International Journal of Computational Intelligence and Applications
Vol. 14, No. 4 (2015) 1550022 (23 pages)

DOI: 10.1142/S1469026815500224

A Cluster-Based Data Balancing Ensemble Classifier for Response Modeling in Bank Direct Marketing

Mohammad Amini

Department of Information Technology
School of Industrial Engineering
Iran University of Science and Technology
Tehran, Postal Code 16846-13114, Iran
moha_amini@ind.iust.ac.ir

Jalal Rezaeenour* and Esmaeil Hadavandi†

Department of Industrial Engineering
School of Technology and Engineering
University of Qom, Alghadir Blvd.
Qom, Postal Code 3716146611, Iran
**J.rezaee@qom.ac.ir*
†*es.hadavandi@aut.ac.ir*

Received 15 June 2014
Revised 13 August 2015
Published 18 December 2015

The aim of direct marketing is to find the right customers who are most likely to respond to marketing campaign messages. In order to detect which customers are most valuable, response modeling is used to classify customers as respondent or non-respondent using their purchase history information or other behavioral characteristics. Data mining techniques, including effective classification methods, can be used to predict responsive customers. However, the inherent problem of imbalanced data in response modeling brings some difficulties into response prediction. As a result, the prediction models will be biased towards non-respondent customers. Another problem is that single models cannot provide the desired high accuracy due to their internal limitations. In this paper, we propose an ensemble classification method which removes imbalance in the data, using a combination of clustering and under-sampling. The predictions of multiple classifiers are combined in order to achieve better results. Using data from a bank's marketing campaigns, this ensemble method is implemented on different classification techniques and the results are evaluated. We also evaluate the performance of this ensemble method against two alternative ensembles. The experimental results demonstrate that our proposed method can improve the performance of the response models for bank direct marketing by raising prediction accuracy and increasing response rate.

Keywords: Direct marketing; response modeling; ensemble; classification; clustering.

1. Introduction

In today's highly competitive business environments, more and more companies are moving toward direct marketing methods. Traditional marketing often involves

decisions based on experience or competitors' actions to provide new marketing plans. In these methods, uniform information is sent to customers without any discrimination. However, the difference in customers' requirements and preferences is a very important factor which must be taken into account. Companies and vendors which apply mass marketing, sending mail catalogues and promotional emails to a very large group of customers, or contacting them via phone calls, will receive a very low rate of response from customers (which is to buy a product or subscribe a service) to their marketing campaigns. Therefore, mass marketing is very expensive, considering the very low average probability of purchase. Direct marketing is a kind of marketing which focuses on customers' differences and tries to find more valuable customers to target. It helps marketers to establish more efficient marketing campaigns and gain more revenues.[42] In direct marketing, those customers who are more likely to respond to a promotional email or phone call must be identified and targeted. For this purpose, response models are very effective since they determine which customers have more likelihood to respond based on their purchase history and other information.

Typically, a response modeling method calculates a score (probability of response) for each customer using his or her history information. When the scores are sorted in descending order, marketing managers can contact customers who are most probable to buy their products or subscribe to their service. A well-designed response model can be beneficial to the marketing campaigns in two ways: First, the total revenue is increased, because that segment of customers, who would not respond unless they are targeted, will be informed about their needs and may be convinced to buy a particular product or subscribe a service. As a result, the amounts of sales for the firms increase. Second, the overall marketing expenses will be lowered since the response model identifies the customers with relatively high response probability. Consequently, the amount of expenses, which would have to be spent on contacting customers who are less interested in buying products, will be significantly diminished.[12,13]

Prediction accuracy of a response model is very important and has a significant effect on the amount of revenue.[33] The impact of a small improvement in response rate is highly considerable and may even change the overall result of a marketing campaign from failure to success.[2,8] Improvement of the response rate can also fortify customer loyalty because properly targeted customers are more likely to be gratified and stay with the company over the longer period.[41] Therefore, it is really worthwhile to develop more powerful response models for marketing experts by means of improved prediction algorithms.

In general, customer response modeling is considered as a binary classification problem in which customers are classified as "respondents" or "non-respondent". If the result of a phone contacts or mail sending marketing program for each customer is desired, two classes can also be "failure" or "success". Various statistical methods and machine learning techniques have been proposed for this problem. Bose

and Chen in Ref. 4 offered a classification of different methods used for response modeling and categorize them into four groups: Basic statistical techniques, advanced statistical techniques, machine learning techniques, and ensemble and hybrid techniques. Logistic regression (LR) has been proposed in many studies because of its simplicity and practicality.[1,21] Some stochastic RFM models[9] and hazard function models[14] have been also proposed by the researchers.

Data mining and machine learning methods have become popular for response modeling recently. Strong classification methods such as decision tree (DT), support vector machine (SVM), and artificial neural networks (ANN) are commonly applied.[20] DT can build classification models which are understandable by human and provide good prediction performance. Haughton and Oulabi applied classification and regression trees (CART) and also Chi-square automatic interaction detector (CHAID) for response modeling.[20] ANN and SVM techniques have more flexibility compared to DT and LR. They can learn complex nonlinear relationships and give more accurate predictions. However, the models built with these two methods are difficult to interpret by market analysts. In Ref. 47, Viaene *et al.* proposed least squares SVM for modeling repeat-purchase behavior. Shin and Cho applied SVM for response modeling in direct marketing.[39] The use of ANNS has reported to give good results for response prediction in some studies.[15,37,46] Kim *et al.* applied ANNs guided by Genetic Algorithm (GA) for customer targeting.[22] However, one common problem of ANNs is their instability and they might require tremendous amount of efforts and computer resources in order to be set up for direct marketing applications.[49]

