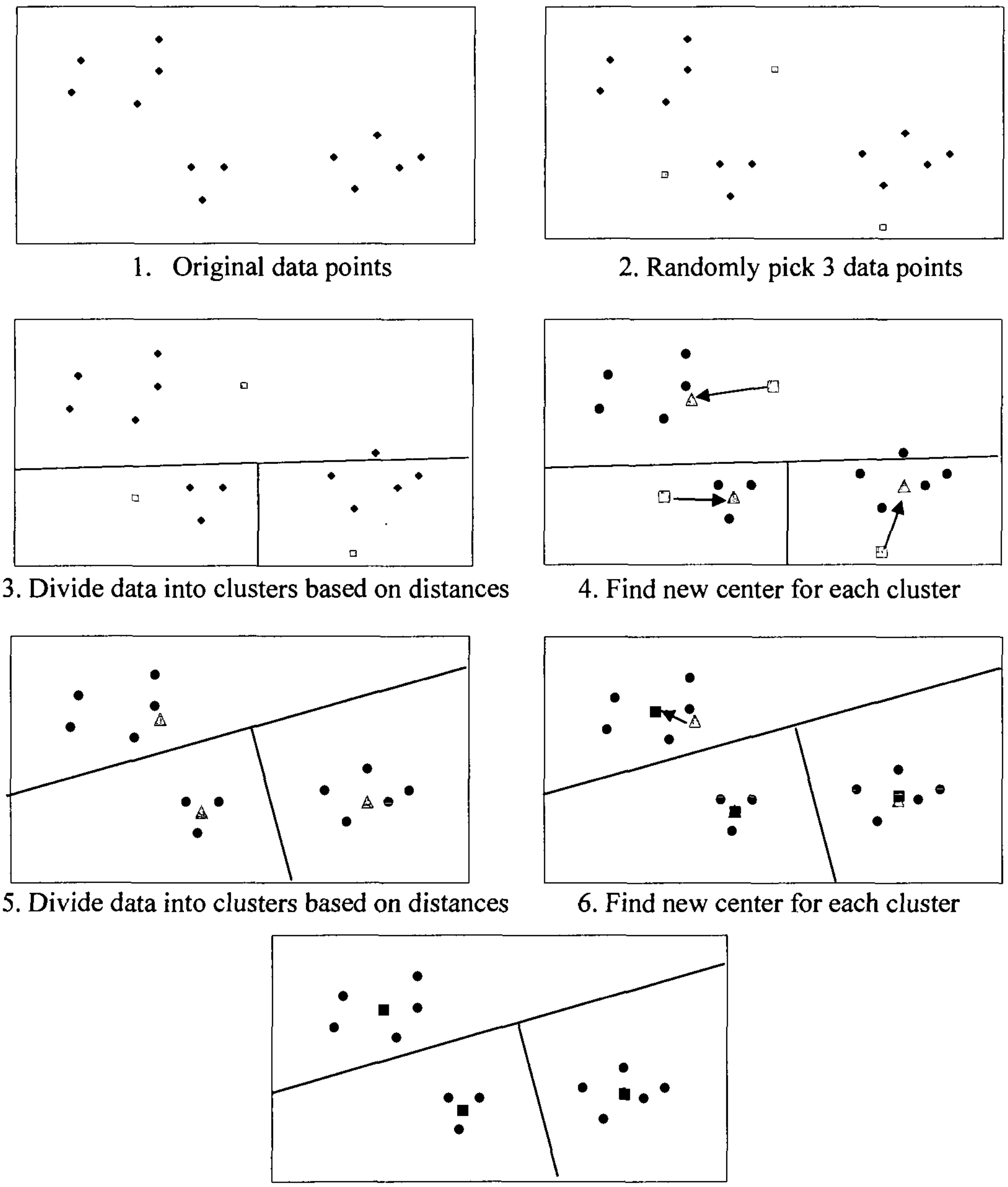
$$d = \|x_i - c_j\|^2 \qquad \text{EQUATION 2-2}$$

where x_i is a data point and c_j is the cluster center, which corresponds to the minimum of d. c_j represents the cluster center that x_i belongs to. Hence, the j is the cluster number that the data point belongs to.

Step4: For each of the k clusters, find out the new cluster center. Suppose that I have n data points (a_1, b_1, c_1), (a_2, b_2, c_2),..., (a_n, b_n, c_n), the new cluster center is the mean of these data points, which is $(\sum a_i / n, \sum b_i / n, \sum c_i / n)$.

Step5: Repeat step 3 to step 5 until the locations of cluster centers have no more changes.

Figure 2-2 graphically illustrates the K-means algorithm.



1. Original data points

2. Randomly pick 3 data points

3. Divide data into clusters based on distances

4. Find new center for each cluster

5. Divide data into clusters based on distances

6. Find new center for each cluster

7. Divide data into clusters based on distances. The solution is stable, so the algorithm terminates

Figure 2-2: K-means algorithm

2.2.2 Kohonen self organizing maps

In 1979, Teuvo Kohonen introduced Kohonen networks based on the Neural Network model [27]. It represents one of the most popular models in the unsupervised studies; also named self-organizing map (SOM) [22]. The goal of SOM is to convert a complex high-dimensional input signal into a simpler low-dimensional discrete map [28]. It consists of an input layer and an output layer (Figure 2-3). The input units are fully connected to output units.

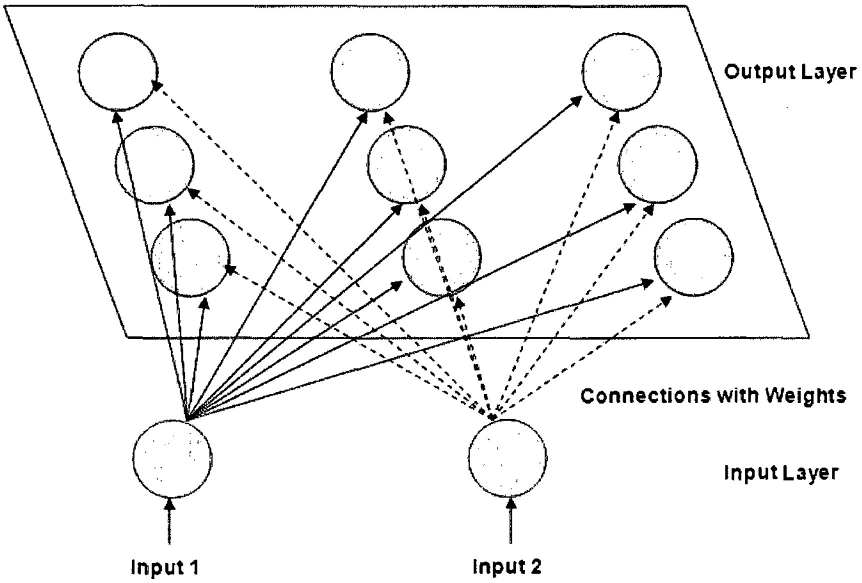


Figure 2-3: Topology of a simple self-organizing map [23]

The algorithm proceeds as follows [22] [23].

Step I: For each output node *j*, use equation 2-3 to calculate the value $d(w_j, x_n)$ of the scoring function while *i* represents input node.

$$d(w_j, x_n) = \sqrt{\sum_i (w_{ij} - x_{ni})^2} \qquad \text{EQUATION 2-3}$$

Find the winning node *J* that minimizes $d(w_j, x_n)$ over all output nodes.

Step2: Adjust the weights as:

$$w_{ij,new} = w_{ij,current} + h_{ck}(j) \times (x_{ni} - w_{ij,current}) \qquad \text{EQUATION 2-4}$$

The $h_{ck}(j)$ represents neighborhood function including the learning rate and χ_{ni} signifies the nth input to node j.

Step 3: Repeat step 1 to step 2 for all the objects. Adjust the neighborhood function.

Step 4: Repeat Steps 1 to 3 for a specified number of epochs.

The input data of SOM needs to be normalized to 0 and 1. The min-max normalization (equation 2-1) can be applied here.

2.3 Classification

Classification is considered to be one of the important techniques in data mining. It is well known as supervised learning which is based on the analysis of a set of predefined data. The output for every data record in training data is known.

Classification is able to describe and separate data into one of several predefined classes [16] [22]. The classification model is trained on a known set of identified target sets. Once the model has been trained, it can be used to divide unknown data into proper classes. Decision tree, Bayesian network, support vector machine (SVM) and neural network are common algorithms used in classification. Classification techniques can help market researches and business decision-making. Some examples of classification tasks in business are shown below:

- ❖ Classifying the customers of a store into two classes based on their spending: customers who spend more than $25 dollars and those who spend less than $25 dollars.

- Determining whether a particular credit applicant is a high risk customer
- Assessing wholesales discount rates for eligible customers
- Identifying whether or not a new product will be beneficial in a particular time period

2.3.1 Decision trees

Decision trees are powerful and popular tools for classification and prediction. They can discover rules to classify data properly through a set of mathematical calculations. Rules can easily be expressed so that humans can understand them [22]. Decision tree has a flow-chart-like tree structure. We will outline a commonly used algorithm named C4.5 in this section.

C4.5 algorithm is developed based on ID3 algorithm [29]. It contains leaf nodes indicating class values and decision notes that specify conditional tests. It can be used to classify an instance by starting at the root of the tree and moving through it down to a leaf node that provides a classification rule. The set of records available for developing such classification methods is generally divided into two disjoint subsets- a *training set* and a *test set*. The *training set* is used for calculating the classifier while the *test set* is used to measure the accuracy of the classifier. The accuracy of the classifier is determined by the percentage of the test examples that are correctly classified.

The C4.5 algorithm uses the concept of *information gain* or *entropy reduction* to select the optimal split [23]. Entropy provides an information-theoretic approach to measure the goodness of a split. Suppose I am given the probability distribution $P = (p_1, p_2, \ldots, p_n)$, then the information communicated by this distribution, called the entropy [22] of P is

$$Entropy(P) = -p_1 \times \log(p_1) - p_2 \times \log(p_2) - \ldots - p_n \times \log(p_n)$$

EQUATION 2-5

For example, if P is (0.5, 0.5), then

$$Entropy(P) = -0.5 \times \log(0.5) - 0.5 \times \log(0.5) = 1$$ EQUATION 2-6

If P is (0.67, 0.33), then

$$Entropy(P) = -0.67 \times \log(0.67) - 0.33 \times \log(0.33) = 0.92$$ EQUATION 2-7

If P is (1, 0), then

$$Entropy(P) = -1 \times \log(1) - 0 \times \log(0) = 0$$ EQUATION 2-8

Equation 2-6, 2-7 and 2-8 indicate the probability distribution is directly proportional to its entropy.

In the context of decision trees, if a set of records T is partitioned into a set of disjoint exhaustive classes C_1, C_2,..., C_n, then the information needed to identify the class of an element of T is

$$Info(T) = Entropy(P)$$ EQUATION 2-9

where P is the probability distribution of the partition C_1, C_2,..., C_n. P is computed based on their relative frequencies, i.e.

$$P = \left[\frac{|c_1|}{|T|}, \frac{|c_2|}{|T|}, \ldots, \frac{|c_n|}{|T|}\right]$$ EQUATION 2-10

If T is partitioned based on a test attribute X, into subsets T_1, T_2,..., T_n, then the information based on the partitioning into subsets by X, is given by $Info(X,T)$:

$$Info(X,T) = \sum_{i=1}^{n}\left(\frac{T_i}{T} \times Info(T_i)\right)$$

EQUATION 2-11

To illustrate how *Info*(*T*) and *Info*(*X,T*) are calculated, consider a dataset that has three distinct classes, C_1, C_2 *and* C_3 and these classes have forty, thirty and thirty objects respectively. The information of the whole data is

$$Info(T) = Entropy(P) = -\frac{40}{100} \times \log\left(\frac{40}{100}\right) - \frac{30}{100} \times \log\left(\frac{30}{100}\right) - \frac{30}{100} \times \log\left(\frac{30}{100}\right) = 1.57$$

EQUATION 2-12

Now, let us consider splitting the dataset into two subsets, T_1 and T_2, with n_1 and n_2 number of records respectively, where $n_1 + n_{2=} n$. If I assume n_1=60 and n_2=40, the splitting is given by:

S_1	C_1	C_2	C_3
60	40	10	10

S_2	C_1	C_2	C_3
40	0	20	20

The information value *Info*(*X,T*) after the partition is:

$$\frac{60}{100} \times \left[-\frac{40}{60} \times \log\left(\frac{40}{60}\right) - \frac{10}{60} \times \log\left(\frac{10}{60}\right) - \frac{10}{60} \times \log\left(\frac{10}{60}\right) \right] + \frac{40}{100} \times \left[-\frac{0}{40} \times \log\left(\frac{0}{40}\right) - \frac{20}{40} \times \log\left(\frac{20}{40}\right) - \frac{20}{40} \times \log\left(\frac{20}{40}\right) \right] = 1.15$$

EQUATION 2-13

The *information gain* due to a split on attribute *X* is as follows:

$$Gain(X,T) = Info(T) - Info(X,T)$$

EQUATION 2-14

Information gain indicates the increase in information produced by partitioning the training data *T* according to *X*. For different attribute *X*, different values of *information gain* are generated. At each decision node, C4.5 chooses the optimal split to be the split with the greatest information gain.

The most attractive aspect of C4.5 is the interpretability of decision trees, especially with respect to construction of decision rules. Decision rules can be constructed from a decision tree simply by traversing any given path from the root node to any leaf. There are two measures, *support* and *confidence* that indicate the usefulness of a rule. *Support* refers to the proportion of records in the datasets that match the rule that particular terminal leaf node represent. *Confidence* represents how likely the rule is true [23].

2.3.2 Bayesian network

Bayesian classifiers are statistical classifiers. They can divide members depending on probabilities, based on Bayes theorem. A simple Bayesian classifier known as the naive Bayesian classifier is presented in this section. It assumes that, in a given class, the effect of an attribute is independent from the values of the other attributes. This assumption is called *class conditional independence* [16].

2.3.2.1 Bayes theorem [16]

Let X be a data sample whose class label is unknown. Let H be some hypothesis, such that X belongs to class C. For classification purposes, I want to determine $P(H|X)$, the probability that the hypothesis H holds given the observed data sample X. Suppose X represents a male customer and H is the hypothesis that X's age is over thirty years old. $P(H|X)$ represents the confidence that X is over thirty given that X is a male customer. On the other hand, $P(H)$ represents the probability that every customer is over thirty regardless of gender. $P(X)$ indicates the probability that X is a male customer while $P(X|H)$ represents the probability that a customer is a male given that the customer's age is over thirty. Hence, the desired parameter, $P(H|X)$ can be estimated by given data as:

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)} \qquad \text{EQUATION 2-15}$$

2.3.2.2 Naïve Bayesian classification

The Naïve Bayesian classifier or simple Bayesian classifier includes five steps [16].

1. Let D be a training set of data and their associated class label. Each data element is represented by an n attribute d given by:

$$D = (d_1, d_2, ..., d_n) \qquad \text{EQUATION 2-16}$$

2. Suppose there are m classes, C_1, C_2,..., Cm. Given an unknown data sample, X (have no class label), the classifier will predict whether X belongs to the highest probability class. By the Bayes theorem:

$$P(C_i|X) = \frac{P(X|C_i) \times P(C_i)}{P(X)} \qquad \text{EQUATION 2-17}$$

3. As $P(X)$ is constant for all classes, only $P(X|C_i) \times P(C_i)$ need be maximized. If the value of $P(C_i)$ is unknown, then it is commonly assumed that the classes are equal, that is, $P(C_1) = P(C_2) = ... = P(C_m)$, and I would therefore maximize $P(X|C_i)$. Otherwise, I maximize $P(X|C_i) \times P(C_i)$.

4. Assume the attributes are conditionally independent, thus:

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i) \qquad \text{EQUATION 2-18}$$

5. In order to classify an unknown sample X, $P(X|C_i) \times P(C_i)$ is evaluated for each class C_i. Sample X is then assigned to the class C_i if and only if

$$P(X|C_i) \times P(C_i) > P(X|C_j) \times P(C_j) \quad \text{for } 1 \leq j \leq m, j \neq i \qquad \text{EQUATION 2-19}$$

In other words, it is assigned to the class C_i for which $P(X|C_i) \times P(C_i)$ is maximum.

2.3.3 Neural network classifier

Neural network is very different from other algorithms in the sense that it does not follow any statistical distribution. It is modeled after the function of human brain. It contains three layers, which are input, hidden and output, and weights associated with each connection (Figure 2-4). The algorithm this thesis describes in the following is called Back Propagation Neural Network.

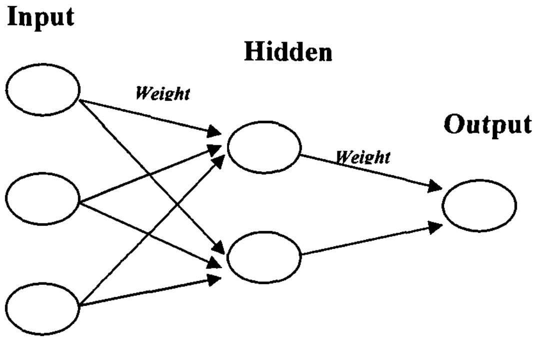


Figure 2-4: The structure of simple Neural Network

The neural network is trained by pre-defined data. During the iterated training, weights can be adjusted by comparing the actual results and output of the network. The training process ends when a predefined minimum error level is reached. Input data of neural network needs to be normalized between 0 and 1. The min-max normalization can be used here.

The first step of Back Propagation Neural Network is to produce a linear combination of the inputs and connection weights into scalar values through a combination function. The scalar values are called *net* [23]. Thus, for a given node *j*,

$$net_j = \sum_{i=0}^{I} W_{ij} x_{ij} \qquad \text{EQUATION 2-20}$$

where x_{ij} represents the ith input to node j, w_{ij} represents the weight associated with the ith input to node j. There are $I+1$ inputs to node j while x_0 represents a constant input. Based on the sigmoid function, the output of net j is computed as follow:

$$Output_j = \frac{1}{1+e^{-net_j}} \qquad \text{EQUATION 2-21}$$

This process is iterated until the final output is calculated. The next step is to determine whether the final output belongs to the right class. Suppose I have two classes whose mean values are 0 (class 0) and 1 (class 1), and the threshold to separate two classes is 0.5. If the final output is larger than 0.5, then I consider the input belongs to class 1. Otherwise, I classify under class 0. This is called the activation function. If the classification result for the input is not correct, then the weights need to be adjusted. The new weights can be derived as follows:

$$W_{ij,new} = W_{ij,current} + \Delta W_{ij} \qquad \text{where } \Delta W_{ij} = \eta \delta_j x_{ij} \qquad \text{EQUATION 2-22}$$

where w_{ij} is the weight of the connection from the ith input to node j; η represents the learning rate and χ_{ij} signifies the ith input to node j. δ which represents the responsibility for a particular error belonging to node j is defined as such:

$$\delta_j = \begin{cases} output_j(1-output_j)(actual_j - output_j) & \textit{for output layer nodes} \\ \dfrac{W_{jk}\delta_k}{\sum W_{ik}} & \textit{for hidden layer nodes} \\ \textit{downstream} & \end{cases}$$

EQUATION 2-23

where W_{jk} is the weight of connection from unit j to a unit k in next higher layer and δ_k is the error of unit k; $\sum W_{jk}$ indicates the sum of all connections' weights which connect to the unit k in next higher layer.

The training stops when [16]:

- All Δw_{ij} in the previous training iteration were below a specified threshold, or
- The percentage of samples misclassified in the previous training iteration is below a specified threshold, or
- A pre-specified number of iteration has expired.

2.3.4 Support vector machines

Support Vector Machines (SVMs) are new classification methods for both linear and nonlinear data. They were proposed by Vapnik to overcome the linearly separable restriction [30][31][32]. They use a nonlinear mapping to transform the original training data into a higher dimension (Figure 2-5). In the new dimension, they classify the data points into two classes with the maximum distance from both classes. Even SVMs are essentially binary classifiers, several methods have been suggested for extending SVMs for multi-classification, including *one-versus-one* (1-V-1), *one-versus-rest* (1-V-R) and DAGSVM [33]. SVMs can be a classification functions or general regression functions. They have been applied in a wide range of applications [34] including handwritten digit recognition [35], face detection in images [36], and text classification [37].

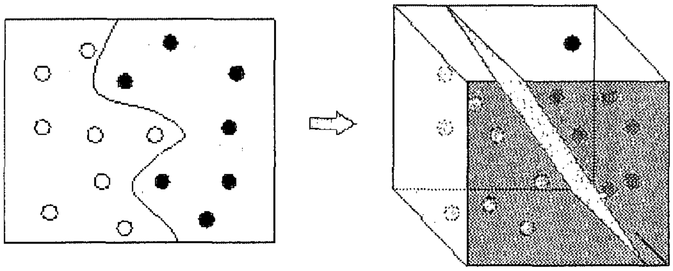


Figure 2-5: SVM Higher dimensions transfer [38]

SVM is able to map inputs into a higher dimension space via a kernel function and use the kernel function to find a hyperplane in the space of possible inputs that split the positive examples from the negative ones. The split will be chosen to have the largest distance from the hyperplane to the nearest of the positive and negative examples (Figure 2-6) [33]. Input data of SVMs needs to be normalized between 0 and 1. The min-max normalization can be used here.

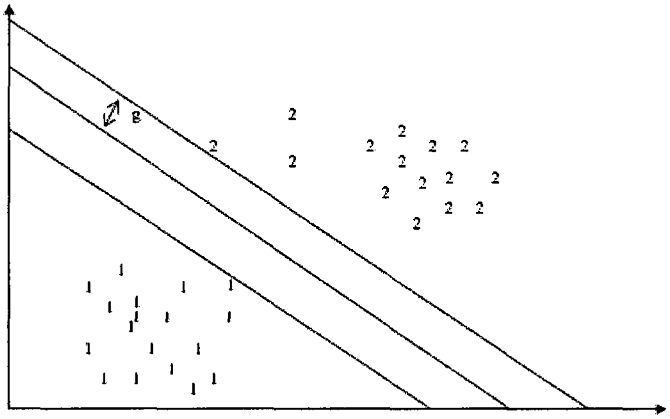


Figure 2-6: Maximizing the margin between two classes [39]

Let x be an input vector in the input space X. Let y be the output in y={-1,+1}. Let $S=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n),\ldots\}$ be the training set used for supervised classification. Let us define the inner product of two vectors x and w as $<x,w>=\sum_j x_j \times w_j$, where x_j and w_j are components of the vectors x and w.

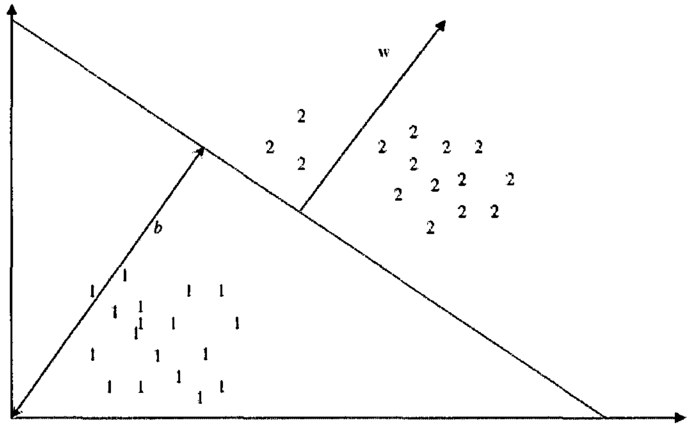


Figure 2-7: Linear separable sample [39]

Figure 2-7 shows a training sample that is linearly separable, i.e. there exists a hyperplane that separates positive and negative objects. If the training set is linearly separable, the learning algorithm will find the vector *w* such that:

$$y \times \left[\langle x, w\rangle + b\right] \geq 0, \quad \text{for all}\,(\text{x}, \text{y}) \in S \qquad \text{EQUATION 2-24}$$

If the training set is not linearly separable, then SVMs will use a mapping *Φ* to transfer the input space to another feature space with higher dimension, as illustrated in Figure 2-5. Therefore, the vector *w* is

$$y \times \left[\langle \Phi(x), \Phi(y)\rangle + b\right] \geq 0 \quad \text{for all}\,(\text{x}, \text{y}) \in S \qquad \text{EQUATION 2-25}$$

In fact, SVMs use a kernel function *K* corresponding to the inner product in the transformed feature space as [33]:

$$K(x, w) = \langle \Phi(x), \Phi(y)\rangle \qquad \text{EQUATION 2-26}$$

Polynomial kernel is one of the more popular kernel functions [39] [40].

2.4 Prediction

Prediction technique is similar to classification. While classification predicts class levels, prediction predicts continuous values. It discovers relationships between inputs and outputs from predefined data. Using the discovered relationships, it predicts future output values based on inputs. Many techniques can be used to this purpose, such as Simple Linear Regression, Multiple Linear Regression, support vector regression (SVR), Neural Network and ARIMA (Autoregressive integrated moving average). Prediction techniques can also be used for market research and business decision-making. However, the precision of the prediction is depended on how to pick a proper technique and build the optimized model. The combinations and comparisons between several techniques should also be explored. A possible task of prediction is to predict the coming year revenue for a bank branch based on historical data. There are several software packages for solving regression problems. Examples include MS Excel (http://office.microsoft.com/), SAS (http://www.sas.com), SPSS (http://www.spss.com) and S-Plus (http://www.mathsoft.com) [16].

2.4.1 Simple linear regression

Simple linear regression is an algorithm that describes the relationship of the inputs and outputs by drawing a straight line between them. Simple linear regression is one of the simplest prediction techniques. Let Y be a response variable and x be a predictor variable, then:

$$Y = \alpha + \beta x \qquad \text{EQUATION 2-27}$$

where α is a constant that indicates the Y-intercept and β is a regression coefficient specifying the slope of the line. Given s samples and data points of the form (x_1,y_1), $(x_2,y_2),\ldots, (x_s,y_s)$, regression coefficients can be estimated with the following equations:

$$\beta = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2} \qquad \text{EQUATION 2-28}$$

$$\alpha = \bar{y} - \beta\bar{x} \qquad \text{EQUATION 2-29}$$

where $\bar{x}$ is the average of $x_1, x_2,\ldots, x_s$ and $\bar{y}$ is the average of $y_1, y_2,\ldots, y_s$ [16].

2.4.2 Multiple linear regression

The Multiple Linear Regression is a stable statistical method that extends on linear regression involving more than one predictor variable [16]. The regression model takes the following form:

$$f(x) = w_0 + \sum_{n=1}^{k}(w_n x_n) \qquad \text{EQUATION 2-30}$$

x_n represents the input values and w_n represents the weights. The weights are calculated from the training data before being applied to predict future output values based on inputs. However, data has to be more than three to five times the number of inputs in order to get a reasonable model. Otherwise, it can cause over-fitting. Another benefit of regression is the ability to measure the standard error using a statistical approach. It provides a range of the predicted value along with confidence level. In the rest of the thesis, I use the term multiple regression for multiple linear regression for simplicity.

2.4.3 Support vector prediction

Support Vector Regression (SVR) is a prediction technique in the support vector family, which is gathering interest among researchers for purposes such as the Stock Market Prediction [41]. SVR extends the traditional linear regression by introducing non-linearity via kernel functions. It employs the margin concept for the regression problems with the help of ε-insensitive loss function [34]. The ε-insensitive loss function adds the notion of tolerance margin. As long as the actual data values are within ε distance from the predicted values, it is assumed that there is no prediction error. Therefore, the predictions from an ε-SVR can be looked at as an ε-tube instead of a curve. Thus, SVR is extremely efficient in handling noisy data.

Let x be an input vector in the input space X. Let y be the output in $y \in \Re$. Let $S=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n),\ldots\}$ be the training set used for supervised prediction. Let us define the inner product of two vectors x and w as $<x,w>=\sum_j x_j \times w_j$, where x_j and w_j are components of the vectors x and w. SVR overcome the linear restriction by using a kernel function to map the input space to another feature space with higher dimension. The mapped vectors are then used to define a function $f(x)$ as a predicted value of y:

$$f(x) = \langle \Phi(x), \Phi(\omega) \rangle + b \qquad \text{EQUATION 2-31}$$

The objective of SVR is to minimize the regression risk ($R_{(f)}$),

$$\mathrm{R(f)} = \frac{1}{2}\langle \Phi(x), \Phi(\omega) \rangle + C\sum_{i=1} l(f(x_i), y_i) \qquad \text{EQUATION 2-32}$$

where C is the cost of error. The first term $\frac{1}{2}\langle\Phi(x),\Phi(\omega)\rangle$ can be seen as the margin in SVMs. The similarity between actual output y and its prediction is given by the loss function $l(f(x), y)$.

