Lingnan University

Digital Commons @ Lingnan University

Theses & Dissertations

Department of Marketing and International
Business

10-2013

Customer selection for direct marketing : bi-objective optimization using support vector machine

Qian WANG

Follow this and additional works at: https://commons.ln.edu.hk/mkt_etd

Part of the [Marketing Commons](#)

Recommended Citation

Wang, Q. (2013). Customer selection for direct marketing: Bi-objective optimization using support vector machine (Master's thesis, Lingnan University, Hong Kong). Retrieved from http://dx.doi.org/10.14793/mkt_etd.15

This Thesis is brought to you for free and open access by the Department of Marketing and International Business at Digital Commons @ Lingnan University. It has been accepted for inclusion in Theses & Dissertations by an authorized administrator of Digital Commons @ Lingnan University.

Terms of Use

The copyright of this thesis is owned by its author. Any reproduction, adaptation, distribution or dissemination of this thesis without express authorization is strictly prohibited.

All rights reserved.

CUSTOMER SELECTION FOR DIRECT MARKETING:
BI-OBJECTIVE OPTIMIZATION USING SUPPORT VECTOR MACHINE

by
WANG Qian

A thesis submitted in partial fulfillment
of the requirements of the Degree of
Master of Philosophy in Business
(Marketing and International Business)

Lingnan University

2013

ABSTRACT

Customer Selection for Direct Marketing: Bi-Objective Optimization Using Support Vector Machine

by

WANG Qian

Master of Philosophy

A major challenge in direct marketing is to build a customer-selection model that can help achieve higher response rate and greater profit at the same time. In this study, I adopt a bi-objective optimization (BOO) approach and propose a two-stage method using support vector machine (SVM) and support vector regression (SVR) to maximize response rate and profit simultaneously. To deal with the difficulty of learning models from imbalanced data, synthetic minority over-sampling technique (SMOTE) is used to generate more balanced datasets. Experiments are conducted on two datasets, a direct marketing dataset and the KDD-98 dataset, to compare the predictive performance of the two-stage BOOSVM with other benchmark methods including logistic regression and the parallel Multi-objective Evolutionary Algorithm (MOEA). The results of decile analysis suggest that the proposed two-stage BOOSVM model with SMOTE method is more effective and efficient than the competing models in improving response rate and profitability.

DECLARATION

I declare that this is an original work based primarily on my own research, and I warrant that all citations of previous research, published or unpublished, have been duly acknowledged.

(WANG Qian)

August 25 2013

CERTIFICATE OF APPROVAL OF THESIS

CUSTOMER SELECTION FOR DIRECT MARKETING:
BI-OBJECTIVE OPTIMIZATION USING SUPPORT VECTOR MACHINE

by

WANG Qian

Master of Philosophy

Panel of Examiners:

SIGNED	(Chairman)
Dr. Hon-kwong LUI
SIGNED	(External member)
Dr. Xubing ZHANG
SIGNED	(Internal member)
Prof. Geng CUI
SIGNED	(Internal member)
Dr. Ling PENG	

Chief Supervisor:

Dr. Geng CUI

Co-supervisor:

Dr. Man-leung WONG

Approved for the Senate:

SIGNED

Prof. Jesús SEADE
Chairman, Postgraduate Studies Committee

03 OCT 2013

Date

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION	1
1.1 DIRECT MARKETING	2
1.2 DIRECT MARKETING DATASET	4
1.3 CUSTOMER-SELECTION MODELS IN DIRECT MARKETING.....	4
1.4 MULTI-OBJECTIVE OPTIMIZATION PROBLEM IN DIRECT MARKETING.....	5
1.5 RESOURCE CONSTRAINT IN DIRECT MARKETING.....	6
1.6 SUPPORT VECTOR MACHINE	8
1.7 RESEARCH GAPS IN EXISTING RESEARCH	9
1.8 ORGANIZATION OF THE THESIS	10
CHAPTER 2. LITERATURE REVIEW	13
2.1 TRADITIONAL STATISTICAL TECHNIQUES FOR DIRECT MARKETING.....	14
2.2 DATA MINING THCHNIQUES USED IN DERECT MARKETING	18
2.3 CUSTOMER-SELECTION MODELS IN DIRECT MARKETING.....	24
2.4 MULTI-OBJECTIVE OPTIMIZATION PROBLEM IN DIRECT MARKETING.....	29
2.5 SUPPORT VECTOR MACHINE IN DIRECT MARKETING	31
2.6 SUMMARY	33
CHAPTER 3. SVM PRINCIPLE AND MODEL DERIVATION	35
3.1 BASIC SUPPORT VESTOR MACHINES THEORY.....	35
3.2 MODEL CONSTRUCTION	42
CHAPTER 4. ALGORITHM AND DATA.....	52
4.1 ALGORITHM	52
4.2 DATASET DESCRIPTION.....	56
4.3 DATA DROCESS.....	58
CHAPTER 5. RESULTS AND COMPARISONS	62
5.1 EXPERIMENTS FOR TESTING THE PROPOSED MODEL.....	62
5.2 RESULTS OF MODELS COMPARISONS: KDD-98 DATASET	71
CHAPTER 6. CONCLUSIONS	75
6.1 MAJOR FINDINGS	75
6.2 IMPLICATIONS	77
6.3 LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH.....	78
APPENDIX A and B.....	80
REFERENCE.....	83

ACKNOWLEDGEMENT

I sincerely acknowledge my supervisors, Prof. Geng Cui and Dr. Manleung Wong, who provide me inspiring ideas and valuable advice with extraordinary patience and consistent encouragement during all the stages of my MPhil studies. Without their strong support, my thesis cannot have been the present form. I would like to also express my gratitude to Dr. Ling Yee Li, Dr. Ling Peng, and Dr. Shing Chung Poon for all the help they provided to me during the last two years. Furthermore, I wish to extend my appreciation to all the seminar participants at Lingnan University for their valuable comments on my research. Finally, I want to thank the staff at the Marketing and International Business Department, especially Ms. Lee Koan Jo, Ms. Kwok Wai Chi and all my dear friends and my family for your support all the time.

Customer Selection for Direct Marketing:

Bi-Objective Optimization Using Support Vector Machine

Chapter 1. Introduction

As marketing has become more competitive and customer needs have become more diversified, corporations face unparalleled challenges. Generally there are two approaches for enterprises to advertise and promote: mass marketing and direct marketing. Mass marketing, which traditionally focuses on the media of radio, television and newspapers, aims at reaching a broad audience without discrimination. It used to be an effective way of promotion when the products were in great demand by the public. However, because nowadays the marketplace is increasingly competitive with an overwhelming variety of products and marketing methods, mass marketing has gradually turned less effective and less popular.

In recent years, a huge amount of customer data is available to companies at an unprecedented speed. The data are kept in databases and contain different types of information including customer purchase history and some demographic information and other customer personal characteristics such as profession, age, sex, etc. These data can be used to select a right subset of customers so as to maximize the profit return as well as to maintain direct relationship with customers. This type of marketing is referred to as direct marketing, which is a channel-agnostic form of advertising that allows businesses and nonprofit organizations to communicate directly with customers. The advertising tools for direct marketing include phone call,

text messaging, email, interactive consumer websites, online display ads, fliers, catalog distribution, promotional letters, and outdoor advertising. The superiorities of direct marketing are obvious in many aspects, for example, shopping convenience, technological advance, the possibility to reduce cost for customers and companies. A more detailed description of direct marketing is as below.

1.1 Direct Marketing

The origin of direct marketing dates back to the nineteenth century and it was firstly used in mail order industry in the United States (Brandweek, 2009). Nowadays, direct marketing plays a broader role to satisfy customers' ever-changing needs and it can be successfully and effectively assisted by various sophisticated techniques. Direct marketing has made it convenient for marketers to communicate directly with their customers, and its feature of measurable result is attractive to many marketers. Meanwhile, changes in tastes and media fragmentation have made conventional mass marketing less effective in getting the response from customers, although mass marketing is still quite significant in sustaining brands (Bose and Chen, 2009).

Direct marketing has become more and more popular these days, evidenced by the data on its remarkable growth. For example, Direct Marketing Association reported in 2010 that a total \$153.3 billion was spent on direct marketing by commercial and nonprofit marketers, and the monetary amount accounted for 54.2% of all advertising expenditures in the USA. In addition, direct marketing accounted for 8.3% of total US Gross Domestic Product in 2010. In the same year, direct

marketing provided job opportunities for 1.4 million employees in USA, as well as another 8.4 million other employees in direct marketing related jobs.

Because of the growing popularity and remarkable performance of direct marketing, academic researchers become more interested in direct marketing. Sophisticated techniques can help marketers formulate effective direct marketing strategies. Elsner et al. (2003) report that direct marketing companies continue to increase their market shares after they adopt a direct modeling approach called Dynamic Multilevel Model.

Some traditional statistical methods and data mining models are popular for developing strategies in direct marketing. Traditional statistical techniques are the most common ones in direct marketing. Regression, a basic statistical technique, is widely used, but it is limited in analytical capabilities. Linear regression models and logistic regression models have been commonly adopted in direct marketing. The former ones can be used to produce continuous scores, while the latter ones can be used to perform binary classification (Bult et al., 1997). Another technique used in direct marketing is data mining, which is narrowly defined as the automated discovery of "interesting" but non-obvious patterns hidden in a database (Peacock, 1998) and the core of the knowledge discovery and data mining (KDD) process (Fayyad et al., 1996). Nowadays, many businesses can generate and collect a huge amount of data in a relatively short period. The explosive growth of data requires a more efficient way to extract useful knowledge. Thus marketing is a major area for applying data mining techniques (Shaw et al., 2001).

1.2 Direct Marketing Dataset

Due to the saturated market and intensive competition, many companies realize that their firm databases are one of their most valuable assets (Athanasopoulos, 2000; Jones et al., 2000). Nowadays, large amounts of data about customers' information are available to companies, and these data can be used to establish and maintain a direct relationship with customers in order to target them individually for specific product offers and services. Companies can also update their databases by making solicitations and transactions with customers, so they regard the communication experience with customers as great opportunities to collect valuable data (Rygielski et al., 2002).

In direct marketing, two types of data about customer characteristics are typically used. The first type of data is referred to as external data (Van der Sheer, 1998), including geographics, lifestyle and demographics; the second type is regarded as customers' interactive behaviors, including customer transaction records, web browsing records and feedback from customers (Bult, 1993).

1.3 Customer-Selection Model in Direct Marketing

For direct marketing companies, sending product offers to all customers in the database is impossible due to budgetary constraints. Thus, it is only necessary to contact only those customers who are the most likely to be interested in the offer, because the costs of such a full-scale mailing campaign are too large to satisfy the budget constraint. In addition, sending many uninteresting product offers to

customers is not only undesirable to customers but also harmful to the relationships between companies and their customers. Hence, customer selection problem should be solved by considering how to select the customers who are interested in their offers.

An effective identification of likely responders involves skillful and powerful analysis of customer database for selecting customers. Thus, a major task in direct marketing is to develop a model for customer selection, and the model performs mapping from attribute vectors characterizing potentials responders to marketing target vectors. An effective customer selection model will increase a firm's profit, while an ineffective one will increase the marketing cost and worsen the customer's relationship with the firm (Potharst et al., 2000). Data availability and effective methodology are important factors for the success of customer selection models. By utilizing the large dataset, statistical or data mining techniques can be used to predict potential customers

1.4 Multi-Objective Optimization Problem in Direct Marketing

The evaluation of customer selection models in direct marketing is usually done on the basis of response rate (accuracy) and profit. Response rate is the ratio of the number of responders over the total number of customers in the sample. A survey's response rate has been viewed as an important indicator of survey quality for many years and the conclusion that higher response rates always ensure more accurate survey results is supported by many studies (Rea and Parker, 1997). However, direct

marketers would like to know not only responders but also profitable customers who will spend more money than others, because profit is the most direct metric to measure return on marketing. High-profit customers are the minority class and it has a smaller effect on response rate but a more significant effect on revenue. If all the low-profit responders are selected, the response rate may be quite high, but the revenue may not be good enough. In other words, good response models may not help generate greater profit. Therefore, besides the classification function, customer models also need regression functions to predict customer profit. Some models can be used for both classification and regression are referred to as two-stage models, which are constructed by the combination of two different models, taking advantage of each model and overcoming the weaknesses of both ones. Usually these two-stage models score the probability of response in their first stage and then score the monetary value of customers in the second stage. Two-stage models can realize the optimization of the both objectives respectively in the two stages. However, up to date, these models view the two objectives as two separate parts and neglect the relationship between them. Thus, a bi-objective (response rate and profit) optimization model to maximize the response rate and return simultaneously for a direct marketing solicitation seems to be necessary and significant.

1.5 Resource Constraint in Direct Marketing

In marketing, most of the revenue and profit is produced by the top deciles of customers and marketing budget for promotion is always limited in that only a

fraction of customers in the dataset can be and also should be selected. Decile analysis is widely used for data analysis at different depths-of-file. In decile analysis, individuals are ranked in a descending order by their respective model scores (higher scores indicating better performance), and then individuals are categorized into 10 groups with the same size. In addition, the results of this series of analysis are evaluated using the measure “lift”. Lift is a measure of the performance of a targeting model against a random choice targeting model, and a higher response rate means that the response rate for the target group is much better than the average response rate for the population as a whole. Lift is simply the ratio of these values: target response divided by average response. In other words, “response lift” is the ratio of the number of true positives to the total number of records identified by the classifier in comparison with that of a random model at a specific decile of the file (Bhattacharyya, 1999). For instance, a lift of 200 indicates that a specified model performs twice as well as a random model. In other words, lift measures the improvement of proposed models in response and/or profitability over a random model at a specific depth of file. The Average Cumulative Lift (ACL) at specific depths of file provides a measure of improvement over a random model, and it is calculated as follows:

$$Average\ Cumulative\ Lift_{decile} = \frac{cumulative\ average\ profit_{decile}}{overall\ average\ profit} \times 100. \quad (1.1)$$

Thus, we compare the performance of models across the top deciles using cumulative lifts. In my study, I use two types of ACLs, which are the average accumulated response rate lift (AARL) and the average accumulated profit lift

(AAPL). AARL and AAPL are both chosen as metrics of model performance in my experiments. The two ACLs are described in detail as follows,

$$AARL_{decile} = \frac{\text{cumulative average response rate lift}_{decile}}{\text{overall average response rate lift}_{decile}} \times 100; \quad (1.2)$$

$$AAPL_{decile} = \frac{\text{cumulative average profit lift}_{decile}}{\text{overall average profit lift}_{decile}} \times 100. \quad (1.3)$$

AARL is simply the ratio of two values: the obtained average response probability of a certain model divided by the average response probability of randomly selection. Similar to that, AAPL means the following ratio: the average profit obtained by using a certain model divided by the average customer profit obtained by randomly selection. Thus, higher AARL and/or AAPL mean it is more effective to select customers by using the model.

1.6 Support Vector Machine

Support Vector Machine (SVM), a supervised learning model, is increasingly attractive because of its solid statistical foundation and effective performance in various areas (Osuna et al., 1997), including object recognition, handwriting recognition, face detection, speaker identification, and text categorization (Burges, 1998). Recent studies have shown SVMs' distinct advantages such as better generalization, increased speed of learning, improved ability to find a global optimum and the ability to deal with linearly non-separable data. In addition, SVMs can handle and evaluate data points fast, without losing the degree of accuracy (Candade, 2004).