Previous research done on response modeling mostly focused on finding the best "single models". However, applying single models have some problems such as low accuracy and high generalization error. As mentioned above some classification methods like ANNs are unstable and may have a large variance in predictions. Some methods like LR tend to have large bias in predictions. It has been demonstrated that an ensemble of classifiers can resolve these issues and reduce generalization error.[34,35] An ensemble system is made by combining the predictions of multiple individual classifiers which have been trained on the same or different data. Bose and Chen applied bagging ensemble method for direct marketing.[4] In bagging method, a group of single classifiers are trained on training data samples obtained by bootstrap sampling technique.[35] In boosting ensemble method, classifier ensemble is built incrementally, adding one classifier at a time. The classifier that joins the ensemble at each step is trained on a dataset selected from the original dataset and has the examples more difficult to classify than the examples in the last step.[38] In Ref. 27, the authors demonstrated that when bagging and boosting are applied on customer churn prediction, it could provide significantly better performance compared to single models. Bagging neural networks was applied by Ha *et al.* as a response model.[16] This method not only improved the prediction accuracy but also stabilized it compared to the single neural network and conventional logistic regression.

Some new methods use clustering as a preliminary step for creating ensembles in which diversity and accuracy can be improved. For instance, Wang *et al.* used fuzzy clustering technique to generate different training subsets and trained different ANN models based on training subsets.[48] In Ref. 36, Rahman and Verma generated an ensemble of classifiers by clustering at multiple layers. The decisions obtained at different layers were fused into a final verdict using majority voting. The authors in Ref. 45 proposed an ensemble method in which the dataset is characterized into multiple clusters and fed to a number of distinctive base classifiers. The base classifiers learned cluster boundaries and produced cluster confidence vectors. A second level fusion classifier combined the cluster confidences and mapped to class decisions. A new method was proposed by Lin *et al.* in which was a hybrid model of ensemble pruning based on k-means clustering and the framework of dynamic selection and circulating in combination with a sequential search method.[28] In Ref. 43, the authors developed a hybrid financial distress model based on a combination of the clustering technique and classifier ensembles. They used two clustering techniques, Self-Organizing Maps (SOMs) and k-means and three classification techniques, logistic regression, multi-layer perceptron (MLP) neural network, and DTs to develop four different types of bankruptcy prediction models.

A very important challenge in response modeling is how to deal with the class imbalance problem. It is very common that the number of customers who respond to a marketing message is much smaller than those who do not respond. As a result, when we build classification models with imbalanced datasets, the classification results tend to have bias toward the larger class, i.e., non-respondents. The response rate in most customer datasets used for research on direct marketing is less than 10% and even lower in real marketing situations.[16,31,39,44] If a single classification algorithm is trained on an imbalanced customer dataset, the resulting model would predict most unknown customers as non-respondent. This is considered a huge opportunity loss. Therefore, devising appropriate methods to resolve the class imbalance issue in the customer data is highly necessary.

There are some data balancing methods by which a new well-balanced training dataset is produced from the original dataset using a sampling strategy. Since these methods are independent from the classification techniques, they can be used with any kind of classifiers. Two major methods are under-sampling and over-sampling. In under-sampling, a subset of the majority class instances is selected. The number of the instances in the majority class is reduced to the extent, in which the proportions of the two class instances in the dataset become relatively equal. This method can be effective in reducing training time. However, if the sampling is performed randomly, the class distribution of the instances will be changed and may not reflect the actual characteristics of the original dataset.[25] The prediction model built on this data does not provide stable results. Over-sampling increases the number of instances in the minority class, by duplicating some instances in this class, to the extent in which the two class sizes show a balance. Although this method can maintain the distribution

of the original data, more time is needed to train the classification algorithm.[7] Moreover, over-sampling in response modeling may provide wrong information about the customers, since "virtual respondents" are created from limited number of actual customers.

In this paper, we propose an ensemble method to deal with the problems above. First, we use a data balancing approach with the combination of clustering, and random under-sampling. We do not use over-sampling because it generates unrealistic data. Since we want to maintain the class distribution of the original dataset, we partition the non-respondent customers into some clusters and take random samples from each cluster. Each cluster forms a group of non-respondents with similar characteristics. The sampled non-respondent instances are added to the whole respondent instances and a new balanced training dataset is formed. By sampling from each cluster, the homogenous groups of non-respondent customers have representatives in the reduced dataset and the information loss will be minimized. Multiple balanced training datasets are created by repeating this procedure and each dataset is given to a base classifier in an ensemble framework. After training the base classifiers, the prediction of the base classifiers are aggregated and form the final prediction. The ensemble method improves the prediction accuracy, reduces variation and provides more stable results. This ensemble framework is implemented by popular classification methods, SVM, ANN, DT, and LR. We evaluate the performance of the proposed ensemble classification system against single methods and some forms of ensemble methods.

Our focus in this paper is on the case of telemarketing for bank customers. Very few research works have been done on the specific field of banking direct marketing. In Ref. 26, the authors investigated the benefits of data mining techniques in Hong-Kong banking sector but they did not test any particular model. A research on customer targeting with pseudo-social networks, based on money transfer relations, was done in Ref. 30. Most recent and significant work is the research of Moro *et al.* in which the authors proposed a response model for predicting the results of phone calls to bank customers in order to sell long term deposits.[31] However, the authors used single classifiers for developing response models. We use this case of bank telemarketing for developing and evaluating our ensemble models.