Vapnik [32] proposed an ε-insensitive loss function:

$$l_\varepsilon(f(x), y) = \max(0, |y - f(x)| - \varepsilon) \qquad \text{EQUATION 2-33}$$

Figure 2-8 demonstrates the ε-insensitive loss function. It differs from the linear loss function in that if the predicted value is within ±ε of the actual value, the prediction is considered lossless. Figure 2-9 shows how the actual values in the margin around the predicted function are considered acceptable or error-free.

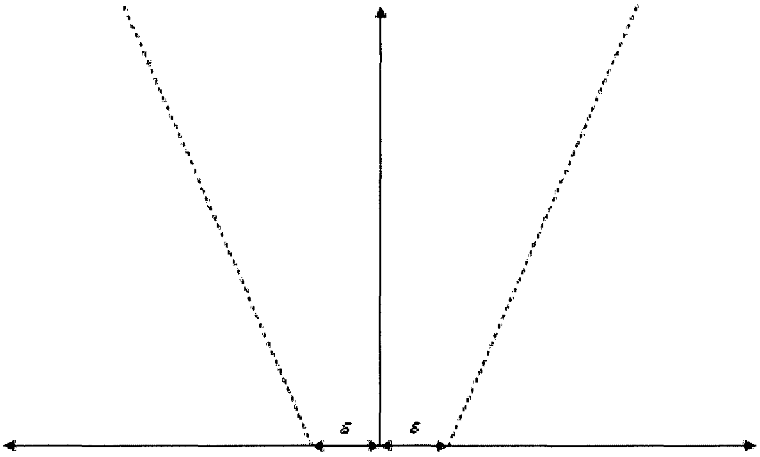


Figure 2-8: ε-insensitive loss function [34]

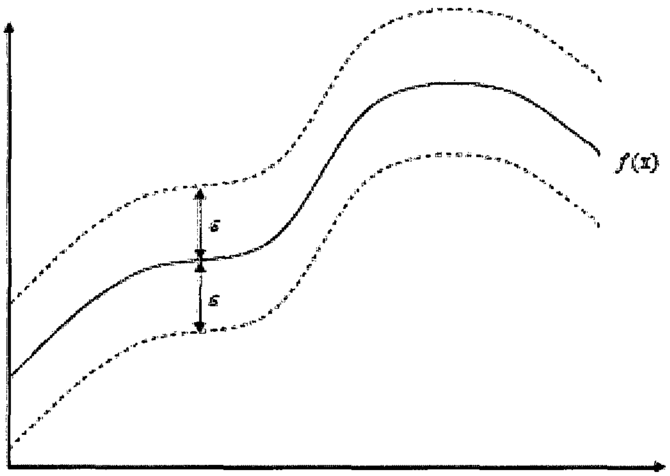


Figure 2-9: Prediction with ε-SVR [34]

The optimal regression function is given by the minimum of the regression risk,

$$\min\ \Gamma(\omega, \mathrm{b}, \xi_{\mathrm{i}}^{(*)}) = \frac{1}{2}\langle \Phi(x), \Phi(\omega) \rangle + C\sum_{\mathrm{i=l}} (\xi_i + \xi_{\mathrm{i}}^{*}) \qquad \text{EQUATION 2-34}$$

subject to:

$$\begin{aligned} b + y_i - \langle \Phi(x), \Phi(w) \rangle &\leq \varepsilon + \xi_{\mathrm{i}} \\ b - y_i + \langle \Phi(x), \Phi(w) \rangle &\leq \varepsilon + \xi_{\mathrm{i}}^{*} \\ \xi_i^{(*)} &\leq 0 \end{aligned} \qquad \text{EQUATION 2-35}$$

The use of (*) implies both the variables with and without asterisks. The solution is given by:

$$\min\ \mathrm{Q}(\alpha^{(*)}) = \frac{\mathrm{I}}{2}\sum_i \sum_j \left((a_i - a_i^{*}) \times (a_j - a_j^{*}) \times \langle \Phi(x_i), \Phi(x_j) \rangle \right) + \sum_i (\alpha_i \times (\varepsilon - y_i)) + \sum_i (\alpha_i^{*} \times (\varepsilon + y_i)) \qquad \text{EQUATION 2-36}$$

subject to:

$$\sum_i (a_i - a_i^{*}) = 0, \quad a_i^{(*)} \in [0, C] \qquad \text{EQUATION 2-37}$$

After solving this QP problem, the regression function is given by:

$$f(x)=\sum_{i}(a_i-a_i^{*})\times\langle\Phi(x),\Phi(x_i)\rangle+b$$

EQUATION 2-38

The value of *b* is calculated as:

$$b=\begin{cases} y_i-\langle\Phi(\omega),\Phi(x_i)\rangle-\varepsilon, & \mathit{for}\ \alpha_i\in(0,\ C)\\ y_i-\langle\Phi(\omega),\Phi(x_i)\rangle+\varepsilon, & \mathit{for}\ \alpha_i^{*}\in(0,\ C)\end{cases}$$

EQUATION 2-39

Where *Φ* maps x into a dimensional space; *C* is a pre-specified value and ε is the variable representing the upper and lower bounds of the outputs. When data points are within the range ±*ε*, they do not contribute to the error [34]. Details of the formulation can be found in [41].

2.4.4 Neural network prediction

As mentioned in section 2.3.3, neural network can also be a classifier. Since neural network is able to produce continuous outputs, it also can be used for prediction. When neural network functions as a predictor, it applies a sigmoid activation. It calculates the outputs of the network and approaches the actual outputs by adjusting the weights of each connection. The network is trained until it either achieves satisfactory results or the pre-specified number of iterations expires. The continuous outputs of the neural network can be used as predicted values.

2.4.5 Autoregression

Autoregression is a popular algorithm for dealing with time series. Time series algorithms handle data collected over a time period or based on other sequence indicator. The general definition of autoregression is given by:

$$x_t = f(x_{t-1}, x_{t-2}, x_{t-3}, x_{t-n}) + \epsilon_t$$

EQUATION 2-40

The value *x* at time *t* is a function of the values *x* at precious times i.e. *t-1, t-2* to *t-n*. If function *f* is linear, the equation 2-40 is

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + a_3 x_{t-3} + \ldots + a_n x_{t-n} + \epsilon_t$$

EQUATION 2-41

Where a_i are the Autoregression coefficients.

2.5 Association

Association is a popular mining method for dependency analysis. It is also known as *market basket analysis* and was formulated by Agrawal et al, in 1993. It extracts information from coincidence. This technique allows the company to discover correlations or co-occurrences of transactions i.e. if a customer purchases product *A*, how likely is he going to purchase product *B*? [23]

Association technique can be used to analyze a sales transaction table and identify those products often sold in the same shopping basket or bought by the same customer. The common usage of association is to identify common sets of items (frequent item sets) and rules for the purpose of cross selling or up-selling. In this section, this thesis describes *a* *priori* algorithm, which is the most accepted algorithm for finding all the frequent rules. It is also known as the *level-wise* algorithm [22].

Let T be the transaction database and σ be the user-specified minimum support. An itemset $X \subseteq A$ is said to be a frequent itemsets in T with respect to σ, if a measure of X named support is bigger than σ. Support measure represents the number of item sets in the entire datasets that contain the combination of items. For example, support = 2% means that itemset appears in at least 2% in the entire transactions. Suppose that $I = \{l_1, l_2, \ldots, l_m\}$ is a set of items and T is a set of transactions, where each transaction t is a set of items. Thus, t is a subset of I. A transaction t is said to support an item l_i, if l_i is present in t. t is said to support a subset of items $X \subseteq I$, if t supports each item l_i in X. An itemset $X \subseteq I$ has a support in T, denoted by $s(X)_T$, if a certain percentage of transactions (e.g. 5%) in T supports X. In the present text, unless specified otherwise, I will assume the support to be %-support. Suppose I have a particular set of items t_1 and another set of items t_2. Then an association rule has the form $t_1 \Rightarrow t_2$. The support for rule $t_1 \Rightarrow t_2$ is

$$Support = P(t_1 \cap t_2) \qquad \text{EQUATION 2-42}$$

The confidence of rule $t_1 \Rightarrow t_2$ is a measure of the accuracy of the rule, which is

$$Confidence = P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad \text{EQUATION 2-43}$$

A priori algorithm works as follows:

- Counts item occurrences to determine the frequent itemsets.
- A subsequent pass, assume pass k, consists of two phases.

  a) The frequent itemsets L_{k-1} found in the $(k-1)^{th}$ pass are used to generate the candidate itemsets C_k, using the *a priori* candidate-generation procedure

b) The database is scanned and the support of candidates in C_k is counted

The candidate-generation procedure works as follows:

Given a set $L_{k\text{-}1}$ that consists of all frequent (K-1)-itemsets, I want to generate a superset of the set of all frequent K-itemsets. The concept behind the *a priori* candidate-generation procedure is that, if an itemset X has minimum support, and then consider all subsets of X [22]. If a subset of (K-1)-itemsets does not match the predefined minimum support; it should be pruned at this point.

Furthermore, another important aspect of association is the ability to analyze the usefulness of a rule through the *importance* function (equation 2-42). Although there is high confidence for a rule, the usefulness of the rule may be considered unimportant. For example, if every item set contains a specific state of an attribute, a rule that predicts the state is trivial even though the probability is very high. The greater the importance, the more important the rule is.

$$\mathit{Importance} = \log\left(\frac{P(B|A)}{P(B|\mathit{not}\ A)}\right) \qquad \text{EQUATION 2-44}$$

2.6 Results evaluation

Many KDD algorithms estimate error and results of estimations are applied to make important business decisions [42]. In particular, several models are built simultaneously to discover useful knowledge. Results from these models will be compared for choosing the optimum one. Without accurate error estimates and effective methods to interpret those estimates, a suboptimal model might be chosen. Many statistical notions can be practically applied in this field.

2.6.1 Clustering results evaluation

Clustering can be evaluated from a length perspective [43]. The length measure can be divided into 2 categories. One is within-cluster scatter, which is the sum of distance values between the mean of a cluster and each data point within that cluster. The other category is the length between-cluster separation. That is the sum of distance between the mean of each cluster [44].

The scatter within the *i*th cluster, denoted by S_i, and the distance between cluster c_i and c_j, denoted by d_{ij}, are defined as:

$$S_{i,q} = \left(\frac{1}{|c_i|} \sum_{x \in c_i} \|x - c_i\|^q \right)^{\frac{1}{q}}$$ EQUATION 2-45

$$d_{ij,t} = \|c_i - c_j\|^t$$ EQUATION 2-46

where c_i is the center of the *i*th cluster. $|c_i|$ is the number of objects in c_i. q and t are integers greater than one that can be selected independently. The Davies-Bouldin index is then defined as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} R_{i,qt}$$ EQUATION 2-47

where $R_{i,qt} = \max_{1 \le j \le k, j \ne i} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\}$.

By maximizing Davies-Bouldin index, the best clustering model can be determined.

2.6.2 Classification results evaluation

For classification problems, the basic method in results evaluation is to use error rates [23]. The classifiers separate data points into classes; if the results are right, that is considered as a success; if not, it is an error. The overall error rate is the sum of the error divided by the total number of data points.

The hypothesis testing is another common method of results interpretation. First, state the default hypothesis, H_0, and the alternative hypothesis, H_1. Then select the level of significance to be used to define the rejection or critical regions. After computing the test distribution by using z test or t test, a statistical decision can be made based on distribution value. A false positive would be considered as a *type I error*, which indicates incorrectly rejected null hypothesis. A false negative error would be considered as a *type П error*, which indicates incorrectly accepted null hypothesis. The false negative error rate can be calculated by dividing the number of false negatives with the total number of negative classifications. Similarly, the false positive error rate can be calculated by dividing the number of false positives with the total number of positive classifications.

2.6.3 Prediction results evaluation

As a conventional prediction-results evaluation function, quadratic loss function can be used to maximize the success rate of predictions by minimizing the outcome sum of this function [43]. Suppose for a single instance there are k possible outcomes. For a given instance, the learning scheme comes up with a probability vector p_1, p_2,..., p_k (where these probabilities sum to 1). The actual outcome for that instance will be only one.

However, it is convenient to express it as a vector a_1, a_2,..., a_k, where ith outcome is one and the rests are 0. Thus, the quadratic loss function is denoted by:

$$\sum_j (P_j - a_j)^2 \qquad \text{EQUATION 2-48}$$

When the test data has several instances, the loss function is summed over them all. By minimizing the sum of all instances' outcomes, the best prediction model can be determined.

2.6.4 Association results evaluation

Support, confidence and importance are the measures to justify rules discovered in association results. The details have been addressed in section 2.5.

Moreover, histograms can be used to summarize results. It counts the frequency of numbers between a specified bin and a previous one. It also displays the cumulative frequency of observations in all of the bins up to the specified bin. The histogram can describe the distribution of the error rate efficiently.

2.7 Application of data mining

In previous sections, this thesis has introduced techniques and algorithms of data mining. Data mining has been applied to many fields which include science research [51], biomedical and DNA data analysis [47] [48], telecommunication services, banking, finance, pharmacology and retailers [9]. It is able to help marketing, system fraud detection, inventory management, multimedia application analysis, effectiveness of sales campaigns analysis, business risks analysis, and cross-selling. Moreover, it assists managers to understand customers, products and suppliers. Many commercial software tools have been developed due to the needs of data mining.

In 1996, Fayyad et al. presented an application of classification applied in the science field. He cataloged and classified Digital Palomar Observatory Sky Survey (POSS-Π) data and developed an automated sky object classification method. This method is still being used in NASA [51].

In [52], Apte et al. demonstrated how classification and prediction techniques can be used in financial industry in 1996. Useful rules were discovered from S&P 500 data (stock market data) and the predictive power of the generated rules was proven. In [64], Safer in 2003 predicted illegal insider trading which usually has higher profits by exploring unusual stock price returns using data mining techniques. Neural networks were applied for classification and prediction. This method is widely in use and helps avoid insider trading. Furthermore, Wuthrich predicted daily stock movements using information contained in articles published on the Web [53]. Carbonara described how data mining can be applied generally in the telecommunication field [54]. Marketing, fraud detection and network fault prediction were the main topics of his study. It covered customer profiling based on calling behavior analysis, customer attrition analysis and Internet usage analysis based on access logs. Visualization techniques have been pointed out as important techniques in the telecommunication field. Roughan and Zhang demonstrated an application of time–series algorithm and some statistical notions. The aim of their study was to predict the current traffic on the Internet [55]. Weiss used prediction technique to investigate telecommunication equipment faults from logs of network alarm messages [56].

In 2002, Lingras et al. demonstrated the usefulness of data mining for supermarkets. Their studies mainly covered analysis of characteristics of supermarket customers based on their loyalty and spending potential [57] [58], and analysis of customer attrition based on clustering results given customers' spending and visit time-series data[59] [60]. Relationship between product-based loyalty and clustering-based loyalty on customers' visits and spending patterns has also been investigated [61]. Through a series of studies, the ability of time-series and Kohonen self organizing maps (SOM) techniques in formulating marketing strategies has been demonstrated. Lingras in 2002 also illustrated the feasibility of rough set clustering for developing user profiles on websites [63]. Moreover, Lawrence et al. in 2001 described a personalized recommendation system designed to suggest new products to supermarket shoppers based on association techniques [62].

#

Chapter 3

Data mining process in wholesale and retail industry

Powerful database systems for collecting and managing data are in use in virtually all companies. All sales transactions generate computer records somewhere [1]. Wholesale and retail companies also hold all of this valuable data. Discovery of trends and patterns from this data, which could be used to improve business decisions and optimize successes, has become an emergent issue. In this chapter I will describe the overview of CRISP data mining process and demonstrate an overview of wholesale and retail industry. Furthermore, the outline of data mining applied in wholesale and retail industry will be addressed. I also introduce two wholesale and retail companies which provided data for this study.

3.1 Overview of the proposed process

The process includes business understanding, data understanding, data preparation, data model construction, and data mining tasks processing with clustering, prediction and association techniques. The process, or a part of it, may need to be repeated several times in order to achieve data mining goals. The data mining goals are always determined by business issues that the company wishes to address.

This study starts by examining the natural business cycle of the company and determining the goals of the data mining project. The data study is conducted simultaneously. At this point, one needs to work very closely with the domain expert. Preliminary data exploration is completed by summarizing available data. A few small sample data files will be created for initial testing. Based on our understanding of the available data, brainstorming and discussions lead to the determination of business issues and interests that can be served by data mining. It is the business issues and interests of the company that ultimately direct the goals of the project.

The data preparation is the first challenge of the project. The quality of data directly determines the quality of outputs. Data preparation includes data extraction, data cleaning, data reduction and data transformation. The main stage of this phase is the data cleaning. Data cleaning must respect the natural business cycle and remove misleading noise and outliers. There are no strict guidelines for this section, however, omitting the record, filling with a default value, filling with average data, or filling with the most probable value are some usual treatments. Data cleaning usually needs to be processed multiple times until the noise is successfully removed.

Clustering should be the first data mining technique applied since it is an unsupervised learning technique, which means that it does not require any labeled data in advance. The algorithm used in this thesis combines K-means algorithm with Standard score, which facilitates the setting of cut-off points and adjustments to the target dataset size. An important feature of clustering is the ability to build models based on data point similarities for further studies, such as prediction. Data mining experts have to build several clustering models based on different attributes of data. Through the examination of clustering results using aggregated histograms, data mining experts are able to profile and discover characteristics related to various aspects of the company, such as revenues and customer numbers of a business unit.

The prediction technique is an attractive topic. Everyone would like to predict the future to prevent possible risks and make effective decisions. On the other hand, it is very difficult part of data mining since the prediction is affected by many business factors and external influences. Prediction study can be divided into two steps. The first step is to detect the optimum data model by applying the clustering results. A popular and handy data mining technique, multiple regression, can be applied to verify data models constructed based on clustering results. Several models need to be built on different data attributes or different business cycles. The best data model can be detected by a comparison of the built models. The next step is to identify the optimum technique for the prediction studies. Comparison of the results of different prediction techniques determines the technique that is the most suitable for the company. Neural Networks and SVR are some of the other typical prediction techniques.

The market basket analysis (association) is another interesting technique in data mining. It should be built based on clustering results, which allow data mining experts to define the target datasets. Association tasks can be done multiple times based on different business focuses. The interpretation of analysis is always a challenge since massive results are generated at this point. The use of support, probability and importance measures can narrow down to the most meaningful results, moreover, batch processing might be another interesting feature of association, and can be run during non-peak times to minimize the effects of computational expense.

In some cases, wholesale and retail companies do not have enough data for KDD project. Therefore they have to purchase external data from government or some organizations such as InformCanada (http://211canada.typepad.com/informcanada/). A potential customer list including demographic data is an example of the purchased external data. Wholesale and retail companies can profile customers on the list based on customer clustering and profiling results derived from existing data mining results on customers' behavior. In most cases, it is not possible to study external individual customer data due to the privacy policies, so group study is common. However, customer surveys can transfer external individual customer data into internal data. Furthermore, based on individuals' responses to customer surveys, a model to predict potential customers' responses is possible. After transforming external data into internal data, our complete data mining process can be applied.

Regardless of what can be learned from data mining, the results are of no value if they are not integrated with company operations. Without managerial support, the company may possess information that could resolve business issues, yet never realize the benefits of data mining studies. Convincing key decision makers to take advantage of the power of data mining may be the biggest challenge of data mining.

Figure 3-1 graphically summarizes the entire data mining process we described above. Steps 1 to 3 are the same as generic CRISP data mining process. The modeling part in CRISP is expanded into 3 sessions: data mining objective determination, data model construction, and data mining technique selection. Data mining has the ability to address many business issues, but the goal of data mining project has to be decided first during step 4. Then several models need to be constructed. The best model will be selected based on comparison of results and be used to select data mining techniques. The results evaluation will be the last step. Figure 3-2 shows the data mining task processing in wholesale and retail industry. It should always start from clustering which can help data cleaning and preparation. Based on clustering results, prediction and associated can beapplied. Each step will need re-application of data cleaning and preparation techniques.

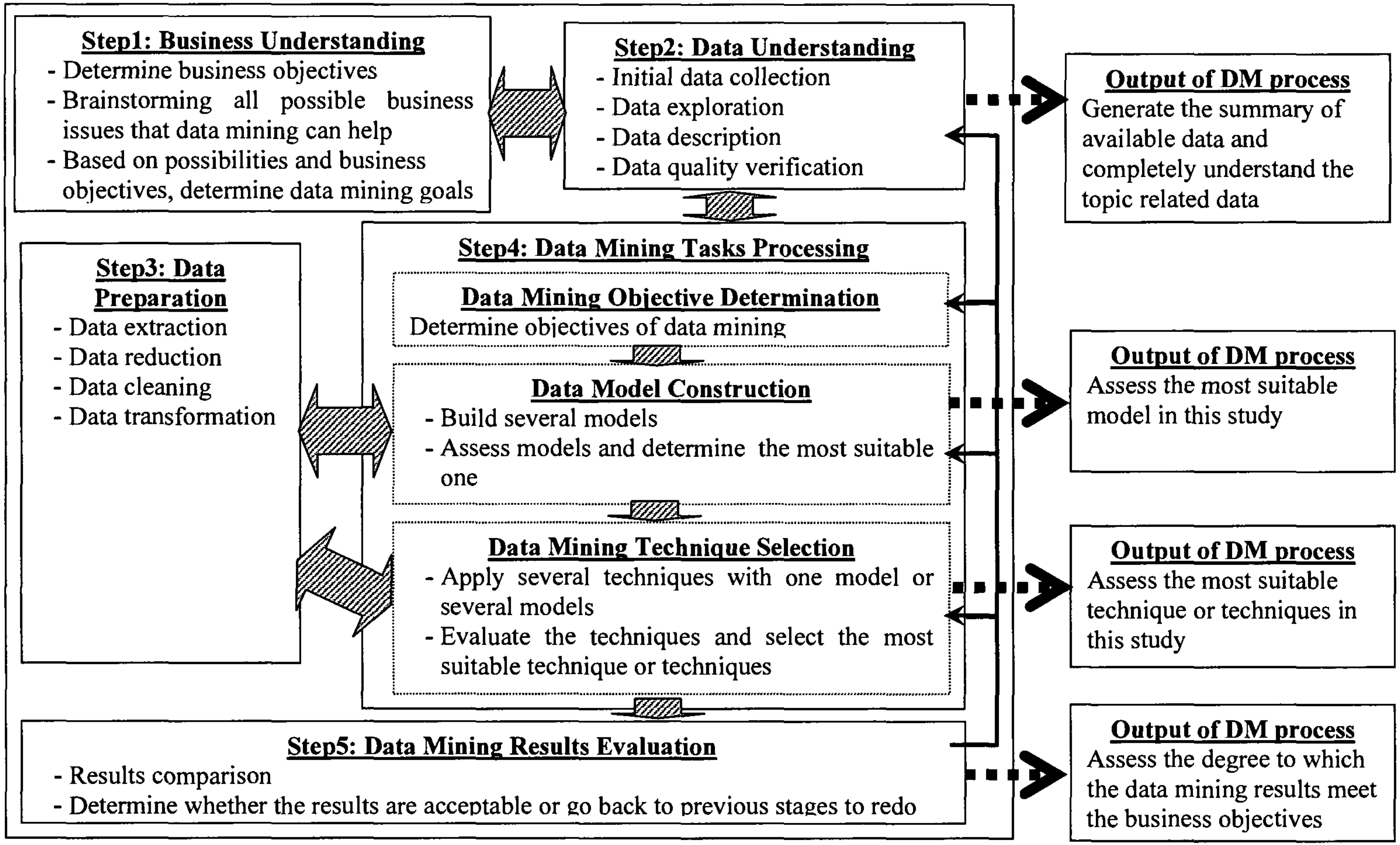
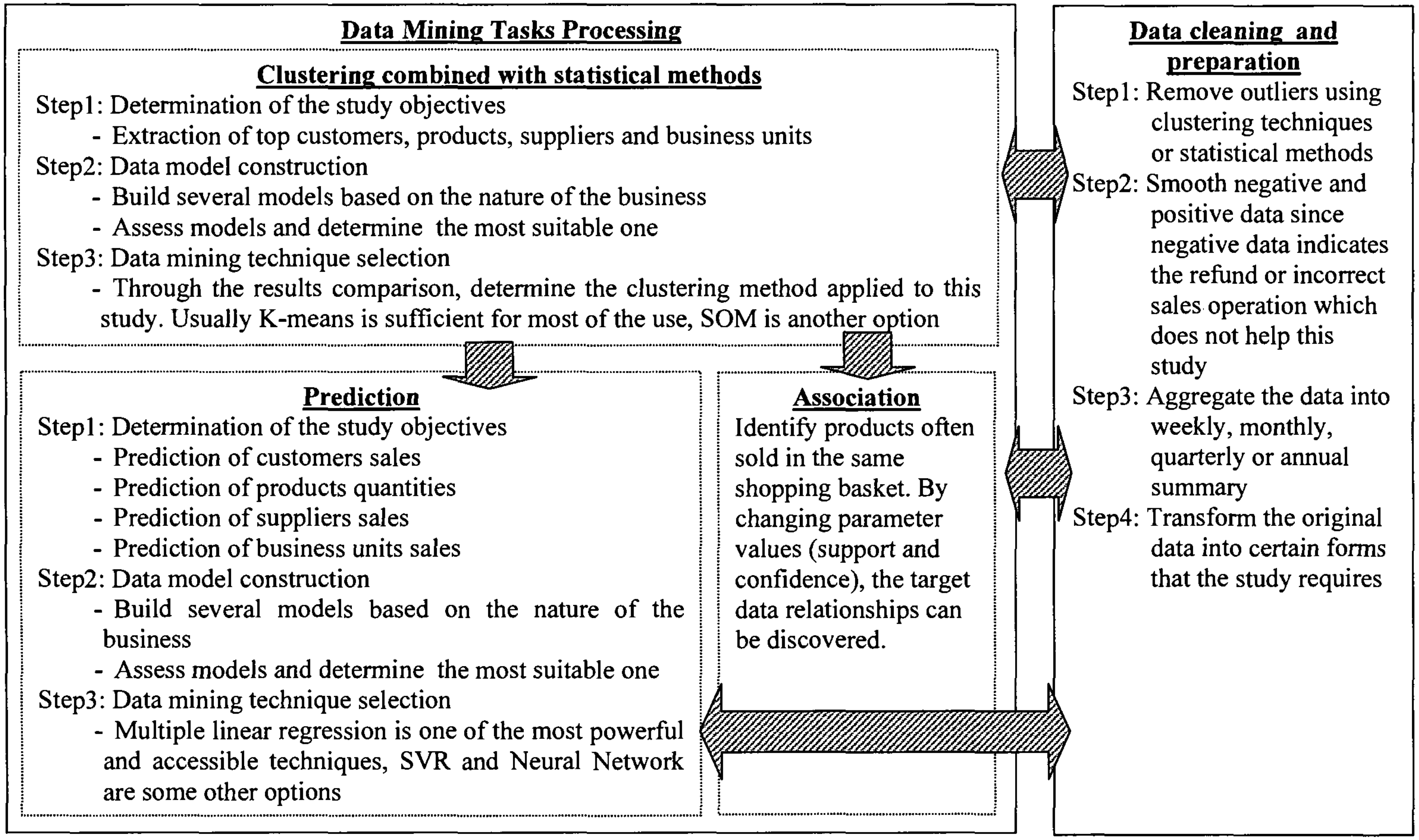

Step1: Business Understanding
- Determine business objectives
- Brainstorming all possible business issues that data mining can help
- Based on possibilities and business objectives, determine data mining goals
Step2: Data Understanding
- Initial data collection
- Data exploration
- Data description
- Data quality verification
Output of DM process
Generate the summary of available data and completely understand the topic related data
Step3: Data Preparation
- Data extraction
- Data reduction
- Data cleaning
- Data transformation
Step4: Data Mining Tasks Processing
Data Mining Objective Determination
Determine objectives of data mining
Data Model Construction
- Build several models
- Assess models and determine the most suitable one
Output of DM process
Assess the most suitable model in this study
Data Mining Technique Selection
- Apply several techniques with one model or several models
- Evaluate the techniques and select the most suitable technique or techniques
Output of DM process
Assess the most suitable technique or techniques in this study
Step5: Data Mining Results Evaluation
- Results comparison
- Determine whether the results are acceptable or go back to previous stages to redo
Output of DM process
Assess the degree to which the data mining results meet the business objectives

Figure 3-1: Summary of data mining process



Data Mining Tasks Processing

Clustering combined with statistical methods

Step1: Determination of the study objectives
- Extraction of top customers, products, suppliers and business units

Step2: Data model construction
- Build several models based on the nature of the business
- Assess models and determine the most suitable one

Step3: Data mining technique selection
- Through the results comparison, determine the clustering method applied to this study. Usually K-means is sufficient for most of the use, SOM is another option

Prediction

Step1: Determination of the study objectives
- Prediction of customers sales
- Prediction of products quantities
- Prediction of suppliers sales
- Prediction of business units sales

Step2: Data model construction
- Build several models based on the nature of the business
- Assess models and determine the most suitable one

Step3: Data mining technique selection
- Multiple linear regression is one of the most powerful and accessible techniques, SVR and Neural Network are some other options

Association

Identify products often sold in the same shopping basket. By changing parameter values (support and confidence), the target data relationships can be discovered.