The current standard soft-margin SVM is originally proposed by Vapnik and Cortes in 1995. Then in 1998 they receive the Paris Kanellakis Theory and Practice Award for their development of support vector machine (SVM), which has proved to be a highly effective algorithm for supervised learning. SVMs belong to a family of generalized linear classifiers and can be interpreted as an extension of the perceptron, which follows a form in that an input is assigned into one of several possible non-binary outputs. An SVM is constructed by seeking the maximum-margin classifiers, which minimize the empirical classification error as well as maximize the geometric margin simultaneously. Comparisons of SVMs with other classifiers have been made by Meyer et al. (2003). In their study, they compare a popular SVM implementation (LIBSVM) with 16 classification methods and 9 regression methods. Among them, SVMs show better performances on both classification and a regression task, however, the superiority is not significant.

The advantages of an SVM in learning patterns from datasets come from three points. First, only two types of free parameters, the upper bound and the kernel parameter, are needed to be chosen in an SVM and this process is also quite practicable. Second, because the training of a SVM is done by solving a linearly constrained quadratic problem, the solution of an SVM is unique and has a better chance of finding the global optimum. As a result, over-fitting problem is unlikely to happen in SVMs. Third, conventional neural networks are based on Empirical Risk Minimization (ERM) principle, so as to minimize misclassification errors and deviation from the optimal fitting of the training data. On the contrary, SVMs' theory

embodies the Structural Risk Minimization (SRM) principle, in which the classifiers can minimize an upper bound of generalization error. The SRM principle can assist SVM to avoid the over-fitting problem by balancing the model's complexity against its success at fitting the training data.

1.7 Research Gaps in Existing Research

The evaluation of customer selection models in direct marketing is usually achieved on the basis of response rate (accuracy) and profit. Response rate is the ratio of the number of responders over the total number of customers in the sample. Till now, many response models have been built with various techniques in direct marketing. These response models can be used for direct marketers to reach higher response rate. However, direct marketers would like to reach not only responders but also profitable customers who will spend more money than others, because profit is the most direct metric to measure performance. Thus, targeting customers only with response models is not enough and cannot ensure greater profitability, since response models tends to ignore high profit customers, which are the most valuable to companies. Considering the importance of profit, some customer selection models in direct marketing are based on profit in that profit is the only metric to measure the performance of a certain customer selection model. However, similar to the response models, these profit models also target a single objective, which are not enough to measure direct marketing performance.

Some other customer selection methods in direct marketing can be used for both

classification and regression, which are referred to as two-stage models. These models are constructed by the combination of two different models. Usually these two-stage models score the probability of response in its first stage and then score the monetary value of customers in the second stage. Two-stage models can realize the optimization of the both objectives respectively in the two stages. However, up to date, these models view the two objectives as two separate parts and neglect the relationship between them. Thus, in direct marketing solicitation, a bi-objective (response rate and profit) optimization model to maximize the response rate and profitability simultaneously is quite necessary. For this reason, I propose a two-stage BOOSVM model to realize the bi-objective optimization problem.

1.8 Organization of the Thesis

Nowadays, direct marketing has proved its benefit given countless successful applications. It has attracted much attention from academics with an increasing number of studies in this area. How to target customers is the core problem in direct marketing. Customer selection models can help to solve this problem properly. Because a customer selection model's performance should be evaluated on both response rate (accuracy) and profit, customer selection problem can be seen as a bi-objective optimization problem. However, most studies on customer selection pay attention to only response rate. Profit models as well as bi-objective optimization models are rare. To solve this problem, this study proposes a bi-objective optimization model for customer selection. Data mining technique can help

marketers extract meaningful knowledge to improve efficiency and quality of their customer selection strategy. Among various data mining techniques, SVM is increasingly attractive for its solid statistical foundation and effective performance in various areas. Therefore, in this study I adopt the SVM method to search for a proper customer selection model to solve the bi-objective (response rate and profit) optimization problem in direct marketing. The remaining of this thesis is organized as follows.

Chapter 2 provides an integrative review of the related work on previous customer selection models and various techniques applied in direct marketing.

Chapter 3 represents an introduction to the basic theory of SVM technique, the deduction process of the two-stage BOOSVM model, and the related techniques used in the two-stage BOOSVM model.

Chapter 4 discusses the algorithm and the datasets used in this thesis.

Chapter 5 contains the experiments with two-stage BOOSVM model and the other two benchmark models: logistic regression and MOEA, and the comparisons of results from two different datasets.

Chapter 6 includes the conclusions and discussion of its implications, practical applications, limitations and directions for future research.

Chapter 2. Literature Review

The Direct Marketing Association (DMA) defines direct marketing as "communications where data are used systematically to achieve quantifiable marketing objectives and where direct contact is made, or invited, between a company and its customers and prospective customers" (Mullin, 2002). The definition obviously implies the importance of customer segmentation and personalized advertising.

The use of direct marketing advertising is growing at a fast pace. The emergency of the Internet has largely decreased the operation cost for direct marketers, which can consolidate the success of direct marketing campaigns. In 2005, the average rate of direct marketing expenditures is 14.6% in Germany, 73.6% in the UK, and 5.5% in France (Barwise and Farley, 2005). The growing popularity of direct marketing has aroused academic interest and research in this area. Abundant and sophisticated direct marketing models can help decision makers build reasonable strategies and assist marketers in conducting effective direct marketing solicitations. It is reported that a direct marketing company, Rhenania, successfully increases its market share after their implication of Dynamic Multilevel Modeling approach (Elsner et al., 2003).

Nash (1984) proposes four successful elements of a direct mail campaign, which are timing, offer, communication, and selection of customers. Among these elements,

selection of customers is critical in determining whether direct marketing activity is successful. Thus, in direct marketing, an important research question is: what techniques should be used for building models for customer selection? Usually traditional statistical and data mining techniques are used for selecting customers in direct marketing.

2.1 Traditional Statistical Techniques for Direct Marketing

Conventionally, many statistical methods have been applied to quantitative modeling, and many of these methods are regression techniques. Traditional statistical techniques include regression models and the RFM model. Table 2.1 shows some of the literature about traditional statistical techniques used in direct marketing and the application details of each technique are described in the following section.

Table 2.1. Direct marketing studies using traditional statistical techniques

Traditional statistical techniques	Representative studies
Linear regression	Levin and Zahavi (1998), Malthouse (1999)
Logistic regression	Bult and Wansbeek (1995), Bult et al. (1997), Levin and Zahavi (1998), Heilman et al. (2003) Bodapati and Gupta (2004)
<i>RFM model</i>	Colombo and Jiang (1999), Yeh et al., (2009), Bult and Wansbeek (1995)

2.1.1 Linear Regression

In statistics, linear regression is an approach to model the relationship between a dependent variable y and an independent variable set X . Results of linear regression are predicted by the linear predictor functions, and the unknown parameters of the

linear predictor functions' are learned from given data. Linear regression is one of the early regression models, which can be used to generate continuous scores. A threshold is usually set before selecting target customers based on the scores given by a linear regression model. If a customer's score is higher than the threshold, then the customer is selected, and vice versa. The proportion of the selected customers from all the customers is referred to as the regression model's response rate (Bose and Chen, 2009). Levin and Zahavi (1998) use a linear regression model to predict customers' response based on three measures: profitability analysis, goodness-of-fit criteria and predictive accuracy. Malthouse (1999) uses another linear regression model to estimate how much a customer will spend on an upcoming offer.

2.1.2 Logistic Regression

An explanation of logistic regression begins with the logistic function, which always takes on values between zero and one:

$$f(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}, \quad (2.1)$$

viewing t as a linear function of an explanatory variable x , one have:

$$\pi(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}, \quad (2.2)$$

which is equivalent to:

$$\frac{\pi(t)}{1 - \pi(t)} = e^{(\beta_0 + \beta_1 x)}. \quad (2.3)$$

A graph of the function f is shown in Figure 2.1.

Figure 2.1. The logistic function, with $\beta_0 + \beta_1 x$ on the horizontal axis on $\pi(x)$ on the vertical axis.

Here the occurring event indicates whether the customers respond to the promotion or not, and the likelihood function is built from the customer dataset. With the function constructed, parameters of all attributes can be obtained according to Maximum Likelihood Estimation (MLE) principle (Linder et al., 2004). Then a logistic function can be built to estimate the response probabilities of each individual.

Most textbooks cover logistic regression as the basic method due to its simplicity and explanatory ability (Hosmer and Lemeshow, 1989). Because the response variable is discrete, it will be much easier to construct models with logistic regression. In direct marketing, logistic regression models are commonly used to predict potential customers. Sen (1990) proposes to use logistic regression to search for a proper number of attributes for analysis in response modeling. Malthouse (1999) suggests combining ridge regression with stepwise regression to search for a proper number of attributes for response modeling. Naik and Tsai (2004) find that, in small databases, marketers could calibrate logistic regression models via maximum-likelihood methods to determine significant variables and assess customers' response probability.

However, direct marketing datasets are updated with new information constantly as more customer information variables are stored. The application of logistic regression method may suffer from some problems. First, the application of Maximum Likelihood Estimation (MLE) may fail in the dataset with a large number of predictor variables due to the lack of convergence, large estimated coefficient variances, poor predictive accuracy, and the reduced power for testing hypotheses concerning model assessment (Genkin et al., 2007). Second, even if such numerical problems can be avoided, the asymptotic properties of maximum likelihood have a tendency to cause the over-fitting problem (Genkin et al., 2007). Third, because of the great possibility to neglect predictor variables during the variable selection process, one may encounter problems of model specification.

2.1.3 RFM Model

A simple Recency–Frequency–Monetary (RFM) value model is originally proposed by Colombo and Jiang (1999). It is a stochastic approach to rank (or score) customers. In the RFM model, recency represents the length of time period since the last purchase; frequency denotes the number of purchases within a specified time period; monetary value means the amount of money spent in this specified time period. In direct marketing, RFM information helps marketers to estimate the probabilities of customers to respond (Colombo and Jiang, 1999) and customer feedback is used by researchers to discover customers' attitudes towards products (Bult et al., 1997; Ha et al., 2005). RFM model is also a behavior-based model to

analyze the behavior of customers and then to make predictions based on the information derived from the data selected beforehand (Yeh et al., 2009).

The RFM model is widely applied in the direct marketing. Bult and Wansbeek (1995) extend the RFM method to Chi-square Automatic Interaction Detection (CH-AID) method in direct marketing. Colombo and Jiang (1999) present a simple stochastic RFM model to solve the problem about how to choose customers in the firm's database to target with an offer, which is a central problem in database marketing.

The RFM model applies quite basic knowledge and it almost does not require any statistical software. But practically, the disadvantages of the RFM model are also quite obvious. In practice, customers' purchase behavior is not determined by only three elements. On the contrary, many more variables may affect the performance of a response model, for instance, the demographic variables. Thus, the variables used in RFM models are not sufficient. In addition, since RFM variables are not necessarily mutually independent, they may have the problem of double counting (Bult and Wansbeek, 1995). Malthouse (1999) considers that RFM models can be susceptible to over-fitting problem.

2.2 Data Mining Techniques Used in Direct Marketing

Data mining is a non-trivial process of discovering novel, implicit, useful and comprehensive knowledge from large amount of data. By using the method of data mining, more non-obvious patterns and hidden knowledge in databases could be

discovered and applied in practice. Data mining techniques are quite efficient in solving classification and regression problems, and they are already used to deal with many marketing problems, such as predicting bankruptcy, loan default, and modeling consumer choice (Hu et al., 1999). In recent years, data mining techniques are also in the spotlight in the field of direct marketing and gradually gain their stable position in the area of targeted marketing. These techniques include but are not limited to decision tree, neural networks, and evolutionary algorithm.

Table 2.2. Literature about data mining techniques

Data mining techniques	Representative studies
Decision tree	Haughton and Oulabi (1997), Ling and Li (1998), Freund and Schapire (1996)
Artificial neural networks	Zahavi and Levin (1997), Gruca et al. (1999), Kim and Street (2004), Ha et al. (2005), Kim et al. (2005), Shin and Sohn (2004)
Evolutionary algorithm	Yu et al. (2005), Bhattacharyya (1998), Bhattacharyya(1999), Bhattacharyya(2000), Dou et al. (2008)

2.2.1 Decision Tree Methods

Decision tree learning is a method commonly used in data mining and its goal is to predict the value of a target variable with a tree-structure model based on several input variables. In a decision tree, internal node represents test on an attribute, a branch represents the outcome of a test and each leaf node stands for a class label. A tree-structure model can be constructed step-by-step by splitting the source set into classifications. This process is repeated on each division step with a so-called recursive partitioning method, which ends until all predefined input variables under a node are considered. This is a top-down induction process and a type of greedy

algorithm (Quinlan, 1986).

Decision trees used in data mining area have two main types. The first type, classification tree analysis, is used when the predicted outcome is the class to which the data belong. The other one is the regression tree analysis in which the range for predicted outcome is a real number (e.g. the price of a house, or a patient's length of stay in a hospital). There are many decision-tree algorithms and the notable ones include the term Classification and Regression Tree (CART) analysis, Iterative Dichotomiser 3 (ID3), a successor of IDS (C4.5), and CHi-squared Automatic Interaction Detector (CHAID). A comparison between CART and CHAID are conducted by Haughton and Oulabi in 1997. They compare the response lifts between the two approaches and find that although the two models are different in their tree-generating mechanisms, no remarkable changes are noted in the response lift. Another comparison between a Naïve Bayes response model and a C4.5 response model is given by Ling and Li (1998).

Various kinds of decision-tree algorithms have been used in direct marketing. Freund and Schapire (1996) conduct experiments on three direct marketing problems to improve the performance in response rate. In their research, they use a C5-based response model to classify responders and improve the classification quality rather than the accuracy of the response model by accommodating a better ranking. Cui et al. (2006) model consumer responses to direct marketing and compare the results of Bayesian networks with those of neural networks, CART, and latent class regression. Haskell (1993) suggests that large decision trees, in which each test sample can

propagate through all paths simultaneously, are typically pruned to provide better classification accuracy when used with test data. By using a stock market prediction example, he concludes that making a large fuzzy tree is an attractive alternative to pruning.

However, there are some limitations for decision tree learning methods. First, practical decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm in which locally optimal decision is made at each node. Such algorithms cannot guarantee that the global optimal decision tree can emerge. Second, decision-tree learners are prone to building over-complex trees, which do not generalize the data well and cause the over-fitting problem. Third, some concepts are hard to learn due to the difficulty in the expression of decision tree learning.

2.2.2 Artificial Neural Networks

An Artificial Neural Network (ANN) is a mathematical model inspired by biological neural networks and it consists of a large number of small computational elements imitating biological neural networks. The building block inside an ANN is called a perceptron/neuron, which computes a weighted sum of its $(n + 1)$ input signals, and then passes the result through a nonlinear activation function. Mathematically, the output of a perceptron can be represented as

$$y = f\left(\sum_{i=1}^n w_i x_i - w_0\right) \quad (2.4)$$

where f is a non-linear activation function, w_i is the weight associated with the i^{th} input, and w_0 is a constant threshold (or bias) value.