The rest of this paper is organized as follows: Sec. 2 provides a description for ensemble system and sampling method. In Sec. 3, we explain the dataset, base models, and performance measures. Section 4 describes the experimental results and analysis. Finally, Sec. 5 draws conclusion and suggests future works.

2. Ensemble System and Data Balancing Method

In this section, we first provide a background on ensemble classification and clustering. Then, we describe the data balancing method and the ensemble framework for response modeling.

2.1. *Ensemble classification*

The main idea behind ensemble classification is that by combining several individual classifiers, it is possible to acquire a classifier which has a higher performance than that of every one of the individual classifiers.[23] The predictions of individual classifiers (or ensemble members) are combined using a proper method, e.g., voting or averaging to produce a final prediction. The benefits given by the combination of redundant and complementary classifiers is to increase accuracy, robustness and overall generalization capability in most applications.[34]

Building an ensemble system consists of three phases: (1) Data sampling and selection, which creates diversity among ensemble members, (2) Training member classifiers which is performed using numerous competing algorithms such as bagging, boosting, stacked generalization, etc., (3) combining member predictions using different rules, e.g., majority voting, simple or weighted average.[35] Dietterich introduces three reason why an ensemble system may perform better than the individual classifiers[10] First, from statistic perspective; the ensemble classifier removes the risk of choosing an incompetent single classifier for the problem. Second, from computational perspective; aggregating the predictions of different classifiers can prevent the system from stopping at the local optima. Third, from representational perspective, if the optimal classifier for the problem does not exist, the ensemble of some non-optimal classifiers may lead to a classifier which is as close to the optimal as possible.

Using the ambiguity decomposition, it is demonstrated that squared error related to the prediction of an ensemble of classifiers is certainly less than the squared error form all individual classifiers[24]

$$(f_{\text{ens}} - d)^2 = \sum_i w_i (f_i - d)^2 - \sum_i w_i (f_i - f_{\text{ens}})^2. \tag{1}$$

In the above equation, f_{ens} is the ensemble prediction and d is the actual target for each instance, f_i is the prediction of ith individual classifier, and w_i is the weight of ith classifier in the ensemble. The first term on the right-hand side of the equation shows bias and the second term indicates the variance of the individual predictions. The decomposition shows that the error of an ensemble depends on two elements: accuracy and diversity. However, the presence of these elements is not sufficient solely. A proper balance between accuracy and diversity will guarantee a better generalization error.[6] An ensemble system is aimed to reduce the generalization error by smoothing either bias or variance.

2.2. *Clustering*

A useful techniques commonly used in customer segmentation is clustering. By clustering, we partition a dataset into several disjoint subsets (or clusters) with homogenous instances, which are meaningful for analysis.[3] Given a dataset with N instances, a clustering method assigns all instances to one of the K clusters

$(C_i, i = 1, 2, \ldots, K)$ based on a criterion, where the union of K clusters is equal to the entire dataset. For clustering the customer data in this paper, we applied the most widely used clustering algorithm, *K-means clustering*.[18] This algorithm creates K clusters by first initializing K centroids, and then minimizing sum of squared error within clusters,

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\| x_j - c_i \right\|^2, \tag{2}$$

where c_i is the centroid of cluster C_i. In K-means algorithms the final clusters are generated by repeating two steps. First, all instances in the dataset are assigned to their nearest centroid. Second, the centroid of each cluster is updated by new cluster members. This repetition is stopped when certain stopping criterion is met.

2.3. *Data balancing method and ensemble*

Generally, in the customer data provided for direct marketing the proportion of the number of respondent customers to that of non-respondents is significantly low. As a result, the dataset is highly imbalanced and not suitable for training a classifier. It is acceptable to consider that all respondents have common characteristics which lead them to buy a product or subscribe service. However, the population of non-respondents might have a high heterogeneity and they may have very different reasons for not responding to a marketing message. Thus, it is reasonable to consider that non-respondent customers have heterogeneous groups inside and a simple sampling of their data may not represent all of their characteristics. Clustering can partition heterogeneous data into more homogenous segments. This is the reason why we decided to cluster the non-respondent group and take sample from each cluster in order to maintain the original distribution of the non-respondent customers. For creating a balanced training dataset for our response model we perform the following procedure:

(a) Divide the original training dataset into respondents and non-respondents.
(b) Partition the non-respondent group into K clusters using *K-means* clustering algorithm.
(c) Draw a sample randomly from each of the K clusters. Each sample size should be proportional to the size of the related cluster. Put together, the total samples should have a size equal to the size of the respondent's group.
(d) Add the total samples created in part (c) to the respondent's data.

By following the procedure above, we can obtain a smaller and balanced training dataset which preserves the information of the original imbalanced dataset. In order to reduce the effect of randomness and decrease variation in the predictions we create N training dataset $(D_1, D_2, \ldots, D_N)$, by repeating the procedure above. Then, each of the N training sets is given to a base classifier in an ensemble framework. The base

classifiers are trained and tuned so that they offer the best generalization error. For classifying any unknown customer, its record is given to all base models. The individual predictions made by base models are aggregated using a combination rule and the final prediction is provided. In this paper, the combination rule is *majority voting*. The class of an unknown customer is determined by the output which is predicted by the majority of the base models.

In order to assess the performance of the ensemble model we use test dataset and provide each test instance to all of the individual classifiers in the ensemble. The output of the ensemble is compared with the actual test targets. The proposed ensemble framework is depicted in Fig. 1. The four steps (a)–(d) for building the balanced datasets are also shown in the Figure.