Data cleaning and preparation

Step1: Remove outliers using clustering techniques or statistical methods

Step2: Smooth negative and positive data since negative data indicates the refund or incorrect sales operation which does not help this study

Step3: Aggregate the data into weekly, monthly, quarterly or annual summary

Step4: Transform the original data into certain forms that the study requires

Figure 3-2: Data mining tasks processing in wholesale and retail industry

3.2 Overview of wholesale and retail industry

The wholesale and retail industry is one of the most important sectors of the Canadian and global economies. In the wholesale and retail industry, companies purchase products in large quantities from manufacturers, importers or wholesalers, and sell small quantities to end users and earn profits through gross margins. In the rest of the thesis, I use the term retail for "wholesale and retail" for simplicity. The retail process includes all activities involved in selling, renting, and providing goods and services to ultimate consumers. It is the final step in the distribution of products. Based on a survey of Retail Council of Canada, the total amount of retail sales surpassed $400 billion in 2006. It accounts for 11.78 percent of employment in Canada. The details of retail sales are described in Figure 3-3.

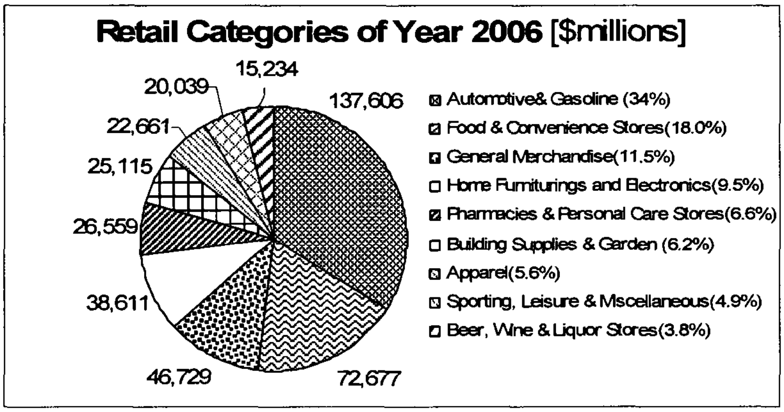


Figure 3-3: Retail sales ($millions) by type of businesses

Based on different attributes of retailers, they can be divided in several ways [2]. By the form of ownership, retailers can be classified into independent retailers, corporate chains, and contractual systems. Independent retailers are owned by individuals. Independents represent 60% of the total retail trade in Canada. Corporate chains involve multiple outlets under common ownership. The Bay and Zellers are examples of chain stores. The purchasing power of chains makes them strong negotiators and allows them to offer consumers more competitive prices. Customers also benefit in dealing with chains because they have consistent management policies, e.g. products refund policies. On the other hand, independent retailers might have various policies; one retailer's policy might be different from another one's. Furthermore, chains benefit from economies of scale. For instance, Wal-Mart has developed a sophisticated inventory management and cost-control system that allows rapid price changes for each product in every store [2]. The third type is contractual system retailers which involves independently owned stores banding together through a central corporation. In this way, members can also benefit from economies of scale that are otherwise only available to chains, such as volume discounts, and consistent management policies. Tim Hortons, McDonald's, and Holiday Inn are some examples of contractual system retailers. The central corporation assists individuals in selecting the store location, setting up the store, and training personnel. Furthermore, the central corporation designs core business processes and procedures that can be applied consistently to member businesses. In return, individual independent retailers pay certain percentages of revenues.

Retailers can also be classified into store retailers and non-store retailers. Television home shopping, online retail, telemarketing, and direct selling are the main forms of non-store retailers. Developing web-based E-commerce is a new and popular trend in retail industry. E-commerce provides more flexibility and accessibility to end users. Nowadays, many retailers have web sites where consumers can make purchases online.

Companies such as Amazon and eBay have proved that successful retailing can be achieved without traditional stores.

3.3 Data mining for retail industry

3.3.1 The needs of retail industry

There are three important needs of retail industry: pricing, customer relationship management (CRM), and inventory management [2].

Pricing is a critical issue for the retail industry. If a price is set too high, then consumers will not buy the product; if the price is too low, then the retailer will have lost profits and the supply may be quickly exhausted. CRM is also essential in identifying and responding to customer needs, and to predict market demands. CRM aims at minimizing the cost and maximizing profit, revenue, and customer satisfaction, which support business strategies. Inventory management is another important factor of retail business. Company resources are underutilized if they are tied up in inventory through inefficient inventory management. Successful retailers attempt to minimize these inventory costs to free up resources that could be applied to profit generation and strategic investments. However, less than necessary inventory may decrease business opportunities.

3.3.2 Retailers data description

Retailers collect huge amounts of data on sales, customers, suppliers, shopping transactions and shipping histories. The quantity of data collected continues to expand rapidly. The data collected mainly consists of time, customer, product, supplier information, order, business unit (BU), employee, and sales information. The print information that indicates whether the invoices have been printed is also stored depending on necessity. An example of basic retailer data constructor is shown in Figure 3-4.

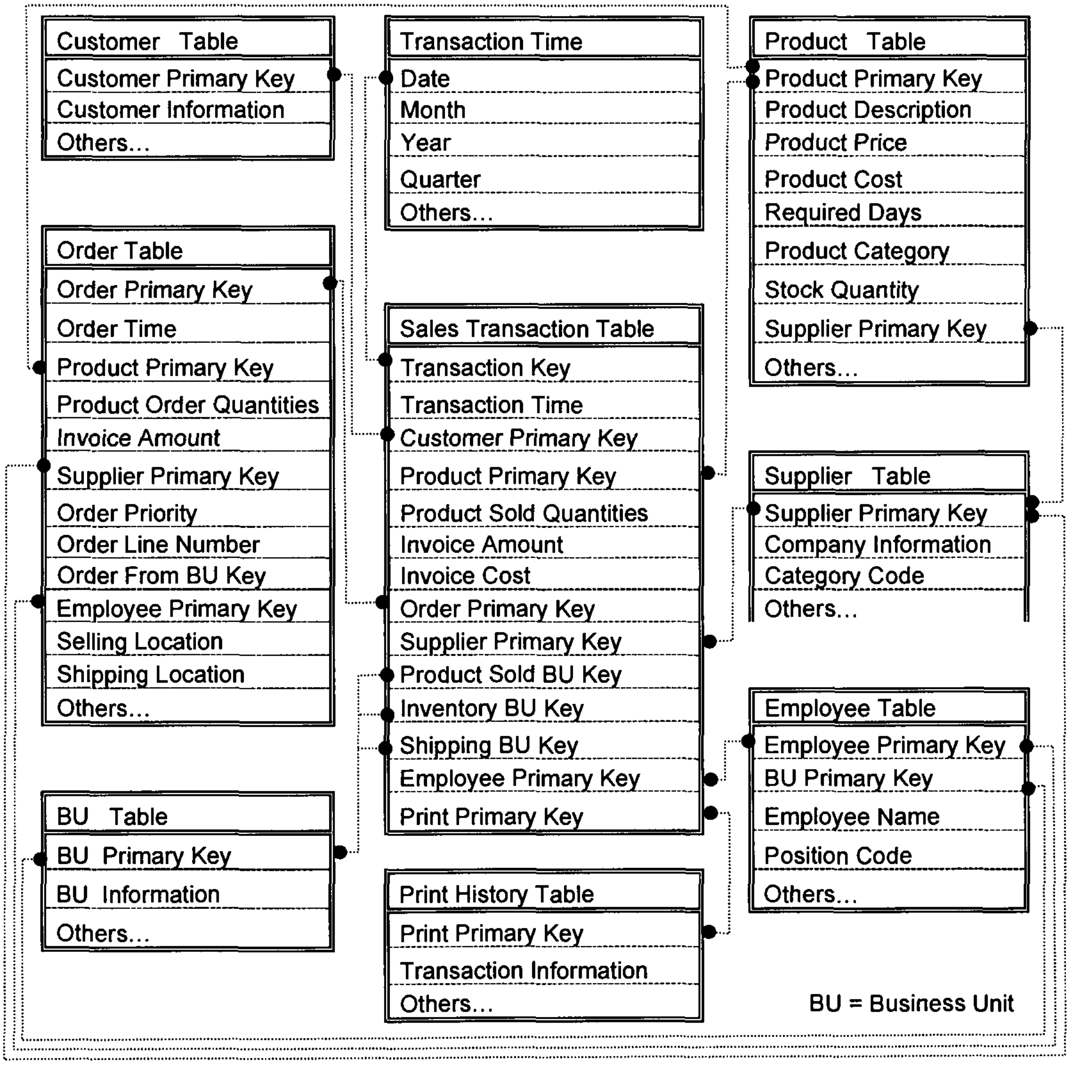


Figure 3-4: An example of basic retailer data constructor

The main table is the transaction table, which refers to all other tables using foreign keys. The transaction table includes all transaction information, e.g. which customer bought which product at what time and how many. The sales amount (or invoice amount) and the sales cost amount of each transaction are also addressed in this table. If there is insufficient inventory on hand to satisfy a customer's purchase, the required stock on backorder is related through the order information which has to be specified in transaction table. A product might be provided by different suppliers. Therefore the supplier information is required. Furthermore, large retailers might have several business units. Even a single transaction might need the cooperation of multiple business units, thus the business unit information about which business unit achieved this transaction, which one offered the product, which one took responsibility to ship the product also need to be specified. To facilitate sales follow-up process and employees' evaluations, each employee who took responsibility to the transactions has to be specified. Print key which refers to print history is also important information which facilitates sales representatives in order to trace the invoices or reprint. Other information can be also included depending on the requirements. However, the transaction table does not include the specific details of customers, products, suppliers, orders, business units, and employees as that information is detailed in their respective tables.

The customer table includes all of the customer information e.g. customer name, address, phone number, employee number. If a customer has special needs or preferences, then it also needs to be recorded in this table.

The product table specifies the product descriptions, price, cost, and category information. The inventory quantities can be included in this table for inventory management. The available suppliers for each product are listed here, but detailed in the supplier table. Due to the nature of the businesses, the changing history of product price has to be recorded somewhere in this table or stored in a separate table. The required days from ordering to receiving will also need to be recorded.

The supplier table indicates the suppliers' name, location, contact information and other detailed information. If retailers categorize their suppliers, then the category information is also required.

The order table identifies the basic information of orders to suppliers: which product has been ordered how many to which supplier, the invoice amount, the business unit information, selling and shipping location also need to be included. A single order invoice may contain several orders to one supplier; therefore the order line number in an invoice is also required. Due to the nature of businesses, the priority of an order is optional information. If not specified, the database will use default values to fill these fields.

The business unit table and the employee table contain basic information e.g. name, location and contact information.

3.3.3 Data mining possibilities for retailers

Retail data can be categorized by five groups: customers, products, suppliers, business units and company internal management information i.e. salaries. From the discussion in section 3.3.2, I know each category has its own characteristics and properties. Figure 3-4 illustrates how these categories are linked to each other. Data mining techniques can be applied for each category in order to achieve different tasks which can help businesses, including CRM and inventory management. Discovery of business issues' solutions is always the main objective of data mining studies. The following list contains some possible ways in which data mining may be helpful for a retail business.

1. **Business unit**

   Business units often carry activities in certain geographic areas. Based on the different attributes of a business unit, the prediction can be applied for different goals. Some data mining possibilities for business units are given below.

   ❖ Business unit annual revenue prediction or quarterly revenue prediction assists with marketing strategy development and distribution of company resources.

   ❖ Based on the clustering results of business units, similar business units can be grouped together. Therefore, a successful strategy of one business unit can be extended to another business unit in the same cluster.

2. **Customers**

   Retail data mining can help CRM to identify customers' purchasing patterns and trends in order to improve qualities of customer service, reduce marketing cost while at the same time increasing revenues, and achieving better customer retention and satisfaction. For instance, if the supermarket manager finds a group of customers who do not come often but spend a large amount of money each time, it is reasonable for him to believe those users are professionals who do not spend much spare time on shopping. Therefore, this manager can encourage those customers to shop more by sending them weekly discount flyers or offering low cost deliveries. Some data mining possibilities for customers are given below.

   ❖ Categorize the loyal customers and potential loyal customers based on sales amounts, frequency of purchases, or latest purchases in order to identify marketing campaign target groups. These customer segments include only a portion of the entire customer that is consistent in their purchasing patterns, and tendency to purchase certain products. By studying these loyal or potential loyal customers and discovering their purchasing patterns, companies can identify the business opportunities which contribute large portion of revenues.

- Categorize high-risk customers and discover their characteristics to reduce the company's business risks. For instance, a credit company wants to discover customers who are most likely not able to pay back monthly payments so the company have to either take the loss or spend more to collect the payment. Such customers are identified as high risk customers who most likely result in a loss to the company's business.
- Categorize customers by their purchase history in order to identify customers' characteristics and purchasing patterns.
- Predict customers' monthly revenues or quarterly revenues. The results can be used for marketing strategy development. For instance, a company may launch a campaign to increase purchasing desire in response to a prediction of shrinking future sales; likewise, a company could profit by accurately predicting and responding to increasing demand.
- Predict customers' lifetime values. More visits from a customer, more beneficial a customer would be. If the customer lifetime values are extremely high during a predetermined length of time, corresponding discounts can be provided to encourage customers' visits. If a customer's lifetime value is very low, the big discount does not contribute much to the company's revenue.

- Predict customers purchasing patterns based on several attributes of customers' purchasing records e.g. purchasing time (morning or evening).

3. **Products**

Retail data mining can help companies to have a clear view of their products. Companies can identify high profit product characteristics, optimize their inventories, detect companies' faults, and benefit by implementing cross-selling strategies.

- Categorize high and low profit products based on prices and costs. These results can be applied for further product development, and through such study the company can identify important products.
- Categorize high risk products, e.g. low profit or post-sales issues (returns, exchanges, and warranty service) based on transactions and order information.
- Optimize inventory management based on order information, inventory on-hand numbers, shipping costs, and shipping required days.
- Optimize inventory management. Inventory management based on product consumption cycles can reduce loss from disposal of expired items.
- Identify top selling products and find other products that customer segments may be interested in, i.e. cross-selling using association mining. Based on association data mining results from transactions, companies can discover that a customer that buys particular brands is also likely to buy another set of products. These results can be applied to cross-selling through product recommendation in order to improve customers' satisfaction.
- Detect companies' possible improvements. For instance, similar customers usually buy 10 products, yet one particular customer does not; here I can

investigate why this customer does not buy it: does not know that item, or buy that item from other companies. Management can use this analysis to detect how the company can improve its business.

- ❖ Discover beneficial procedures in order to help employees to set appropriate sales pricing based on customer segments.

4. **Suppliers**

Suppliers are an important component of the retail industry. The company can capitalize on the strengths of each supplier through the use of efficient supplier relationship management (SRM). Each retailer may receive competitive advantages by using suppliers that respond effectively to the business needs for products that meet quality, price, and inventory requirements. Data mining techniques can help the study of suppliers and increase business performance.

- ❖ Categorize suppliers based on delivery time responsiveness. The length of time required for suppliers to deliver will affect a retail business, particularly in whether capital is tied in inventory that is in transit or sitting unsold on a shelf. Using a supplier that can efficiently balance costs and delivery speed will minimize the quantity of stock required in inventory. Savings from low stock requirements allows capital to be reallocated to more productive uses while simultaneously enabling the retailer to be more responsive to its customer's demands. Periodic review should be conducted to ensure only the most effective and responsive suppliers are used.

Categorize suppliers based on the number of important products that a supplier provides. The company needs to keep good relationships with those top suppliers since they are the base of the company's business. New products' development could also involve those top suppliers.

- Categorize suppliers based on the risk of poor supplier performance. Indicators of poor performance can be decided by management and would likely include factors such as the number of product returns and the products' profits related to each supplier. High product return rates or a low contribution to profit would suggest further investigation is required by managers to determine whether company goals can be achieved more efficiently through alternative suppliers.

3.3.4 Motivation for data mining in retail industry

As I described in section 3.3.3, retail data mining can help develop business strategy on business unit level, discover customers' purchasing patterns and trends to improve customer service, reduce marketing cost while at the same time increase revenues, and achieve better customer retention and satisfaction. Retail data mining can also help optimize inventories and capitalize on the strengths of each supplier. Many commercial applications are adopted to address above business challenges. Examples are intelligent miner from IBM [10]; SAS Enterprise Miner from SAS (http://www.sas.com/); Clementine from SPSS (http://www.spss.com/); SQL Server Business Intelligence Development Studio from Microsoft [11]. A review of the work by a number of researchers is presented to address their contributions to retail data mining.

In customer relationship management (CRM) field, Buckinx applied Logistic regression and Neural Network to analysis of customers' loyalty in 2003 [67]. In 2004, Ester et al. demonstrated customer-oriented segmentation [68]. In 2005, Chen et al. discussed personalized marketing campaigns through analysis of customer behaviours' changes [69]. In 2008, Giering outlined a retail sales prediction and product recommendation system that was implemented for a chain of retail stores [70].

In product related studies, association data mining is a popular technique. Brjis et al. presented a product selection model PROFSET which can "select interesting products from a product assortment based on their cross-selling potential given some retailer defined constraints" [71]. In 2005, Chen et al. proposed a method to handle association data mining analysis across multiple stores [72].

Knowledge discovery for E-commerce is another major component of retail data mining which conducts on-line transaction knowledge discovery. In 2000, Dhond explained key data mining principles that are mainly used in E-commerce and also illustrated two case studies in on-line transaction knowledge discovery [73]. In 2004, Kohavi et al. addressed his lessons and challenges from mining retail E-commerce data from two dimensions: business-level vs. technical [74].

Despite the availability of mature commercial applications and research contributing to retail data mining, there are still only a few retail companies that have succeeded in practical use of data mining. The challenge is to assist domain experts to practically apply data mining techniques to their data. The reason for limited use of data mining is because it needs significant investment of company resources [2]. Companies using data mining are mostly limited to international major companies, i.e. Master Card, Wal-Mart and Burger King [75]. Therefore, commercial data mining application vendors, i.e. IBM, SAS and SPSS lunched data mining outsourcing services. However, as I described in section 1.4, these companies do not elaborate on the modeling and evaluation of the data mining processes. Only data mining results are available for end-users. Once a wholesale and retail company has signed up with the consultant, then the association will necessarily have to continue for a long period of time, because data mining is an evolving process and knowledge changes as time goes by. The cost of such data mining consultancy can be prohibitive for small to medium wholesale and retail companies. Moreover, using outsourcing services raises company and customer privacy concerns.

On the other hand, the research articles cited in this section focus on particular data mining techniques and its evaluations. Such articles do not provide details of the complete model development process from data collection, data preparation, identification of model parameters, and evaluation of results. One of the useful researches in this aspect is the work by Kohavi, et al. [74], which describes the complete data mining process for an e-commerce business from a customer-centric point of view. The work presented in this thesis expands on Kohavi, et al.'s theme. The treatment in this thesis is based on consultation with domain experts in a retail and a wholesale business. The scope of data mining is governed by the problems that were identified to be of importance by these experts. In addition to the customer-centric view described by Kohavi, et al., the thesis addresses other issues such as analysis and prediction of activities related to business units and products.

This thesis describes an elaborate generic enterprise-wide data mining process based on the CRISP model for wholesale and retail businesses. The focus of this thesis is on the development and evaluation of models used in the knowledge discovery. Description of the underlying models will help the decision makers better understand the quality and limitations of the knowledge discovery process. The goal of this thesis is to guard domain experts to process data mining tasks using existing data mining tools such as SQL Server 2005 BI application. This thesis covers the complete data mining process from business issues identification to results evaluation and interpretation with businesses.

3.4 Outline of retail organizations

There are many opportunities for data mining to help retailers' businesses. Selecting the right data mining tasks and proper techniques always starts from data exploration and data study. First of all, the number of business units, customers, products and suppliers are the basic information that data mining experts need to know.

There are many ways to summarize company data. For example, one can summarize data based on business unit. In this case, average, minimum and maximum values are some of the useful attributes. Also these values can be collected in different time scales, such as weekly, monthly, quarterly or yearly. If one needs to summarize data based on other attributes such as customer, product or suppler, one also needs to consider similar values and scales. An example of a possible summary of companies' data can be the revenue that a business unit has per week, the revenue that a customer has per month, the revenue that a product has per quarter, the revenue that a supplier achieves per year.

Different attributes can be combined to understand the company. No matter whether a retailer is an independent retailer, a corporate chain retailer or a contractual system retailer, the business unit, customer, product, supplier, employee, order information, transaction information, revenue, and time are some of the typical attributes found in all three forms of retailers. Plotting this data forms a high dimensional cube which summarizes the company. A three dimensional cube combining business unit, customer and product information is shown in the Figure 3-5.

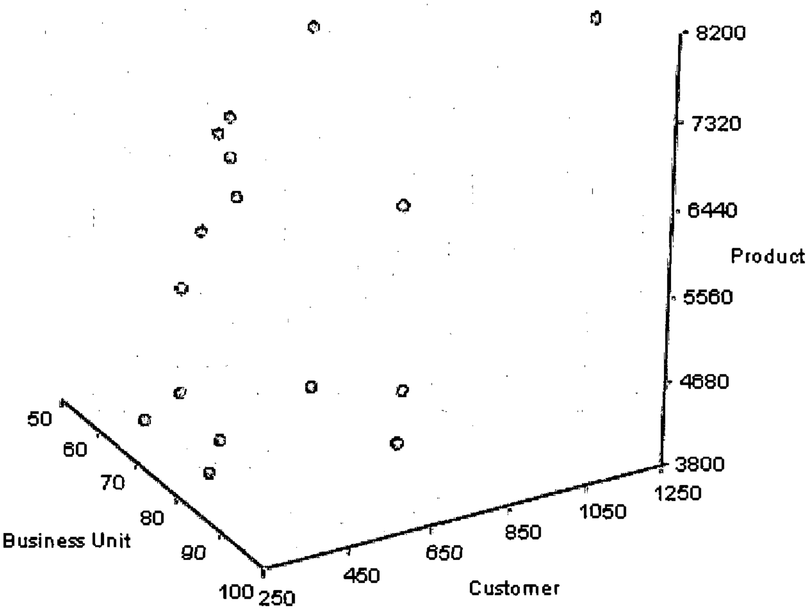


Figure 3-5: A three dimension cube example

3.5 Overview of multinational retailer

3.5.1 Nature of the business

A multinational retailer that provided the data for this study is one of the largest companies in the world. It is a corporate chain retailer that operates in more than 30 countries and has over 2600 business units across the world. These business units provide various products in many fields. Most customers are commercial. Together, all of the globally distributed business units contribute more than $10 billion in annual sales. The subject of our study is one region of this multinational retailer. Because of the sensitive nature of the analysis performed in this project, specific names and numeric details have been disguised in the following descriptions.

The company receives orders from customers primarily by telephone or online; traditional brick and mortar retail sales are a minor part of their business. Representatives process orders and check inventory. If the products are in inventory, the shipment is sent immediately. If products are in short supply, representatives have to procure the products in question from their vendors or secure them from other business units. Key customers and suppliers have dedicated representatives to address these orders. Both the customer and the supplier enjoy a long-standing relationship. The current process is common in the industry.

3.5.2 Available data

More than 40 GB of sales transactions are recorded every year. This data is available for this study in flat format (CSV format). In this study I am analyzing one region of this multinational retailer. The data consists of transactions from January 2004 to December 2007 and included all of the basic data this thesis discussed in section 3.4. About 15.9 million transactions, 60,000 customers, and 277,000 products are included in this data. In only 2004, about 23,000 customers contributed to over $480 million of revenue.

3.5.3 Goals of the data mining

Data mining has to focus on solving business issues. The domain expert identified prediction of revenues as an important business issue. The prediction of revenues was identified as an important part of our complete data mining process. However, the main aim of this study is to offer an improved understanding of business units' products, customers and suppliers. The study of suppliers has less priority for the multinational. Hence, the enterprise-wide data mining mainly focused on relationships between products, customers, and business units. The detailed outline of this project is shown below.

Part 1: Business unit based analysis

1. Business unit profiling

Based on business units' different features, they can be divided into several groups such as:

- ❖ Revenue based profiling
- ❖ Customer number per business unit based profiling
- ❖ Product number per business unit based profiling.

Furthermore, the revenue based profiling includes monthly revenue based profiling and annual revenue based profiling. Since the multinational retailer has many business units across multiple regions, a successful strategy in one business unit can apply to other business units that are in the same group in order to increase benefits for entire company. The results of this section offer an overview of the company's business units to help the company's marketing strategy development.

2. Prediction of the revenue at the business unit level

Because the company's business scale is extremely big, the company's limited resources have to be divided in an optimized way for the further growth. The next year's business strategies and projects are planned and determined in the fall; this is due to the requirements of resource preparation and distribution. Therefore the prediction of next year's annual revenue for each business unit becomes a critical and extremely important issue. Prediction techniques are applied to meet this need.

3: Benchmark error rate prediction

Currently the multinational retailer is using a benchmark algorithm to calculate next year's annual revenue. The details will be described in chapter 4. The benchmark results' accuracy is unreliable and must be verified. Therefore, the company aspires to use data mining to investigate the accuracy of the benchmark results.

Part 2: Customer based analysis

Due to the vast number of customers that the multinational retailer serves, knowing customers and targeting the right groups are essential for project management and business development. They are able to reduce marketing cost while at the same time helping to increase the company's benefits by developing form-fitting strategies for each customer's group.

Part 3: Products based analysis

1: Products Profiling

The company is handling about 277,000 products. Some products contribute a large portion of the company's revenue and profits. Some products are only sold several times and produce small benefits. Therefore, product profiling is important for business management and further product development. Based on different product features, the products can be divided into several groups such as:

- Revenue based profiling

- Purchased customers' number based profiling

For the top products that contribute to a large portion of the company's revenue, the sub-profiling will be used for the further study of these top products.

2: Association mining

Cross-selling is one of the most important tools to increase the benefits of a company. Based on products profiling results, the relationship between target products will be discovered by applying association data mining technique.

3.6 Overview of local specialty grocery store

3.6.1 Nature of the business

Another data provider for this study is a local specialty grocery store which is an independent retailer. Customers have special needs that may require them to spend significant amount of money in the store. Most customers are from the neighborhood area. Few customers travel a significant distance to visit the store looking for specific brands and specialty products that are difficult to find elsewhere. Customers visit the store regularly to purchase products for personal uses. Compared to the multinational retailers, the value of the individual sales is much smaller. Online shopping is the next business development direction.

Similar to many local grocery stores, customers walk into stores and select products that are of interest to them. Rarely do customers order products by telephone. Weekly flyers and consumer shows are the main techniques to attract customers. Direct mail or E-mail will be the next challenge in order to further grow the business.

3.6.2 Available data

The data available for this study involved a 33 month period starting from January 2005 to September 2007 and it included all of basic data this thesis discussed in section 3.4. It is in MS Access format. More than 17,000 of customers; 10,900 products and 411,000 transactions are involved in this study. In 2005, over 7,000 customers contributed to $1.8 million of revenue.

3.6.3 Goals of the data mining

The goal of data mining is to facilitate business decisions. The main purpose of this study is to help inventory management and apply cross-selling in order to reduce the business risks and simultaneously increase the benefits. As described in section 3.3.1, inventory management is the key in reducing business costs. The relationship discovery between each product is used to extract suitable product candidates for flyers to implement cross-selling. Cross-selling can help to attract customers while at the same time increasing profits. Moreover, clustering techniques are applied to provide a better overview of products and customers. Suppliers' studies are omitted in this project.