The activation function is also referred to as a squashing function, which limits the amplitude of the output of a perceptron to a closed interval, e.g. [0, 1]. The operation of this function can be interpreted as a mapping from a linear space to a nonlinear space. So far, the most common form of the activation function is sigmoid function, which is defined as a mathematical function having an "S" shape (sigmoid curve). It is a strictly increasing function that exhibits smoothness and asymptotic properties. Usually, the sigmoid function refers to the special case of the logistic function, which is defined by the followed formula:

$$f(x) = \frac{1}{1 + e^{-ax}}, \quad (2.5)$$

where a is the slope parameter and is used to determine the fuzziness of the decision boundary.

ANN is good at learning nonlinear relationships between the input and output and it is the most popular data mining technique used in direct marketing. Zahavi and Levin (1997) employ ANNs in direct marketing with no prior knowledge or assumption about the error distribution. They prove that in some instances neural network model can improve the response rate up to 95% in direct marketing. Besides response prediction, ANN is also used for other purposes in direct marketing. Shin and Sohn (2004) combine ANN with data envelopment analysis techniques to predict the profitability of each customer.

However, there are also some shortages for ANNs. It is proved that neural networks sometimes do not outperform simple logistic regression models. Indeed, it is often the case that a simple logistic regression predicts better than a neural network

(Zahavi and Levin, 1997). One major reason for this phenomenon is that a neural network model has to be built with great care and its performance is sensitive to its complexity, which is determined by the number of synapses or weight parameters. If a network is more complex than the network embedded in the dataset, then the network learns not only the underlying function but also the noise (Hansen and Salamon, 1990). In that case, the learning process will lead to the over-fitting problem, in which the learned function performs quite well on the training data, but fails to predict using the unseen testing data.

2.2.3 Evolutionary Algorithm (EA)

Another type of data mining techniques that continuously gain popularity in direct marketing is Evolutionary Algorithms (EA), such as Genetic Algorithms (GA), Evolutionary Programming (EP), and Genetic Programming (GP). EAs are inspired by biological evolution including selection, crossover, and mutation (Goldberg, 1989). Fitness function determines the quality of the solutions. Evolution of the population then takes place after the repeated application of the above operators. GA, GP, and EP are suitable for optimization problems such as the selection of solicitation targets. EA technique is proved to be effective in solving a variety of problems that are difficult for other techniques. The application of EAs in direct marketing dates back to long time ago. Bhattacharyya (1998) presents a GA-based approach for developing response models to maximize performance at the desired mailing depth. Then Bhattacharyya (1999) presents a GA based approach for obtaining models with

explicit consideration of the mailing depth and performs experiments based on a real-life data set. Bhattacharyya (2000) proposes the use of evolutionary computation based procedures for obtaining a set of non-dominated models with respect to multiple stated objectives.

However, there are several limitations for the use of an EA compared with alternative optimization algorithms. First, it will have some segment limitations in the part of repeating fitness function evaluation for complex problems. In complex high dimensional data, finding the optimal solutions may need quite expensive fitness function evaluations, which is time-consuming. For example, in reality, several hours may be required to deal with only one single function evaluation for the structural optimization problems. Second, EA-based approach is not good at dealing with complexity. Large number of elements that are exposed to mutation will often bring about an exponential increase in the size of search space. Moreover, the stop criterion of an EA-based approach is often not clear. Therefore, EAs are prone to converge towards local optimal or even arbitrary points rather than the global optimum.

2.3 Customer-Selection Models in Direct Marketing

2.3.1 Response Models

Identifying likely customers of a product or service is an important issue in direct marketing. A survey's response rate has been viewed as an important indicator of survey quality for many years. It has been proved that higher response rates

always ensure more accurate survey results (Rea and Parker, 1997). To increase response rate, response models can help scholars and managers to understand how consumers respond to marketing activities individually and collectively (Hanssens et al., 2005). In a response model, the dependent variable is a two-dimensional variable representing response or not. Response models usually belong to classification models, aiming at improving predicative accuracy by inducing classifiers for a given dataset. The prediction is realized by scoring customer probabilities to purchase so that higher scored customers can be indicated with greater target preference. Good response models can select more responders in a firm's database to reduce overall marketing cost without losing opportunities.

Response models are seen as a baseline model in customer selection. Up to now, various statistical and machine learning techniques have been applied to response modeling. Traditional statistical models are popular for response modeling because of their simplicity, explaining ability and availability (Hosmer and Lemeshow, 1989). For example, Colombo and Jiang (1999) suggest a Recency–Frequency–Monetary (RFM) model to estimate the likelihood of future responses; Malthouse (2002) proposes a logistic regression approach to build a response model with considering the dollars spent in response to an offer.

Recently, data mining techniques have been extensively used to identify potential customers for a new product. By analysis on historical purchase dataset, data mining can help marketers to build response models to increase customer selection accuracy. Various response models are built with different kinds of data

mining techniques in direct marketing: Haughton and Oulabi (1997) design response modeling with two decision trees algorithms, which are Classification and Regression Tree (CART) and Chi-Square Automatic Interaction Detector (CHAID); Zahavi and Levin (1997) propose a response model combining neural networks and statistical method in order to find the more effective technique for response modeling; Ling and Li (1998) compare a Naive Bayes response model and a C4.5 response model on three direct marketing problems, which are loan product promotion, life insurance product campaign and bonus program for analysis; Chiu (2002) builds a Case-Based Reasoning response model with Genetic Algorithm (GA) targeting at higher response rate (identification accuracy); Deichmann et al. (2002) construct a decision tree and linear regression combined response model. Response models have some practical problems, as indicated in some literature. Hanssens et al. (2005) point out that usually response models cannot gain top management attention, and they propose short-term and long-term effect response modeling while considering capital markets to deal with the problem; Kim et al. (2006) point out that existing response models only predict classification scores rather than predicting the total amount of money spent, for which they employed a pattern selection method designed for SVR.

2.3.2 Profit Models

The second type of models in direct marketing is profit model, in which the customer features are used as independents to simulate customers' profit variables. Customers' probabilities of purchase can be estimated by a response model for

getting a powerful customer selection strategy. However, there may be an inverse or non-significant correlation between customers' likelihood to buy and their dollar amount to spend. Meanwhile, the greater dollar amount a customer will spend, the more cautious he will be in making a purchase decision. Thus, a direct marketing promotion equipped with a better response model may not lead to more revenue. For example, customers' purchase probabilities to buy may have a negative correlation with the dollar amount to spend in the widely-used KDD98 dataset task in KDD98 Cup. As a result, profit models are necessary to predict customers' monetary amount to purchase.

Customer profitability plays a significant role in the development of customer-selection strategies and some techniques are proposed for profit modeling in direct marketing. Bult and Wansbeek (1995) propose a profit maximization approach in which profit and cost factors are included in the estimation function for logistic models. For a fixed budget, the profit maximization approach to customer selection, which includes all the customers with a positive expected marginal profit, is not realistic. Rao and Steckel (1995) suggest a model that combines the predicted probabilities and the expected profit to re-rank the customers and to float the high value customer to the top of a file, so as to improve profitability. Another Logic-based profit maximization approach is suggested by Van der Sheer (1998), whose estimation function involves profit and cost factors. Meanwhile, researchers have increasingly adopted the lifetime value (LTV) approach to select high value customers based on their long-term contribution of profit to a company in the future

(Venkatesan and Kumar 2004). Otter et al. (2002) use the probability model mainly as a benchmark considering a practice: select a household if the probability of response exceeds the ratio of cost to average yield. They take into account profit and cost parameters for estimation of model related coefficients.

2.3.3 Cost-Sensitive Learning Models

Another type of models in direct marketing is cost-sensitive learning. Response models can be applied to improve customer selection accuracy. However, accuracy is insufficient to guarantee good customer selection models because direct marketing datasets are always markedly imbalanced. For example, a direct marketing dataset may contain 98% non-response individuals and only 2% response individuals. Predictive accuracy will reach 98% if none of the non-responders are missed. In addition, it is assumed that all classification errors have the same cost, while practically misclassification errors of responders and non-responders involve different costs. This implies that customer selection does not merely involve a classification problem. Cost-sensitive learning methods can help to improve the profitability of customer selection models (Cui et al., 2012) by addressing classification problems with non-uniform costs. They typically place greater weight on the minority class, i.e., the responders.

Cost-sensitive learning techniques have been widely used in direct marketing. Bult (1993) propose a discriminate analysis to conduct cost-sensitive learning, in which he combines cost factors with score estimation under two situations: symmetric situations (false negatives cost equals to false positives cost) and

asymmetric loss situations (false negatives cost is different from false positives cost). Considering the unbalanced class and skewed profit distribution character of direct marketing datasets, Cui et al. (2012) integrate cost-sensitive learning via priority sampling technique into a forecasting model to improve the return on investment under resource constraint. Pan et al. (1997) combine cost-sensitivity and data imbalance problem into a data preprocessing operation to explain cost information in the data-preprocessing stage. In addition, to deal with the difficulty of labeling data in direct marketing, Wang (2012) proposes two classification strategies to learn cost-sensitive classifier from training datasets with both labeled and unlabeled data, based on Expectation Maximization (EM).

2.4 Multi-Objective Optimization Problem in Direct Marketing

Predictive models in direct marketing seek to identify individuals who are most likely to respond to promotional solicitations or other intervention programs (Bhattacharyya, 2000). Standard customer selection modeling approaches embody a single objective. However, in reality, decision problems often seek multiple performance measures and marketing decision makers intend to seek the strategies that can optimize on multiple goals. This case can be seen as a kind of multi-objective optimization problem, which is concerned with mathematical optimization problems in which more than one objective functions have to be optimized simultaneously.

In previous literature, more than one objective can be optimized with a single

model, for example, Wong and Cui (2010) propose to implement a parallel MOEA on a real-life direct marketing problem. In their study, a scoring function is used to divide customers into responders and non-responders: if a customer's score is greater than all the other ones' in non-responders class, then the customer is predicted as a responder. They converse a constrained optimization problem into an unconstrained multi-objective (response rate and profit) optimization problem and use MOEAs to find the scoring function and the threshold value. By comparing the models generated by the parallel MOEA, the parallel HGA and the DMAX approach on a large real-life direct marketing dataset, they prove that the parallel MOEA significantly outperform other approaches.

In some other multi-objective optimization used methods, two simple statistical models are combined to take the advantages of each model while overcoming the weaknesses of both of them, and such models are two-stage models. Usually, a two-stage model tries to score the probability of response and monetary value a customer might spend in its two stages respectively (Bose and Chen, 2009). Thus more than one objective can be optimized. Some two-stage models have been used in previous literature. Levin and Zahavi (1998) combine logistic regression models and linear regression models to estimate customer response and monetary value respectively. Colombo and Jiang (1999) use a beta distribution model to predict the customer response probability and a gamma distribution model to predict profit. Baumgartner and Hruschka (2005) adopt a profit model to distinguish different kinds of customers and apply nonlinear regression models to estimate customers' profit and

the returned products.

Besides two-stage models, some other methods and techniques have been used to solve multi-objective optimization problems in direct marketing. Bhattacharyya (1999) presents a GA-based approach for developing response models aimed at maximizing performance at the desired mailing depth. Response rates at different depth-of-files are considered to test model performance. In his study, tradeoffs among these objectives are explored and revamping approaches to control the over-fitting of training data are also investigated. This type of constrained optimization can be considered as indirect multi-objective optimization. Bi (2003) suggests using SVMs to solve a multi-objective program (MOP) by setting two competing goals to minimize the empirical risk and maximize the model capacity.

Generally, multi-objective optimization problem is a practical problem in direct marketing. However, in direct marketing, comparing with single objective optimization models, multi-objective optimization models are still under-developed, and relatively fewer studies have attempted to optimize more than one objective in a single model.

2.5 Support Vector Machine in Direct Marketing

Support Vector Machines (SVMs), have great generalization performance by employing Structural Risk Minimization (SRM) principle (Vapnik, 1995). Although SVMs have shown excellent generalization performance in a wide range of areas, including evaluation of consumer loans, image classification, automatic face

authentication, production values estimation, and text categorization, but their applications in marketing are rather scarce (Cui and Curry, 2005).

SVMs can be used for classification in direct marketing to build response models, and the classification vision of SVM is Support Vector Classification (SVC). Coussement and Poel (2006) apply SVMs to obtain a response model in a newspaper subscription context. They prove that SVMs show better generalization performance than logistic regression and Random Forests when they are applied to noisy marketing data. Cui and Curry (2005) investigate the ability of SVMs to predict outcomes in emerging environments of marketing and they develop a framework to position SVMs against current modeling techniques. Lessmann et al. (2009) improve the SVM approach: genetic algorithms and support vector machines combined model (GA-SVM model). They find that GA-SVM with more accurate and stable results has greater effect than standard SVM with radial kernel by constructing SVMs in a complex real-world scenario of direct marketing.

SVM can be also used to do regression and the regression vision of SVM is Support Vector Regression (SVR), which is originally proposed by Druker et al. (1997) for the linear regression function. SVR adopts slack variables and kernel functions and has been successful in various domains (Druker et al., 1997). SVR also has applications in direct marketing. Kim et al. (2006) apply SVRs for response modeling and solve the two drawbacks of conventional customer selection models in direct marketing. The first drawback is that conventional response models predict only classification scores, but not the total amount of money spent. The second is the

training data is too large.

2.6 *Summary*

In the above sections, most important statistical and data mining techniques used in direct marketing are provided. For each technique, I give a brief theoretical introduction and a literature review of its implication in direct marketing. This chapter also includes a summary of different types of customer selection models. However, in general, there are some problems in the existing models. First, some models only consider a limited part of customer purchase features, which are too simple to achieve satisfactory results. Second, some of these techniques neglect to consider the uncertainty and make the same assumptions for the whole population, which will result in the over-fitting problem that the learning models performs well on training set but poorly on the testing set. Third, single objective models including response models and profit models only consider the optimization on either responses rate or the profit in direct marketing modeling. Good performance on a single objective may not be enough to anticipate an overall good performance in practical direct marketing promotions. Forth, in direct marketing, it is easy to understand that response rate and profit are closely related. However, techniques used in existing two-stage models neglect the relationship between them. The above questions of existing customer selection models inspire me to design a better model for targeting customers. SVM has solid statistical foundation and has also shown effective performance in various areas. In addition, this technique is an approach

which can be used for both classification and regression. Thus I adopt the SVM approach and the two-stage model to improve the performance of customer selection models by optimizing response rate and profit simultaneously in each stage.

Chapter 3. The SVM Principle and Model Derivation

3.1 Support Vector Machines Theory

3.1.1 Optimization Problem

The goal of SVMs is to devise a computationally efficient way of learning ‘good’ separating hyperplanes in a high dimensional feature space. Optimization theory provides the mathematical techniques necessary to find the proper hyperplanes.

The general form of the optimization problem is to find the maximum or minimum of a function, typically subject to certain constraints.

Given functions $f(x)$, $g_i(x)$ ($i=1, \dots, k$) and $h_i(x)$ ($i=1, \dots, m$) defined on a domain $\Omega \in R^n$,

$$\begin{aligned} & \text{minimise} && f(w), && w \in \Omega, \\ & \text{subject to} && g_i(w) \leq 0, && i=1, \dots, k, \\ & && h_i(w) = 0, && i=1, \dots, m, \end{aligned} \tag{3.1}$$

where $f(x)$ is called the objective function, and the remaining relations are called inequality and equality constraints respectively. The optimal value of the objective function is the optimum solution.