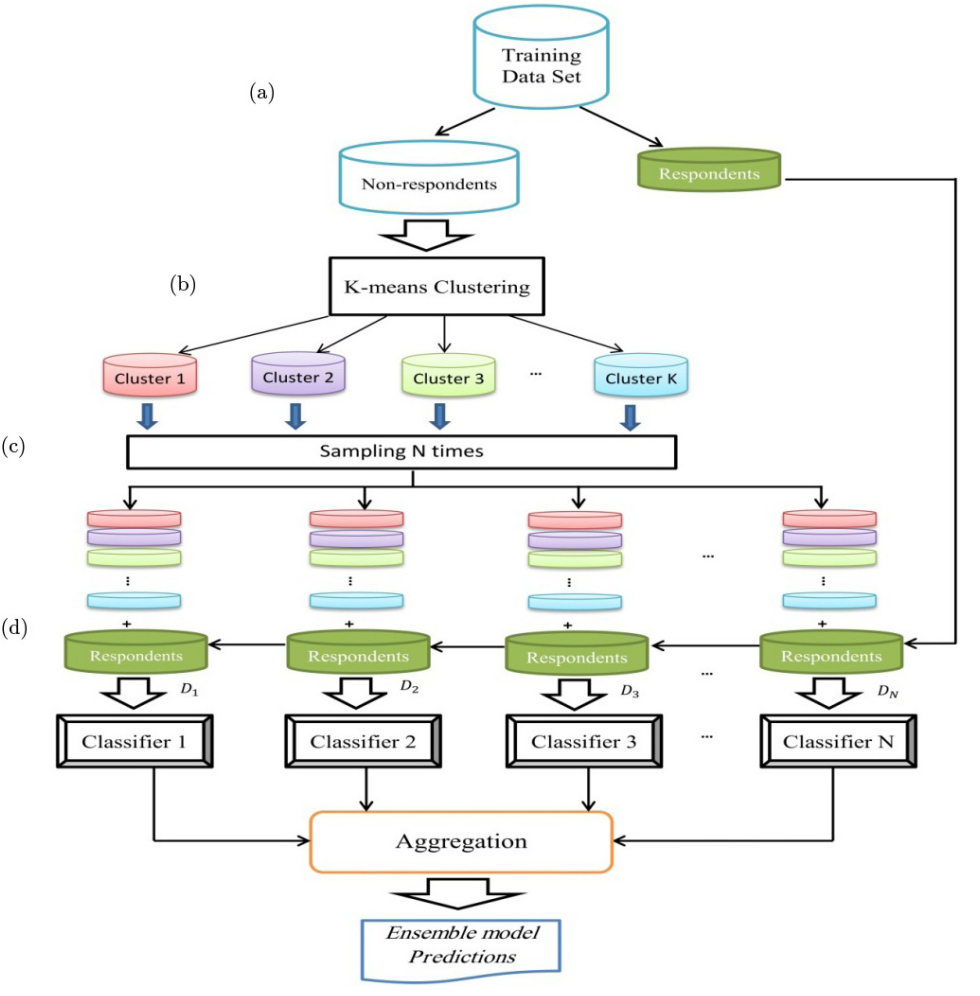

Fig. 1. The Ensemble framework for response modeling.

The benefit provided by our proposed ensemble method is two-fold: First, it maintains the original structure of the non-respondent data and reduces the sampling bias. Sampling bias is a major cause of performance variation. As a result, the prediction performance of our model will be stable and more reliable for marketing experts. Second, the prediction accuracy is improved by our model. As mentioned before, an ensemble system can increase the prediction performance given by single classifiers in many cases. At least, we are sure that it does not worsen the prediction error compared to the single classifiers.

3. Building the Response Model

In this paper, we build a response model for targeting customers in bank telemarketing campaigns. In a telemarketing campaign, customers are contacted through phone calls by the bank call center agents and are offered to deposit their money in the bank for a particular condition and benefit. If the customer is persuaded and subscribes for the term deposit, the result of the call is "success"; otherwise, it is "failure". This will lead to a binary classification problem, in which the goal is to predict if a particular customer will subscribe to a term deposit or not. In this section, we elaborate on the bank data and the base classifiers used for ensemble model construction. We also introduce the performance measures which we use for evaluating the models.

3.1. *Bank telemarketing data*

Our research uses the dataset collected from direct marketing campaigns of a Portuguese retail bank, provided by Moro *et al.*, in their research on using data mining for bank direct marketing.[31,32] The dataset contains 45,211 records for phone calls registered from May 2008 to November 2010. Obviously, this dataset is imbalanced and only 3937 records show successful result (number of respondents). The records are ordered by the date of the contacts. Therefore, the dataset can be used to reflect the temporal characteristics of the customers. In order to do this, we selected data for an interval of 20 months (form May 2008–December 2009) for training the model, and for an interval of 11 months (form January 2010–November 2010) for testing the model performance. The training and test datasets contain 42,591 and 2620 instances, respectively.

Each record consists of 16 features and 1 target value (the outcome of phone call). The 16 features are related to the customer information (age, job, ...), phone call information (date, duration, ...) and history (number of previous contacts, ...). The target value shows if the customer has subscribed the term deposit or not ("yes" or "no"). Table 1 shows detailed information about the dataset features and the type of characteristics they present.

Table 1. Bank telemarketing dataset features and descriptions.