Part 1: Customer based analysis

1. Customers profiling

As described in section 3.3.1, knowing customers and targeting the right groups are essential for project management and business development. These points not only suit corporate chains but also benefit independent retailers. Customers can be divided into several groups based on annual revenue and annual profits. The results can be used in future promotional campaigns.

2. Prediction at customer level

Customers purchasing patterns can be used to identify prospective customers. Based on these results, the grocery store can recommend products based on product profiling results. For instance, consider customers who purchased product A in first month. If I know that customers who buy product A usually buy product B with a high confidence, then the recommendation of product B can be sent to prospective customers who bought product A. Based on customers' purchase histories, the purchase patterns can be extracted. These discovered patterns can be used to predict future revenues and purchased products' number.

Part 2: Products based analysis

1. Products Profiling

As described in section 3.3.3, products profiling provides the focus of products for business management and further product development. Based on products' revenues and purchased times, the profiling can provide a blueprint about the grocery store's important products.

2. Prediction of products sold quantities

As addressed previously, the inventory management is one of the main topics of this study. Therefore, predicting the quantity of products that will be sold is essential. The results allow managers to order necessary products from suppliers and optimize products quantities in stock. For instance, consider product A whose expiry date is one month from now. I know that consumers need one week to consume A and the store has one hundred units of product A in stock. If the grocery store realizes only thirty product A will be sold in the next month by prediction results, then a discount campaign of product A has to be considered.

3. Association mining

Cross-selling is one of the most important techniques to increase company profits. Therefore, the relationship between target products will be discovered by applying association data mining techniques based on transaction data of those products. The results are used to implement effective cross-selling campaigns.

#

Chapter 4

Application of the proposed process to multinational retailer

In this chapter, I describe and demonstrate a complete data mining process with the help of a multinational retailer. To solve the company's business issues, different models using clustering, prediction and association will be illustrated. The most suitable model and technique will be selected through the results comparison.

4.1 Business understanding and data study

4.1.1 Business understanding

Understanding the company's natural business is the only way that data mining experts can define proper goals for a data mining project and lead the business in the right directions. The cooperation between data mining experts and domain experts is indispensable. The first step in the data mining process is to brainstorm all possible business issues where data mining can help. Data mining in retail industry is primarily used to support four interconnected categories: the business unit, the customer, the product, and the supplier. The company's internal management, i.e. salaries, is not considered in this step. Some possible tasks were described in section 3.3.3.

The next step is to ascertain business issues through phone meetings and online meetings with a domain expert, an IT manager of the company. Therefore the goals of this project were determined based on the company's business issues and interests. The goals of this project were presented in chapter 3. Up to this stage, the domain expert worked with us to formulate a clear blueprint of tasks for this study.

4.1.2 Data study

During the project goal-definition process, data study process is also performed. The available data is the foundation of the entire project. It affects the problem solving ability of data mining.

The purpose of this stage is to collect all initial data and proceed with some activities in order to understand the available data. The data study includes initial data collection, data exploration, data description, and data quality verification.

The data from the multinational retailer is in CSV format, which is directly exported from the company's database. Since the data size is much bigger than 2GB, which is the limit of an MS Access database file, Microsoft Sequence Sever (SQL Server) is chosen as the primary database. Data Transformation Services (DTS) is applied to data transportation. DTS is a tool included in the SQL server package that allows automation of data import or transformation to an SQL database. The collected data includes all necessary subjects which were described in section 3.3. It contains the business unit, customer, supplier, product, order and sales information. The details of available data have been described in section 3.5.2.

The next step is to explore original data in order to determine the available data attributes. Data mining experts have to spend time to explore data and use graphs and histograms to estimate available data. Since the original data is extremely large, several CSV files were created in order to facilitate the data exploration. Through careful examination of the original data, I have a clear view of available data attributes. Therefore, a data schema description report was generated to list all available attributes in the original data. Data exploration conducted by the domain expert confirmed the data quality and verified that the original data contains everything I need for this study.

Attribute Name		Year 2004	Year 2005	Year 2006
Customer	Number	23,723	23,602	23,824
	Minimum Annual Sales	-19,574	-23,254	-32,210
	Maximum Annual Sales	737,101,822	13,676,672	13,272,264
	Average Annual Sales	54,770	26,583	30,094
Business Unit	Number	104	109	114
	Minimum Annual Sales	-9,605	0	0
	Maximum Annual Sales	749,674,958	41,253,597	50,309,321
	Average Annual Sales	12,493,298	5,756,137	6,289,055
Product	Number	81,142	86,185	85,469
	Minimum Annual Sales	-150,889	-235,566	-360,241
	Maximum Annual Sales	736,507,800	17,015,086	18,440,587
	Average Annual Sales	16,012	7,280	8,388
Supplier	Number	2,593	2,727	2,648
	Minimum Annual Sales	-76,097	-35,532	-209,387
	Maximum Annual Sales	739,075,667	116,860,047	86,851,751
	Average Annual Sales	501,081	230,077	270,752

Figure 4-1: The overview of the original data

Figure 4-1 shows the overview of the original data. It provides customer, business unit, product, and supplier's basic data including number, minimum annual sales, and maximum annual sales. The average of annual sales is also shown in Figure 4-1.

4.2 Data preparation

The quality of data determines the success of the entire process. After having a clear view of the project goal and available data, the next step is the data preparation. As described in section 1.3, data preparation includes data extraction, data reduction, data cleaning, data integration and data transformation. In one study, data preparation process might need to be repeated several times. In the worst case, for each model in the study, the data preparation may have to be revisited.

4.2.1 Data extraction

Different data mining tasks use different parts of company's available data. Since the data cleaning and data transformation process might change the original data, task oriented data should be extracted as separate files in order to distinguish the original data from the modified data, or one sub-project data from another sub-project data. Furthermore, data separation also facilitates future reviews of data mining tasks. In this study, I extracted three kinds of files. Each includes business unit oriented, customer oriented and product oriented data.

4.2.2 Data reduction

As described in section 1.3 and section 2.1, the target datasets have to be reduced to subject oriented data. For instance, the sales table contains 73 fields; the order table contains 134 fields; most of them are not related to our objectives. Moreover, some fields have incomplete and inconsistent data. Therefore, removal of the unrelated data not only saves physical space, but also facilitates the data preparation process itself and further data mining processes. Based on the data study results and the domain expert's advice, I decided to select the attributes which are shown in Figure 4-2.

NO	Field Name		
1	Transaction Key	2	Transaction Time
3	Customer Primary Key	4	Product Primary Key
5	Product Sold Quantities	6	Invoice Amount
7	Invoice Cost	8	Order Primary Key
9	Supplier Primary Key	10	Product Sold BU Key
11	Inventory BU Key	12	Shipping BU Key
13	Employee Primary Key	-	-

Figure 4-2: Selected attributes list

The attributes shown in Figure 4-2 are all selected from the transaction table. As described in section 3.3.2, the transaction table is the main table in the entire dataset, which refers to all other tables using foreign keys. The transaction table enables us to focus on customers, products, and suppliers and employees that were active during the period relevant to the study. For instance, the records of former customers that made purchases ten years ago can be considered as unimportant data because they are most likely irrelevant to the current business issues.

4.2.3 Data cleaning

Data from the real world usually are usually incomplete, inconsistent or noisy. The data cleaning algorithms introduced in chapter 2 can be applied to handle these issues. The company's data quality was found to be good. There was no missing data in selected attributes (refer to section 4.2.2). However, for customers who did not purchase any products during the study period or for products that were not sold during this period, the default value of 0 will be used to fill the original datasets, if necessary.

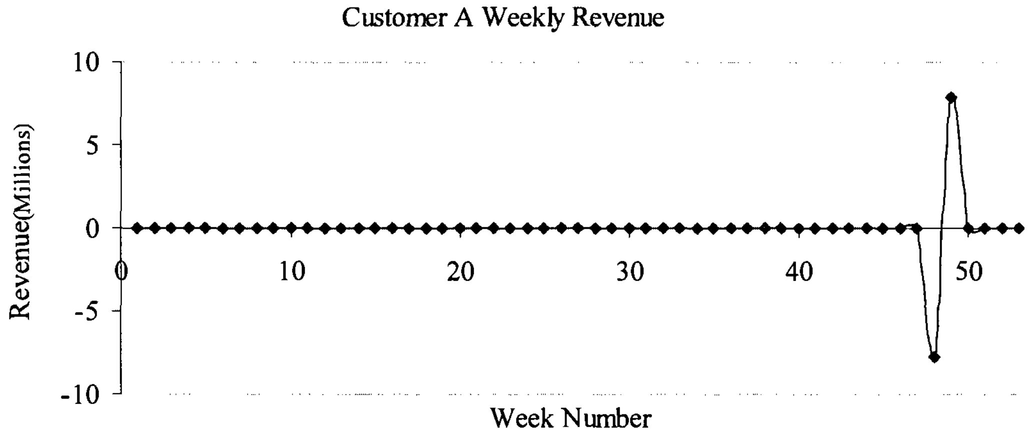


Figure 4-3: Noisy data pattern – negative data in weekly revenues

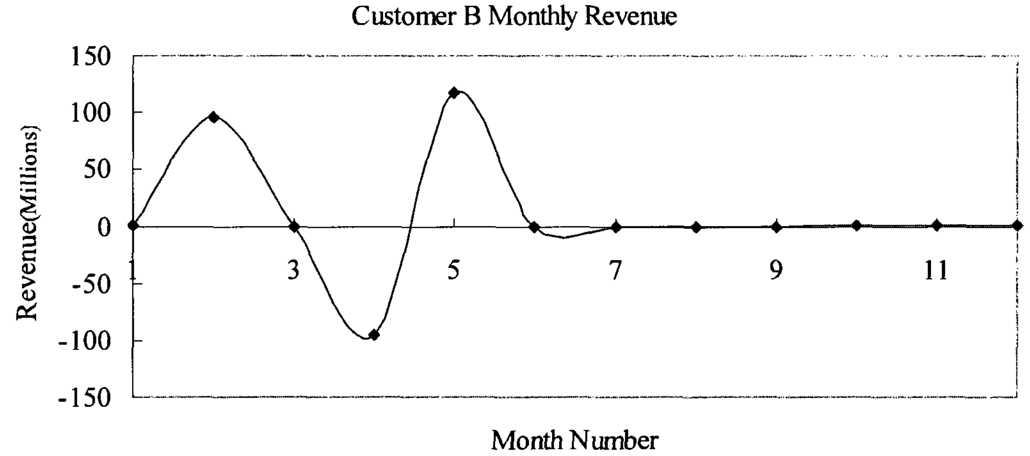


Figure 4-4: Noisy data pattern – negative data in monthly revenues

Figure 4-3 and Figure 4-4 show the noisy data for this study, which has negative values for revenues. The negative values might be caused by inappropriate operations or products rebates and refunds. In the Figure 4-3, customer A seemed to have an unusual transaction. From the exploration of the data, I found that customer A's purchase range is around twelve thousand dollars (the median number is $11,983). In the 48th week, there is one extremely big negative revenue record (about $7.79 millions); and in the 49th week, there is an equally positive big revenue record (about $7.82 millions). We consider this pattern to be an operator input error. Operator made a mistake in the first week and corrected it in the following week. The operator errors do not represent any retail activities; therefore, they have to be removed from target datasets. Customer B in Figure 4-4 represents the products refund pattern in monthly revenues. Customer B makes an order in February around $96,000, and returns exactly same amount in April. This kind of retail activities does not contribute to company's revenues. It is useful for company's fraud detection analysis; however, it is not the main subject of this study. It is necessary to smooth negative and positive data.

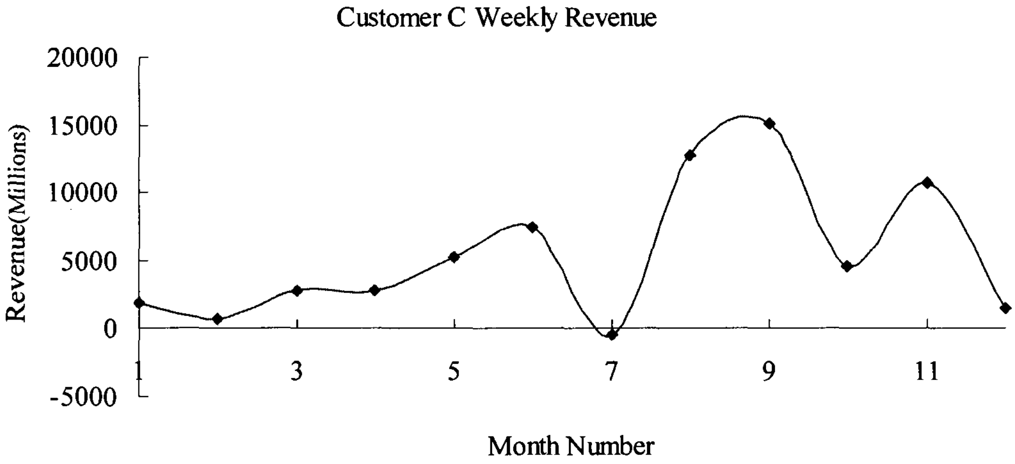


Figure 4-5: Product refunds or rebates pattern

Figure 4-5 shows another type of refund or rebate activity. Customer C continually buys products from the multinational retailer, yet in July there is a small negative sales value. Even though a negative value occurred, it still affected to the company's revenue. This kind of negative value needs to be kept in the dataset.

The criterion of smoothing data decision is:

- Extract monthly revenue data for every business unit, customer and product.
- For every negative monthly revenues data, check previous two month revenues data. If there is a positive value which is only 10% different from the negative value, then replace the positive value with the difference of the positive value and the negative value; replace the negative value to 0.
- For every negative monthly revenues data, check two proceeding months revenues data. If there is a positive value which is only 10% different from the negative value, then replace the positive value with the difference of the positive value and the negative value; replace the negative value to 0.

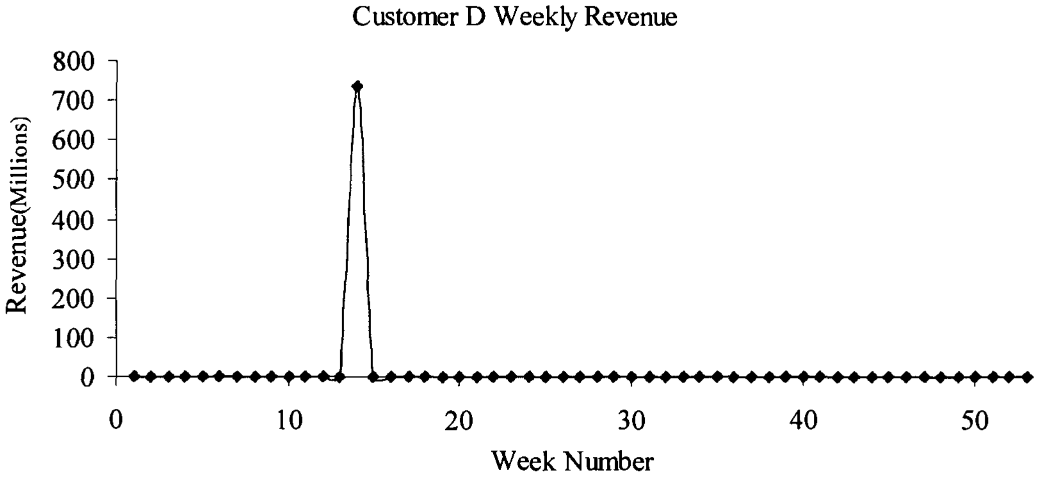


Figure 4-6: Noisy data pattern – outlier data in weekly revenues

Figure 4-6 shows the noisy data which are outliers in the target datasets. The median value of customer D is $11,000; however only one value is $736.5 million, which is about 67,000 times bigger than the median value. In the business world, such a growth rate can be considered as extremely unusual. Therefore, it can be called an outlier and must be removed from the dataset. Figure 4-7 shows an outlier in the monthly data.

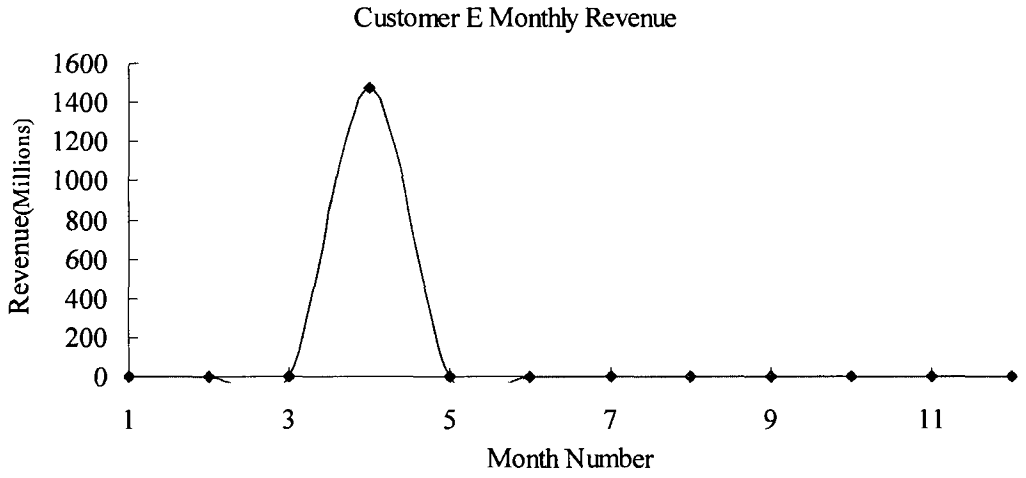


Figure 4-7: Noisy data pattern – outlier data in monthly revenues

In individual studies, the clustering technique can be applied to check outliers. If a small cluster whose mean is far away from others, it can be considered as an outlier and removed from the study to enhance the accuracy of data mining results. For instance, Figure 4-8 shows a clustering result for a group of customers. The clustering is based on 12 monthly revenues of individual customers.

Cluster Number	Number of Data	Annual Revenues (Million)		
		Minimum	Maximum	Mean
Cluster 1	1	720	720	720
Cluster 2	19550	-0.019	0.022	0.003
Cluster 3	837	0.104	0.5	0.240
Cluster 4	3335	0.008	16	0.183

Figure 4-8: Noisy data pattern – outlier data in annual revenues

Minimum monthly revenues, maximum monthly revenues and the mean of the cluster are shown in Figure 4-8. As we can see from Figure 4-8, cluster 1 only has one member and the mean of cluster 1 is extremely bigger than other clusters', therefore I consider that the only member of cluster 1 is an outlier.

4.2.4 Data transformation

In this project, it is not necessary to combine data from multiple data sources such as external data. Therefore, data transformation is the final stage of data preparation. The data needs to be transferred into certain formats so that the data mining technique can be applied. In this study, I summarize the data to monthly, quarterly and annual revenues and profits.

4.3 Clustering and profiling

To discover the characteristics of the data, unsupervised learning clustering is the first technique that should be applied to the data mining project. In this study, K-means algorithm is combined with a statistical measure called standard score (Z score), which can discover how far away the variable is from the average. The definition of standard score is:

$$SD = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

$$z = \frac{x-\bar{x}}{SD}$$

EQUATION 4-1

where $\bar{x}$ is the mean of input and n is the number of inputs.

First, based on the business unit, customer and product sales transaction data, the standard score will be calculated for each data point. Then the standard scores are used to segment data points into several groups. The next step is to apply K-means algorithm to cluster data points into several clusters. The overlap of the segmentation and the clustering results are considered as the final data profiling results. The details are shown in section 4.3.1, 4.3.2 and 4.3.3.

4.3.1 Clustering and profiling business units

Several clustering models are considered for this exercise. Models are built based on the data such as annual revenue, monthly revenue, customer numbers of a business unit, and product numbers of a business unit. Moreover, monthly revenues percentages of annual revenue are also used as a measure in order to profile and re-profile business units. The percentage model is used to extract business units that have the same sales pattern, but different purchase scales. For instance, assume a business unit in Toronto and a business unit from Halifax have the same sales pattern of contributing 6% to the annual revenue (total is 54%) during each month from January to September; the sales from November to December contribute 46% to the annual revenue. However, the absolute sales amount in Toronto might be several times bigger than the absolute sales amount in Halifax. The percentage model is able to discover the similarity of these two business units. Clustering technique is applied to percentage model twice while the first clustering process is used to remove outliers.

Microsoft (MS) Excel is selected to calculate standard scores. One of the goals of this study is to demonstrate the practical use of data mining to users that are not familiar with the data mining field. Currently MS Excel is one of the most popular calculation and graphing applications on the market. Most people are comfortable to use MS Excel for daily work. The demonstration of the practicality of using MS Excel for data mining will encourage people to actually apply data mining techniques to solve daily business issues. For the same reason, the popular statistics software, SPSS, is chosen for applying K-means clustering technique. The cluster number should be reasonably small in order to facilitate results interpretation. In this study, the number of clusters is initially assigned to 5 or 8. Based on the graphical interpretation of clustering results, the better one will be chosen. The details of each model are shown below.

Model 1: Annual Revenue based model

The standard score of each business unit is calculated based on annual revenues. Based on calculated standard scores, the business units are segmented into 3 groups. At the same time, K-means algorithm is applied for the clustering task.

Standard Score	Number of BU	Annual Revenues (Million)		
		Minimum	Maximum	Mean
Z>=0	7	15.7	750	127.9
0>Z>=-1	28	5	13	8.3
-1>Z	69	-0.009	4.8	2.5
Total	104	BU = Business Unit		

Figure 4-9: Segmentation results of 2004 based on annual revenue (Million)

Cluster Number	Number of BU	Annual Revenues (Million)		
		Minimum	Maximum	Mean
Cluster 1	14	8	17	22.8
Cluster 2	4	23	37	28.2
Cluster 3	52	-0.009	3.5	3.8
Cluster 4	1	750	750	750
Cluster 5	33	3.8	8	10.9
Total	104	BU = Business Unit		

Figure 4-10: Clustering results of 2004 based on annual revenues (Million)

The results, based on company's 2004 transaction data, are shown in Figure 4-9 and Figure 4-10. The minimum, maximum and mean indicates the respective values in each group. Based on the observation of Z score distribution, I separated data at Z=0 and Z=-1. Business units can be distinguished based on segmentation results (Figure 4-9) and clustering results (Figure 4-10). As described in section 4.2.3, cluster 4 is a small cluster whose mean is extremely bigger than others, so I consider it an outlier. This assumption is confirmed by a domain expert, who recognizes that it is a virtual business unit for a special project. If a business unit in cluster 2 has a standard score bigger than 0, it can be confirmed as a top sales business unit. If a business unit belongs to cluster 1 or 2, and the standard score is bigger than 0, then it is considered as a high profit business unit. Similarly, inefficient business units can be extracted using the overlap of a small standard score and clusters whose mean is smaller than others.

	Top Revenue BU	High Revenue BU	Entire Company
Count	4	18	103
%	4%	17%	100%
Sales Revenue (million)	226	545	1100
%	21%	50%	100%

Figure 4-11: Overview of customers based on annual revenues

Figure 4-11 shows the clustering results for the top business units and high profit business units along with the overview of entire business units. As top business units, four business units contribute 21% of sales for the entire company. As beneficial business units, eighteen business units provide 50% of the company's revenues.

Model 2: Customer number based model

Based on the customer numbers of business units, business units can be segmented into several groups using standard scores. The segmentation results are combined with

K-means clustering results in order to identify the business units that have more customers than other business units.

Standard Score	Number of BU	Customer Number		
		Minimum	Maximum	Mean
Z>=10	13	740	1,730	1,000
10>Z>=0	32	445	740	570
0>Z	59	1	435	250
Total	104			

BU = Business Unit

Figure 4-12: Segmentation results of 2004 based on customer number·

Cluster Number	Number of BU	Customer Number		
		Minimum	Maximum	Mean
Cluster 1	21	435	585	510
Cluster 2	7	940	1730	1,200
Cluster 3	18	610	850	700
Cluster 4	32	270	430	360
Cluster 5	26	1	230	105
Total	104			

BU = Business Unit

Figure 4-13: Clustering results of 2004 based on customer number

Figure 4-12 shows the segmentation results and Figure 4-13 shows the clustering results. The business units whose standard scores are bigger than 10 and belong to cluster 2 are considered as units that have large amount of customers. On the other hand, the business units that belong to cluster 5 are the ones which do not have businesses with many customers.

The results of model 2 can also be combined with model 1. For instance, assume I discovered a business unit has more customers than other units and does not belong to the top revenue business units; promotional campaigns, such as cross-selling sales campaigns, can be applied to increase the revenues of these units. Based on the observation of Z score distribution, I separated data at Z=10 and Z=0.

Model 3: Product number based model

Based on the product numbers, business units can be segmented into several groups using standard scores. Similar to model 1 and model 2, the results are combined with K-means clustering results in order to identify the business units that handle more products than other business units.

Standard Score	Number of BU	Product Number		
		Minimum	Maximum	Mean
Z>=10	11	7,460	15,360	9,160
10>Z>=0	40	4,570	6,990	5,675
0>Z	53	1	4,525	2,760
Total	104	BU = Business Unit		

Figure 4-14: Segmentation results of 2004 based on product number

Cluster Number	Number of BU	Annual Revenue (Million)		
		Minimum	Maximum	Mean
Cluster 1	16	1,760	3,360	2,790
Cluster 2	10	7,680	15,360	9,330
Cluster 3	31	5,125	7,458	6,040
Cluster 4	34	3,590	4,940	4,200
Cluster 5	13	1	1,390	480
Total	104	BU = Business Unit		

Figure 4-15: Clustering results of 2004 based on product number

Figure 4-14 shows the segmentation results and Figure 4-15 shows the clustering results. The business units in cluster 2 whose standard score is bigger than 10 are considered as units that have a volume of products sold. On the other hand, the business units in cluster 5 are the ones that do not have many products sold in 2004.

The results of model 3 can also be combined with model 1. For instance, assume I discovered a business unit that handles more products than other units, yet does not belong to the top revenue business units; explanations can include the following: refund costs or sales costs that include shipping fees are too high; gross margins are too low; the quantity of sold products is small. Through the identification of the reason, the business process of the business unit can be improved and optimized.

Model 4: Monthly revenue based model

Clustering techniques are not only used to discover the data characteristics based on one or multiple attributes, but also to discover the suitable patterns of attributes that help future model building. Models 4 and 5 are used to discover suitable prediction models for further prediction studies. In model 4, K-means algorithm is applied based on monthly sales data. The training set consisted of around 100 data points, where each data point was a 12-dimensional vector <x_1, ..., x_12>, where x_i represents the profit in month i for unit j, 1<= j <= the number of data points. The output of the model gave a single number, y, which represented the predicted average profit for the next month for an average business unit. Once the model is trained/optimized, the input given to the model was a vector V of 12 revenue values (one per month), and the output, y=f(V), was a predicted revenue value for the next month. The error was calculated as the difference between y and the actual average profit for around 100 business units that were considered. The focus of this study is to separate business units by the pattern of sales movement. For instance, the business units whose sales revenues decrease in the first half year and increase from July to December need to be grouped together; the business units whose sales revenues increase in first half year and decrease from July to December need to be grouped together. Model 4 facilitates prediction techniques to stand out the focus and discover the most suitable patterns.

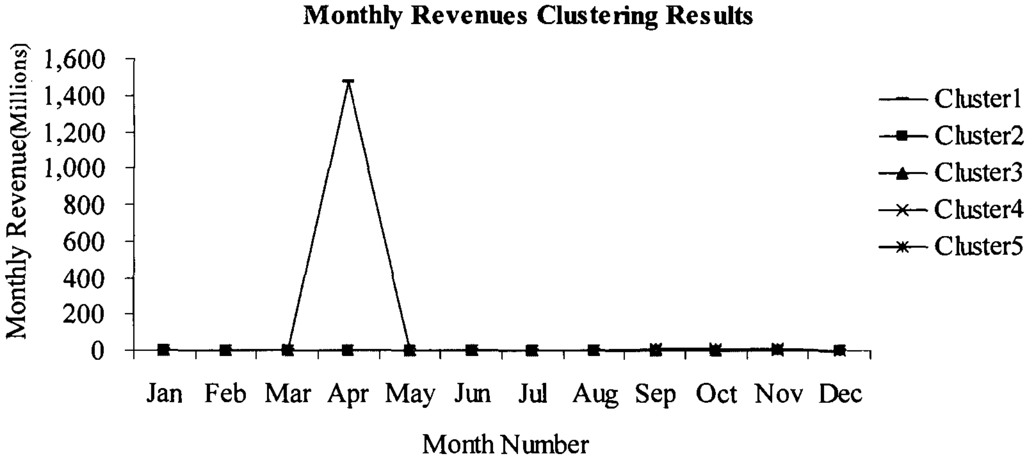


Figure 4-16: Monthly revenues clustering results with outliers

The cluster results are shown in Figure 4-16. The absolute values in cluster 1 are extremely bigger than other clusters. It can be considered as an outlier. Cluster 3 only has one member and it has a significant growth in November. At the same time, the data in cluster 3 also has a negative value in December which is only 10% different from November's positive revenue. As already discussed in section 4.2.3, this pattern can be removed from our studies.