Specially, if the objective function $f(x)$ is convex over a convex domain, and g_i , h_i are affine, the optimization problem is called convex optimization problem.

Given the optimization problem (3.1) over domain $\Omega \in R^n$, the Lagrangian function is defined as

$$L(w, \beta) = f(w) + \sum_{i=1}^k \alpha_i g_i(w) + \sum_{i=1}^m \beta_i h_i(w) \quad (3.2)$$

where the coefficients α_i and β_i are called Lagrange multipliers.

Given an optimization problem (3.1) over the convex domain $\Omega \in \mathbb{R}^n$, $f(x)$ is convex and g_i, h_i are affine (Karush–Kuhn–Tucker conditions). Then the Lagrangian function of the optimization problem is formulated as 3.2. The necessary and sufficient conditions for a normal point w^* to be an optimum are the existence of α^*, β^* which satisfy

$$\begin{aligned} \frac{\partial L(w^*, \alpha^*, \beta^*)}{\partial w} &= 0 \\ \frac{\partial L(w^*, \alpha^*, \beta^*)}{\partial \beta} &= 0 \\ \alpha_i^* g_i(w^*) &= 0 \quad i = 1, \dots, k, \\ g_i(w^*) &\leq 0 \quad i = 1, \dots, k, \\ \alpha_i^* &\geq 0 \quad i = 1, \dots, k, \end{aligned} \quad (3.3)$$

3.1.2 Support Vector Classification (SVC)

The simplest model of SVM is the maximal margin classifier, which forms the strategy of SVM. The goal of SVM is to search for the maximal margin hyperplane in a proper-chosen kernel-induced feature space.

In SVC, searching maximal margin classifier is equivalent to minimizing the norm of the weight vector by setting the functional margin to 1. In detail, if w is the weight vector realizing a functional margin of 1 on the positive point x^+ and the negative point x^- , then a functional margin of 1 implies

$$\begin{aligned}\langle \mathbf{w} \cdot \mathbf{x}^+ \rangle + b &= +1, \\ \langle \mathbf{w} \cdot \mathbf{x}^- \rangle + b &= -1,\end{aligned}\tag{3.4}$$

The geometric margin γ is then the functional margin of the resulting classifier

$$\begin{aligned}\gamma &= \frac{1}{2} \left(\left\langle \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \cdot \mathbf{x}^+ \right\rangle - \left\langle \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \cdot \mathbf{x}^- \right\rangle \right) \\ &= \frac{1}{2\|\mathbf{w}\|_2} (\langle \mathbf{w} \cdot \mathbf{x}^+ \rangle - \langle \mathbf{w} \cdot \mathbf{x}^- \rangle) = \frac{1}{\|\mathbf{w}\|_2}.\end{aligned}\tag{3.5}$$

Hence, the resulting geometric margin equals $1/\|\mathbf{w}\|_2$. Maximizing margin hyperplane γ equals minimizing $\|\mathbf{w}\|$.

Given a linearly separable training sample $S = ((x_1, y_1), \dots, (x_l, y_l))$, the hyperplane $(\mathbf{w}, b)$ that solves the optimization problem is as follows

$$\begin{aligned}\text{minimize} \quad & \frac{1}{2} \langle \mathbf{w} \cdot \mathbf{w} \rangle \\ \text{subject to} \quad & y_i (\langle \mathbf{w}^T \cdot \mathbf{x} \rangle_i + b) \geq 1, \\ & i = 1, \dots, l\end{aligned}\tag{3.6}$$

realizes the minimization of $\|\mathbf{w}\|$. The primal Lagrangian is

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle - \sum_{i=1}^l \alpha_i [\gamma (\langle \mathbf{w}, \mathbf{x} \rangle_i + b) - 1]\tag{3.7}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers. According to Kuhn–Tucker conditions, the hyperplane $(\mathbf{w}, b)$ to realize the minimization of $\|\mathbf{w}\|$ satisfies

$$\begin{aligned}
\frac{\partial L(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} &= \mathbf{0} \\
\frac{\partial L(\mathbf{w}, b, \alpha)}{\partial b} &= 0 \\
\alpha_i [y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1] &= 0 \quad i = 1, \dots, l, \\
y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 &\leq 0 \quad i = 1, \dots, k, \\
\alpha_i &\geq 0 \quad i = 1, \dots, k,
\end{aligned} \tag{3.8}$$

One then brings 3.8 into 3.7 to obtain

$$\begin{aligned}
L(\mathbf{w}, b, \alpha) &= \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle - \sum_{i=1}^l \alpha_i [y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1] \\
&= \frac{1}{2} \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^l \alpha_i \\
&= \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle
\end{aligned} \tag{3.9}$$

Furthermore, the optimal hyperplane can be expressed in the dual representation in terms of this subset of the parameters

$$f(\mathbf{x}, \alpha_i^*, b^*) = \sum_{i=1}^l y_i \alpha_i^* \langle \mathbf{x}_i, \mathbf{x} \rangle + b^* = \sum_{i \in \text{SV}} y_i \alpha_i^* \langle \mathbf{x}_i, \mathbf{x} \rangle + b^*. \tag{3.10}$$

The decision function $f(\mathbf{x})$ can be written as

$$f(\mathbf{x}) = \text{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle + b) = \text{sgn} \left[\sum_{i=1}^l \alpha_i^* y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b^* \right] \tag{3.11}$$

where, $\text{sgn}()$ is a sign function.

In order to find a hyperplane in a high-dimension space, kernel trick is brought into the original optimization function to create nonlinear classifiers (Aizerman et al., 1964). It is used to find maximal margin hyperplane. By the application of kernel trick, SVC usually takes a linear decision function as follows

$$f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + b, \tag{3.12}$$

where $\mathbf{w}$ and b are the model parameters needed to be learned; $\Phi: R^d \rightarrow R^f$ is an

implicit function mapping the original data from R^d to R^f .

The application of a kernel trick allows the algorithm to fit the maximal margin hyperplane in a transformed feature space. The transformation may be nonlinear and the transformed space is in high dimension. Though the classifier is a hyperplane in the highly dimensional feature space, it may be nonlinear in the original input space. Then the primal optimization problem of SVC to maximize the margin width is changed into the following Quadratic Programming (QP) problem:

$$\begin{aligned} & \text{minimise} \quad \frac{1}{2} \|w\|^2 \\ & \text{subject to} \quad y_i (w^T \Phi(x_i) + b) \geq 1 \end{aligned} \quad (3.13)$$

The above formulation requires the training data to be separable, which is too strict in real applications and is likely to overfit the data. Slack variables are introduced for the hinge loss function to measure the classification error and to resolve the issue when the data are not separable. This yields the following QP problem:

$$\begin{aligned} & \text{minimise} \quad \frac{1}{2} \|w\|^2 + C_1 \sum_{i=1}^m \xi_i \\ & \text{subject to} \quad \begin{cases} y_i (w^T \Phi(x_i) + b) \geq 1 - \xi_i, \\ \xi_i \geq 0 \end{cases} \end{aligned} \quad (3.14)$$

where $C_1 \geq 0$ is a non-negative constant used to balance the margin width and errors.

Figure 3.1. Illustration of SVC

In Figure 3.1, individuals are supposed to belong to two different classifications. SVC searches for the maximal margin hyperplane, which is the distance between two dotted lines. The full line is the proper hyperplane and the circled dots are support vectors. This SVC satisfies a set slack variable $\{\xi_i\}$.

3.1.3 Support Vector Regression (SVR)

The Support Vector method can also be applied for regression use, which is called Support Vector Regression (SVR). SVR follows the maximal margin hyperplane strategy, in which a non-linear function is learned through a linear learning machine by introducing a kernel trick. The capacity of the operation system is controlled by a parameter that does not depend on the dimensionality of the space, because in the classification case, the learning algorithm minimizes a convex function and its solution is sparse. They rely on defining a loss function that ignores errors within a certain distance of the true value. This type of function is referred to as an ε -insensitive loss function.

Figure 3.2 gives a description of SVR from a MATLAB-based plot on the same dataset in Figure 3.1, and it shows an example of a one-dimensional linear regression function with an ε -insensitive band. The variable ξ measures the cost of the errors on the training points.

Figure 3.2. Illustration of a one dimensional linear regression

The idea of applying the ε -insensitive loss function has enormous computational advantages as it can guarantee the existence of a global minimum.

Consider a regression function

$$f(x) = \mathbf{w} \cdot x + b \quad (3.15)$$

where $\mathbf{w}, x \in R^N, b \in R$ and $\{(x_1, y_1), \dots, (x_l, y_l)\} \in R^N \times R$. Training patterns are learned by moving $f(x)$ inside ε -insensitive tube. SVR is formulated to balance trade-off among empirical error, the flatness, ε -insensitive tube size, and slack variables.

As in SVC, a parameter C is introduced to measure the trade-off between the complexity and losses. SVR can be defined as follows:

$$\begin{aligned} \text{minimise} \quad & \|\mathbf{w}\|^2 + C \sum_{i=1}^l (\xi_i + \hat{\xi}_i) \\ \text{subject to} \quad & (\langle \mathbf{w} \cdot x_i \rangle + b) - y_i \leq \varepsilon + \xi_i, i = 1, \dots, l, \\ & y_i - (\langle \mathbf{w} \cdot x_i \rangle + b) \leq \varepsilon + \hat{\xi}_i, i = 1, \dots, l, \\ & \xi_i, \hat{\xi}_i \geq 0, i = 1, \dots, l. \end{aligned} \quad (3.16)$$

It can also be solved by Lagrange multipliers and Kuhn–Tucker conditions.

3.2 Model Construction

3.2.1 Bi-Objective Optimization Support Vector Machine (BOOSVM) Model

Next, I will consider a more difficult problem: minimize the classification error and regression error together.

On one hand, in SVC, the optimization function is constructed by finding a proper hyperplane to obtain a maximum-margin between the support vectors with the hyperplane, and the geometric margin γ equals to $1/\langle w \cdot w \rangle$, thus finding the maximum-margin equals to finding the minimum $\langle w \cdot w \rangle$. Generally, SVC searches for a hyperplane (w, b) to realize the minimal $\langle w \cdot w \rangle$, according to which the optimization problem in SVC can be constructed, that is

$$\begin{aligned} \min_{w, b} \quad & \frac{1}{2} \langle w \cdot w \rangle \\ \text{subject to} \quad & y_i (\langle w^T \cdot x \rangle_i + b) \geq 1, \\ & i = 1, \dots, l \end{aligned} \quad (3.17)$$

Then, get the primal Lagrangian

$$L(w, b, \alpha) = \frac{1}{2} \langle w, w \rangle - \sum_{i=1}^l \alpha_i [y_i (\langle w, x \rangle_i + b) - 1] \quad (3.18)$$

where $\alpha_i \geq 0$ are the Lagrange multipliers.

On the other hand, SVR is constructed by finding a linear function to give a best interpolates to a given set S of training points. Geometrically this corresponds to searching a proper hyperplane to fit the given points. A solution to solve the

regression construction problem in SVR is ridge regression (Hoerl and Kennard, 1970), the algorithm of which is as follow:

$$L(w, b) = \lambda \langle w \cdot w \rangle + \sum_{i=1}^l \left(\langle w \cdot x_i \rangle + b - y_i \right)^2, \quad (3.19)$$

in which the parameter λ controls a trade-off between low square loss and low norm of the solution. This algorithm presenting a complex cost function is made of two parts, one controlling the 'complexity' of the hypothesis, and the other controlling its accuracy on the training data. The point is that this algorithm is analogous to the maximal margin algorithm in SVC case. Thus, SVC can be combined with SVR in a uniform framework.

According to the principles of SVC and SVR, the optimization problem in the combination of SVC and SVR can be written as:

$$\begin{aligned} \text{minmise} \quad & \frac{1}{2} \|w\|^2 + C_1 \sum_{i=1}^m \xi_i + C_2 \sum_{i=1}^m (\eta_i + \zeta_i) \\ \text{subject to} \quad & \begin{cases} y_i (w^T \Phi(x_i) + b) \geq 1 - \xi_i \\ (w^T \Phi(x_i) + b) - z_i \leq \varepsilon + \eta_i \\ z_i - (w^T \Phi(x_i) + b) \leq \varepsilon + \zeta_i \\ \xi_i \geq 0 \\ \eta_i \geq 0 \\ \zeta_i \geq 0 \end{cases} \end{aligned} \quad (3.20)$$

The optimization problem 3.20 is used to find a common hyperplane

$$f(x) = \langle w \cdot x \rangle + b$$

for both SVC and SVR, where the w is shared by these two SVM algorithms. In order to find maximum-margin hyperplane in SVC and minimize the 'complexity' of

the hypothesis in SVR, the common $\|w\|$ has to be minimized in 3.20. In addition, slack variables $C_1 \sum_{i=1}^m \xi_i$ need to be minimized to optimize SVCs' performance, and $C_2 \sum_{i=1}^m (\eta_i + \zeta_i)$ are used to find a low square loss in SVR to improve the accuracy on the training data. In order to solve the optimization problem in (3.20), I introduce the corresponding Lagrange multipliers, that is

$$\begin{aligned}
L(w, b, \xi, \eta, \zeta, \alpha, \beta, \gamma, \mu^\xi, \mu^\eta, \mu^\zeta) = & \frac{1}{2} \|w\|^2 + C_1 \sum_{i=1}^m \xi_i + C_2 \sum_{i=1}^m (\eta_i + \zeta_i) \\
& - \sum_{i=1}^m \alpha_i \left[y_i (w^T \Phi(x_i) + b) + \xi_i - 1 \right] \\
& - \sum_{i=1}^m \beta_i \left[\varepsilon + \eta_i + z_i - (w^T \Phi(x_i) + b) \right] \\
& - \sum_{i=1}^m \gamma_i \left[\varepsilon + \zeta_i - z_i + (w^T \Phi(x_i) + b) \right] \\
& - \sum_{i=1}^m (\mu_i^\xi \xi_i + \mu_i^\eta \eta_i + \mu_i^\zeta \zeta_i)
\end{aligned} \tag{3.21}$$

where $\alpha, \beta, \gamma, \mu^\xi, \mu^\eta, \mu^\zeta \in \mathbf{R}_+^m$ are the Lagrange multipliers. According to Kuhn-Tucker conditions, the corresponding dual is achieved by differentiating with respect to w, b, ξ, η and ζ , obtaining

$$\begin{cases} \frac{\partial L}{\partial w} = w - \sum_{i=1}^m (\alpha_i y_i - \beta_i + \gamma_i) \Phi(x_i) = 0 \\ \frac{\partial L}{\partial b} = -\sum_{i=1}^m \alpha_i y_i + \sum_{i=1}^m \beta_i - \sum_{i=1}^m \gamma_i = 0 \\ \frac{\partial L}{\partial \xi_i} = C_1 - \alpha_i - \mu_i^\xi = 0 \\ \frac{\partial L}{\partial \eta_i} = C_2 - \beta_i - \mu_i^\eta = 0 \\ \frac{\partial L}{\partial \zeta_i} = C_2 - \gamma_i - \mu_i^\eta = 0 \end{cases}, \tag{3.22}$$

which are transformed into

$$\left\{ \begin{array}{l} \mathbf{w} = \sum_{i=1}^m (\alpha_i y_i - \beta_i + \gamma_i) \Phi(x) \\ \sum_{i=1}^m \alpha_i y_i = \sum_{i=1}^m \beta_i - \sum_{i=1}^m \gamma_i \\ \mu_i^\varepsilon = C_1 - \alpha_i \\ \mu_i^\eta = C_2 - \beta_i \\ \mu_i^\eta = C_2 - \gamma_i \end{array} \right. . \quad (3.23)$$