Type of feature	Name	Description
Personal customer information	Age	Age at the contact date (Numeric ≥18)
	Marital	Marital Status (categorical: married, single, divorced)
	Job	Type of job (categorical: management, blue-collar, ...)
	education	Type of education (categorical: elementary, secondary, ...)
Bank customer information	Default	Does customer have credit in default? (binary: "yes", "no")
	Balance	Customer average yearly balance, in euros (numeric)
	Housing	Does customer have housing loan? (binary: "yes", "no")
	Loan	Does customer have personal loan? (binary: "yes", "no")
Last contact information	Contact	contact communication type (categorical: "unknown","telephone"," cellular")
	Day	last contact day of the month (numeric)
	Month	last contact month of year (categorical)
	Duration	last contact duration, in seconds (numeric)
History information	Campaign	number of contacts performed during this campaign and for this customer (numeric, includes last contact)
	Pdays	number of days that passed by after the customer was last contacted from a previous campaign (numeric)
	Previous	number of contacts performed before this campaign and for this customer (numeric)
	Poutcome	outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

3.2. *Base classifiers*

We build our ensemble system using single classifiers as base models. An individual classifier can be made using any good classification method. For this study, we applied popular classification methods which have been widely used in response modeling studies.

These methods include LR, DT, ANN, and SVM.

- Logistic Regression: This method is well-known for its simplicity and interpretability. It can estimates the probability of response for any instance x by performing a logit transformation on a multiple regression model as follows:[21]

$$p(y=1) = \frac{1}{1+\exp(\beta_0 + \sum_1^n \beta_i x_i)}, \tag{3}$$

where $\boldsymbol{x}$ has n features and β_i indicates the coefficient for the ith feature and is adjusted through the learning process. The final values for coefficients can be determined by either maximum likelihood estimation (MLE) or expectation–maximization (EM) algorithm. LR has a strong theoretical foundation and allows us to interpret the model and examine how much a change of an input variable affects the classification output. However, this method is still not sufficiently capable of modeling complex nonlinear relationships.

- Decision Tree: This method builds a tree structure which classifies each object by a set of conditions established on the tree nodes. The conditions examine the value of a particular feature and direct the object to the lower branches based on the feature value.[17] This hierarchical structure can be translated into a set of IF-THEN rules which are easy to interpret for marketing experts. Various versions of decision tree exist such as C4.5, C5, CHAID, and CART.
- Artificial Neural Networks: This method is really preferred for nonlinear complex models. Two forms of ANNs are widely used for classification: MLP and radial basis function (RBF) networks.[11]

MLP networks have one input layer, one output layer, and one or more hidden layers. All neurons in the hidden and output layers have their own activation function. MLP networks usually are trained with back propagation (BP) algorithm. The BP algorithm is used to compute the weights between the input layer and the first hidden layer, between hidden layers and between the last hidden and the output layers. This algorithm usually uses the gradient decent technique to adjust the weights. The number of hidden layers determines the network complexity. For a regular three-layer MLP network on response modeling problem, there are n input nodes for n features, h hidden nodes in the hidden layer, and one output node for predicting the target. The value of the jth hidden node a_j, is determined by the following formula:

$$a_j = \sum_{i=0}^{n} w_{ij} x_i, \tag{4}$$

where w_{ij} is the weight of the connection between ith input node and jth hidden node. The value of the output node (the probability estimate of the target class) is the weighted sum of the activated values from all hidden nodes.

$$p(y = 1) = \sum_{i=0}^{h} w'_j g(a_j), \tag{5}$$

where w'_j is the connection weight between jth hidden node and the output node, and g is the activation function. The activation function is usually the logistic sigmoid function (as in our study). But, it can be the step function or hyper-tangent function as well.

The RBF neural networks consist of three constant layers. The input layer, one hidden layer, and the output layer. The nodes are completely connected, but the difference here is that just the weights between the hidden layer and output layer are adjusted by training. The activation of hidden layer nodes is computed using radial basis function. The most widely used form of RBF is the Gaussian Kernel function which is calculated as follows:

$$g_i(x) = exp\left(\frac{-\|x - v_i\|^2}{2\sigma_i^2}\right), \tag{6}$$

where x is the input vector and v_i is the vector denoting the center of the receptive field unit (hidden layer node) g_i with σ_i as its unit width parameter. This way, if we have - nodes in the hidden layer, the output of RBF network would be calculated by:

$$p(y=1) = \sum_{i=1}^{h} w_i g_i(x), \tag{7}$$

where w_i is the connection weight between the ith hidden node and the output node, and g_i is the value of ith receptive field unit.

The method of training RBF networks is a hybrid approach in two phases: unsupervised learning and supervised learning. First an unsupervised clustering algorithm is utilized to obtain the parameters of radial basis functions, i.e., width and the centers. Next, a supervised algorithm using least mean square error is performed to compute the weights of the connections between hidden nodes and the output nodes.

- Support Vector Machine: This classifier makes a nonlinear mapping from input space into a higher dimensional feature space. Then it tries to find the best hyper plane which maximizes the margin between two classes in the new space. Given a training set of m instance-target pairs (x_i, y_i), SVM solves the following optimization problem:[17]

$$\min_{w,b,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i$$

$$s.t \quad y_i(w^T \varphi(x_i) + b \geq 1 - \xi_i, \quad \xi_i \geq 0. \tag{8}$$

  φ is the mapping function and $C > 0$ is the penalty parameter of the error term. For nonlinear mapping, a kernel trick is used in which:

$$K(x_j, x_i) = \varphi(x_j)^T \varphi(x_i), \tag{9}$$

  where x_j is a support vector.