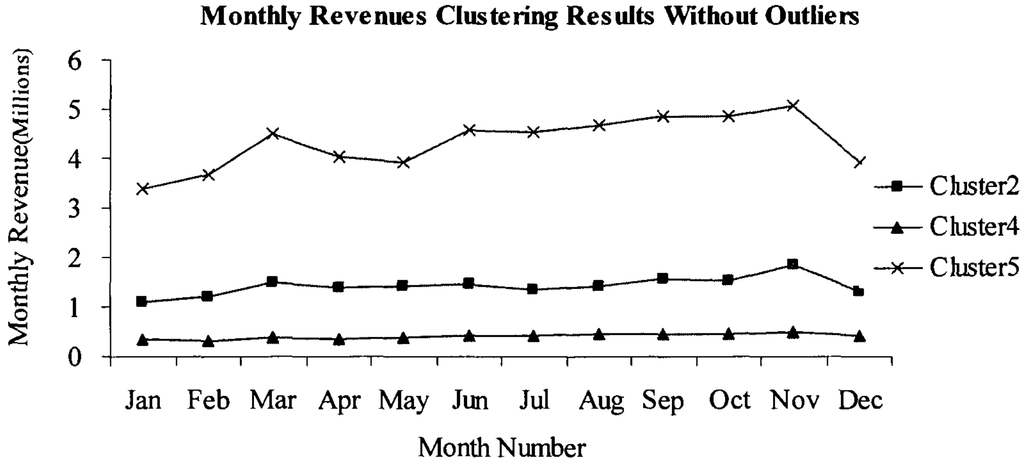


Figure 4-17: Monthly revenues clustering results without outliers

Figure 4-17 shows the clustering results without outliers. It indicates that the business units can be divided into 3 groups based on monthly revenue data. The monthly revenue average of cluster 2 is about $1.4 million; the average of cluster 4 is about $0.4 million; the average of cluster 5 is about $4.3 million. This result helps the company to better distribute its resources.

Model 5: Monthly sales percentage based model

As described at the beginning of this section, I do not use the absolute values of monthly revenues' in this model. The monthly revenues percentages of annual sales are applied for the clustering in order to discover the similarity of sales pattern regardless of the scale of business units. In this model, the clustering technique will be applied twice. The first application identifies the outliers. After removing the outliers from the raw data, a second clustering discovers the suitable grouping for further prediction studies.

Cluster Number	Number of Business Unit	Percentages
1	38	37%
2	4	4%
3	45	43%
4	3	3%
5	14	13%
Total BU	104	100%

Figure 4-18: The overview of clustering results based on monthly revenues

Figure 4-18 is the overview of the clustering results. Percentages represent total revenues ratio that clusters have. The distributions for each cluster are shown from Figure 4-19 to Figure 4-23.

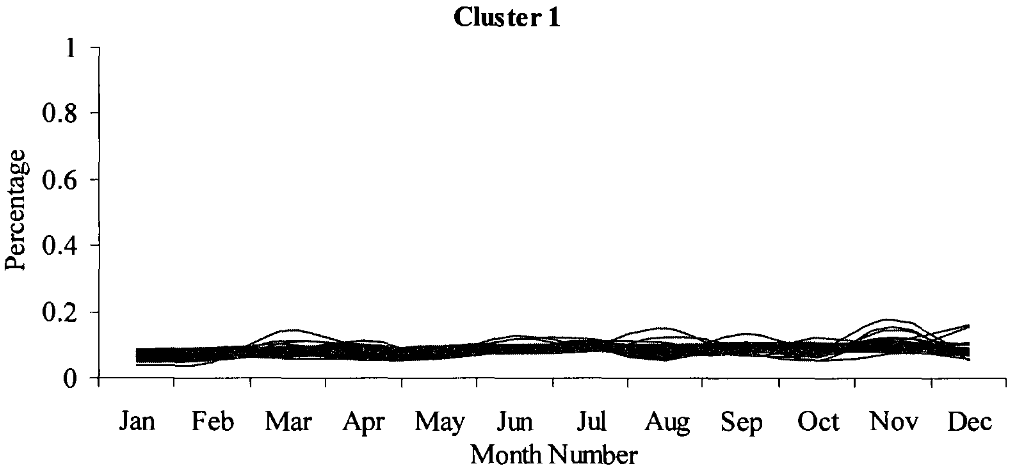


Figure 4-19: Clustering results based on monthly revenues percentage (cluster 1)

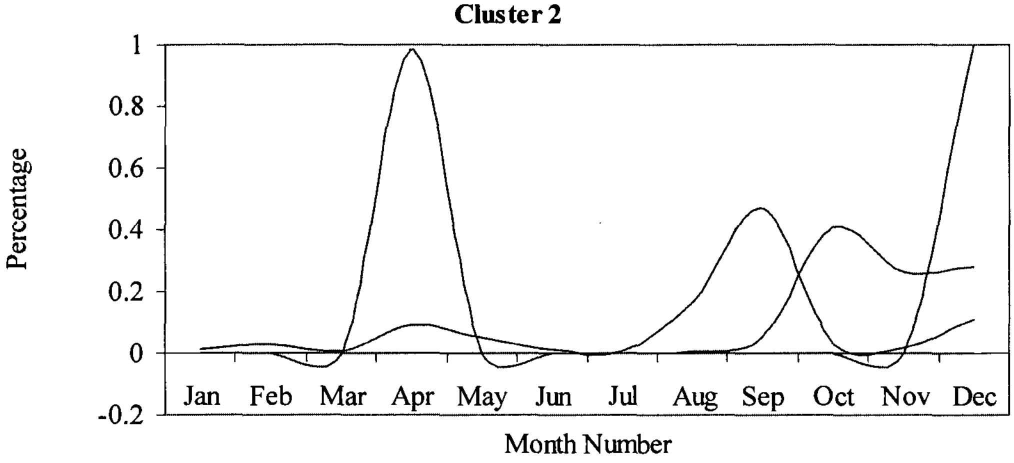


Figure 4-20: Clustering results based on monthly revenues percentage (cluster 2)

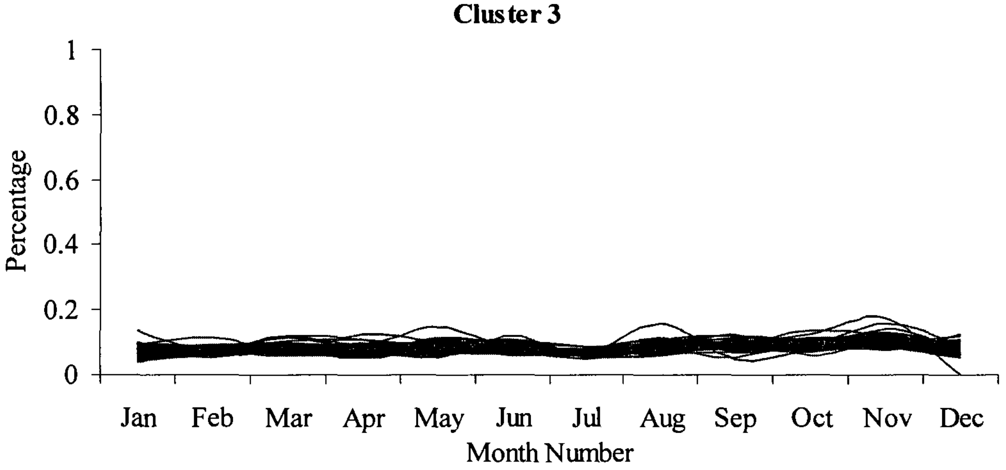


Figure 4-21: Clustering results based on monthly revenues percentage (cluster 3)

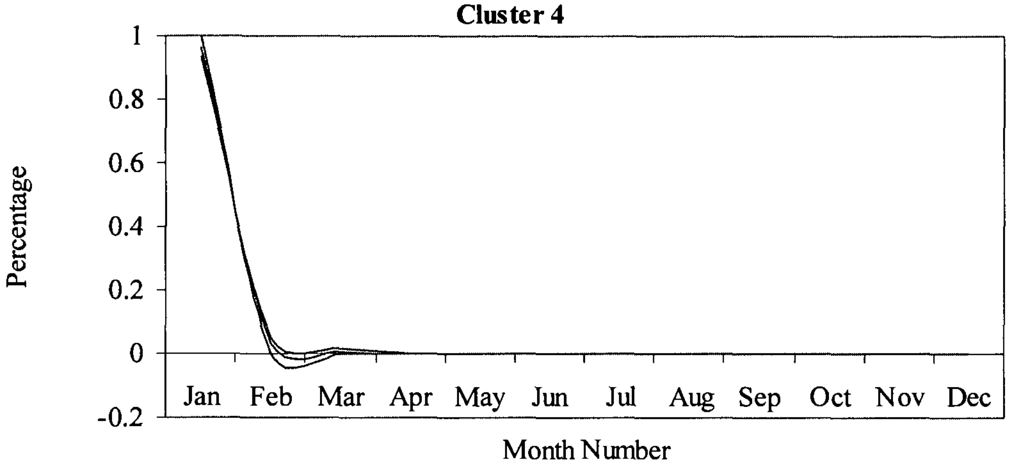


Figure 4-22: Clustering results based on monthly revenues percentage (cluster 4)

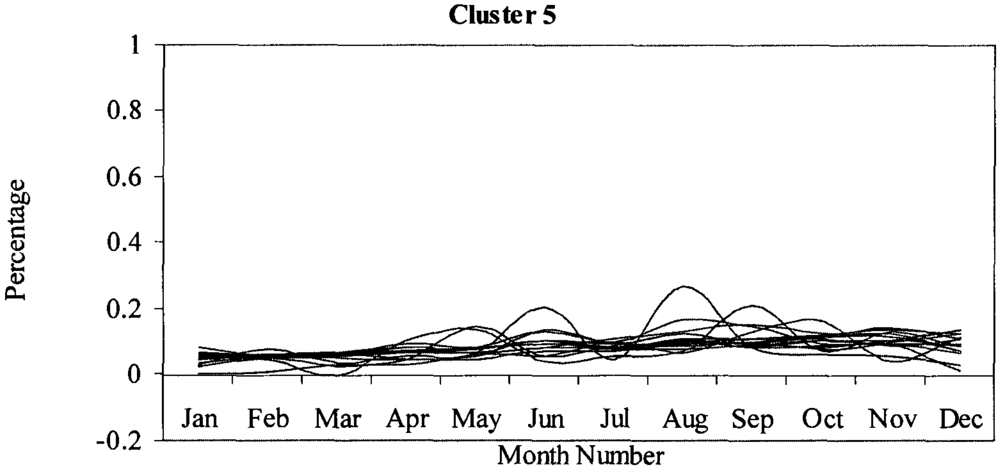


Figure 4-23: Clustering results based on monthly revenues percentage (cluster 5)

The group of figures (Figures 4-19 to 4-23) shows objects from different clusters of branches based on a ratio of monthly and annual revenues. The clustering seems to separate stable and volatile patterns, and helps improve the prediction of annual revenues for a branch. The cluster-1, cluster-3 and cluster-5 consist of stable patterns. The sales amounts remain almost same for each month. The cluster-2 consists of unstable patterns. The cluster-4 consists of patterns which only have business activities in the first quarter. From Figure 4-20 and Figure 4-22, I can easily recognize the uncertain and unusual patterns found in cluster 2 (four business units) and cluster 4 (three business units). Moreover, the data in cluster 1 and cluster 3 was not separated successfully. Therefore I remove these two clusters from our studies and focus on the remaining 97 business units (clusters 1, 3, 5). The re-clustering results based on the percentages of monthly revenues in the annual sales are shown in Figure 4-24 to Figure 4-31. The results of re-clustering will be used as prediction models in further prediction studies.

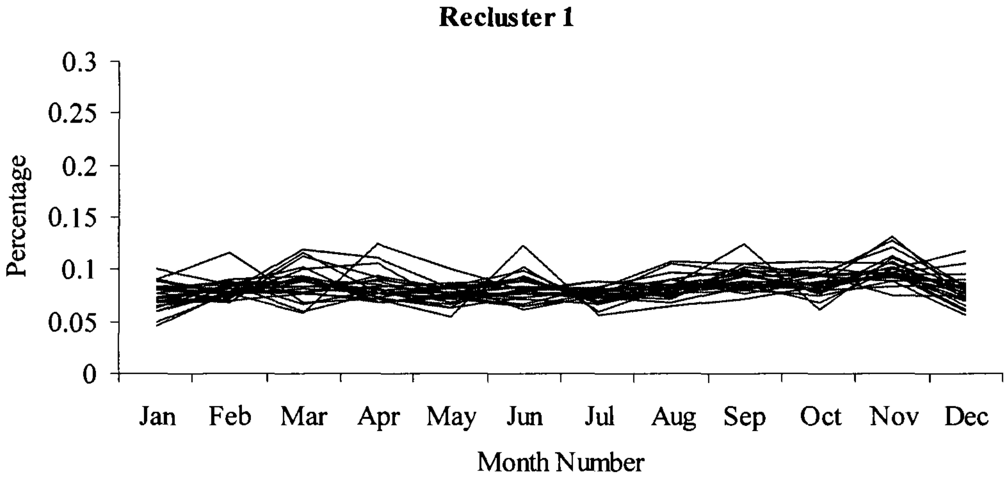


Figure 4-24: Re-clustering results based on monthly revenues percentage (cluster 1)

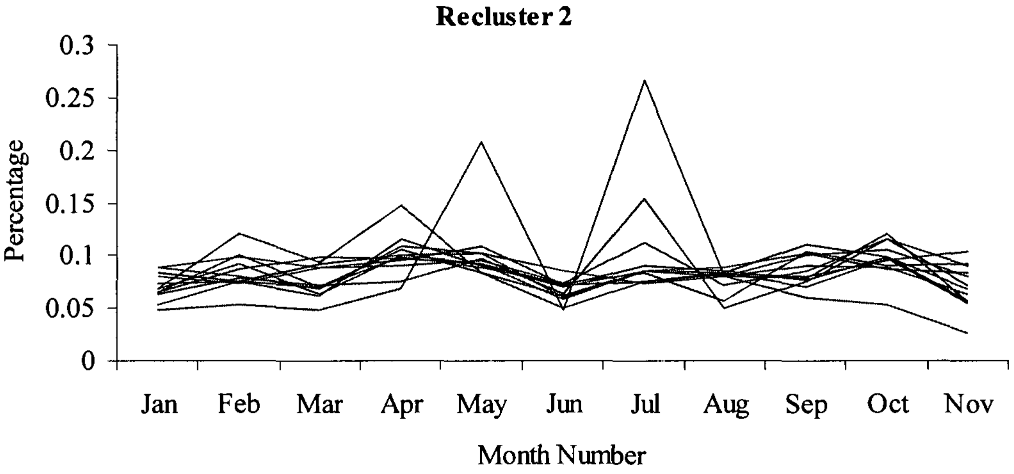


Figure 4-25: Re-clustering results based on monthly revenues percentage (cluster 2)

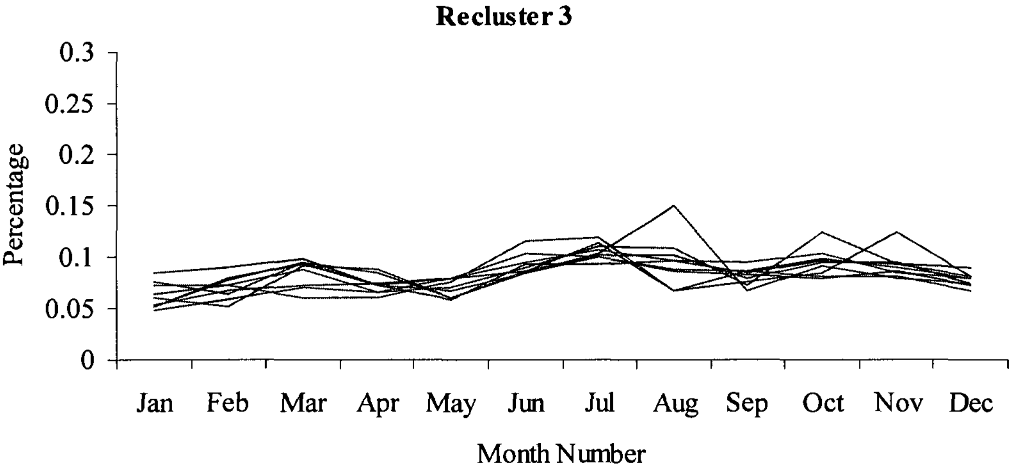


Figure 4-26: Re-clustering results based on monthly revenues percentage (cluster 3)

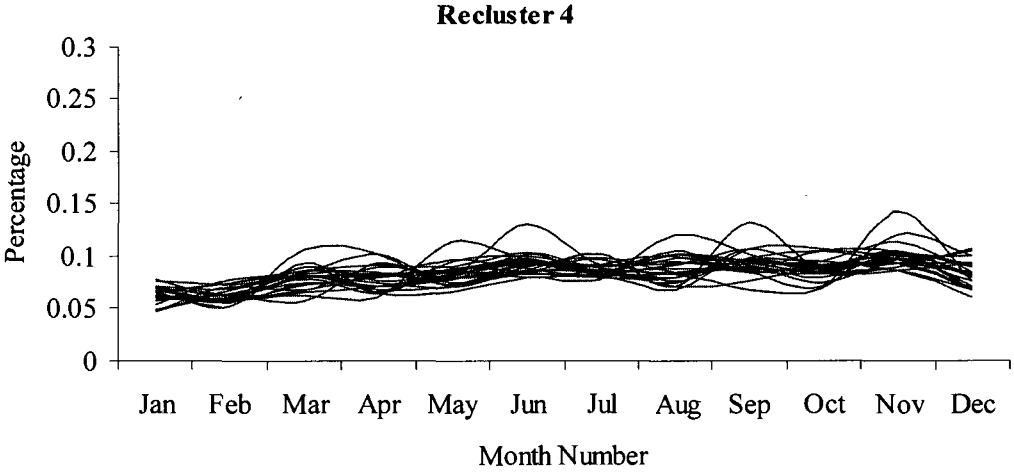


Figure 4-27: Re-clustering results based on monthly revenues percentage (cluster 4)

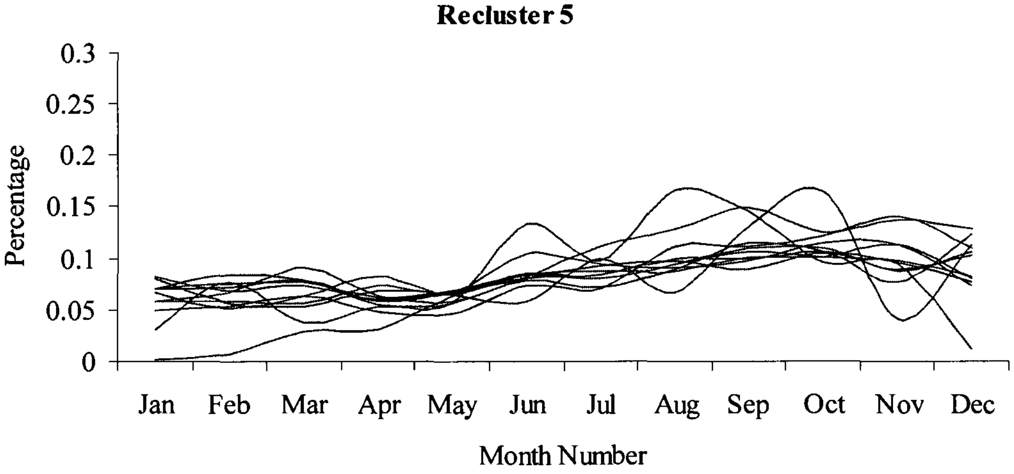


Figure 4-28: Re-clustering results based on monthly revenues percentage (cluster 5)

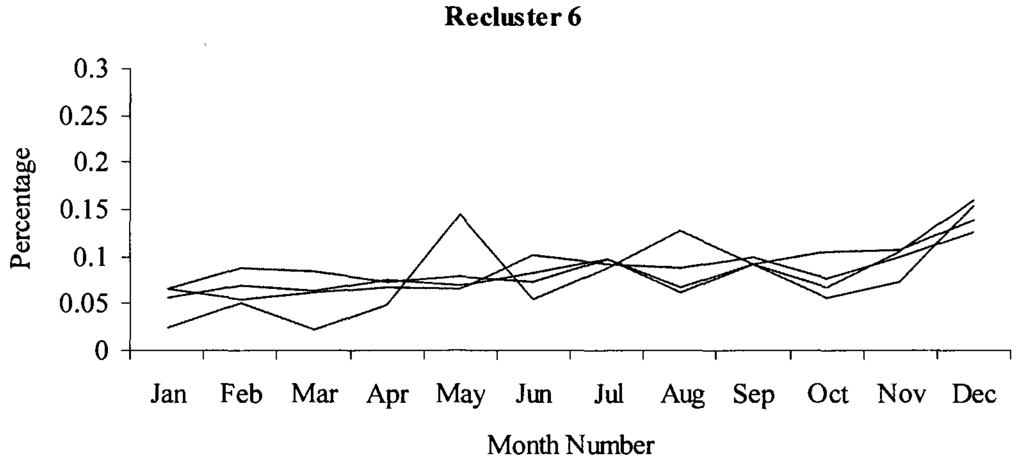


Figure 4-29: Re-clustering results based on monthly revenues percentage (cluster 6)

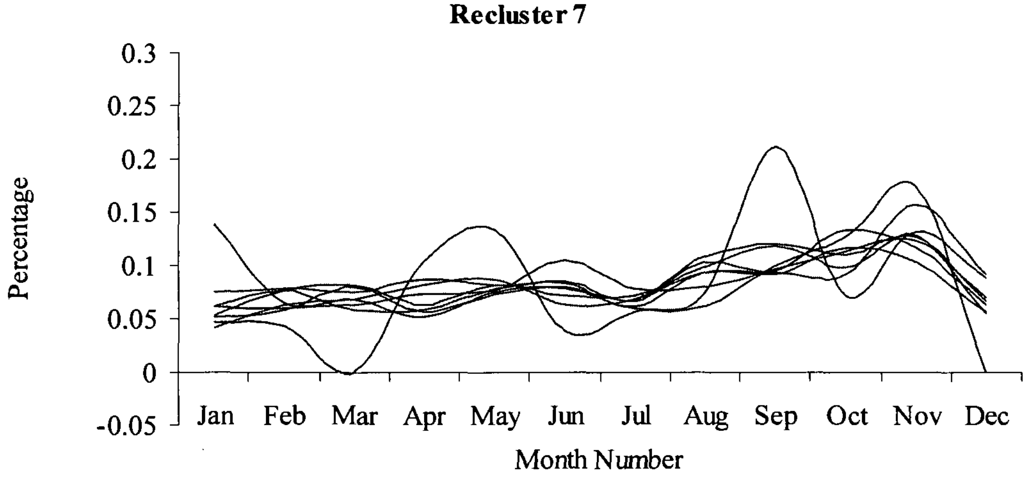


Figure 4-30: Re-clustering results based on monthly revenues percentage (cluster 7)

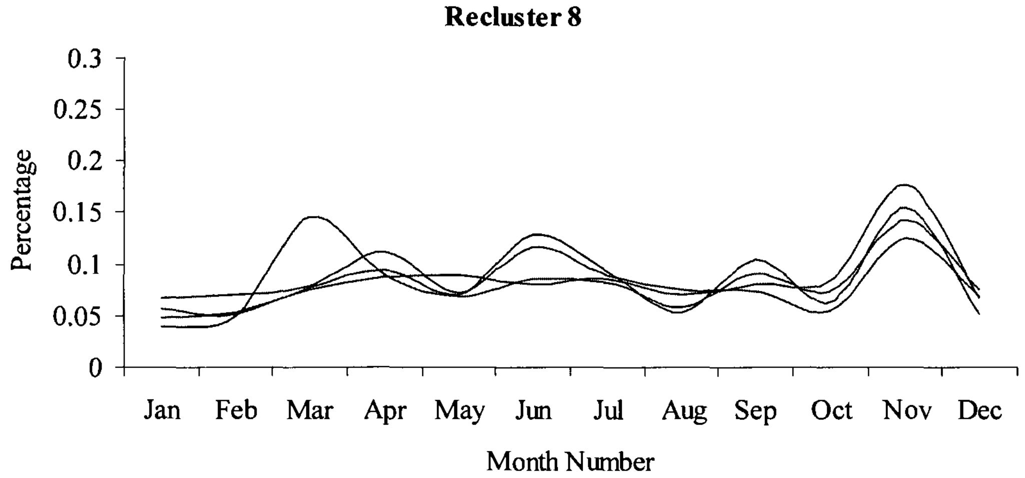


Figure 4-31: Re-clustering results based on monthly revenues percentage (cluster 8)

The group of figures (Figures 4-24 to 4-31) shows re-clustering results based on a ratio of monthly and annual revenues. The recluster-1 consists of stable patterns and seems to be similar to recluster-3. The recluster-2 consists of unstable patterns which have obvious up and down from April to August. The recluster-4 consists of stable patterns. Sales amounts of this cluster slightly increase from month to month. The recluster-5 consists of unstable patterns. For most business units in this cluster, sales amounts slowly increase from month to month before November and slightly drop in December. The recluster-6 consists of unstable patterns. The recluster-7 consists of unstable patterns and seems to be similar to recluster-5. However, it is more volatile than cluster 5 and always has a big drop the end of the year. The recluster-8 consists of stable patterns. The re-clustering results can be used for further prediction.

4.3.2 Clustering and profiling customers

Similar to the study for business units, the combination of K-means algorithm and standard scores are used to cluster and profile customers. The following algorithm is applied to cluster customers. The purpose of this clustering is to discover top and high revenue and profit customers. We use 2004 transaction data to demonstrate the algorithm. Based on domain expert's advice, several fictitious customers which are used to facilitate the company's internal resource transfer were excluded from this study. Therefore the total customers' number in 2004 transactions is 23,613.

Step 1:

- Segmentation based on standard score of annual revenue for all of customers in 2004 transaction data.
- Segmentation based on standard score of annual profits for all of customers in 2004 transaction data.
- Select the customers that have high standard scores on both annual revenue and annual profits.

Figure 4-34 shows the segmentation results of 2004 customers. In the Figure 4-34, the total revenue and total profit of each group are addressed.

Step 2:

- Clustering based on annual revenue for all of customers in 2004 transaction data.
- Clustering based on annual profits for all of customers in 2004 transaction data.
- Select the customers that are in high revenue group and high annual profits group.

Step 3: Take overlap of previous two steps' results as top customers.

	Top Customers	High Profit	Total
Customer Number	599	2,292	23,613
%	2.54%	9.71%	100%
Total Revenue*	248.0	375.9	486.4
%	50.99%	77.28%	100%
Total Profit*	40.0	64.1	86.7
%	46.19%	74.05%	100%

Figure 4-32: Clustering and profiling results of 2004 top customers and high profit customers (*: Unit =Million)

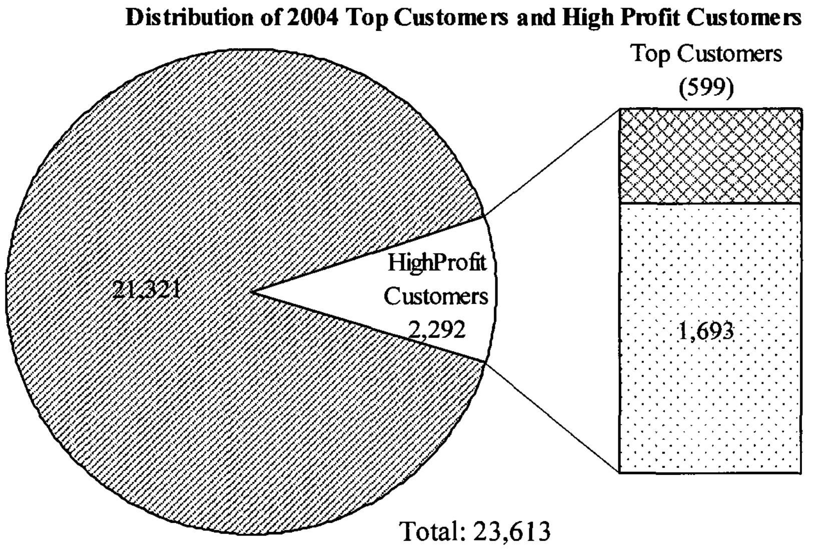


Figure 4-33: Distribution of 2004 top customers and high profit customers

Figure 4-32 shows the clustering and profiling results of top customers and high profit customers. According to Figure 4-32, high profit customers (2,292) represent 10% of all customers and contributed about 77% of revenues and 74% of profits. Top customers (599) represent only 2.54% of all customers and contributed about 51% of revenues and 46% of profits. Figure 4-33 graphically shows clustering and profiling results of top customers and high profit customers among all the 2004 customers.