Since $\mu_i^\varepsilon, \mu_i^\varepsilon, \mu_i^\eta \geq 0$, thus I have

$$\alpha_i \in [0, C_1], \quad \beta_i \in [0, C_2], \quad \gamma_i \in [0, C_2].$$

By plugging them into the Lagrange formulation (3.21), I obtain the dual form of (3.20) as follows:

$$\begin{aligned} \underset{\alpha, \beta, \gamma}{\text{maximise}} \quad & W = -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m (\alpha_i y_i - \beta_i + \gamma_i) (\alpha_j y_j - \beta_j + \gamma_j) k(x_i, x_j) \\ & + \sum_{i=1}^m \alpha_i - \sum_{i=1}^m (\varepsilon + z_i) \beta_i - \sum_{i=1}^m (\varepsilon - z_i) \gamma_i \\ \text{subject to} \quad & \left\{ \begin{array}{l} \sum_{i=1}^m \alpha_i y_i = \sum_{i=1}^m \beta_i - \sum_{i=1}^m \gamma_i \\ \alpha_i \in [0, C_1] \\ \beta_i \in [0, C_2] \\ \gamma_i \in [0, C_2] \end{array} \right. \end{aligned} \quad (3.24)$$

In (3.24), I set $\beta_i - \gamma_i = \theta_i$ to obtain

$$\begin{aligned} \underset{\alpha, \theta}{\text{maximize}} \quad & W = -\frac{1}{2} \sum_{i,j=1}^m (\alpha_i y_i - \theta_i) (\alpha_j y_j - \theta_j) K(x_i, x_j) + \sum_{i=1}^m \alpha_i - \varepsilon \sum_{i=1}^m |\theta_i| + \sum_{i=1}^m z_i \theta_i \\ \text{subject to} \quad & \left\{ \begin{array}{l} \sum_{i=1}^m \theta_i = \sum_{i=1}^m \alpha_i y_i \\ \alpha_i \in [0, C_1] \\ \theta_i \in [-C_2, C_2] \end{array} \right. \end{aligned} \quad (3.25)$$

where $K(\cdot, \cdot)$ is the RBF-kernel. The above formula is the optimization problem which

can be used for classification and regression simultaneously, and it is the optimization problem for a model called BOOSVM model. The decision function $f(x)$ can be written as

$$f(x) = \text{sgn} \left(w \cdot x + b \right) = \text{sgn} \left[\sum_{i=1}^m (\alpha_i y_i \theta) \Phi(x) \right] \quad (3.26)$$

where, $\text{sgn}()$ is a sign function.

The mechanism of BOOSVM model can be described in Figure 3.3. In Figure 3.3, a common hyperplane can be used for classifying all the binary-classification dots, as well as fitting all the dots. The solution of BOOSVM model is to find such a proper hyperplane, which is determined by w and b .

Figure 3.3. Illustration of BOOSVM model

3.2.2 Gaussian RBF Model

The original optimal hyperplane algorithm suggested by Vapnik (1995) is a linear classifier. Boser et al. (1992) propose a way to create nonlinear classifiers by applying the kernel trick to search maximum margin hyperplanes. In their research, the proposed algorithm is formally similar to previous classifiers, except that every dot product is replaced by a nonlinear kernel function. Kernel trick allows the

algorithm to fit the maximum margin hyperplane in a transformed feature space. The transformation may be nonlinear and the transformed space can be high-dimensional, implying that it may be nonlinear in the original input space, although the classifier is in the high-dimensional feature space. Kernels trick can be regarded as generalized dot products. Linear kernel can only handle the case in which the relationship between the class labels and attributes is linear. Some common nonlinear kernels include: Polynomial (Homogeneous and inhomogeneous), Gaussian radial basis function (RBF kernel), and Hyperbolic tangent. Among them, RBF kernel is the most popular kernel function used in SVM (Chang, 2010).

SVM classifier uses inner product as the metric. If there is any dependent relationship among pattern's attributes, such relationship can be realized by a mapping (Scholkopf and Smola, 2002) $\Phi : X \rightarrow H; x \rightarrow \Phi(x)$, where H is called feature space. The inner product similarity is computed through $\langle \Phi(x), \Phi(y) \rangle$ for pattern x and y . In the SVM literature, the above course is realized through kernel function

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle. \quad (3.19)$$

A Gaussian RBF kernel is formulated as

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \quad (3.20)$$

and rewritten in the new form of

$$k(x, y) = \exp(-\gamma \|x - y\|^2), \quad (3.21)$$

where the variation of parameter γ can change the width of RBF kernel in SVM.

SVM with Gaussian RBF kernel is sensitive to the kernel width. Small kernel widths may cause over-fitting, and large kernel width may cause under-fitting. The so-called optimal kernel width is merely selected based on the tradeoff between under-fitting loss and over-fitting loss. So, it is necessary to adjust the parameter λ to eliminate the inconsistency of coexisting over-fitting and under-fitting in SVM learning.

3.2.3 Two-Stage BOOSVM Model

The data in direct marketing area have some special characters. For one aspect, direct marketing data have many more non-responders than responders, where non-responders contribute no profit and their profit equals to zero. In my research, when predict profit with SVR, a regression function is learnt to describe the possible relationships between dependent variables and independent variables, and geometrically, regression function is an overall simulation of all the individuals. If the profit regression model is built based on all the customers, the model will be greatly influenced by the overwhelming numbers of non-responders, and the learnt regression function are prone to including the majority of non-responders in geometrics. However, forecasting the profit of non-responders is meaningless since their profit is a given value. Thus, the response rate prediction is based on the classification approach, while the profit forecasting based on regression uses different groups of customers to learn patterns.

In my proposed BOOSVM model, all the customers in the testing set are used to predict response rate and profit simultaneously. As for this problem, I propose a

two-stage method to improve the prediction accuracy of the BOOSVM model and enhance the profit prediction. Since non-responders have no profit or negative profit and only responders have positive profit, one may say that profit is dependent on response. Therefore, it is quite logical and appealing to adopt the two-stage method. The so-called two-stage method is easy to understand in that the final results are obtained from two stages, and the basic framework can be declared by the following figure. In the first stage, I learn a first BOOSVM model to produce a score to each customer in test set. In the second stage, by adjusting parameters in BOOSVM model, I strengthen the regression function in BOOSVM, and use only the responders in the train set to train the second model. The learned second model is only used to generate another score for individuals who have positive scores in the first stage. The result from multiplying the two scores is taken as the final score for a certain customer, and only the responders are given such updates to their first scores. Then we can rank the customers in the testing set in a descending order according to their final scores, and then use decile analysis to calculate the AARL and AAPL.

The design of two-stage method attributes to the necessity of using different groups of customers to predict response rate and profit. Previous literature also provides support for this. For example, Cui et al. (2012) only use positive cases in the training process with profit as the dependent variable to predict profit lift in their comparison studies. In the research of Kim et al. (2006), since a classification model is not their research interest, when they apply SVR for response modeling to predict total amount of money spent of each respondent, they select a new dataset consisting of only the respondents, which is only a small part of all the customers in dataset. In addition, many multi-objective optimization models use a cutoff point to reduce the influence of non-responders on profit prediction. In the MOEA method proposed by Wong and Cui (2010), the cutoff point method is used to include only responders for profit prediction. In the second step of model proposed by Bult (1995), the firm's profit is maximized by determining an optimal cutoff point.

3.2.4 Summary of Model Structure

To deal with the problem of optimization to both response rate and profit in direct marketing, I propose a two-stage BOOSVM model. The two-stage BOOSVM is built by combining both Support Vector Classification and Support Vector Regression based on SVM theory. According to SVM principle, this model is a convex, thus I propose to use CVX, an effective modeling system, to construct and solve this problem. The following content is an introduction to the algorithm of

realizing two-stage BOOSVM and the software used is MATALB.

Chapter 4. Algorithm and Data

4.1 Algorithm

4.1.1 CVX

CVX is an effective modeling system to construct and solve disciplined convex programs. CVX has wide applications of many standard problems, including linear and quadratic programs, second-order cone programs, and semi-definite programs. Its formulation to solve convex norm programs is feasible and easy to handle. The modeling system of CVX is MATLAB-based. With the help of CVX, convex optimization problems can be easily solved with standard MATLAB expressions.

Special requirements for the content and format of the CVX algorithm is as follow: all variables in constraints or objective functions must be declared using the specified form of variable commands beforehand; the objective minimization function must be convex (the maximization objective function must be concave); operators permitted in CVX constraints are limited, and permitted ones are "==" equality constraints, "<=" less-than constraints and ">=" greater-than constraints.

Generally, the standard format of CVX is described in following table:

Table 4.1. Standard format of CVX

```
cvx_setup;  
cvx_begin  
    variables;  
    maximize (optimization function);  
    subject to    Equality Constraints;  
                 Inequality Constraints;  
cvx_end
```

The optimization problem in SVM model is a convex optimization problem (Bennett and Campbell, 2000), and BOOSVM is SVM-based. Thus, CVX is an ideal choice to solve the optimization problem in BOOSVM and it is applied by an insert function form in the whole MATLAB code.

4.1.2 Grid Search Algorithm

In data mining model construction, parameter selection processes cannot be avoided. A practical and standard way to perform parameter optimization is grid search, which is guided by some performance metrics, for example, cross-validation (Hsu, 2010). In a grid search algorithm, parameters should be set bounds and conducted discretization manually in advance according to machine learners' parameter value spaces.

In a RBF kernel SVM, grid search is a reasonable first choice for parameter selection (Hsu et al., 2010). In a RBF kernel equipped SVM, two types of parameters, the regularization parameter C and the kernel parameter γ , should be selected. The exponentially growing sequences

$$\begin{aligned} C_i &\in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\} \\ \gamma &\in \{0.001, 0.01, 0.1, 1, 10, 100\} \end{aligned} \quad (4.2)$$

are applied in grid search algorithm for parameter selection in BOOSVM. In this grid search algorithm, the BOOSVM model which is equipped with different cross-product parameters is applied to the training sets to obtain results respectively. Different pairs of C and γ can be used to obtain different values of AARL and AAPL. According to the Pareto efficiency principle, the AARL and AAPL pairs that satisfy Pareto efficiency are retained. Next, comparing the pairs of AARL and AAPL

which are in the Pareto frontier, the AARL and AAPL pair with the maximal AAPL is selected as the best solution. Meanwhile, the C and γ pair corresponding to the best solution will be chosen for the next step.

4.1.3 Algorithm Framework

Table 4.2 outlines the algorithm framework of two-stage BOOSVM model:

Table 4.2. Algorithm Framework

The overall algorithm

Input: The learning set

Output: Scores for individuals in testing set, AARL and AAPL

Step 1: SMOTE Processing;

Step 2: Parameters Generating;

Step 3: Model Building;

Step 4: Scoring;

Step 5: Two-Stage BOOSVM;

Step 6: Model Testing

Step 1: SMOTE Processing

The over-sampling amount I used in SMOTE is 600%. With this step the amount of response individuals is six times as many as the original.

Step 2: Parameters Generating

The parameters need to be generated include C_1, C_2 and γ , where C_1 and C_2 are parameters in the optimization constraint, and γ is used in kernel function. Specifically, when C_1 equals zero, and the α in the optimization function equals to zero, and then BOOSVM model can be transferred into a SVR-used model. Similarly, when C_2 equals zero, BOOSVM model is equivalent to a SVC-used model. So, the magnitude of C_1 and C_2 determines the application degree of classification and regression in the BOOSVM model.

Different (C_1, C_2, γ) combinations are used in form of cross-product of the three sets

$$C_1 \in \{-1000, -100, -10, 0, 10, 100, 1000\}; C_2 \in \{-1000, -100, -10, 0, 10, 100, 1000\}; \\ \gamma \in \{-100, -10, 0, 10, 100\},$$

and are evaluated in terms of performance by grid search.

Step 3: Model Building

I use the CVX method to solve the dual optimization problem and get proper α^* and θ^* in 3.22. Then BOOSVM model-1 can be obtained.

Step 4: Scoring

Bring α^* and θ^* into 3.23 to get $f^*(x)$, which is a scoring function. Then use $f^*(x)$ to give a score to each customer in testing set.

Step 5: Two-Stage BOOSVM

By setting $C1=0$, use grid search to find a proper pair $(C2, \gamma)$. Then CVX learns a model on responders in training set to get a BOOSVM model-2 and use the BOOSVM model-2 to give another score to each customer who is predicted as a responder after step 3. The result from multiplying the two scores is taken as the final score to the certain customer.

Step 6: Model Testing

The customers are ranked according to the scores in a descending order. Decile analysis is performed according to the descending-order scores to generate the AARL and AAPL on 10 different depth-of-files. The algorithm of this thesis is explained by the following figure.

4.2 Data Description

Selection of target customers is the core activity in direct marketing. There are various sophisticated techniques that can help marketers formulate effective direct marketing strategies. Practical budget for promotions is always limited, thus only a fraction of the customers can be selected. Decision makers do not deal with a single fixed objective. In addition to striving for a higher response rate, decision makers always pursue higher profit.

Support Vector Machine (SVM) technique is valued in data mining areas for its complete and systematic theory. In my thesis, I adopt a two-stage BOOSVM model to learn patterns on SMOTE-processed dataset (two-stage BOOSVM (SMOTE)) to deal with the bi-objectives optimization problem. The two objectives are response rate and profit. I use two datasets for testing two-stage BOOSVM (SMOTE) model's performance and compare its results with two benchmark techniques, which are logistic regression and parallel Multi-Objective Evolutionary Algorithm (MOEA).

4.2.1 Dataset One

The first dataset (DM dataset) comes from a U.S.-based catalog direct marketing company that sells multiple product lines of general merchandise ranging from gifts and apparel to consumer electronics. The DM dataset contains information of 106,284 consumers selected from a perennial promotion. The dataset only uses the most recent version, which lasts for six months and achieves a 5.4% response rate with 5,740 responders. Seventeen independent variables are finally selected by

applying the forward selection criterion, which is an automatic procedure for carrying out predictive variable selection. In addition, DM dataset contains only one dependent variable (profit), and the variable can also be transformed to represent customer response. If the sign of the profit of a certain customer is positive, the customer is a responder, otherwise, the customer is a non-responder. These seventeen independent variables and the one dependent variable are summarized in the Appendix A.

4.2.2 Dataset Two

The second dataset is KDD-98 dataset. This dataset is a well-studied dataset firstly used in the 1998 Knowledge Discovery and Data mining competition. The competition task in 1998 is a regression problem whose goal is to estimate the return from a direct mailing in order to maximize donation. I select the KDD-98 dataset as the second dataset since it is related to direct marketing and widely used in data mining research.