Some possible kernels are linear, polynomial, Gaussian, and hyper-tangent. Gaussian kernel is calculated as follows:

$$K(x_j, x_i) = \exp(-\gamma \left\| x_j - x_i \right\|^2), \gamma > 0. \tag{10}$$

The predicted target for instance x is generated by:

$$y(x) = F\left[\sum_{j=1}^{l} y_j \alpha_j K(x_j, x) + b \right], \tag{11}$$

where F is the sign function, l is the number of support vectors, $y_j \in \{-1, 1\}$ is the target for a binary classification, b and α_j are coefficients of the model determined through optimization process.

3.3. *Performance measures*

In order to evaluate the performance of the classification methods, we consider four basic concepts:

True Positive (TP): number of respondents *correctly* classified as respondents,

False Positive (FP): number of non-respondents *incorrectly* classified as respondents,

True Negative (TN): number of non-respondents *correctly* classified as non-respondents,

False Negative (FP): number of respondents *incorrectly* classified as non-respondents.

Accuracy is the most widely used measure for evaluating the performance of classification systems. This measure is as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{12}$$

However, when the data is imbalanced, this metric is not sufficient to reliably assess the classification performance. Here, TN is significantly greater than TP. For example, if there are 10 respondents in a test dataset with 1000 customer records, a classifier which predicts all 1000 customers not to respond, will have an accuracy of 99%. However, this classifier is not usable because it cannot detect any of the respondents. Therefore, we need a measure to evaluate the ability of predicting instance targets in both majority and minority classes.

The measure used in this study is balanced correction rate (BCR). If we consider true positive rate (TPR = TP/(TP + FN)) and true negative rate (TNR = TN/(TN + FP)), then BCR measure is the geometric mean of TPR and TNR as follows:

$$\text{BCR} = \sqrt{\text{TPR} \times \text{TNR}}. \tag{13}$$

Therefore, if the prediction accuracy in any of the two classes is low, it will result in a low BCR.

The receiver operating characteristic (ROC) curve demonstrates the performance of a binary classifier based on the relative changes of TPR and FPR values. A useful measure for evaluating the classification models is the area under ROC curve (AUC) which shows the degree of discrimination obtained from a particular model.[40] This measure is also independent from the imbalanced distribution in the two classes. For an ideal classifier AUC is equal to 1, while a random classifier will produce AUC = 0.5.

In direct marketing domain, a popular method for evaluating the performance of the response models is *Lift analysis*. Lift is a measure of effectiveness of a predictive model calculated as the ratio between the results obtained with or without the model.[29] We use cumulative gain and Lift charts as visual tools to assess the response models built in this study. Both charts consist of a lift curve and a baseline. A better model has a greater area between lift curve and the baseline. After producing the prediction outputs for the test data, the test instances are ordered in descent based on their predicted response probability. The ordered set is divided into deciles.

The cumulative gain chart plots the percentage of actual positive responses (y-axis) versus the percentage of customer population size in deciles (x-axis). The baseline shows a straight line in which X% customers will produce X% of the total responses. The Lift chart shows the actual Lift and is plotted by calculating the ratio between the results predicted by our model and the results using no model. For example, by choosing 10% of customers, using no model, we should receive 10% of respondents and using a good classifier, we may receive 30% of respondents. The y-value of the lift curve at $x = 10\%$ is $30/10 = 3$. In real situations, the cumulative gain chart and the lift chart for the response model help us determine how effectively we can reduce costs by selecting a relatively small number of customers and receiving a large portion of the respondents.

4. Experiments and Results

4.1. *Model implementation*

We implemented the proposed ensemble framework on the five type of classifiers introduced in the previous sections; DT, LR, MLP neural network, RBF neural networks, and SVM. The performance of the ensemble classification system is compared to that of the single classifiers, using the performance measures. In addition, the single models are implemented using random under-sampling methods, since the performance of classifiers with no balancing method will be worse. All models were executed 20 times and the average performance results are obtained. The experiments were performed in Matlab R2012a environment on a PC with Core i5 CPU and 4 GB of RAM and a 64 bit windows operating system.

Before training the models, we performed a data preprocessing scheme on both training and test data. First, the categorical features were converted to numeric. Then, we linearly scaled the feature values in the range [0, 1], in order to avoid attributes in greater numeric ranges dominating those in smaller ranges.

For some classifiers, there are a few parameters which should be determined by the user. In MLP networks, the number of hidden nodes (considering that we only used one hidden layer in this study) and in SVM, the Kernel function and misclassification penalty parameter, C should be selected. The best number of hidden nodes for MLP was selected by 10-fold cross validation from a set of candidate numbers (4, 6, 8, 10, 12, 14, 18) and was set to 10. For SVM, we used the Gaussian Kernel since it fits our problem, needs fewer parameters and has fewer numerical difficulties. A grid-search approach was performed on C and γ using 10-fold cross-validation. Various pairs of (C, γ) values were tried and the one with the best cross-validation accuracy was chosen. The best C and γ were found to be 100 and 2^{-4}, respectively.

Our proposed ensemble method was implemented by $N = 10$ base classifiers and the number of clusters for K-means clustering was determined based on the SD validity index and reasonably was set to $K = 5$.

4.2. *Results and analysis*

Table 2 shows the performance of the proposed Ensemble method using each of the five classification algorithms. The ensemble method implemented by each algorithm has been compared against a single classifier using the same algorithm. The single models were trained by normal random under sampling technique. We used Accuracy, BCR, and AUC for the main performance measures. Since the prediction of respondents in direct marketing is much more important than prediction of non-respondents, we also used TPR as a measure for evaluating the prediction ability of the models.