	Z<-0.1	-0.1=<Z<0	0=<Z<0.1	0.1=<Z<1	1=<Z<10	Z>=10	Total
Number	19,294	162	151	929	2,325	752	23,613
%	81.71%	0.69%	0.64%	3.93%	9.85%	3.18%	100%
Total Revenue*	57.8	2.9	2.9	21.2	136.7	264.86	486.4
%	11.89%	0.59%	0.60%	4.37%	28.10%	54.46%	100%
Total Profit*	11.7	0.6	0.6	4.3	25.4	44.2	86.7
%	13.52%	0.67%	0.65%	4.96%	29.26%	50.94%	100%

Figure 4-34: Segmentation results of 2004 customers (*: Unit =Million)

Figure 4-34 shows the segmentation results of 2004 customers. In the Figure 4-34, the total revenue and total profit of each group are addressed.

	Z<-0.1	-0.1=< Z <0	0=<Z <0.1	0.1=<Z<1	1 =<Z<100	Z>=100	Total
Product	8	75923	3143	1874	193	1	81,142
%	0.01%	93.57%	3.87%	2.31%	0.24%	0.00%	100%
Total *	-0.9	127.0	80.2	168.9	187.2	736.5	1299.5
%	-0.03%	4.89%	3.08%	6.50%	7.21%	28.34%	100%
Total Profit *	-0.9	25.41	14.0	25.1	23.2	736.5	823.7
%	-0.06%	1.54%	0.85%	1.52%	1.41%	44.70%	100%

Figure 4-35: Segmentation results of 2004 products (*: Unit =Million)

Figure 4-35 shows the segmentation results of 2004 products. In the Figure 4-35, the total revenue and total profits for each group are addressed.

4.3.3 Clustering and profiling products

We use a similar algorithm with customers' clustering to discover the top products. The 2004 transaction data is used to demonstrate the algorithm. Therefore the total number of products in 2004 transactions is 81,142.

Step 1:

- Segmentation based on standard score of annual revenue for all the products in 2004 transaction data.
- Segmentation based on standard score of annual profits for all the products in 2004 transaction data.
- Select the products that have high standard score on both annual revenue and annual profits.

Figure 4-35 shows the segmentation results of 2004 products. In the Figure 4-35, the total revenue and total profits for each group are addressed. Based on the description in section 4.2.3, the one member in Z>=100 group can be excluded from this study as confirmed by the domain expert.

Step 2:

- Clustering based on annual revenue for all the products in 2004 transaction data.
- Clustering based on annual profits for all the products in 2004 transaction data.
- Select the products that are in high revenue group and high annual profits group.

Step 3:

- Take overlap of previous two steps' results as top products.

	Top Products	Total
Product Number	130	81141
%	0.16%	100%
Total Revenue*	326.4	1125.6
%	29.00%	100%
Total Profit*	41.5	174.5
%	23.78%	100%

Figure 4-36: Clustering and profiling results for top products (*: Unit =Million)

Figure 4-36 shows the clustering and profiling results of top products. According to Figure 4-36, top products (130), which represent only 0.16% portion of the entire product line, contributed about 29% of revenues and about 24% of profits. However, the domain expert pointed out that there are 16 temporary products included in the discovered top products. These temporary products exist only for some temporary transactions; therefore they need to be excluded from the top products list.

	Top Products (Temp Products Removed)	Total
Product Number	114	81125
%	0.14%	100%
Total Revenue*	162.3	961.6
%	16.89%	100%
Total Profit*	21.2	154.2
%	13.75%	100%

Figure 4-37: Clustering and profiling results for top products without temporary products (*: Unit =Million)

Figure 4-37 shows clustering and profiling results of top products after removing the 16 temporary products.

4.4 Sales predictions for business units

The data mining tasks have to serve the company's business needs. In this project, the main business issue is to predict the next year's revenue for business planning and decision making. Therefore, the prediction part in this project will focus on annual revenue prediction at business unit level.

Prediction is completed by two steps: data model construction and data mining techniques selection. Data models are constructed following the nature of the business, the goals of the data mining activities and the character of data mining techniques. Several data models need to be created and compared, and the most suitable model will be chosen from results comparison. All results are evaluated based on error rates and summarized by histogram. Furthermore, the most suitable data mining technique is selected by comparing data mining results based on the selected model.

The prediction section starts from the existing practice. Multiple regression technique will be used to discover the best model for this study. In this section, we will compare three models: Model 1: 2006 annual revenue prediction is using nine monthly sales data as predictors, Model 2:2006 annual revenue prediction is using three quarterly sales data as predictors, Model 3: 2007 half annual revenue prediction is using seven quarterly sales data as predictors. The difference between three models is the time period that is used for colleting and summarizing data. The details will be explained in the section 4.4.2. After discovering the best model, several prediction techniques such as SVR and Neural Networks, will be applied to the best model. The results comparison is also addressed.

4.4.1 Existing practice

The multinational retailer currently uses a benchmark algorithm to predict the annual revenue of next year. We demonstrate its algorithm based on the business units clustering results (refer to section 4.3.1). The outliers I discovered using clustering are excluded from this study.

Based on total revenue from January to September of previous year (2004) and this year (2005), the annual rate of growth can be calculated as:

$$\mathrm{GR} = \frac{\sum_{i=1}^{9} \text{Monthly Sales Amount of This Year} - \sum_{i=1}^{9} \text{Monthly Sales Amount of Previous Year}}{\sum_{i=1}^{9} \text{Monthly Sales Amount of Previous Year}}$$

GR= GROWTH RATE EQUATION 4-2

This growth rate can be used to predict revenue from October to December of this year (2005). It also indicates that annual revenue of this year (2005) is able to be predicted based on:

$$\text{Sales Amount of Oct to Dec, 2005} = \left(\sum_{i=10}^{12} \text{Monthly Sales Amount of First Year}\right) \times (1 + \mathrm{GR})$$

EQUATION 4-3

As a result, the annual revenue of next year (2006) can be predicted based on predicted annual revenue of this year (2005).

$$\text{Annual Sales Amount of 2006} = (\text{Annual Sales Amount of 2005}) \times (1 + \mathrm{GR})$$

EQUATION 4-4

The error in this report will be calculated as equation 4-5.

$$\text{Error Rate} = \frac{|\text{Predicted Value - Actual Value}|}{\text{Actual Value}}$$

EQUATION 4-5

Error	Number of Business Units	Cumulative
0.1	35	36.08%
0.2	26	62.89%
0.3	17	80.41%
0.4	7	87.63%
0.5	5	92.78%
0.6	1	93.81%
0.7	1	94.85%
0.8	0	94.85%
0.9	3	97.94%
1	1	98.97%
More	1	100%
Total	97	-

Figure 4-38: 2005 October to December business unit revenue prediction error histogram

Figure 4-38 shows the prediction results of 2005 October to December's revenue using the error histogram.

Error	Number of Business Units	Cumulative
0.1	38	39.18%
0.2	17	56.70%
0.3	10	67.01%
0.4	13	80.41%
0.5	4	84.54%
0.6	3	87.63%
0.7	4	91.75%
0.8	3	94.85%
0.9	2	96.91%
1	1	97.94%
More	2	100%
Total	97	-

Figure 4-39: 2006 business unit revenue prediction error histogram

Figure 4-39 shows the prediction results of the 2006 revenue using the error histogram.

4.4.2 Multiple regression

Multiple regression is firstly applied to all business units without using clustering results (refer to section 4.3.1). The results are shown in Figure 4-40.

Error	Number of Business Units	Cumulative
0.1	48	49.48%
0.2	25	75.26%
0.3	11	86.60%
0.4	4	90.72%
0.5	1	91.75%
0.6	2	93.81%
0.7	1	94.85%
0.8	1	95.88%
0.9	0	95.88%
1	0	95.88%
More	4	100.00%
Total	97	-

Figure 4-40: 2006 business unit revenue prediction error histogram

About 86% percent of errors from the multiple regression model are less than 30%. However, application of all available data is very computational expensive, which is not practical for large datasets. Therefore, clustering results are combined with prediction techniques. By using cluster method, we may obtain equally good or better results from partial source data depending on the quality of resource. Based on clustering results (refer to section 4.3.1), multiple regressions can be applied to discover relationships between inputs and outputs for each group. The clustering techniques are able to group similar data together to enhance the similarity within each cluster. The clustering techniques are used to prepare for the prediction study. We apply multiple regression with several models. The most suitable model will be chosen for applying other prediction techniques.

Model 1: 2006 annual revenue prediction using nine months sales data

This model uses nine monthly sales data as predictors, which is based on monthly revenue data from January 2004 to September 2004, a suitable formula is calculated for each group to predict the annual revenue of 2005. The formula can then be used to predict annual revenue of 2006 based on the monthly sales data from January 2005 to September 2005. However, since this model uses nine variables, it can be only applied to clusters which contain more than ten data points. Therefore, 76 business units out of 104 business units have prediction results.

Model 2:2006 annual revenue prediction using three quarters sales data

This model uses three quarterly sales data as predictors, which is based on quarterly revenue data from January 2004 to September 2004 and annual revenue of 2005. This model can be used to predict annual revenue of 2006 based on sales data of January 2005 to September 2005.

Model 3: 2007 half annual revenue prediction using seven quarters sales data

This model uses seven quarterly sales data as predictors, which is based on quarterly revenue data from January 2004 to September 2005. The half annual revenue of next year is the target of prediction. The characteristic of this model is to use seven quarters (21 months) data to train the model. Then this model is applied to predict first half annual amount of 2007 based on quarterly sales data of January 2005 to September 2006.

Figure 4-41 shows the evaluation results of built models. One business unit is absorbed by another business unit, so the total number of business units for seven quarters model is 96.

Error	9 Months Model Error		3 Quarters Model Error		7 Quarters model Error	
	Frequency	Cumulative	Frequency	Cumulative	Frequency	Cumulative
0.1	53	69.74%	44	45.36%	65	67.71%
0.2	17	92.11%	22	68.04%	14	82.29%
0.3	4	97.37%	17	85.57%	9	91.67%
0.4	1	98.68%	5	90.72%	3	94.79%
0.5	0	98.68%	4	94.85%	4	98.96%
0.6	1	100%	2	96.91%	1	100%
0.7	0	100%	0	96.91%	0	100%
0.8	0	100%	1	97.94%	0	100%
0.9	0	100%	0	97.94%	0	100%
1	0	100%	0	97.94%	0	100%
More	0	100%	2	100.00%	0	100%
Total	76 Business units		97 Business units		96 Business units	

Figure 4-41: Three prediction models' evaluation

Figure 4-41 indicates that there is no one significantly bad model of the three models. Each model is able to describe the 85% relationship between inputs and outputs within a 30% margin of error, which indicates the three models' usability has to be determined in prediction results evaluation. However, in seven quarters model, overfitting occurs in two clusters since the number of variables is close to the number of training data set points. Usually the number of data points in the training data needs to be more than 3-5 times bigger than the number of variables in order to avoid overfitting. If number of variables is large, it will increase the chances of overfitting.

The comparison of three multiple regression models

Figure 4-42 shows the evaluation results of predictions using the three models I described previously including benchmark model results and the model without using clustering results. Three quarters model provides slightly better results than others including the benchmark model that the company is currently using. Seventy-eight percent of errors from the linear regression model are less than 30%, while the sixty-seven percent of errors from the benchmark model are less than 30%. Two out of 97 business units are mores than 100%, while five out of 97 business units from the model without using clustering results are more than 100%. Moreover, as I described before this model is not practical for large datasets. Using nine months and seven quarters model leads to results that are similar to the benchmark. Based on the comparison results, the three quarters model technique is the best among the considered models to predict future annual revenue.

The reason that there was not significant improvement from existing benchmark model's prediction results is shown below.

> The quantity of training data was not enough to discover the optimum relationship between first year revenue and second year revenue. Due to the limitation of training data, the model could only be trained once, which was not enough to achieve a high precision. Furthermore, the second year's annual revenue is volatile; it depends on many business factors. Therefore, only one time training could not deduce the effect of the natural fluctuation in annual sales.

Error	Benchmark		No Clustering model		9 Months Model Error		3 Quarters Model Error		7 Quarters model Error	
	Frequency	Cumulative	Frequency	Cumulative	Frequency	Cumulative	Frequency	Cumulative	Frequency	Cumulative
0.1	38	39.18%	43	44.33%	19	25.00%	36	37.11%	33	34.38%
0.2	17	56.70%	24	69.07%	23	55.26%	25	62.89%	19	54.17%
0.3	10	67.01%	8	77.32%	6	63.16%	15	78.35%	9	63.54%
0.4	13	80.41%	8	85.57%	7	72.37%	7	85.57%	9	72.92%
0.5	4	84.54%	4	89.69%	9	84.21%	5	90.72%	9	82.29%
0.6	3	87.63%	3	92.78%	4	89.47%	2	92.78%	5	87.50%
0.7	4	91.75%	2	94.85%	4	94.74%	2	94.85%	3	90.63%
0.8	3	94.85%	0	94.85%	1	96.05%	2	96.91%	0	90.63%
0.9	2	96.91%	0	94.85%	1	97.37%	1	97.94%	1	91.67%
1	1	97.94%	0	94.85%	2	100.00%	0	97.94%	2	93.75%
More	2	100.00%	5	100.00%	0	100.00%	2	100.00%	6	100.00%
Total	97 Business units		97 Business units		76 Business units		97 Business units		96 Business units	

Figure 4-42: Prediction results evaluation

Figure 4-42 shows the prediction results evaluation of different models I described previously including benchmark model and the model without using clustering.

4.4.3 SVR

Since the three quarters model is selected as best one, it is the one used in this study. The input data is the revenue data from January 2004 to September 2004 and the output data is the annual revenue data of 2005. The input and output data must to be normalized to between 0 and 1. The min-max normalization is used here. SVR is processed by using MATLAB 7.0 and SVM package which is created by Steve Gunn (Image Speech and Intelligent Systems Group, University of Southampton). This model is able to predict the annual revenue for 2006 based on three quarterly revenues of 2005. The cubic function (power=3) is chosen here.

The results of this model will be shown in section 4.4.5.

4.4.4 Neural networks

The neural networks technique is also applied based on the most suitable model – three quarters data based model, which is discovered by using multiple regression. Three quarterly revenue data of 2004 and annual revenue data of 2005 are used as training data. A software JavaNNS is applied to advance this model. The data needs to be normalized between 0 and 1 before actually using it. The min-max normalization is used here. This model is able to predict annual revenue of 2006 based on the sales data of January 2005 to September 2005.

The results of this model will be shown in the next section.

4.4.5 Results comparison

Figure 4-43 shows the evaluation results of the SVR model and neural network model.

Error	Neural Networks		SVR	
	Frequency	Cumulative	Frequency	Cumulative
0.1	18	18.56%	18	18.56%
0.2	17	36.08%	17	36.08%
0.3	13	49.48%	14	50.52%
0.4	11	60.82%	7	57.73%
0.5	11	72.16%	11	69.07%
0.6	6	78.35%	4	73.20%
0.7	1	79.38%	4	77.32%
0.8	1	80.41%	4	81.44%
0.9	1	81.44%	1	82.47%
1	4	85.57%	2	84.54%
More	14	100.00%	15	100.00%
Total	97 Business Units			

Figure 4-43: SVR and neural network comparison results

From the Figure 4-43, I know that the SVR and Neural Networks methods each produce similarly accurate results. Therefore, I pick SVR as our nonlinear prediction method since the operation of SVR is simpler than Neural Networks'. No normalization is needed in SVR.

Error	Benchmark		Linear Regression		SVR	
	Frequency	Cumulative	Frequency	Cumulative	Frequency	Cumulative
0.1	38	39.18%	36	37.11%	20	20.62%
0.2	17	56.70%	25	62.89%	17	38.14%
0.3	10	67.01%	15	78.35%	7	45.36%
0.4	13	80.41%	7	85.57%	10	55.67%
0.5	4	84.54%	5	90.72%	3	58.76%
0.6	3	87.63%	2	92.78%	6	64.95%
0.7	4	91.75%	2	94.85%	3	68.04%
0.8	3	94.85%	2	96.91%	1	69.07%
0.9	2	96.91%	1	97.94%	1	70.10%
1	1	97.94%	0	97.94%	1	71.13%
More	2	100%	2	100%	28	100%
Total	97 Business Units					

Figure 4-44: 2006 business unit revenue prediction results

Figure 4-44 shows the predictions results for the benchmark model, the linear regression model and the SVR model. From the Figure 4-44 I know the linear regression is more accurate than the benchmark model that the multinational retailer is currently using. Both the benchmark model and the linear regression model achieved better results than the SVR model. However, there is not enough improvement to recommend using linear regression once the cost of applying the technique is considered. Therefore prediction results can only be a reference measure since accuracy of the results is not stable enough to be an independent indicator in real world business.

4.4.6 Inventory prediction

Product inventory management is another important aspect of the company's business. The quantities of products sold relate directly to the revenues of business units. As products are sold, the inventory must be replaced efficiently in order to maximize revenues and minimize costs. Thus inventory prediction assists inventory management by identifying how much stock must be purchased to respond to future sales demands. However, as I described in previous chapters, data mining is merely a technique applied in response to business issues and interests. The domain expert of our multinational retailer currently considers inventory prediction as low priority issue. While inventory prediction in an important part of our proposed enterprise-wide data mining process with many potential benefits, I have not applied this technique to the multinational retailer. Inventory prediction may be highly effective for other retailers depending on their unique business issues and interests. In fact, the inventory prediction is an important issue for the specialty grocery store. Therefore, a detailed description of the process will appear in chapter 5.

4.5 Market basket analysis

Market basket analysis is another important feature of data mining. It is able to determine the relationships between products. Through consultations with the domain expert, I confirmed that the multinational retailer highly values information that reveals the relationships between products in order to benefit from cross-selling of complementary goods. To be effective, these relationships must also be related to customer clusters that have also been uncovered through clustering techniques. An interesting topic for future consideration may be the application of these techniques to reveal valuable insight into the supply-side issues of the company. Since the association technique is computationally expensive, this study is focusing on the top customers and products discovered in section 4.3. The SQL Server 2005 Data Mining Add-ins for Office 2007 is used for this study.

4.5.1 Association

The first approach is to apply association technique based on all the transactions involving top products purchased by top customers (about 2,300). The products are linked by customers' information. For each time a customer purchases two (or more) products together, I infer that a relationship exists between those products; the frequency of finding these products purchased together indicates the strength of their relationship. These relationships may be refined by further considering the relationship strength within each subgroup of the top customers.

	Entire	Subgroup 1		Subgroup 2	
The Number of Transactions	191,926	19,564	10%	25,098	13%
The Number of Customers	2033	35	2%	80	4%
Minimum Support	20%	29%		44%	
Minimum Probability	0.71	1		0.92	
Minimum Importance	0.85	1.09		1	

Figure 4-45: Overview of the dataset and association parameters

Figure 4-45 shows an overview of the target datasets along with the association parameters. The “entire” in Figure 4-45 indicates all top customers’ transactions data while subgroup 1 has the transactions belonging to the customers whose total revenue is bigger than two million; subgroup 2 shows the transactions belonging to customers whose total revenue is between one and two million. The minimum support, the minimum probability and the minimum importance values are also identified.

Itermsets	Support		
	Entire	Subgroup 1	Subgroup 2
235642, 235641	23%	29%	49%
235641, 207496	22%	29%	44%
232201, 237661	20%	-	44%
235642, 207496	20%	29%	-
233449, 235641	20%	29%	45%
237661, 235641	20%	-	48%

Figure 4-46: Examples of itemsets

Figure 4-46 shows some examples of discovered itemsets. For instance, from all top customers’ transactions data, I discovered that 23% of them bought product 235642 and product 235641 together; on the other hand, 29% of customers in subgroup 1 and 49% of customers in subgroup 2 purchased product 235642 and product 235641 together.

Discovered Rules	Probability	Importance
235642 -> 235641	92 %	1.13
235641 -> 235642	82 %	1.44
235642 -> 207496	79 %	0.85
235641 -> 207496	78 %	0.93
207496 -> 235641	77 %	0.94
235641 -> 233449	71 %	1.20
207496 -> 235642	71 %	0.98
235641 -> 237661	71 %	1.27

Figure 4-47: Examples of association rules

Figure 4-47 shows the probability and importance of discovered rules from all top customers' transactions data. For instance, the first line in Figure 4-46 indicates a 92% probability of buying product 235641 if that customer is purchasing product 235642; similarly, the second line indicates that the customer who buys product 235641 will simultaneously purchase product 235642 with 82% probability.

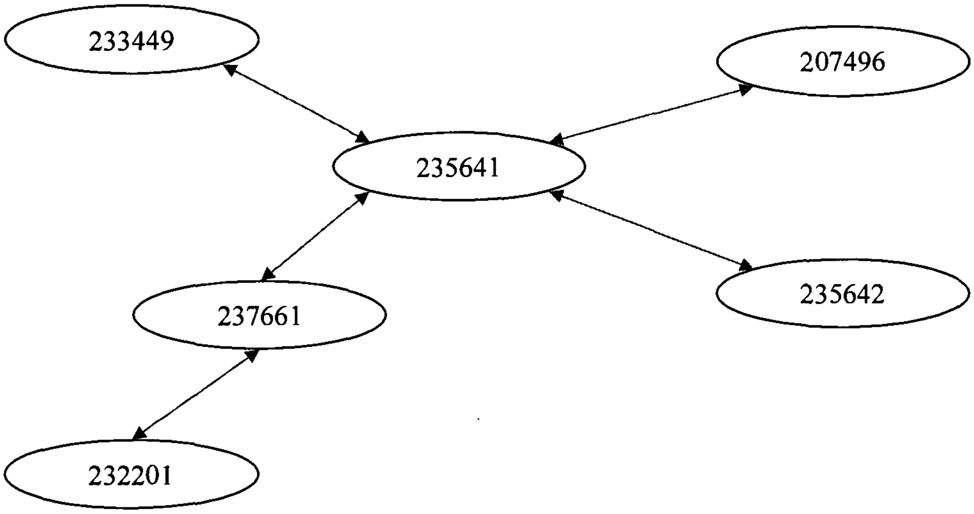


Figure 4-48: Examples of dependency networks

Figure 4-48 shows the discovered relationships of products graphically. The user can interpret association rules from this figure easily and directly.

4.5.2 Targeted marketing campaigns

The association tasks can be conducted multiple times to discover the relationships between products for different target datasets. Since the association jobs are computationally expensive, they can be done during non-peak hours (i.e. every Sunday morning) with the results stored as a file or a table in a database. Store owners or managers can access this resource to find interesting products related rules and apply them to business.

The association technique can also be used for product recommendation. After association rules are discovered, the customers who did not follow these rules can be the target of suggested selling. Some of the customers might just not realize existence of related products. The recommendation process can bring related products to them and create business opportunities for the company. In addition, the association results can be applied to the organization of items and shelves. Related products can be placed together for cross-selling.

Chapter 5

Application of the proposed process to smaller specialty retailer

This chapter provides further validation of the enterprise-wide data mining process I introduced in chapter 3, and illustrated in Chapter 4. The same process will be applied to a different type of business, and its usability in other fields is confirmed.

5.1 Business understanding and data study

5.1.1 Business understanding

As described in section 3.6, the target of the second part of this study is a smaller specialty grocery store. As part of the enterprise-wide data mining process for this retailer, this thesis will focus on inventory management and cross-selling, which were identified as two important objectives by the owner. The analysis is divided into two sections: customer based analysis and product based analysis. Each section describes the profiling of customers and products, as well as prediction of revenues associated with customers and products. The market basket analysis is addressed at the end of products based analysis section.

5.1.2 Data study

The data received from the company is an Access file; therefore the data transportation and target database creation steps are completed simultaneously. Through careful data exploration and discussions with the grocery store, some tables that were unrelated to the project goals were removed such as employee table and blank tables. The data quality and quantity of the remaining fields were confirmed to be sufficient to complete the study. Figure 5-1 shows the overview of the original data.

Attribute Name		*Year 2004*	*Year 2005*	*Year 2006*
Customer	Number	7,377	9,825	9,375
	Minimum Annual revenue	0.20	1.30	0.20
	Maximum Annual revenue	6,572	8,138	8,911
	Average of Customer Visiting Times	13.86	14.39	12.86
	Average Annual revenue	25.14	26.69	27.00
Product	Number	5,716	7,336	7,473
	Minimum Annual revenue	1.69	1.29	1.00
	Maximum Annual revenue	28,416	36,653	24,399
	Average Annual revenue	450	514	436

Figure 5-1: 2ó˘ßóĚ˘Ö [illegible]

5.2 Data preparation

Since the original data is contained in a Microsoft (MS) Access file, the data extraction step is to copy the original file as a base work database. Then data attribute reduction is executed to remove extraneous data. Based on the data study results, all unnecessary tables are removed. Therefore the base work database can be deployed for use in each sub-project included in this study and each data analysis model I am going to develop. All of subject related data attributes are collected from several tables into one table to have a clear overview of all available data. Theoretically, this is not an efficient way to store data because of data repetition. However, due to the small data size, availability of cheap data store space, and high computation ability, a united table has been created to summarize daily data into monthly data. Based on the nature of the grocery store's business, the shoppers who did not complete a purchase are recorded in the study with a sales value of 0. Outliers and negative values do not need to be removed, because their presence may originate from the perfectly normal occurrence of refunds, or employee input errors that are corrected within each month. The small data size and naturally clean data is another reason that I am able to skip additional data cleaning. The selected data attributes are shown in Figure 5-2.

NO	*Filed Name*		
1	Transaction Key	2	Transaction Time
3	Customer Primary Key	4	Product Primary Key
5	Product Sold Quantities	6	Invoice Amount
7	Invoice Cost	8	Product Quantity in Stock
9	Business Unit Number	-	

Figure 5-2: Selected attributes list

5.3 Clustering and profiling

In this study, the same clustering approach is applied to discover the characteristics of the data. K-means algorithm is combined with a statistical measure called standard score (Z score) in order to extract top customers, products, or target datasets. The annual revenue and annual profits are the attributes that were selected for analysis. The entire process is completed by using SPSS and MS Access which are both widely available.

5.3.1 Top customers discovery

The following methods are applied to cluster customers. The purpose of this clustering is to discover top customers for each year.

Step 1:

- Segmentation based on standard score of annual revenue amount for all of customers in 2005 transactions.
- Segmentation based on standard score of annual profits for all of customers in 2005 transactions.
- Select customers that have high standard scores on both annual revenue and annual profits.

Step 2:

- Clustering based on annual revenue for all of customers in 2005 transaction data.
- Clustering based on annual profits for all of customers in 2005 transaction data.
- Select the customers that are in high revenue group and high annual profits group.

Step 3: Take overlap of previous two steps' results as top customers.

Step4: Repeat the step 1, step 2 and step 3 for 2006 transaction data.

Step5: Repeat the step 1, step 2 and step 3 for 2007 transaction data.

Figure 5-3 shows the clustering results for top customers.