The KDD-98 dataset contains 191,779 records from the purchase contact history of the last year's mailing campaign, which is a charitable organization sends out regular mailings to its target customers to solicit donations. Similar to the KDD-98 dataset used in the research by Cui et al. (2012), I used the 44 independent variables suggested in SAS Enterprise Miner Tutorial for model building. These variables and the two dependent variables are summarized in the Appendix B. The dataset has been pre-split in a fixed form, with 95,412 records for training set, and

96,367 records in the testing set. The KDD-98 dataset is special for its feature that there is an inverse correlation between customers' probabilities to purchase and the dollar amount to spend. This feature may result from the practical case that customers will be more cautious to make a purchase decision when they intend to spend more money. In this sense, a model obtaining a higher response rate on KDD-98 dataset may not lead to greater profit.

4.3 Dataset Process

4.3.1 Synthetic Minority Over-sampling Technique (SMOTE)

Real-world datasets are often predominately composed of “normal” examples with only a small percentage of “abnormal” or “interesting” examples (Chawla et al., 2002). This means that imbalanced datasets, the negative instances far outnumbering the positive instances, are the common cases in real classification datasets, such as the datasets in gene profiling, medical diagnosis, and credit card fraud detection. Even in some circumstances, the ratio of positive to negative instances can be smaller than 0.01.

SVMs obtain effective applications in many areas; however the remarkable performance of SVMs is negatively affected by imbalanced datasets. The phenomenon can be explained by the Occam's razor principle in which classifiers are designed to generalize sample data and to output the simplest hypothesis performing best to fit the data. Thus, in general cases, classifiers will perform much poorer on the imbalanced dataset

A successful and effective technique, SMOTE, is good at overcoming the difficulties of data mining approaches in analyzing imbalanced data. SMOTE is inspired by an application in handwritten character recognition (Ha & Bunke, 1997) and it is also particularly relevant to the study of Ling and Li (1998), in which under-sampling of the majority class is combined with over-sampling of the minority class. Over-sampling the minority class is processed by creating “synthetic” examples: for each sample in the minority class, its nearest k -minority (k is a parameter) class neighbors are added to it. Along the line segments between the sample and its neighbors, synthetic examples are introduced by multiplying the difference between the sample and its nearest neighbor by a random number between 0 and 1. Then depending upon the amount of over-sampling required, neighbors from the k -nearest neighbors can be randomly chosen.

Data of direct marketing are usually very imbalanced as the responders account for only a small part of non-responders, thus the SVM-based BOOSVM model may perform badly on direct marketing dataset. To solve the problem, I propose to process the data by SMOTE method and make the training set more balanced. The class of responders is the minority class and it is over-sampled by introducing k synthetic examples along the line segments of each sample between its k nearest neighbors. k is the over-sampling amount, the selection of k can influence the results (AARL and AAPL). I try 9 different values of k ($k = 200\%, \dots, 1000\%$) on subsets of DM datasets, and get 9 different pairs of AARL and AAPL. Then based on the trial and error, the obtained AAPL and AARL are both the highest when the

over-sampling amount (k) equals to 600%. Thus, I select 600% as the over-sampling amounts in my thesis.

4.3.2 Cross-validation

Cross-validation is a standard technique used for minimizing the over-fitting problem (Hsu et al., 2010). In cross-validation, the training data set is divided into k subsets of equal size. Then each subset is used as a testing set to evaluate the performance of the model trained on the remaining ($k-1$) subsets. The process comes to an end when each instance of the whole training set has been used once as test set and the cross-validation result is the average of all runs.

Figure 4.1. Illustration of 10-fold cross-validation

In this thesis, I use 10-fold cross-validation in DM dataset to avoid the

over-fitting problem. In 10-fold cross-validation, the original sample is randomly partitioned into 10 equal size subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are combined and used as training data. Such process is then repeated 10 times, with each of the 10 subsamples used exactly once as the testing data. The results from the 10 folds then can be averaged to produce a single estimation. The framework of 10-fold cross-validation is explained by Figure 4.1.

Chapter 5. Results and Comparisons

5.1 Experiments for Testing the Proposed Model

In order to make the prediction more accurate, I use two methods to assist BOOSVM model, which are SMOTE and two-stage method. In order to prove the merits of these methods, I conduct four experiments as described in Figure 5.1. First, BOOSVM is a combined method of SVC and SVR. To prove that BOOSVM performs better than SVC and SVR in selecting customers, I compare the results of BOOSVM, SVC and SVR. Second, to test the necessity of SMOTE method for the BOOSVM model, I use the original training set and SMOTE-processed training set respectively as the training set for BOOSVM model. Third, to prove the merit of the two-stage method, I compare the results of BOOSVM model using the two-stage method with those of the BOOSVM model without the two-stage method. Fourth, to verify the effectiveness of the SMOTE method, I compare the proposed model with logistic regression and MOEA models. The datasets used in four experiments are from the DM dataset.

5.1.1 BOOSVM, SVC and SVM

In my study, BOOSVM is a SVM-based model, combining SVC and SVR to search optimization of response rate and profit. To prove that BOOSVM can perform better in customer selection than original SVC and SVR, results (AARL and AAPL)

of BOOSVM, SVC, and SVR are compared. A 10-fold cross-validation using the DM dataset is performed in this study. The results are shown in Table 5.1.

Table 5.1. AARL and AAPL of BOOSVM, SVC and SVR on original DM dataset.

Model /decile	AARL of SVC	AARL of SVR	AARL of BOOSVM	AAPL of SVC	AAPL of SVR	AAPL of BOOSVM
1	301.7(36.1)	190.3(19.7)	289.1(13.6)	358.8(29.1)	403.2(20.6)	417.2(24.7)
2	260.8(25.3)	178.7(19.0)	261.3(9.4)	301.0(12.0)	310.5(15.2)	305.9(13.6)
3	215.3(16.3)	163.8(16.9)	204.5(5.7)	232.1(11.3)	251.5(12.4)	243.5(10.0)
4	177.6(10.2)	152.1(14.8)	180.2(3.6)	199.3(8.2)	226.8(8.8)	212.7(6.7)
5	142.5(8.7)	134.9(6.5)	150.8(3.5)	183.9(7.3)	186.6(7.9)	183.4(4.1)
6	119.6(5.9)	125.2(5.5)	120.1(2.2)	152.7(4.8)	170.4(4.3)	160.4(4.7)
7	116.2(5.6)	120.1(2.5)	118.5(4.0)	134.8(3.4)	153.1(5.1)	143.3(2.4)
8	109.5(2.3)	111.4(2.2)	114.2(1.8)	120.3(3.5)	126.8(2.8)	132.9(1.6)
9	105.2(1.4)	105.8(1.7)	103.5(1.2)	108.6(1.9)	112.1(1.6)	115.3(1.1)
10	100.0	100.0	100.0	100.0	100.0	100.0

Table 5.1 (continued) T-test results of AARL differences

Model /Decile	AARL of SVC	AARL of BOOSVM	Difference	T-value (P-value)
1	301.7 (36.1)	289.1 (13.6)	-12.6	1.03 (0.34)
2	260.8 (25.3)	261.3 (9.4)	0.5	0.06 (0.95)
Model /Decile	AARL of SVR	AARL of BOOSVM	Difference	T-value (P-value)
1	190.3 (19.7)	289.1 (13.6)	98.8***	13.05 (0.00)
2	178.7 (19.0)	261.3 (9.4)	82.6***	12.32 (0.00)

Table 5.1(continued) T-test results of AAPL differences

Model /Decile	AAPL of SVC	AAPL of BOOSVM	Difference	T-value (P-value)
1	358.8 (29.1)	417.2 (24.7)	58.4***	4.84 (0.00)
2	301.0 (12.0)	305.9(13.6)	4.9	0.85 (0.41)
Model/ Decile	AAPL of SVR	AAPL of BOOSVM	Difference	T-value (P-value)
1	403.2 (20.6)	417.2 (24.7)	14.0	1.38 (0.19)
2	310.5 (15.2)	305.9 (13.6)	-4.6	1.40 (0.18)

*, **, *** indicate the 10%, 5%, and 1% significance levels, respectively

Comparing the results of SVC and BOOSVM, I find the AARLs of BOOSVM (289.1 and 261.3) are roughly equal to those of SVC (301.7 and 260.8) in the top two deciles and the t-tests (t-value: 1.03 and 0.06) show no significant difference between them. But the AAPL of BOOSVM (417.2) is significantly (t-value: 4.84) higher than that of SVC (358.8) in the first decile. The comparison suggests that BOOSVM model is better than SVC model. When comparing the results of SVR and BOOSVM, the t-tests (t-value: 1.38 and 1.40) show that the AAPLs of BOOSVM (417.2 and 305.9) are at a comparable level with those of SVR (403.2 and 310.5) in the top two deciles, while the AARL of BOOSVM (289.1) is significantly (t-value: 13.05) higher than that of SVR (190.3) in the first decile. As a result, BOOSVM is also better than SVR in selecting customers.

5.1.2 Two-Stage BOOSVM with and without SMOTE Comparison

In my thesis, I proposed a two-stage BOOSVM method to seek the optimization of both response rate and profit. There are two datasets which are used in this study, namely the DM dataset and the KDD98 dataset. The two data sets are both quite unbalanced, and the ratios of responders in all the customers (5.4% and 5.1%) are both quite small. Because SVM is weak at dealing with unbalance dataset, the SMOTE method is used to deal with this problem. To test the necessity of the SMOTE method in two-stage BOOSVM model, I use the original training set and the SMOTE-processed training set respectively as the training set for BOOSVM model. 10-fold cross-validation DM dataset is used in this study. For the SMOTE-processed

model, SMOTE is only used for the training dataset in each validation. The results are showed in Table 5.2.

Table 5.2. AARL and AAPL of two-stage BOOSVM on original DM dataset and two-stage BOOSVM on SMOTE-processed DM dataset.

Model /deciles	AARL of two-stage BOOSVM (original)	AARL of two-stage BOOSVM (SMOTE)	AAPL of two-stage BOOSVM (original)	AAPL of two-stage BOOSVM (SMOTE)
1	289.1 (13.6)	352.2 (16.1)	417.2 (24.7)	628.2 (32.1)
2	261.3 (9.4)	270.8 (10.9)	305.9 (13.6)	403.4 (16.8)
3	204.5 (5.7)	220.4 (6.3)	143.5 (10.0)	276.5 (9.3)
4	180.2 (3.6)	180.3 (4.2)	212.7 (6.7)	214.4 (8.8)
5	150.8 (3.5)	154.0 (3.1)	183.4 (4.1)	180.5 (5.5)
6	120.1 (2.8)	130.7 (1.6)	160.4 (4.7)	149.7 (6.1)
7	118.5 (4.0)	114.5 (2.2)	143.3 (2.4)	131.2 (4.9)
8	114.2 (1.8)	110.2 (1.5)	132.9 (1.6)	114.2 (4.7)
9	103.5 (1.2)	104.7 (0.9)	115.3 (1.1)	108.9 (1.9)
10	100.0	100.0	100.0	100.0

Table 5.2 (continued). T-test results of the differences (original data and SMOTE-processed data)

Model/Decile	AARL of two-stage BOOSVM (original)	AARL of two-stage BOOSVM (SMOTE)	Difference	T-value (P-value)
1	289.1 (13.6)	352.2 (16.1)	63.1***	9.47 (0.00)
2	261.3 (9.4)	270.8 (10.9)	9.5*	2.09 (0.06)
Model/Decile	AAPL of two-stage BOOSVM (original)	AAPL of two-stage BOOSVM (SMOTE)	Difference	T-value (P-value)
1	417.2 (24.7)	628.2 (32.1)	211***	16.47 (0.00)
2	305.9 (13.6)	403.4 (16.8)	97.5***	14.26 (0.00)

*, **, *** indicate the 10%, 5%, and 1% significance levels, respectively

The above table indicates that the model obtained from SMOTE-processed training set performs much better than the model learned from original dataset. In the top 20% deciles, the AARLs of two-stage BOOSVM (SMOTE) (352.2 and 270.8) are significantly (t-value: 9.47 and 2.07) higher than the AARLs of the two-stage BOOSVM (original) (289.1 and 261.3). The AAPLs of two-stage BOOSVM (SMOTE) (628.2 and 403.4) are also significantly (t-value: 16.47 and 14.26) bigger than the values for AAPL of two-stage BOOSVM (original) (417.2 and 305.9). So far, it has been proven that the SMOTE method can help to improve the performance of BOOSVM model in terms of model training on imbalanced data.

5.1.3 One-Stage BOOSVM and Two- Stage SMOTE Comparisons

To prove that the proposed two-stage BOOSVM model can contribute to the prediction of high-profit customers in order to increase the profit lift, I make the following comparisons between BOOSVM model and the two-stage BOOSVM model. I use 10-fold cross-validation for the DM dataset in this study. The training sets used in this experiment are all processed by SMOTE and SMOTE is only used for the training dataset in each validation. The comparisons results are showed in Table 5.3.

Table 5.3. AARL and AAPL of BOOSVM on SMOTE-processed DM dataset and two-stage BOOSVM on SMOTE-processed DM dataset

Model /deciles	AARL of BOOSVM (SMOTE)	AARL of two-stage BOOSVM (SMOTE)	AAPL of BOOSVM (SMOTE)	AAPL of two-stage BOOSVM (SMOTE)
1	354.6 (15.6)	352.2 (16.1)	601.7 (29.8)	628.2 (32.1)
2	271.2 (8.3)	270.8 (10.9)	389.2 (18.4)	403.4 (16.8)
3	219.7 (5.9)	220.4 (6.3)	274.8 (10.1)	276.5 (9.3)
4	180.3 (4.2)	180.3 (4.2)	214.4 (8.7)	214.4 (8.8)
5	154.0 (4.5)	154.0 (3.1)	180.5 (8.2)	180.5 (5.5)
6	130.7 (2.1)	130.7 (1.6)	149.7 (5.0)	149.7 (6.1)
7	114.5 (1.2)	114.5 (2.2)	131.2 (6.2)	131.2 (4.9)
8	110.2 (1.0)	110.2 (1.5)	114.2 (1.5)	114.2 (4.7)
9	104.7 (0.3)	104.7 (0.9)	108.9 (1.1)	108.9 (1.9)
10	100.0	100.0	100.0	100.0

Table 5.3 (continued). T-test results of the differences (BOOSVM and two-stage BOOSVM)

Model /Decile	AARL of BOOSVM (SMOTE)	AARL of two-stage BOOSVM (SMOTE)	Difference	T-value (P-value)
1	354.6 (15.6)	352.2 (16.1)	-2.4	0.34 (0.73)
2	271.2 (8.3)	270.8 (10.9)	-0.4	0.09 (0.92)
Model /Decile	AAPL of BOOSVM (SMOTE)	AAPL of two-stage BOOSVM (SMOTE)	Difference	T-value (P-value)
1	601.7 (29.8)	628.2 (32.1)	16.5*	1.99 (0.07)
2	389.2 (18.4)	403.4 (16.8)	14.2*	1.80 (0.10)

*, **, *** indicate the 10%, 5%, and 1% significance levels, respectively

The AARLs (352.2, 270.8 and 220.4) and AAPLs (628.2, 403.4 and 276.5) are only changed in the top three deciles, from AARLs (354.6, 271.2 and 219.7) and AAPLs (601.7, 389.2.4 and 274.8) respectively. It is because that 20% to 30% percent of customers are forecasted as responders in the first stage BOOSVM and only these customers are used the two-stage method to give a profit prediction. Between the BOOSVM and the two-stage BOOSVM using SMOTE, there are no significant (t-value: 0.34 and 0.09) differences in their AARLs (352.2 and 270.8 vs. 354.6 and 271.2). However, the AAPLs of the two-stage BOOSVM are significantly (t-value: 0.07 and 0.10) higher than those of BOOSVM (628.2 and 403.4 vs. 601.7 and 389.2). The results indicate the superiority of the of two-stage method.