As can be seen in the table, the ensemble method combined by the data balancing procedure we proposed has made a significant improvement on the single models' performance. TPR for all ensemble classifiers has higher values than that of the single models. The amount of improvement in TPR for SVM (17%) and LR (12%) is significant. The ensemble method has raised the prediction accuracy of the models. All models show a relative improvement in accuracy, but the increase in accuracy for SVM is considerably more (about 7%). This can be explained by the TPR increase in the SVM ensemble. The level of improvement in BCR for SVM and RBF-NN is significantly high (about 6% and 4%, respectively). MLP-NN and DT ensemble slightly increase BCR compared to their single models, while BCR for LR does not change very much (actually it shows a very small reduction by 0.35% which is negligible). This is because ensemble of LR classifiers decreases the TNR and the level of BCR cannot grow. The AUC values for the classifiers have also been increased by the ensemble method. This demonstrates that ensemble of classifiers for response modeling can produce better classifiers. Obviously, the ensemble of classifiers shows a high improvement in the AUC values for SVM, RBF-NN, and DT (for about 9%, 6%, and 8%, respectively). Generally, SVM performance is increased by the ensemble method much more than the rest of the models. The MLP-NN seems to be less influenced by the proposed ensemble method. Nevertheless, the positive effect of cluster-based sampling method and ensemble framework for predicting responsive customers is obviously observable.

Table 2. Performance comparison of the proposed ensemble method using five classification algorithms.

Classifier	Method	TPR	Accuracy	BCR	AUC
SVM	Single	66.11	66.84	65.46	0.71
	Ensemble	83.23	73.61	71.52	0.80
MLP-NN	Single	83.03	63.93	60.14	0.69
	Ensemble	85.28	64.73	60.43	0.71
RBF-NN	Single	82.47	63.4	59.59	0.68
	Ensemble	86.95	67.16	64.12	0.74
DT	Single	79.59	60.08	55.91	0.59
	Ensemble	82.91	61.45	56.55	0.67
LR	Single	69.6	66.95	65.8	0.72
	Ensemble	81.54	67.46	65.45	0.74

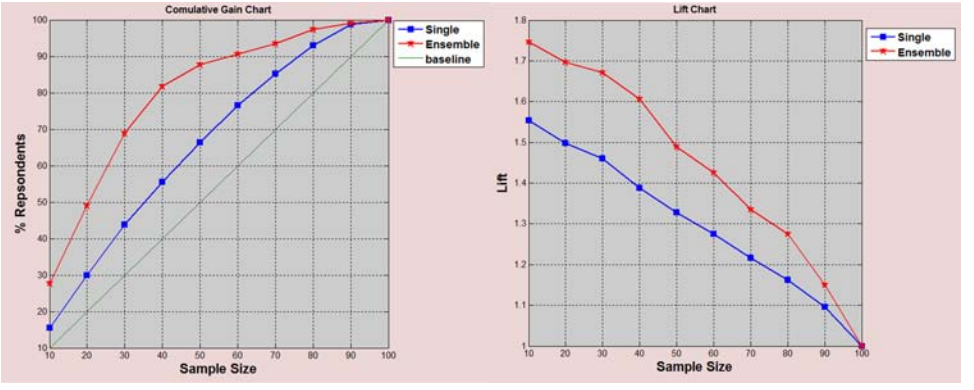

Fig. 2. Cumulative gain and Lift values for single and ensemble models using SVM.

For evaluating the effectiveness of response models built by the ensemble method we compared the single and ensemble models using their cumulative gains and Lift values. Figures 2–6 show the comparison of single and ensemble models using the cumulative gain charts and Lift charts.

A good response model has a cumulative gain chart leaning to the left of the figure and a Lift chart going to the top of the figure for the first deciles. This will help marketers to choose a smaller sample for telemarketing and catch more respondents for less cost.

Although SVM is a good classifier, the ensemble method has made most improvement on its prediction ability. Considering a sample size of 40% for contacts, SVM ensemble can target more than 80% of the respondents, while the single SVM only targets 56% of them, giving a 24% improvement. On this sampling size, the amount of improvement for predicting correct respondents in MLP-NN and RBF-NN is 17%, in DT is 7% and in LR is 16%. DT is the weakest classifier among the five models and the ensemble method has less effect on its performance.

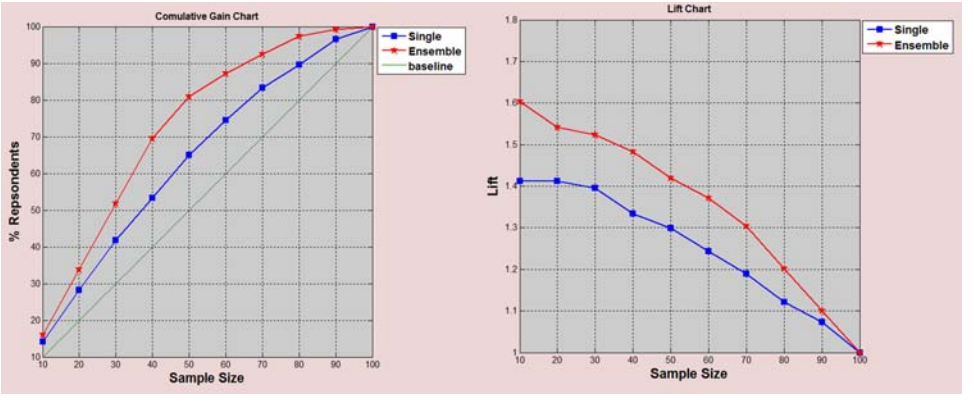

Fig. 3. Cumulative gain and Lift values for single and ensemble models using MLP-NN.