	Top Customers			*Entire Customers*		
Year	*Customers NO*	*Annual Revenue*	*Annual Profits*	*Customers NO*	*Annual Revenue*	*Annual Profits*
2005	97	277,890	151,106	7,377	1,847,248	940,980
%	1.3%	15.0%	16.1%	100%	100%	100%
2006	133	448,779	234,995	9,825	2,742,646	1,374,026
%	1.4%	16.4%	17.1%	100%	100%	100%
2007	104	324,043	169,898	9,375	2,417,396	1,192,608
%	1.1%	13.4%	14.2%	100%	100%	100%

Figure 5-3: Top customers' distribution

Figure 5-3 confirmed that the clustering method of using the combination of K-means and standard score has ability to extract top customers. In 2007, only 1.1% of customers contributed 13.4% of all revenues and about 14.2% of all profits. This result can be used to further customer oriented campaign i.e. DM promotions. Depending on the needs of the business, the same approach can also be applied to extract multiple years' top customers, i.e. previous three years' top customers.

5.3.2 Top products discovery

The approach used to extract top customers is also applied to extract the top products. The same steps were executed based on products' annual revenue amount and annual profits.

Figure 5-4 shows the clustering results for top products.

	Top Products			*Entire Products*		
Year	*Products NO*	*Annual Revenue*	*Annual Profits*	*Products NO*	*Annual Revenue*	*Annual Profits*
2005	41	503,829	285,908	5,716	2,572,025	1,317,487
%	0.72%	19.59%	21.70%	100%	100%	100%
2006	48	803,531	443,964	7,336	3,773,445	1,893,563
%	0.65%	21.29%	23.45%	100%	100%	100%
2007	49	689,641	406,014	7,473	3,254,878	1,618,120
%	0.66%	21.19%	25.09%	100%	100%	100%

Figure 5-4: Top products' distribution

The results in Figure 5-4 shows the store only needs to manage less than 1% of products to achieve the products management that contributes to about 20% of store incomes. The analysis of top products can assist new products development and inventory management.

5.3.3 Target customers and target products discovery

The clustering technique also has the ability to extract other interesting datasets. In the market basket analysis, this thesis only focuses on customer datasets containing customers who visit the store regularly and products datasets containing products which are purchased most. Based on preliminary data exploration results and customer behaviors, the target customers for cross-selling are separated using annual revenue amount and annual profits; the target products for cross-selling are separated using purchase times. The preliminary data exploration has been done for the discovery of mainstream customers based on the annual revenues and annual profit. Figure 5-5 to Figure5-10 show the exploration results.

Annual Revenue	*Number of Customers*	*Cumulative*
100	3542	48.01%
200	1596	69.65%
300	688	78.98%
500	627	87.47%
1000	546	94.88%
2000	280	98.67%
3000	72	99.65%
4000	15	99.85%
5000	3	99.89%
6000	6	99.97%
7000	2	100%
Total	7377	-

Figure 5-5: Frequency distribution of 2005 customers' annual revenues

The group of figures (Figures 5-5 to 5-10) shows the exploration results of mainstream customers based on the annual revenue and annual profit. Figure 5-5 shows the frequency distribution based on 2005 customer revenues.

Annual Profit	*Number of Customers*	*Cumulative %*
100	5152	69.84%
200	1072	84.37%
300	423	90.10%
500	358	94.96%
1000	247	98.31%
2000	109	99.78%
3000	13	99.96%
4000	2	99.99%
5000	1	100%
Total	7377	-

Figure 5-6: Frequency distribution of 2005 customers' annual profits

The group of figures (Figures 5-5 to 5-10) shows the exploration results of mainstream customers based on the annual revenue and annual profit. Figure 5-6 shows the frequency distribution based on 2005 customer profits.

Annual Revenue	*Number of Customers*	*Cumulative %*
100	4491	45.71%
200	2020	66.27%
300	988	76.33%
500	940	85.89%
1000	832	94.36%
2000	404	98.47%
3000	87	99.36%
4000	34	99.70%
5000	13	99.84%
6000	7	99.91%
7000	5	99.96%
More	4	100%
Total	9825	-

Figure 5-7: Frequency distribution of 2006 customers' annual revenue

The group of figures (Figures 5-5 to 5-10) shows the exploration results of mainstream customers based on the annual revenue and annual profit. Figure 5-7 shows the frequency distribution based on 2006 customer revenues.

Annual Profit	*Number of Customers*	*Cumulative*
100	6616	67.34%
200	1455	82.15%
300	678	89.05%
500	539	94.53%
1000	334	97.93%
2000	172	99.68%
3000	18	99.87%
4000	11	99.98%
5000	2	100%
Total	9825	-

Figure 5-8: Frequency distribution of 2006 customers' annual profits

The group of figures (Figures 5-5 to 5-10) shows the exploration results of mainstream customers based on the annual revenue and annual profit. Figure 5-8 shows the frequency distribution based on 2006 customer profits.

Annual Revenue	*Number of Customers*	*Cumulative*
100	4413	47.07%
200	1895	67.29%
300	934	77.25%
500	908	86.93%
1000	747	94.90%
2000	374	98.89%
3000	62	99.55%
4000	24	99.81%
5000	8	99.89%
6000	5	99.95%
7000	3	99.98%
More	2	100%
Total	9375	-

Figure 5-9: Frequency distribution of 2007 customers' revenue (up to Sep 30)

The group of figures (Figures 5-5 to 5-10) shows the exploration results of mainstream customers based on the annual revenue and annual profit. Figure 5-9 shows the frequency distribution based on 2007 customer revenues (up to September 30).

Annual Profit	*Number of Customers*	*Cumulative*
100	6414	68.42%
200	1508	84.50%
300	541	90.27%
500	459	95.17%
1000	288	98.24%
2000	148	99.82%
3000	13	99.96%
4000	2	99.98%
5000	1	99.99%
6000	1	100%
Total	9375	-

Figure 5-10: Frequency distribution of 2007 customers' profits (up to Sep 30)

The group of figures (Figures 5-5 to 5-10) shows the exploration results of mainstream customers based on the annual revenue and annual profit. Figure 5-9 shows the frequency distribution based on 2007 customer profits (up to September 30).

Purchase Times	*Number of Products*	*Cumulative*
1	871	15.24%
2	610	25.91%
3	463	34.01%
5	666	45.66%
6	265	50.30%
7	221	54.16%
8	224	58.08%
9	146	60.64%
10	158	63.40%
20	855	78.36%
30	415	85.62%
40	214	89.36%
50	135	91.72%
100	305	97.06%
200	132	99.37%
300	24	99.79%
400	5	99.88%
500	1	99.90%
1000	5	99.98%
More	1	100%
Total	5716	-

Figure 5-11: Frequency distribution of 2005 products purchase times

Frequently bought products are also extracted by applying this approach. The results are shown in Figure 5-11 to Figure 5-13. Figure 5-11 shows the frequency distribution based on products purchased times of 2005.

Purchase Times	*Number of Products*	*Cumulative*
1	1216	16.58%
2	813	27.66%
3	587	35.66%
5	853	47.29%
6	300	51.38%
7	272	55.08%
8	240	58.36%
9	217	61.31%
10	162	63.52%
20	1083	78.29%
30	464	84.61%
40	285	88.50%
50	180	90.95%
100	403	96.44%
200	200	99.17%
300	37	99.67%
400	12	99.84%
500	3	99.88%
1000	8	99.99%
More	1	100%
Total	7336	-

Figure 5-12: Frequency distribution of 2006 products purchase times

The group of figures (Figures 5-11 to 5-13) shows the exploration results of mainstream products based on purchase times. Figure 5-12 shows the frequency distribution based on products purchased times of 2006.

Purchase Times	*Number of Products*	*Cumulative*
1	1400	18.73%
2	849	30.10%
3	622	38.42%
5	861	49.94%
6	305	54.02%
7	290	57.90%
8	231	60.99%
9	187	63.50%
10	194	66.09%
20	1131	81.23%
30	449	87.23%
40	247	90.54%
50	162	92.71%
100	352	97.42%
200	158	99.53%
300	19	99.79%
400	8	99.89%
500	1	99.91%
1000	7	100%
Total	7473	-

Figure 5-13: Frequency distribution of 2007 products purchase times (up to Sep 30)

The group of figures (Figures 5-11 to 5-13) shows the exploration results of mainstream products based on purchase times. Figure 5-13 shows the frequency distribution based on products purchased times of 2007 (up to September 30).

Based on the results shown from Figure 5-5 to Figure 5-13, the customers whose annual profits are more than 100 and less than 1000 will be in the target customer group. The products which are purchased over 50 times will be in target product group.

Figure 5-14 shows the overview of target customers.

Year	*Target Customers*			*Entire Customers*		
	Customers NO	*Annual Revenue*	*Annual Profits*	*Customers NO*	*Annual Revenue*	*Annual Profits*
2005	2100	1,120,498	559,341	7,377	1,847,248	940,980
%	28.50%	60.70%	59.40%	100%	100%	100%
2006	3005	1,649,056	802,063	9,825	2,742,646	1,374,026
%	30.60%	60.10%	58.40%	100%	100%	100%
2007	2796	1,492,728	710,012	9,375	2,417,396	1,192,608
%	29.80%	61.70%	59.50%	100%	100%	100%

Figure 5-14: Target customers' distribution

Figure 5-15 shows the overview of target products.

Year	*Target Products*			*Entire Products*		
	Products NO	*Products Revenue*	*Annual Profits*	*Products NO*	*Annual Revenue*	*Annual Profits*
2005	473	1,343,446	670,037	5,716	2,572,025	1,317,487
%	8.28%	52.23%	50.86%	100%	100%	100%
2006	664	2,175,606	1,083,883	7,336	3,773,445	1,893,563
%	9.05%	57.66%	57.24%	100%	100%	100%
2007	545	1,683,328	854,781	7,473	3,254,878	1,618,120
%	7.29%	51.72%	52.83%	100%	100%	100%

Figure 5-15: Target products' distribution

5.4 Prediction

The purpose of this study is to predict customers' monthly revenue and quantities of products that will be sold to assist the marketing strategy development. The prediction section starts from target data selection. Multiple regression technique will be used to discover the best model for this study. After discovering the best model, several prediction techniques such as SVR and Neural Networks, will be applied to the best model. The results comparison is also provided.

5.4.1 Target customers selection

Prediction techniques extract useful patterns from data. The target customers in this activity are customers who visited stores very often, so the regression formula is able to obtain enough training data. Based on previous clustering results, 5 customers who visit stores most often were chosen as our preliminary target customers.

5.4.2 Customers monthly revenue prediction

The two prediction models are examined in this project.

Model 1: prediction using three monthly sales data

Based on the previous three monthly revenues and revenue from the same month of the previous year, the expected revenue for the next month can be predicted. For example, the revenue for 2006 January can be predicted based on revenues of previous three months, i.e. October to December 2005, and revenue from January 2005.

Model 2: prediction using half years sales data

Based on previous half years revenue and revenue from the same month of previous year, the expected revenue for the next month can be predicted.

I randomly pick two customers from the five selected customers and process the comparison of the two models. The comparison results are shown in Figure 5-16 and 5-17.

Prediction Error Rates	*Model1*	*Model2*
0.1	33.33%	16.67%
0.2	42.86%	38.89%
0.3	57.14%	50.00%
0.4	66.67%	77.78%
0.5	80.95%	77.78%
0.6	85.71%	94.44%
0.7	90.48%	94.44%
0.8	90.48%	94.44%
0.9	90.48%	94.44%
1	90.48%	94.44%
More	100%	100%

Figure 5-16: The model comparison results of customer A

Figure 5-16 shows the comparison results of the two models explained previously based on the data of customer A.

Prediction Error Rates	*Model1*	*Model2*
0.1	14.29%	0.00%
0.2	19.05%	16.67%
0.3	23.81%	22.22%
0.4	33.33%	50.00%
0.5	61.90%	50.00%
0.6	66.67%	55.56%
0.7	71.43%	66.67%
0.8	71.43%	66.67%
0.9	76.19%	66.67%
1	76.19%	66.67%
More	100%	100%

Figure 5-17: The model comparison results of customer B

Figure 5-17 shows the comparison results of the two models explained previously based on the data of customer B.

Based on the comparison results, model 1 is better than model 2. For customer A, about 57% of customers' prediction error is below 30% in model 1; only about 50% of customers' prediction error is below 30% in model 2. Therefore, model 1 is chosen and the rest of evaluation results of model 1 are shown in Figure 5-18.

Prediction Error Rates	*Customer C*	*Customer D*	*Customer E*
0.1	14.29%	14.29%	47.62%
0.2	19.05%	28.57%	61.90%
0.3	42.86%	33.33%	76.19%
0.4	61.90%	57.14%	90.48%
0.5	80.95%	66.67%	90.48%
0.6	85.71%	76.19%	100%
0.7	85.71%	80.95%	100%
0.8	85.71%	85.71%	100%
0.9	85.71%	85.71%	100%
1	85.71%	85.71%	100%
More	100%	100%	100%

Figure 5-18: Prediction results evaluation

The accuracy of predictions in above charts varies among customers. For example, customer E has reasonably high prediction accuracy with 76% of prediction errors under 30%. On the other hand, errors for only 40-60% of predictions for customer A and C were under 30%. The prediction accuracy for customer B and D was worst with 33% predictions had accuracy of 30% or less.

5.4.3 Target products selection

The target products in this activity are products that were purchased frequently. The regression formula is able to obtain enough training data to achieve these aims. Based on the previous clustering results, five products purchased at a high frequency were chosen to be our preliminary target products.

5.4.4 Products quantity prediction

The purpose of prediction in this study is to predict quantities of products that will be sold to help inventory management.

5.4.4.1 Existing practice

The company is currently using monthly average value of year-to-date sales data as predicted value. The average data is used to calculate products order quantity. The average can be calculated as:

$$\text{Average} = \frac{\text{The Total of Year - to - Date Sales Amount}}{12}$$

EQUATION 5-1

The order quantity is the difference between average and on hand product quantity. Therefore, the goal of this project becomes to detect a better algorithm.

5.4.4.2 Multiple regression

As described in chapter 4, I first tested the data model with multiple regression to determine which of the following two models is better. Then I applied the data model to several prediction techniques to discover the best performance technique.

Thereby, I build two models at this point.

Model 1: previous three months data based model

Based on previous three months data of product sold quantity and the quantity data of product sold from the same month of previous year, the expected quantity of products sold for the next month is predicted. For example, the quantity of product sold for 2006 January can be predicted based on quantity of the previous three months, i.e. October to December 2005, and quantity from January 2005.

Model 2: previous three quarters data based model

The only difference between model 1 and model 2 is to use previous three quarterly data instead of using three months data of product sold quantity. The other part will follow model 1's algorithm.

The prediction results are shown in Figure5-19 to Figure 5-23. Model 1 can predict 21 data points from January 2006 to September 2007 based on January 2005 to September 2006. Model 2 is able to cover 24 data points from January 2006 to December 2007 based on January 2005 to December 2006. Model 2 can predict October, November and December of 2007 since the latest quarterly data is known (July to September of 2007). However, Model 1 needs November 2007 data to predict December 2007 and November 2007 data that is not available.

Error Rate	*Store Method*		*Model 1*		*Model 2*	
	Frequency	*Cumulative*	*Frequency*	*Cumulative*	*Frequency*	*Cumulative*
0.1	11	45.83%	12	57.14%	10	41.67%
0.2	8	79.17%	6	85.71%	12	91.67%
0.3	4	95.83%	3	100%	2	100%
0.4	1	100%	0	100%	0	100%

Figure 5-19: Model comparison results of product A

The group of figures (Figures 5-19 to 5-23) shows model comparison results based on products data. Figure 5-19 shows model comparison results of product A.

Error Rate	*Store Method*		*Model 1*		*Model 2*	
	Frequency	*Cumulative*	*Frequency*	*Cumulative*	*Frequency*	*Cumulative*
0.1	5	20.83%	7	33.33%	6	25.00%
0.2	4	37.50%	5	57.14%	8	58.33%
0.3	2	45.83%	5	80.95%	4	75.00%
0.4	3	58.33%	3	95.24%	1	79.17%
0.5	3	70.83%	1	100%	3	91.67%
0.6	3	83.33%	0	100%	2	100%
0.7	0	83.33%	0	100%	0	100%
0.8	1	87.50%	0	100%	0	100%
0.9	1	91.67%	0	100%	0	100%
1	1	95.83%	0	100%	0	100%
More	1	100%	0	100%	0	100%

Figure 5-20: Model comparison results of product B

The group of figures (Figures 5-19 to 5-23) shows model comparison results based on products data. Figure 5-20 shows model comparison results of product B.

Error Rate	*Store Method*		*Model 1*		*Model 2*	
	Frequency	*Cumulative*	*Frequency*	*Cumulative*	*Frequency*	*Cumulative*
0.1	3	12.50%	2	9.52%	7	29.17%
0.2	2	20.83%	4	28.57%	7	58.33%
0.3	1	25.00%	6	57.14%	6	83.33%
0.4	0	25.00%	4	76.19%	0	83.33%
0.5	2	33.33%	0	76.19%	1	87.50%
0.6	3	45.83%	3	90.48%	0	87.50%
0.7	2	54.17%	1	95.24%	1	91.67%
0.8	4	70.83%	1	100%	1	95.83%
0.9	1	75.00%	0	100%	0	95.83%
1	1	79.17%	0	100%	0	95.83%
More	5	100%	0	100%	1	100%

Figure 5-21: Model comparison results of product C

The group of figures (Figures 5-19 to 5-23) shows model comparison results based on products data. Figure 5-21 shows model comparison results of product C.

Error Rate	*Store Method*		*Model 1*		*Model 2*	
	Frequency	*Cumulative*	*Frequency*	*Cumulative*	*Frequency*	*Cumulative*
0.1	12	50.00%	14	66.67%	13	54.17%
0.2	7	79.17%	5	90.48%	7	83.33%
0.3	4	95.83%	2	100%	4	100%
0.4	1	100%	0	100%	0	100%

Figure 5-22: Model comparison results of product D

The group of figures (Figures 5-19 to 5-23) shows model comparison results based on products data. Figure 5-22 shows model comparison results of product D.

Error Rate	*Store Method*		*Model 1*		*Model 2*	
	Frequency	*Cumulative*	*Frequency*	*Cumulative*	*Frequency*	*Cumulative*
0.1	1	4.17%	5	23.81%	4	16.67%
0.2	1	8.33%	6	52.38%	3	29.17%
0.3	2	16.67%	4	71.43%	3	41.67%
0.4	5	37.50%	2	80.95%	1	45.83%
0.5	1	41.67%	1	85.71%	2	54.17%
0.6	4	58.33%	1	90.48%	0	54.17%
0.7	0	58.33%	2	100%	3	66.67%
0.8	0	58.33%	0	100%	0	66.67%
0.9	0	58.33%	0	100%	1	70.83%
1	1	62.50%	0	100%	1	75.00%
More	9	100%	0	100%	6	100%

Figure 5-23: Model comparison results of product E

The group of figures (Figures 5-19 to 5-23) shows model comparison results based on products data. Figure 5-23 shows model comparison results of product E.

Consequently, from Figure 5-19 to Figure 5-23, the three months model and three quarters model lead to better results than the method that is used by the store. Both regression models have similar results. The accuracy of the results is dependent on the product itself. The accuracy usually varies from one product to another. However, the three quarters model covers the 9 months sales and three months model covers only the previous three months. These techniques demonstrate that the three quarters model is able to build more stable model that is less affected by temporary upward or downward swings in sales. Therefore, the three quarters model is chosen to be the default data model.

5.4.4.3 SVR

SVR uncovered the same patterns revealed by multiple regression. One important factor in data analysis is the growth rate, which is calculated by dividing one year's sales quantity by the previous year's quantity. To obtain values in the same scale, the data of 2006 and 2007 is rescaled to 2005 scale by using the growth rate. For example, for product A, the quantity of product sold for 2006 is 1000 and the quantity of this product sold for 2005 is 500, therefore the growth rate is 1000/500, which is 2. Thus all quantity data for products sold in 2006 will be divided by 2 to reduce it to the 2005's scale. We call it the GR factor.

5.4.4.4 Autoregression

Autoregression is another algorithm we applied to predict products sold quantities. Based on previous two years sale histories of products, autoregression is able to build a model and predict future data.

5.4.4.4 Results comparison

Figure 5-24 to Figure 5-28 indicate how well that each algorithm is able to describe data based on model 2. In these figures, *Fre* and *Cum* represent *Frequency* and *Cumulative*, respectively.

	Store		*Regression*		*Regression(GR)*		*SVR*		*SVR(GR)*		*Auto Regression*	
Error	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*
0.1	11	45.83%	10	41.67%	13	54.17%	11	45.83%	14	58.33%	9	37.50%
0.2	8	79.17%	12	91.67%	10	95.83%	9	83.33%	6	83.33%	10	79.17%
0.3	4	95.83%	2	100%	1	100%	4	100%	4	100%	4	95.83%
0.4	1	100%	0	100%	0	100%	0	100%	0	100%	1	100%
0.5	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%
0.6	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%
0.7	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%
0.8	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%
0.9	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%
1	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%
More	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%

Figure 5-24: The prediction methods evaluation for product A

The group of figures (Figures 5-24 to 5-28) shows the prediction methods evaluation results of different algorithms based on model 1.

Figure 5-24 shows the comparison results of product A.

	Store		*Regression*		*Regression(GR)*		*SVR*		*SVR(GR)*		*Auto Regression*	
Error	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*
0.1	5	20.83%	8	33.33%	6	25.00%	6	25.00%	8	33.33%	4	16.67%
0.2	4	37.50%	5	54.17%	6	50.00%	11	70.83%	8	66.67%	9	54.17%
0.3	2	45.83%	4	70.83%	6	75.00%	5	91.67%	6	91.67%	5	75.00%
0.4	3	58.33%	2	79.17%	3	87.50%	1	95.83%	0	91.67%	2	83.33%
0.5	3	70.83%	3	91.67%	0	87.50%	1	100%	2	100%	3	95.83%
0.6	3	83.33%	2	100%	1	91.67%	0	100%	0	100%	1	100%
0.7	0	83.33%	0	100%	1	95.83%	0	100%	0	100%	0	100%
0.8	1	87.50%	0	100%	1	100%	0	100%	0	100%	0	100%
0.9	1	91.67%	0	100%	0	100%	0	100%	0	100%	0	100%
1	1	95.83%	0	100%	0	100%	0	100%	0	100%	0	100%
More	1	100%	0	100%	0	100%	0	100%	0	100%	0	100%

Figure 5-25: The prediction methods evaluation for product B

The group of figures (Figures 5-24 to 5-28) shows the prediction methods evaluation results of different algorithms based on model 1.

Figure 5-25 shows the comparison results of product B.

	Store		*Regression*		*Regression(GR)*		*SVR*		*SVR(GR)*		*Auto Regression*	
Error	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*
0.1	3	12.50%	10	41.67%	9	37.50%	5	20.83%	5	20.83%	2	8.33%
0.2	2	20.83%	5	62.50%	5	58.33%	8	54.17%	5	41.67%	0	8.33%
0.3	1	25.00%	3	75.00%	4	75.00%	3	66.67%	5	62.50%	1	12.50%
0.4	0	25.00%	2	83.33%	3	87.50%	2	75.00%	3	75.00%	1	16.67%
0.5	2	33.33%	2	91.67%	1	91.67%	0	75.00%	2	83.33%	3	29.17%
0.6	3	45.83%	0	91.67%	0	91.67%	1	79.17%	0	83.33%	2	37.50%
0.7	2	54.17%	0	91.67%	1	95.83%	2	87.50%	3	95.83%	3	50.00%
0.8	4	70.83%	1	95.83%	0	95.83%	1	91.67%	0	95.83%	1	54.17%
0.9	1	75.00%	0	95.83%	0	95.83%	1	95.83%	0	95.83%	2	62.50%
1	1	79.17%	0	95.83%	0	95.83%	0	95.83%	1	100%	1	66.67%
More	5	100%	1	100%	1	100%	1	100%	0	100%	8	100%

Figure 5-26: The prediction methods evaluation for product C

The group of figures (Figures 5-24 to 5-28) shows the prediction methods evaluation results of different algorithms based on model 1.

Figure 5-26 shows the comparison results of product C.

	Store		*Regression*		*Regression(GR)*		*SVR*		*SVR(GR)*		*Auto Regression*	
Error	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*
0.1	12	50.00%	12	50.00%	15	62.50%	11	45.83%	13	54.17%	16	66.67%
0.2	8	83.33%	7	79.17%	7	91.67%	9	83.33%	6	79.17%	5	87.50%
0.3	4	100%	4	95.83%	2	100%	4	100%	3	91.67%	3	100%
0.4	0	100%	1	100%	0	100%	0	100%	1	95.83%	0	100%
0.5	0	100%	0	100%	0	100%	0	100%	1	100%	0	100%
0.6	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%
0.7	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%
0.8	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%
0.9	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%
1	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%
More	0	100%	0	100%	0	100%	0	100%	0	100%	0	100%

Figure 5-27: The prediction methods evaluation for product D

The group of figures (Figures 5-24 to 5-28) shows the prediction methods evaluation results of different algorithms based on model 1.

Figure 5-27 shows the comparison results of product D.

Error	Store		Regression		Regression(GR)		SVR		SVR(GR)		Auto Regression	
	Fre	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*	*Fre*	*Cum*
0.1	1	4.17%	4	16.67%	8	33.33%	1	4.17%	2	8.33%	6	25.00%
0.2	1	8.33%	6	41.67%	5	54.17%	2	12.50%	3	20.83%	4	41.67%
0.3	2	16.67%	3	54.17%	3	66.67%	5	33.33%	4	37.50%	3	54.17%
0.4	5	37.50%	1	58.33%	4	83.33%	1	37.50%	1	41.67%	3	66.67%
0.5	1	41.67%	2	66.67%	0	83.33%	4	54.17%	4	58.33%	2	75.00%
0.6	4	58.33%	3	79.17%	0	83.33%	2	62.50%	3	70.83%	1	79.17%
0.7	0	58.33%	0	79.17%	2	91.67%	1	66.67%	1	75.00%	0	79.17%
0.8	0	58.33%	0	79.17%	0	91.67%	1	70.83%	1	79.17%	2	87.50%
0.9	0	58.33%	1	83.33%	0	91.67%	2	79.17%	2	87.50%	0	87.50%
1	1	62.50%	0	83.33%	0	91.67%	2	87.50%	3	100%	1	91.67%
More	9	100%	4	100%	2	100%	3	100%	0	100%	2	100%

Figure 5-28: The prediction methods evaluation for product E

The group of figures (Figures 5-24 to 5-28) shows the prediction methods evaluation results of different algorithms based on model 1.

Figure 5-28 shows the comparison results of product E.

Based on the above figures, results from autoregression is worse than results of other algorithms. Therefore we remove autoregression from our candidates. SVR and Multiple Regression have similar performances with that SVR usually delivers better results. Both deliver significant improvements over the existing algorithm. The models that applied the GR factor clearly perform better. Furthermore, the SVR model based on data adjusted with the GR factor produces the best results. The accuracy of predictions in the previous charts, however, varies from product to product. For example, products A and D have reasonably high prediction accuracy i.e. 80-95% with an error rate under 20%. On the other hand, errors for only 50-70% of the predictions for product B and C were under 20%. The prediction accuracy for E was worst i.e. 30-66% predictions with an accuracy of 30%. Considering the computation cost and data preparation cost, the regression using the GR factor is the superior method of representing the data.

Figure 5-29 to Figure 5-33 show the prediction results applied by each model. Since the data I analyzed ended on September 30 of 2007, I predicted the purchase times of products in each of the following three months, i.e. October, November, and December. In these figures, PV represents the *Predicted Value*; ER represents *Error Rates*, which indicates the difference between actual values and predicted values.

Year	*Month*	*Actual*	*Store*		*Regression*		*Regression(GR)*		*SVR*		*SVR(GR)*	
		Value	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*
2007	10	55	57.58	4.70%	60.82	10.58%	52.59	4.38%	58.17	5.77%	52.59	11.55%
2007	11	47	57.63	22.62%	60.07	27.81%	51.79	10.19%	57.72	22.81%	51.79	2.70%
2007	12	48	58.27	21.39%	62.10	29.37%	53.97	12.45%	58.92	22.75%	53.97	2.65%

Figure 5-29: The prediction results comparison for product A

The group of figures (Figures 5-29 to 5-33) shows the products quantity prediction results of different algorithms based on model 1. Figure 5-29 shows the prediction results comparison for product A.

Year	*Month*	*Actual*	*Store*		*Regression*		*Regression(GR)*		*SVR*		*SVR(GR)*	
		Value	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*
2007	10	60	117.58	95.97%	94.59	57.65%	102.27	70.45%	86.63	44.38%	89.66	49.44%
2007	11	68	118.97	74.95%	79.67	17.16%	86.60	27.35%	54.70	19.56%	56.62	16.74%
2007	12	158	121.88	22.86%	115.65	26.80%	124.40	21.27%	130.79	17.22%	135.38	14.32%

Figure 5-30: The prediction results comparison for product B

The group of figures (Figures 5-29 to 5-33) shows the products quantity prediction results of different algorithms based on model 1. Figure 5-30 shows the prediction results comparison for product B.