5.1.4 Results of Models Comparisons: DM Dataset

BOOSVM model constructs a common hyperplane for SVC and SVR to optimize response rate lift and profit lift with a single model. I compare my method with two other methods, which are the logistic regression and multi-optimization algorithm (MOEA). Logistic regression is used as a baseline comparison model. It is a discrete selection model, and the model solution speed is fast and convenient for probability expression of its significant characteristics. In addition, another competing method, MOEA, has been proved excellent in customer selection area in the literature (Wong and Cui, 2010).

In this study, I use SMOTE method to assist all the three models. A 10-fold cross-validation using the DM dataset is used in this study. For SMOTE-processed

model, SMOTE is only used for the training dataset in each validation. Results are showed in Table 5.4 and Table 5.5.

Table 5.4. AARL of logistic regression with SMOTE-processed DM dataset, MOEA with SMOTE-processed DM dataset and two-stage BOOSVM with SMOTE-processed DM dataset

Model/ Decile	AARL of logistic regression (SMOTE)	AARL of MOEA (SMOTE)	AARL of two-stage BOOSVM (SMOTE)
1	365.9 (23.1)	284.6 (53.2)	352.2 (16.1)
2	254.4 (11.9)	213.5 (29.2)	270.8 (10.9)
3	210.0 (7.1)	178.4 (21.1)	220.4 (6.3)
4	183.6 (3.7)	155.6 (14.7)	180.3 (4.2)
5	160.6 (3.7)	140.4 (12.1)	154.0 (3.1)
6	144.6 (2.5)	128.4 (7.5)	130.7 (1.6)
7	128.9 (1.5)	118.0 (5.0)	114.5 (2.2)
8	117.1 (1.1)	110.5 (2.6)	110.2 (1.5)
9	109.7 (0.8)	105.3 (1.4)	104.7 (0.9)
10	100.0	100.0	100.0

Table 5.4 (continued). T-test results of the differences

Model/Decile	AARL of logistic regression (SMOTE)	AARL of two-stage BOOSVM (SMOTE)	Difference	T-value (P-value)
1	365.9 (23.1)	352.2 (16.1)	-13.7	1.54 (0.15)
2	254.4 (11.9)	270.8 (10.9)	16.4***	3.21 (0.01)
Model/Decile	AARL of MOEA (SMOTE)	AARL of two-stage BOOSVM (SMOTE)	Difference	T-value (P-value)
1	284.6 (53.2)	352.2 (16.1)	67.6***	3.99 (0.00)
2	213.5 (29.2)	270.8 (10.9)	57.3***	5.45 (0.00)

*, **, *** indicate the 10%, 5%, and 1% significance levels, respectively

Table 5.5. AAPL of logistic regression with SMOTE-processed DM dataset, MOEA with SMOTE-processed DM dataset and two-stage BOOSVM with SMOTE-processed DM dataset

Model/Decile	AAPL of logistic regression (SMOTE)	AAPL of MOEA (SMOTE)	AAPL of two-stage BOOSVM (SMOTE)
1	594.3 (40.8)	498.8 (91.9)	628.2 (32.1)
2	365.7 (20.9)	321.5 (43.0)	403.4 (16.8)
3	262.3 (12.3)	243.5 (26.7)	276.5 (9.3)
4	215.1 (7.1)	199.2 (16.5)	214.4 (8.8)
5	176.9 (5.4)	168.4 (14.1)	180.5 (5.5)
6	142.2 (4.8)	147.2 (10.1)	149.7 (6.1)
7	124.2 (3.8)	129.6 (7.7)	131.2 (4.9)
8	112.8 (2.1)	117.3 (3.6)	114.2 (4.7)
9	115.9 (1.5)	108.2 (2.0)	108.9 (1.9)
10	100.0	100.0	100.0

Table 5.5 (continued). T-test results of the differences

Model/Decile	AAPL of logistic regression (SMOTE)	AAPL of two-stage BOOSVM (SMOTE)	Difference	T-value (P-value)
1	594.3 (40.8)	628.2 (32.1)	33.9*	2.06 (0.06)
2	365.7 (20.9)	403.4 (16.8)	37.7***	4.45 (0.00)
Model/Decile	AAPL of MOEA (SMOTE)	AAPL of two-stage BOOSVM (SMOTE)	Difference	T-value (P-value)
1	498.8 (91.9)	628.2 (32.1)	129.4***	4.20 (0.00)
2	321.5 (43.0)	403.4 (16.8)	81.9***	5.61 (0.00)

*, **, *** indicate the 10%, 5%, and 1% significance levels, respectively

In Table 5.4, it is observed that in the first deciles, AAPL of MOEA (SMOTE) (284.6) is significantly lower (t-value: 3.99) than the result of BOOSVM method (352.2). In the second decile, AAPL of logistic regression (SMOTE) method (254.4) and MOEA method (213.5) are both lower than that of BOOSVM method (270.8). In Table 5.5, the AAPLs of MOEA method (498.8 and 321.5) are significantly (t-value:

4.20 and 5.61) lower than those of BOOSVM method (628.2 and 403.4) in the first and second deciles. Meanwhile, the AAPL of logistic regression (365.7) is significantly (t-value: 4.02) lower than that of BOOSVM method (403.4) in second decile.

The results in Table 5.4 and Table 5.5 exclude the sole influence of SMOTE method. In the following comparison experiments, two-stage BOOSVM methods are all learnt models on SMOTE-processed dataset and other comparison models are learned from the original training sets.

5.2 Results of Models Comparisons: KDD-98 Dataset

I use another dataset, KDD-98 dataset, to test the performance of the two-stage BOOSVM model. There are two reasons to choose KDD-98 dataset in my thesis. First, DM dataset is not commonly used and two-stage BOOSVM model's excellent performance on DM dataset is insufficient to prove its efficiency, while KDD-98 dataset is widely available to the public and has already been used in many studies. Second, in common cases customers' probabilities to purchase have positive relationship with the dollar amount to spend and a higher response rate can lead to more profit (for example, DM dataset). However, the likelihood of customer to response may also be inversely correlated to the dollar amount spent, and a higher response rate cause a reduction of profit (for example, KDD-98 dataset) (Wang et al., 2005). Thus, applying both DM dataset and KDD-98 dataset can prove the two-stage BOOSVM model's efficiency in different relationship situations between two target

variables (response rate and profit).

In this study, logistic regression, MOEA and two-stage BOOSVM model all employ the SMOTE-processed KDD-98 dataset. Compared results are shown in Table 5.6.

Table 5.6. AARL of logistic regression with KDD-98 dataset, MOEA with KDD-98 dataset and two-stage BOOSVM with SMOTE-processed KDD-98 dataset

Model /Decile	AARL of logistic regression (SMOTE)	AARL of MOEA (SMOTE)	AARL of two-stage BOOSVM (SMOTE)
1	138.7	116.0	175.6
2	132.3	113.5	164.4
3	126.9	110.5	150.3
4	120.1	108.6	145.7
5	113.8	108.3	137.1
6	110.1	107.0	131.2
7	107.2	105.5	123.5
8	104.6	104.3	116.2
9	102.2	102.7	105.5
10	100.0	100.0	100.0

Table 5.7. AAPL of logistic regression with KDD-98 dataset, MOEA with KDD-98 dataset and two-stage BOOSVM with SMOTE-processed KDD-98 dataset

Model /Decile	AAPL of logistic regression (SMOTE)	AAPL of MOEA (SMOTE)	AAPL of two-stage BOOSVM (SMOTE)
1	514.1	482.1	714.3
2	313.0	367.6	538.4
3	227.9	296.5	327.2
4	180.4	243.6	279.3
5	166.1	216.9	220.5
6	153.6	187.2	169.3
7	145.4	162.3	146.1
8	135.3	143.5	132.2
9	120.6	122.5	112.9
10	100.0	100.0	100.0

Table 5.6 gives the results of average accumulate response rate lift (AARL) obtained by logistic regression, MOEA method and the two-stage BOOSVM method, respectively. Both the logistic regression and MOEA method are based on the KDD-98 dataset, while the two-stage BOOSVM method is based on SMOTE-processed KDD-98 dataset. It is showed in table 5.6 that when selecting the first or the second decile customers, the AARLs (138.7 and 132.3) obtained by logistic regressions are much lower than those (175.6 and 164.4) obtained by two-stage BOOSVM. Similarly, the AARLs (116.0 and 113.5) obtained by MOEA in the first and second deciles are also much lower than those (175.6 and 164.4) obtained by two-stage BOOSVM. Thus, the above comparisons of the results indicate that, at budget constraint situation where only 10% or 20% customers can be selected, two-stage BOOSVM method can generate higher response rate than logistic regression or MOEA when they are used to select customers.

Table 5.7 contains the results of average accumulate profit lift (AAPL) obtained by logistic regression, MOEA method and the two-stage BOOSVM method, respectively. Both the logistic regression and MOEA method are based on the KDD-98 dataset, while the two-stage BOOSVM method is based on SMOTE-processed KDD-98 dataset. Table 5.7 indicates that in the top two deciles, the AAPLs of two-stage BOOSVM (714.3 and 538.4) are much higher than those (514 and 313.0) of logistic regression. In the top two deciles, the AAPL of two-stage BOOSVM (714.3 and 538.4) are also much higher than those (482.1 and 367.6) of MOEA. The above comparisons indicate that in budget constraint situation where

only ten percent or twenty percent customers are to be chosen, applying two-stage BOOSVM method to select customers can bring more profit to the companies than applying logistic regression or MOEA.

Chapter 6. Conclusions

6.1 Major Findings

Direct marketing becomes increasingly important in retail, banking, insurance and fundraising industries. Customer-selection model is a core element for direct marketing advertisings. Customer's response probability and profitability play significant roles in the development of customer-selection models. In this study, I suggest a customer selection model to realize bi-objective optimization and make a contribution in methodologies to solve the difficulties in model construction.

One of the difficulties to construct such a model is that a better prediction model should optimize on both response rate and profit, which are not independent but have a close relationship with each other. Two-stage models are commonly used to perform multi-objective optimization in direct marketing. However, in existing two-stage models, technique in each stage neglects the relationship of response rate and profit, and so the combined results in two stages is not theoretically sufficient. I suggest an SVM-based BOOSOM model that can take advantage of both SVC and SVR. By searching a common maximum-margin hyperplane, the BOOSVM model can be used to predict response rate and profit at the same time. The well developed theory and significant practical applicability of SVM guarantee the feasibility of BOOSVM. It also fills in the gaps of optimization of two objectives in a single SVM-based model in direct marketing.

Another contribution of the thesis is the proposed method to handle the unbalanced dataset in direct marketing. Commonly, SVM is not good at learning patterns in unbalanced dataset and two datasets used in my thesis are both significantly unbalanced. In that case, the SVM-based BOOSVM model cannot function well. By applying a SMOTE approach to assist BOOSVM model, the performance improvement of BOOSVM model is prominent. In other research areas, many studies have used SMOTE to help SVM function, however few have adopted SMOTE in direct marketing. The successful application of BOOSVM model on SMOTE-processed dataset can not only confirm the existing conclusion that SMOTE can assist SVMs, but also is one of the original applications of SMOTE method in direct marketing area.

Lastly, I propose a method to solve the problem arising from the special features of direct marketing dataset. To obtain higher response rate, BOOSVM model should be learnt from the whole training set. To generate a more powerful profit prediction model, BOOSVM model should be learnt from only the responders. The paradox inspires me to use a two-stage method to enhance the profit prediction of the BOOSVM model. Thus, I design the two-stage BOOSVM model. Specifically, I use a regression-based model to give another prediction of potential customers' profit, and then I combine the two results from each stage together as the final prediction values. The two-stage BOOSVM model give an enhancement to profit prediction and overcome the problem that the majority class of non-responders have no effective to profit modeling.

Two datasets (DM dataset and KDD-98 dataset) are applied for comparisons in this study. According to previous literature, DM dataset has a positive relationship between response rate and profit, while KDD-98 dataset has a negative relationship between two objectives. Testing models on both two datasets can help prove the practical applicability of the proposed model. In addition, the KDD-98 dataset, an authoritative dataset used in a competition, is widely used by many researchers. Thus, the good prediction performance of two-stage BOOSVM model on both datasets is sufficient to prove its ability to select customers.

6.2 Implications

The thesis' findings may provide important implications to companies. In my thesis, a two-stage BOOSVM model is proposed to optimize response rate and profit. SMOTE method is used as a data processing method to assist BOOSVM. I show a comprehensive framework of customer selection models and the techniques used for customer selection in direct marketing. So this study can be utilized as a guideline of customer selection models in direct marketing. Meanwhile I also describe the SVM theory in detail, thus my thesis can also be used as a guide for SVM. The BOOSVM model is obtained from a precise mathematic derivation, and this provides a new idea to optimize the accuracy of classification and regression.

The realization of bi-objective optimization in direct marketing provides direct marketers a novel and effective strategy to select customers. By communicating and interacting with customers, companies build their huge firm databases. Direct marketers can get better customer targeting by applying the two-stage BOOSVM

model on their firm databases than by applying traditional customer-selection models. By the assistance of two-stage BOOSVM model, marketers can obtain more responders as well as more profit simultaneously, thus maintain a balanced trade-off between the two objectives. In addition, it can also assist direct marketers to reduce promotion cost and avoid enraging the customers. The proved remarkable performance of two-stage BOOSVM model can enhance the efficiency and profitability that direct marketers can obtain more profit by selecting customers with a higher accuracy. In addition, better performance in selecting customers in direct marketing help companies accumulate more customer information and build more sophisticated models.

6.3 Limitations and Directions for Future Research

First, the evaluation results of these direct marketing models are usually done on the basis of response rate (accuracy) and profit. However, in practical direct marketing strategy, marketers may be concerned with some other issues, for example, customer loyalty. This implies that in addition to response rate and profit, other objectives should be taken into consideration by marketers in some special situations. My thesis only focuses on two popular objectives and realizes optimization of them. In future studies, researchers can perform more detailed analysis to marketer's concerns by combining response rate and profit with customer management. In addition, researchers can also combine bi-objective optimization in direct marketing with time series analysis, for example, the effect of bi-objective optimization to improve companies' revenue in the short term as well as in long term.

Second, this research does not fully capture the computational complexity problem (the number of support vectors). Computational complexity can be optimized with effective software to a certain degree. Moreover, reducing the number of support vectors in SVMs can improve operation speed and reduce operation time effectively, especially for some large datasets in direct marketing. Researchers may also address the problem of computational complexity in future study.

Third, grid search approach is used to find optimal parameters in my thesis. Although grid search is appropriate for the problem of SVM, it is still exhaustive and time consuming. In some other studies, random search has been found to be more effective in high-dimensional spaces than exhaustive search. In addition, grid search is prone to suffer from the curse of dimensionality, but it is often embarrassingly parallel because typically the hyper-parameter settings are independent of each other. In future studies, researchers may explore some other methods, such as genetic algorithm, to search for proper parameters and to balance the complexity and precision.