Year	*Month*	*Actual*	*Store*		*Regression*		*Regression(GR)*		*SVR*		*SVR(GR)*	
		Value	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*
2007	10	82	162.08	97.66%	91.40	11.47%	111.12	35.51%	134.78	64.37%	134.78	64.37%
2007	11	90	167.76	86.40%	57.91	35.65%	80.15	10.94%	109.48	21.64%	109.48	21.64%
2007	12	307	176.65	42.46%	326.87	6.47%	328.81	7.10%	331.34	7.93%	331.34	7.93%

Figure 5-31: The prediction results comparison for product C

The group of figures (Figures 5-29 to 5-33) shows the products quantity prediction results of different algorithms based on model 1. Figure 5-31 shows the prediction n results comparison for product C.

Year	*Month*	*Actual*	*Store*		*Regression*		*Regression(GR)*		*SVR*		*SVR(GR)*	
		Value	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*
2007	10	81	63.58	21.50%	57.38	29.16%	77.56	4.25%	83.07	2.56%	92.13	13.74%
2007	11	82	63.88	22.10%	54.83	33.13%	74.24	9.46%	82.52	0.64%	91.53	11.62%
2007	12	66	64.87	1.71%	58.98	10.64%	79.63	20.65%	83.42	26.40%	92.52	40.19%

Figure 5-32: The prediction results comparison for product D

The group of figures (Figures 5-29 to 5-33) shows the products quantity prediction results of different algorithms based on model 1. Figure 5-32 shows the prediction results comparison for product D.

Year	*Month*	*Actual*	*Store*		*Regression*		*Regression(GR)*		*SVR*		*SVR(GR)*	
		Value	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*	*PV*	*ER*
2007	10	13	40.33	210.26%	90.32	594.78%	34.05	161.94%	48.81	275.45%	13.83	6.39%
2007	11	15	36.94	146.30%	82.68	451.21%	30.86	105.71%	41.91	179.43%	11.88	20.82%
2007	12	24	34.69	44.54%	86.73	261.36%	32.55	35.62%	45.55	89.79%	12.91	46.22%

Figure 5-33: The prediction results comparison for product E

The group of figures (Figures 5-29 to 5-33) shows the products quantity prediction results of different algorithms based on model 1. Figure 5-33 shows the prediction results comparison for product E.

The above charts indicate that the accuracy of predictions differ from product to product. However, regression and SVR with the GR factor provide significantly better accuracies than the store's method does. For some products the regression model achieves better results than SVR; in some cases SVR is better than the other. For instance, in the regression model that applies the GR factor, product A and D have very high prediction accuracies with error rate less than 5%. For product B and E, SVR applied GR lead better results.

Although the regression model and SVR model's results are significantly improved than the store model, the accuracies are still not good enough to put into practical use. The stability of models is still a challenge.

5.5 Market basket analysis

The target customers and target products in this activity are extracted based on number of visits and amount of purchases. In order to collect and cover more data points, 1,120 customers were designated as target customers because they had visited the store over fifty times from January 2005 to September 2007. The products which were purchased over one hundred times were designated as preliminary target products, which resulted in 848 products.

5.5.1 The association data preparation

The 848 items in the preliminary target product list contained many similar products. Before conducting market basket analysis, the data aggregation process is required. For example, product "sports drink A 15ML" and product "sports drink A 30ML" was considered to be the same product: “sports drink”. After merging all similar products, the list was reduced to 510 merged products. In this thesis, the merged products will be called meta-products. The aggregation process is based on product description and department information. It was only executed between same department’s products. The 145 meta-products which were purchased over 300 times are extracted as our target products.

5.5.2 The association data mining

Based on the selected association target data, the association technique is applied to the transaction data (63,035 records).

Name	Parameters
Transaction Records Number	63,035
Customer Number	1120
Meta-Product Number	145
Mini Support	5%
Mini Probability	40%
Mini Importance	0.10

Figure 5-34: Overview of the dataset and association parameters

Figure 5-34 shows the details of the target data and association parameters.

Products	Support
Nu100, Si100	14.5%
Or100,Si100	12.1%
Ge100, Nu100	11.8%
Ge100, Si100	11.2%
Or100,Nu100	10.3%
Ud100, Si100	10.3%
Ud100,Nu100	9.9%
Ja100,Si100	9.5%
Or100,Ge100	9.2%

Figure 5-35: Examples of linked products

Figure 5-35 shows the examples of discovered linked products. Due to the sensitive nature of business, the specific names of products have been disguised. Figure 5-35 shows discovered relationships between products, i.e. 14.5% of customers purchase Nu100 and Si100 together. These discovered links can be applied to cross-selling or products recommendations.

Figure 5-36 shows the probability and importance of discovered rules.

Probability	Importance	Rule
60 %	0.27	Ja100, Nu100 -> Si100
55 %	0.24	Ja100, Si100 -> Nu100
52 %	0.23	Ja100 -> Si100
55 %	0.23	Or100, Nu100 -> Si100
54 %	0.22	A0100 -> Si100
44 %	0.22	Bi100 -> Or100
54 %	0.22	Nat100 -> Si100

Figure 5-36: Examples of association rules

For instance, the first line in Figure 5-36 indicates a 60% probability of buying product Si100 if that customer is purchasing product Ja100 and Nu100. The importance of these relationships was discussed in section 2.5.

The discovered relationships of products can also be graphically represented for the benefit of decision makers. Figure 5-37 shows the links between products.

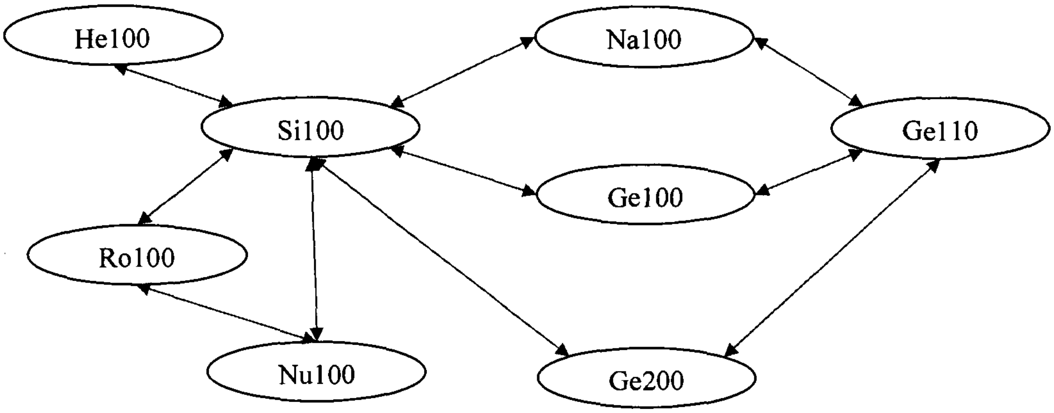


Figure 5-37: Association dependency network

5.6 Revisit the data mining process

This chapter confirms the usability of a complete data mining process using data from a smaller specialty grocery store. The process is exactly the same as that demonstrated in chapter 4, which describes the business understanding, data study, data preparation, data model construction, and data mining tasks processing with clustering, prediction and association techniques. The evaluation of data mining results is also addressed. The process, or a part of it, needs to be repeated several times in order to achieve data mining goals. As described in the previous chapters the data mining goals are always determined by business issues the company wishes to address.

#

#

#

#

#

#

Chapter 6

Concluding remarks and future works

6.1 Thesis summary

This thesis demonstrated an elaborate enterprise-wide data mining process for the retail industry. While the entire data mining process needed to be described for the sake of completeness, the focus of the thesis was on the modeling and evaluation phases. Modeling and evaluation tends to be a commercially guarded secret and it can be a barrier in helping decision makers understands the promises and limitations of the knowledge discovery process. The process was illustrated to be effective for two businesses: a retailer with significantly different business profiles. The process includes business understanding, data study, data preparation, modeling and evaluation of knowledge discovery. We clearly described what data mining can do and how data mining can be done in each stage addressing challenges and solutions.

Daily transaction data is provided by two businesses: a multinational wholesaler with some retail business and a small specialty retailer. The transaction data consists of basic data including time, customer, product, supplier, business unit (BU), order, and sales information. Customer, product, supplier, order and sales information are considered as five essential elements of data mining for retail businesses. In these five elements, sales information is the focus in the entire study which unifies the other elements.

A successful data mining process is premised upon a fundamental understanding of the business being examined. All data mining activities have to follow the unique business challenges identified in order to develop solutions specifically designed to address these challenges. The data study is conducted simultaneously, which provides a clear view of available data and determines the data mining possibilities. Up to this point, cooperation with the company is very important. The business issues and interests of the company will be clarified and goals of data mining will be determined by the company and data mining experts.

Based on the companies' business needs and available data, proper data preparation has been pursued which covers data extraction, data reduction, data cleaning and data transformation. Data cleaning might need to be repeated several times depending on the needs of the study. We spent most of time on data cleaning in this step in order to enhance the quality of target data and remove noise i.e. smoothing negative values. The summary will be addressed in section 6.2.

Data mining tasks are standardized into two steps: data model construction and data mining techniques selection. Data models are constructed following the nature of the business, the goals of the data mining activities and the character of data mining techniques. Several data models need to be created and compared, and the most suitable model will be chosen from results comparison. All results are evaluated based on error rates and summarized by histogram. For the multinational retailer, we constructed three models: Model 1: 2006 annual revenue prediction using nine months sales data, Model

2:2006 annual revenue prediction using three quarters sales data, Model 3: 2007 half annual revenue prediction using seven quarters sales data. The difference between three models is the time period that is used for collecting and summarizing data. Based on results comparison, model 2 was selected. Furthermore, the most suitable data mining technique is selected by comparing data mining results based on the selected model.

Clustering, prediction and association are data mining techniques used in our activities. Clustering technique should be the first technique used in data mining, since the clustering results can be used for other activities. K-means and Kohonen SOM are common techniques which can also be combined with statistical methods. Clustering and profiling of business units, customers and products determined important business units, loyal customers and mainstream products. For the multinational retailer, business units were separated into 8 clusters based on the monthly revenues; 2.5% of customers which contribute over 50% of the revenues are discovered; 0.16% of products which contribute 24% of the profits are specified. Clustering based on the contribution percentage instead of using absolute values was introduced to remove affects from business scales of different regions. These results can be used to optimize the companies' resources. The details appear in section 6.3.

Prediction is another powerful technique applied to data mining activities. Simple linear regression, multiple linear regression, support vector regression, Auto-regression and neural network are some general prediction techniques. A new factor called GR was introduced here to remove affects from yearly changes of businesses. Reasonable results have been achieved in sales prediction and inventory predictions (refer to section 4.4 and section 5.4). Results by my methods are better than that of the currently used method.

However, prediction results can only be a reference measure since accuracy of the results is not stable enough to be an independent indicator in real world business. Results can be more accurate if there is more data available for analysis. Massive data quantities can reduce the effect of noisy data for extracting and observing the true customer behavior and product patterns of retailers' business.

The association technique effectively discovers relationships between products. A priori algorithm is the most well-known association algorithm. The results can be easily applied to cross-selling in order to increase benefits to retailers (refer to section 4.5 and section 5.5). However, since the association technique is computationally expensive, the target of the association technique should be focused on specified datasets, i.e. for specialty retailer, I focus on customers who visit stores regularly and frequently purchased products.

6.2 Summary of data mining process

6.2.1 Data acquisition and cleaning

Data preparation is a key data mining step, which directly affects the quality of the results. The main task of this phase is data cleaning. Data cleaning must respect the natural business cycle and remove misleading noises and outliers. As described in chapter 4, there are no strict guidelines for data mining. However, some typical treatments include omitting outliers, assigning default values, filling with average data, or filling with the most probable values. In the retail industry, refunds and miss-inputs are two major patterns of noisy data. Refunds can be identified and smoothed (refer to section 4.2.3). Miss-inputs data can also be revealed using clustering techniques (refer to section 4.2.3).

Refunds are negative values of revenues that do not represent any positive retail activities. Therefore, based on the companies refund policies, a refundable time period can be discovered first. The difference between negative and positive values within a range during the refundable time period can replace the positive value, while the negative value is set to 0. Moreover, the miss-inputs data is usually different from other data points; therefore the clustering technique can isolate noisy data and replace it with reasonable values such as 0.

6.2.2 Summary of the business based on data analysis

Each retailer has unique business challenges. A better understanding of each business' needs will not only help companies re-recognize their own business, but also help data mining experts establish goals for data mining activities that yield effective solutions. The business can be summarized based on the five essential elements of data mining for retails. Clustering techniques can profile customer, product, supplier, order and sales information, i.e. discovery of top customers, top products and top suppliers, frequency of orders and amount of revenues. These five elements construct a five dimensional space. The values within this space describe the business. Depending on the companies' unique challenges, other attributes, such as business unit information, can be added to this space to better describe the business. The maximum, minimum and average values are always interesting measurements and describe the business better.

6.3 Data mining possibilities in business

6.3.1 Clustering in business

6.3.1.1 Application of the clustering results

Understanding the customers is an important component in marketing analysis for retailers. Proper usages of customers' valuable information can streamline business processes and increase profits. Any information that customers provide to retailers can help the company understand purchase behavior patterns. Managers who understand patterns of customer behaviour can increase sales, reduce costs, and target other strategic goals.

Clustering technique can help the customer studies and determine the corresponding business target groups. An example will be a direct mail campaign. The company can only mail target customer groups instead of mailing every customer, but still receive almost same marketing benefits but save a large amount of the cost.

The output of the clustering technique yields data useful for profiling customers, products, and suppliers. The top customers' information can be used to enhance the relationship between the company and customers. The top products list can be used to manage resources and inventory efficiently, that is to focus efforts on smaller numbers of products that produce more income. This information can also be used for further product development.

In addition, clustering techniques provide fundamental information about the entire data mining study, i.e. clustering results can be applied to discover necessary number of data models for different type of target groups. Clustering techniques can also be used as part of the data cleaning process. Without clustering, further data mining results will be adversely affected.

6.3.1.2 Clustering business opportunities

Clustering can be applied to make discoveries and create opportunities in many business areas, such as marketing, inventory management and fraud detection. Some details will be discussed here.

Determination of target customers, products and suppliers

Customers, products and suppliers can be divided using different aspects of data, i.e. divide customers by visiting frequency; divide products by purchased quantities; divide suppliers by products numbers.

Inventory management

Products can be divided into groups based on monthly revenues or annual sales. The most sold products group need to be considered in inventory management. Products with a large proportion of the overall value but a small percentage of the overall volume and products with a small proportion of the overall value but a large percentage of the overall volume can be separately identified.

Fraud detection

Clustering techniques can be used to detect outliers, which could be suspicious data. Clustering is a well known technique applied in the data cleaning field.

6.3.2 Prediction in business

Many business opportunities can be created by applying prediction techniques. It can be applied in many fields i.e. marketing, inventory management, investment and fraud detection. The details will be discussed in this section.

Revenue prediction at the customer level

These techniques help the company to predict the life time value of customers. Some potential profitable customers may be discovered during the process.

Revenue prediction for suppliers

Suppliers are an important part of the business process. This prediction can help the company plan purchase activities in advance.

High risk customers and products prediction

Based on existing high risk customers and products' data, the potential high risk customers and products can be predicted. The company can use these results to build up a suitable strategy, either eliminating them or providing suitable services.

Optimum inventory number prediction

The optimum inventory number prediction can help inventory management and save the company's resources.

Investment prediction

When a new product is developed and before it is actually put into market, the company can predict future sales amounts and profits by using data of similar products. Therefore, the company will be able to build up a suitable strategy in advance.

Fraud detection

The company is able to detect frauds by comparing the results of prediction and actual sales amounts.

6.3.3 Association in business

6.3.1.1 Application of the association results

Association technique is applied to discover links between products at customer level. The company can increase turnover by applying cross-selling using association results.

Related products recommendation based on association results is a major application. The recommendation process brings related products to customer and creates business opportunities for the company. Currently, many online stores are using association technique to apply cross-selling. Companies can discover profitable product relationships by applying the association results to their operations. Ultimately, the company gains the ability to recommend proper products to the right customers to improve company's service quality. Amazon bookstore is a typical example. When a user buys book A through Amazon, it always shows related books so that the customer can easily find out other books he/she might be interested in. A discount for buying related books at the same time of buying book A is able to encourage consumption.

The same method can be applied for other online stores such as grocery stores or electrical stores. Even search engines such as Google can use association to improve their services. If a user typed London museum as search keywords, Google can provide a list of all popular museums the user might be interested in London based on association results of web logs.

Association rules can also be applied in businesses to design direct mail campaigns or re-organize store layouts. Locating related items together is a common way to apply cross-selling. The benefits of cross-selling are the most valuable results of this technique.

6.4 Future work

We have demonstrated a practical approach to data mining for retail industry. The usages of data mining results have been addressed in the previous chapters. However an accurate, stable and easy access prediction algorithm is still an open issue.

For multinational retailers, a powerful and accessible data mining tool which can assist the unique business challenges is a vital need. The tool needs to be able to interpret the existing databases, and automatically process new datasets which contain updated data, then process intelligent database queries. The coverage of basic data mining techniques - clustering, prediction and association - is essential. Moreover, the accessibility of the data mining tool is the critical factor. Accessibility not only refers to the accessibility of data mining techniques for non data mining experts, but also refers to the visualization of data mining results. The use of this tool should not require significant efforts from domain experts in order to provide user customizable reports that allow the results to be easily applied to the business.

For the small specialty retailer, automated inventory management based on consumption cycle can be a powerful tool in future business opportunity development. The retailer can predict next visits of customers with a high confidence by tracking consumption dates in the real time. The prediction results can be used for consolidating customers' loyalty and applying cross-selling. The retailer can build up a consumption history for the convenience of customers, which covers purchased products history and expiration information. At the same time, the retailer also receives benefits by managing the customers' information and consolidating relationships with customers. Product recommendation can also be applied based on this information. A simulation with predicted demand and available inventory data may be helpful in identifying the impact of the predictions on profits.

Moreover, convincing the businesses the importance of data mining is still a challenge.

[1] Fayyad UM, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery: An Overview. Advances in knowledge discovery and data mining table of contents 1996; 1-34.

[2] Fayyad UM, Piatetsky-Shapiro G, Uthurusamy R. Summary from the Kdd-03 Panel: Data Mining: The Next 10 Years. ACM SIGKDD Explorations Newsletter 2003; 5: 191-196.

[3] Goebel M. A Survey of Data Mining and Knowledge Discovery Software Tools. ACM SIGKDD Explorations Newsletter 1999; 1: 20-33.

[4] Bishop CM. Pattern Recognition and Machine Learning. Springer. 2006.

[5] Chen MS, Han J, and Yu PS. Data Mining: An Overview From a Database Perspective. IEEE Trans Knowled Data Eng 1996; 8: 866-883.

[6] Groebner DF. A Course in Business Statistics. Pearson Prentice Hall. 2006.

[7] Klosgen W, Zytkow JM. Handbook of Data Mining and Knowledge Discovery. Oxford University Press, Inc. New York, NY, USA.2002.

[8] Fayyad U, Piatetsky-Shapiro G, Smyth P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. Proc.2nd Int.Conf.on Knowledge Discovery and Data Mining, Portland, OR 1996; 82-88.

[9] Westphal CR, Blaxton T. Data Mining Solutions. Wiley New York. 1998.

[10] Tkach DS. Information Mining With the IBM Intelligent Miner Family. An IBM Software Solutions White Paper 1998; 1-29.

[11] Tang ZH, MacLennan J. Data Mining with SQL Server 2005. John Wiley & Sons. 2005.

[12] Seifert JW. CRS Report for Congress. 2004.

[13] Chapman P, Clinton J, Kerber R. CRISP 1.0 Process and User Guide. 2000. CRISPDM Consortium.

[14] Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0: Step-By-Step Data Mining Guide. SPSS Inc 2000; 78.

[15] Reinartz T. Stages Of The Discovery Process. Handbook of data mining and knowledge discovery table of contents 2002; 185-192.

[16] Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann.2006.

[17] Engels R, Lindner G, Studer R. A Guided Tour through the Data Mining Jungle. Proceedings of the 3nd International Conference on Knowledge Discovery in Databases.Newport Beach, CA 1997.

[18] Pawlak Z, Wong S, Ziarko W. Rough Sets: Probabilistic Versus Deterministic Approach. International Journal of Man-Machine Studies 1988; 29: 81-95.

[19] Lingras P. Rough Neural Networks. Proceedings of Sixth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Granada, Spain 1996; 1445–1450.

[20] Lingras P. Applications of Rough Patterns. Rough Sets in Data Mining and Knowledge Discovery 1998; 2: 369–384.

[21] Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence Menlo Park, CA, USA. 1996.

[22] Akerkar R, Lingras P. Building an Intelligent Web: Theory and Practice. 2008: 326.

[23] Larose DT. Discovering Knowledge in Data: An Introduction to Data Mining. Wiley-Interscience. 2005.

[24] Ming Zhu. Data Mining. University of Science and Technology of China. 2002.

[25] Weiss SM, Indurkhya N. Predictive Data Mining: A Practical Guide. Morgan Kaufmann. 1998.

[26] John GH. Enhancements to the Data Mining Process. 1997.

[27] Kohonen T. Self-Organizing Maps. Springer. 2001.

[28] Haykin S. Neural Networks: A Comprehensive Foundation. Prentice Hall PTR Upper Saddle River, NJ, USA. 1994.

[29] Quinlan JR. C4. 5: Programs for Machine Learning. Morgan Kaufmann. 1993.

[30] Cortes C, Vapnik V. Support-Vector Networks. Mach Learning 1995; 20: 273-297.

[31] Vapnik VN. Statistical Learning Theory. Wiley New York. 1998.

[32] Vapnik VN. The Nature of Statistical Learning Theory. Springer. 2000.

[33] Lingras P, Butz C. Rough Set Based 1-V-1 and 1-Vr Approaches to Support Vector Machine Multi-Classification. Inf Sci 2007; 177: 3782-3798.

[34] Lingras P, Butz C. Rough Support Vector Regression. 2007.

[35] Boser BE, Guyon IM, Vapnik VN. A Training Algorithm for Optimal Margin Classifiers. Proceedings of the fifth annual workshop on Computational learning theory 1992; 144-152.

[36] Osuna E, Freund R, Girosi F. Support Vector Machines: Training and Applications. 1997.

[37] Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Springer. 1997.

[38] Milenova BL, Yarmus JS, Campos MM. SVM in Oracle Database 10g: Removing the Barriers to Widespread Adoption of Support Vector Machines. Proceedings of the 31st international conference on Very large data bases 2005; 1152-1163.

[39] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press. 2000.

[40] Hoffmann A. Learning Theory and Support Vector Machines. 2003.

[41] Yang H. Margin Variations in Support Vector Regression for the Stock Market Prediction. 2003.

[42] Jensen DD. Knowledge Evaluation: Statistical Evaluations. In: Klosgen W and Zytkow JM eds. Handbook of Data Mining and Knowledge Discovery. New York: Oxford University Press, 2002: 475-489.

[43] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA. 2000.

[44] Dunn JC. Well Separated Clusters and Optimal Fuzzy Partitions. Journal of Cybernetics 1974; 4: 95-104.

[45] John GH, Langley P. Static versus Dynamic Sampling for Data Mining. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining 1996; 367370.

[46] Kivinen J, Mannila H. The Power of Sampling in Knowledge Discovery. Proceedings of the thirteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems 1994; 77-85.

[47] Gusfield D. Algorithms on strings, trees, and sequences. Cambridge University Press New York. 1997.

[48] Waterman MS. Introduction to computational biology. Chapman & Hall New York, NY. 1995.

[49] Hudson S, Ritchie B. Understanding the Domestic Market Using Cluster Analysis: A Case Study of the Marketing Efforts of Travel Alberta. Journal of Vacation Marketing 2002; 8: 263.

[50] Sung TK, Chang N, Lee G. Dynamics of Modeling in Data Mining: Interpretive Approach to Bànkruptcy Prediction. J Manage Inf Syst 1999; 16: 63-85.

[51] Fayyad UM, Djorgovski SG, Weir N. Automating the Analysis and Cataloging of Sky Surveys. Advances in knowledge discovery and data mining table of contents 1996; 471-493.

[52] Apte C, Hong SJ. Predicting Equity Returns from Securities Data with Minimal Rule Generation. Advances in Knowledge Discovery and Data Mining 1996; 514–560.

[53] Wuthrich B, Cho V, Leung S, Permunetilleke D, Sankaran K, Zhang J. Daily Stock Market Forecast from Textual Web Data. Systems, Man, and Cybernetics, 1998.1998 IEEE International Conference on 1998; 3.

[54] Carbonara L. Telecommunications. Handbook of data mining and knowledge discovery table of contents 2002; 781-787.

[55] Roughan M, Zhang Y. Secure Distributed Data-Mining and its Application to Large-Scale Network Measurements. ACM SIGCOMM Computer Communication Review 2006; 36: 7-14.

[56] Weiss GM. Predicting Telecommunication Equipment Failures from Sequences of Network Alarms. Handbook of Knowledge Discovery and Data Mining, Oxford University Press, June 2002; 891-896.

[57] Lingras P, Hogo M, Snorek M, West C. Temporal Analysis of Clusters of Supermarket Customers: Conventional Versus Interval Set Approach. Inf Sci 2005; 172: 215-240.

[58] Lingras P, Adams G. Selection of Time-Series for Clustering Supermarket Customers. 2002.

[59] West C, Jain A, Lingras P, Leonard B. Supermarket Customer Attrition Analysis Based on Cluster Membership Patterns. Proceedings of the First Indian International Conference on Artificial Intelligence 2003; 1132-1140.

[60] Lingras P, Hogo M, Snorek M, Leonard B. Clustering Supermarket Customers Using Rough Set Based Kohonen Networks. Proceedings of Fourteenth International Symposium on Methodologies for Intelligent Systems, Lecture Notes in Artificial Intelligence Series 2003; 2871: 169–173.

[61] West C, MacDonald S, Lingras P, Adams G. Relationship between Product Based Loyalty and Clustering Based On Supermarket Visit and Spending Patterns 2004; 2: 85-100.

[62] Lawrence RD, Almasi GS, Kotlyar V, Viveros MS, Duri SS. Personalization of Supermarket Product Recommendations. Data Mining and Knowledge Discovery 2001; 5: 11-32.

[63] Lingras P. Rough Set Clustering for Web Mining. fuzzy systems, 2002.FUZZ-IEEE'02.Proceedings of the 2002 IEEE International Conference on 2002; 2.

[64] Safer AM. A comparison of two data mining techniques to predict abnormal stock market returns. Intelligent Data Analysis 2003; 7: 3-13.

[65] Goebel M. A survey of data mining and knowledge discovery software tools. ACM SIGKDD Explorations Newsletter 1999; 1: 20-33.

[66] Berkowitz EN, Crane FG, Kerin RA, Hartley SW, and Rudelius W. Marketing, fifth Canadian edition. 2003.

[67] Buckinx W. Using predictive modeling for targeted marketing in a non-contractual retail setting. 2005.

[68] Ester M, Ge R, Jin W, and Hu Z. A microeconomic data mining problem: Customer-oriented catalog segmentation. 2004; 557-562.

[69] Chen MC, Chiu AL, and Chang HH. Mining changes in customer behavior in retail marketing. Expert Syst Appl 2005; 28: 773-781.

[70] Giering M. Retail sales prediction and item recommendations using customer demographics at store level. 2008.

[71] Brijs T, Goethals B, Swinnen G, Vanhoof K, and Wets G. A data mining framework for optimal product selection in retail supermarket data: The generalized PROFSET model. 2000; 300-304.

[72] Chen YL, Tang K, Shen RJ, and Hu YH. Market basket analysis in a multiple store environment. Decis Support Syst 2005; 40: 339-354.

[73] Dhond A, Gupta A, and Vadhavkar S. Data mining techniques for optimizing inventories for electronic commerce. 2000; 480-486.

[74] Kohavi R, Mason L, Parekh R, and Zheng Z. Lessons and challenges from mining retail e-commerce data. Mach Learning 2004; 57: 83-113.

[75] Ahmed S. Applications of data mining in retail business. 2004; 2.