Last but not the least, the process to select variables has to be addressed. My study selects 46 independent variables from 307 of variables for the KDD98 dataset. As the number of independent variables is too large, the methodology to select a suitable number of independent variables from a large number of variables has to be developed. The validity of the independent variables needs to be tested while relevant variables have to be controlled.

Appendix A. Variables for the DM Dataset

Variable name	meaning
prord85	NUMERIC: prior order from promotion 85
salcat	NUMERIC: sales category (categorical)
crcpr85	NUMERIC: prior responses to promotion 85
totord	NUMERIC: total number of orders
totsale	NUMERIC: total amount of sale
hcrd	NUMERIC: use of house credit card
cash	NUMERIC: cash payment
tele	NUMERIC: order by telephone
salflg	NUMERIC: sales flag
pred_r1	NUMERIC: predicted recency
resid_r1	residual of predicted recency
pred_f1	predicted frequency
resid_f1	residual of predicted frequency
pred_m1	predicted money
resid_m1	residual of predicted money
SALCLS3	NUMERIC: sales from class 3 products
SSORCLS3	NUMERIC: orders from class 3 products
targpbfo	NUMERIC: profit before fix cost

Appendix B. Variables for the KDD98 Dataset

Variable name	meaning
MONTHS_SINCE_ORIGIN	Elapsed time since first donation
IN_HOUSE	Is in house donor?
OVERLAY_SOURCE	M = Metromail, P = Polk, B = both
DONOR_AGE	Age as of June 1997
DONOR_GENDER	Actual or inferred gender
PUBLISHED_PHONE	Published telephone listing
HOME_OWNER	Is home owner?
MOR_HIT	Mail order response hit rate
CLUSTER	54 socioeconomic cluster codes
SES	5 socioeconomic cluster codes
INCOME	7 income group levels
MED_HOUSEHOLD_INCOME	Median income
PER_CAPITA_INCOME	Income per capita
WEALTH	10 wealth rating groups
MED_HOME_VALUE	Median home value
PCT_OWNER_OCCUPIED	Percent owner occupied housing
URBANICITY	U = urban, C = city, S = suburban, T.=.town, R.=.rural, M = unknown
PCT_MALE_MILITARY	Percent male military in block
PCT_MALE_VETERANS	Percent male veterans in block
PCT_VIETNAM_VETERANS	Percent Vietnam veterans in block
PCT_WWII_VETERANS	Percent World War II veterans in block
NUMBER_PROM_12	Number of promotions in the last 12 months
CARD_PROM_12	Number of card promotions in the last 12 months
FREQ_STATUS_97NK	Frequency status, June 1997
RECENCY_STATUS_96NK	Recency status, June 1996
LAST_GIFT_AMT	Amount of the most recent donation
RECENT_RESPONSE_COUNT	Recent response count
RECENT_RESPONSE_PROP	Recent response proportion
RECENT_AVG_GIFT_AMT	Recent average gift amount
RECENT_STAR_STATUS	Recent STAR status (1 = yes, 0 = no)
CARD_RESPONSE_COUNT	Response count since June 1994
CARD_RESPONSE_PROP	Response proportion since June 1994
CARD_AVG_GIFT_AMT	Average gift amount since June 1994
PROM	Total number of promotions
GIFT_COUNT	Total number of donations
AVG_GIFT_AMT	Overall average gift amount
GIFT_AMOUNT	Total gift amount
MAX_GIFT	Maximum gift amount
GIFT_RANGE	Maximum less minimum gift amount

MONTHS_SINCE_FIRST	First donation date from June 1997
MONTHS_SINCE_LAST	Last donation date from June 1997
PEP_STAR	STAR status ever (1 = yes, 0 = no)
CARD_PROM	Number of card promotion
MIN_GIFT	Minimum gift amount
TARGET_D (dependent variable)	Response amount to June 1997 solicitation
TARGET_B (dependent variable)	Response to June 1997 solicitation

References

- Aizerman, M., Braverman, E., Rozoner, L., 1964. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control* 25, 821–837.
- Athanassopoulos, A.D., 2000. Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of Business Research* 47, 191–207.
- Barwise, P., Farley, J.U., 2005. The state of interactive marketing in seven countries: Interactive marketing comes of age. *Journal of Interactive Marketing* 19, 67–80.
- Bennett, K.P., Campbell, C., 2000. Support vector machines: hype or hallelujah? *SIGKDD Explor. Newsl.* 2, 1–13.
- Bhattacharyya, S.K., 1998. The international Warwick manufacturing group. *Assembly automation* 18.
- Bhattacharyya, A., 1999. Deterministic retrieval of ionospheric phase screen from amplitude scintillations. *Radio Science* 34, 229–240.
- Bhattacharyya, S., 2000. Evolutionary algorithms in data mining: Multi-objective performance modeling for direct marketing. In: *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining*, Boston, MA, USA, pp. 465–473.
- Bi, J., 2003. Multi-objective programming in SVMs, in: *Machine Learning, Proceedings of the 20th International Conference (ICML 2003)*, AAAI. AAAI Press, pp. 35–42.
- Bodapati, A., Gupta, S., 2004. A direct approach to predicting discretized response in target marketing. *Journal of Marketing Research* 41 (1), 73–85.
- Bose, I., Chen, X., 2009. Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research* 195, 1–16.
- Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers, in: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. ACM Press, pp. 144–152.
- Brandweek, 2009. The Next generation of direct marketing. *Academic Search Complete*, EBSCO host, 50, p.6.
- Bult, J.R., 1993. Semiparametric versus parametric classification models: An application to direct marketing. *Journal of Marketing Research* 30 (3), 380–390.
- Bult, J.R., Wansbeek, T., 1995. Optimal selection for direct mail. *Marketing Science* 14, 378–394.

- Bult, J.R., Van der Scheer, H., Wansbeek, T., 1997. Interaction between target and mailing characteristics in direct marketing, with an application to healthcare fund raising. *International Journal of Research in Marketing* 14 (4), 301–308.
- Burges, C.J.C., 1998. A Tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* 2, 121–167.
- Candade, N., 2004. Application of support vector machines and neural networks in digital mammography: A comparative study. Graduate School Theses and Dissertations.
- Chang, Y.-W., Hsieh, C.-J., Chang, K.-W., Ringgaard, M., Lin, C.-J., 2010. Training and testing low-degree polynomial data mappings via linear SVM. *J. Mach. Learn. Res.* 99, 1471–1490.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16, 321–357.
- Chiu, C., 2002. A case-based customer classification approach for direct marketing. *Expert Systems with Applications*, 22, 163–168
- Colombo, R., Jiang, W., 1999. A stochastic RFM model. *Journal of Interactive Marketing* 13, 2–12.
- Coussement, K., Van den Poel, D., 2008. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Syst. Appl.* 34, 313–327.
- Cui, D., Curry, D., 2005. Prediction in Marketing Using the Support Vector Machine. *Marketing Science* 24, 595–615.
- Cui, G., Wong, M.L., Lui, H.-K., 2006. Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. *Management Science* 52(4), 597–612.
- Cui, G., Wong, M., Wan, X., 2012. Cost-Sensitive Learning via Priority Sampling to Improve the Return on Marketing and CRM Investment. *J. Manage. Inf. Syst.* 29(1), 341–374.
- Deichmann, J., Eshghi, A., Haughton, D., Sayek, S., Teebagy, N., 2002. Application of multiple adaptive regression splines (MARS) in direct response modeling. *Journal of Interactive Marketing* 16, 15–27.
- Dou, W., Hu, J., Hirasawa, K., Wu, G., 2008. Quick response data mining model using genetic algorithm, in: *SICE Annual Conference, 2008*. Presented at the SICE Annual Conference, 2008, pp. 1214–1219.
- Elsner, R., Krafft, M., Huchzermeier, A., 2003. Optimizing Rhenania's mail-order

- business through dynamic multilevel modeling. *Interfaces* 33 (1), 50–66.
- Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), 1996. *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence, Menlo Park, CA, USA.
- Freund, Y., Schapire, R. E., 1996. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 148–156.
- Genkin, A., Lewis, D., Madigan, D., 2007. Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics* 49, 291–304.
- Goldberg, D.E., 1989. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Longman Publishing, Boston.
- Gruca, T.S., Klemz, B.R., Petersen, E.A.F., 1999. Mining sales data using a neural network model of market response. *SIGKDD Explor. Newsl.* 1, 39–43.
- Hansen, L.K., Salamon, P., 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12, 993–1001.
- Hanssens, D.M., Leeflang, P.S.H., Wittink, D.R., 2005. Market response models and marketing practice: Research Articles. *Appl. Stoch. Model. Bus. Ind.* 21, 423–434.
- Ha, K., Cho, S., Maclachlan, D., 2005. Response models based on bagging neural networks. *Journal of Interactive Marketing* 19 (1), 17–30.
- Ha, T.M., Bunke, H., 1997. Off-line, handwritten numeral recognition by perturbation method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 535–539.
- Haughton, D., Oulabi, S., 1997. Direct marketing modeling with CART and CHAID. *Journal of Direct Marketing* 11 (4), 42–52.
- Haskell, F., 1993. *History and Its Images: Art and the Interpretation of the Past*. Yale University Press.
- Heilman, C.M., Kaefer, F., Ramenofsky, S.D., 2003. Determining the appropriate amount of data for classifying consumers for direct marketing purposes. *Journal of Interactive Marketing* 14 (3), 5–28.
- Hsu, C., Chang, C., Lin, C., 2010. *A practical guide to support vector classification*. Department of Computer Science and Information Engineering, National Taiwan University.
- Hu, P.J., Chau, P.Y.K., Sheng, O.R.L., Tam, K.Y., 1999. Examining the technology acceptance model using physician acceptance of telemedicine technology. *J.*

- Manage. Inf. Syst. 16, 91–112.
- Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.
- Hosmer, D.W, Jr. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley & Sons, Inc.
- Jones, G.A., Thornton, C.A., Langrall, C.W., Mooney, E.S., Perry, B., Putt, I.J., 2000. A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning* 2, 269–307.
- Kim, D., Lee, H., Cho, S., 2008. Response modeling with support vector regression. *Expert Systems with Applications* 34, 1102–1108.
- Kim, K.I., Franz, M.O., Schölkopf, B., 2005. Iterative kernel principal component analysis for image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 2005.
- Kim, Y., Street, W.N., 2004. An intelligent system for customer targeting: A data mining approach. *Decision Support Systems* 37 (2), 215–228.
- Lessmann, C., 2009. Fiscal decentralization and regional disparity: evidence from cross-section and panel data. *Environment and Planning A* 41, 2455-2473.
- Linder, R., Geier, J., Kölliker, M., 2004. Artificial neural networks, classification trees and regression: Which method for which customer base? *Journal of Database Marketing & Customer Strategy Management* 11, 344–356.
- Ling, C., Li, C., 1998. Data mining for direct marketing problems and solutions. In *proceedings of the fourth international conference on knowledge discovery and data mining (KDD-98)*. New York, NY: AAAI Press.
- Levin, N., Zahavi, J., 1998. Continuous predictive modeling – a comparative analysis. *Journal of Interactive Marketing* 12 (2), 5–22
- Malthouse, E.C., 1999. Ridge regression and direct marketing scoring models. *Journal of Interactive Marketing* 13 (4), 10–23.
- Malthouse, E.C. 2002. Performance-based variable selection for scoring models. *J. of Interactive Marketing* 16(4), 37–50.
- Meyer, D., Leisch, F., Hornik, K., 2003. The support vector machine under test. *Neurocomputing* 55, 169–186.
- Mullin, R., 2002. *Direct marketing: a step-by-step guide to effective planning and targeting*. Kogan Page, London.
- Naik, P.A., Tsai, C.-L., 2004. Isotonic single-index model for high-dimensional

- database marketing. *Computational Statistics & Data Analysis* 47, 775–790.
- Nash, E.L., 1984. *The direct marketing handbook*. McGraw-Hill, New York.
- Osuna, E.E., Freund, R., Girosi, F., 1997. Support vector machines: Training and applications. A.I. MEMO 1602, MIT A. I. LAB.
- Quinlan, J.R., 1986. Induction of decision trees. *Mach. Learn* 81–106.
- Rea, L., & Parker, R., 1997. *Designing and conducting survey research* (2nd Edition). San Francisco: Jossey-Bass.
- Pan, J., Yang, Q., Yang, Y., Li, L., Li, F.T., Li, G.W., 2007. Cost-Sensitive-Data Preprocessing for Mining Customer Relationship Management Databases. *IEEE Intelligent Systems* 22, 46–51.
- Peacock, P. R. , 1998, "Data Mining in Marketing: Part I," *Marketing Management*, 1, (Winter), 9-18.
- Potharst, R., Bioch, J.C., 2000. Decision trees for ordinal classification. *Intelligent Data Analysis* 4, 97–111.
- Rygielski, C., Wang, J.-C., Yen, D.C., 2002. Data mining techniques for customer relationship management. *Technology in Society* 24, 483–502.
- Scholkopf, B., Smola, A.J., 2002. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge, MA, USA.
- Sen, A., 1990. Development as capability expansion. *Human development and the international development strategy for the 1990s*, London: Macmillan, pp.41-58
- Schindler, M., Baumgartner, B., Hruschka, H., 2007. Nonlinear Effects in Brand Choice Models: Comparing Heterogeneous Latent Class To Homogeneous Nonlinear Models. *Schmalenbach Business Review* 59, 118–137.
- Shaw, M.J., Subramaniam, C., Tan, G.W., Welge, M.E., 2001. Knowledge management and data mining for marketing. *Decis. Support Syst.* 31, 127–137.
- Shin, H.W., Sohn, S.Y., 2004. Multi-attribute scoring method for mobile telecommunication subscribers. *Expert Systems with Applications* 26 (3), 363–368.
- Van der Sheer, H.R., 1998. Quantitative approaches for profit maximization in direct marketing, PhD thesis, University of Groningen, Netherlands.
- Venkatesan, R., Kumar, V., 2004. A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. *Journal of Marketing* 68, 106–125.
- V. Vapnik., 1995. *The nature of statistical learning theory*. Springer-Verlag, New

York.

- Wang, T., Qin, Z., Zhang, S., Zhang, C., 2012. Cost-sensitive classification with inadequate labeled data. *Inf. Syst.* 37, 508–516.
- Wong, M.-L., Cui, G., 2010. Data mining using parallel Multi-Objective Evolutionary algorithms on graphics hardware. Presented at the 2010 IEEE Congress on Evolutionary Computation (CEC), pp. 1–8.
- Yeh, S.-W., Kug, J.-S., Dewitte, B., Kwon, M.-H., Kirtman, B.P., Jin, F.-F., 2009. El Niño in a changing climate. *Nature* 461, 511–514.
- Yu, C.C.W., Sung, R.Y.T., So, R.C.H., Lui, K., Lau, W., Lam, P.K.W., et al., 2005. Effects of strength training on body composition and bone mineral content in children who are obese. *Journal of Strength and Conditioning Research* 19, 667–672.
- Zahavi, J., Levin, N., 1997. Issues and problems in applying neural computing to target marketing. *Journal of Direct Marketing* 11, 63–75.