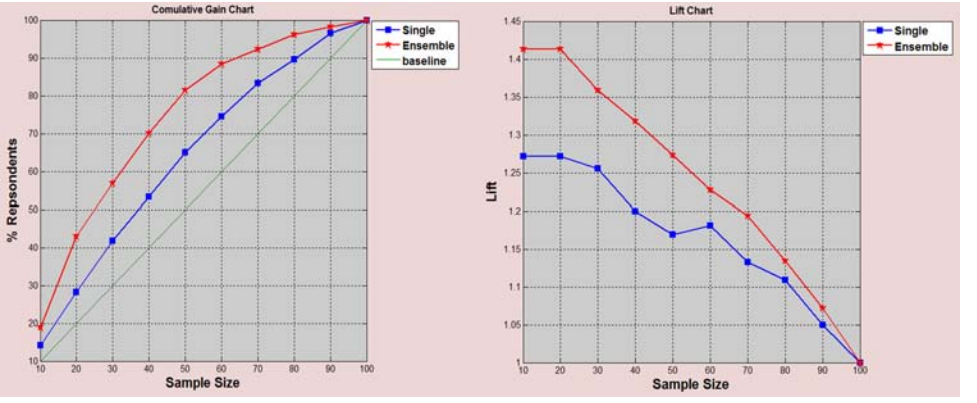

Fig. 4. Cumulative gain and Lift values for single and ensemble models using RBF-NN.

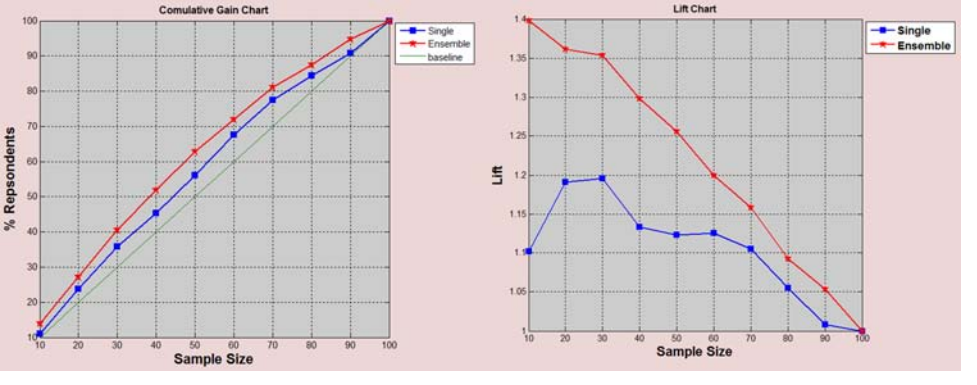

Fig. 5. Cumulative gain and Lift values for single and ensemble models using DT.

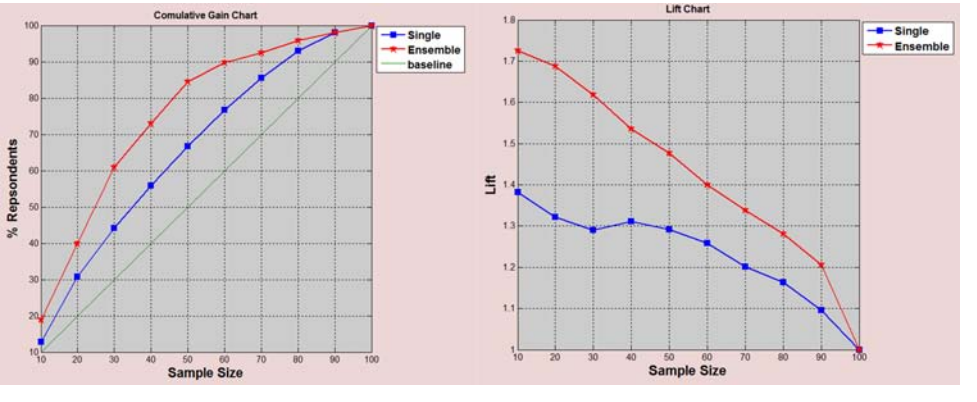

Fig. 6. Cumulative gain and Lift values for single and ensemble models using LR.

The same effect can be observed in the Lift charts. On the same sampling size 40%, the most amount of improvement is from SVM and LR ensembles with about 0.3 and 0.23, respectively. Ensemble of DTs provides 0.2 improvements on this sample size. It can be seen that the Lift Chart for DT has some fluctuations, but the ensemble Lift chart is smooth. Therefore, the ensemble system can increase the possibility of predicting real respondents for less numbers of customers contacted. This will lead to obtaining more profit with less contacting costs.

In order to propose the best response model using the examined ensemble methods, we provide cumulative gain and Lift charts of five ensemble classification systems built by these five base classifiers introduced so far. Figures 7 and 8 demonstrate the comparative cumulative gain and Lift charts for five models built by the ensemble of the single classifiers. Ensemble of SVM classifiers makes the strongest response model among the five ensembles. For the sample size less than 60%, SVM ensemble has better prediction performance and more respondents can be detected by this model. The two neural network models have a similar performance, although for a sample size less than 40% RBF ensemble performs slightly better. As can be seen, DT makes the weakest ensemble for respondent prediction. This happens generally, because DT is a simple classifier and cannot model complex relationship to a satisfactory degree. The data balancing method and ensemble cannot overcome the inherent weakness of DT.

On the Lift chart, SVM and LR show a better prediction performance compared to the other tree. From the sample size 20% up to 50% SVM is more advantageous than LR. Therefore, SVM ensemble can be proposed as the most effective and efficient classification system for response modeling in bank direct marketing.

The classifier ensemble proposed in this paper has better classification performance, stability, and response prediction compared to the single models. We used a

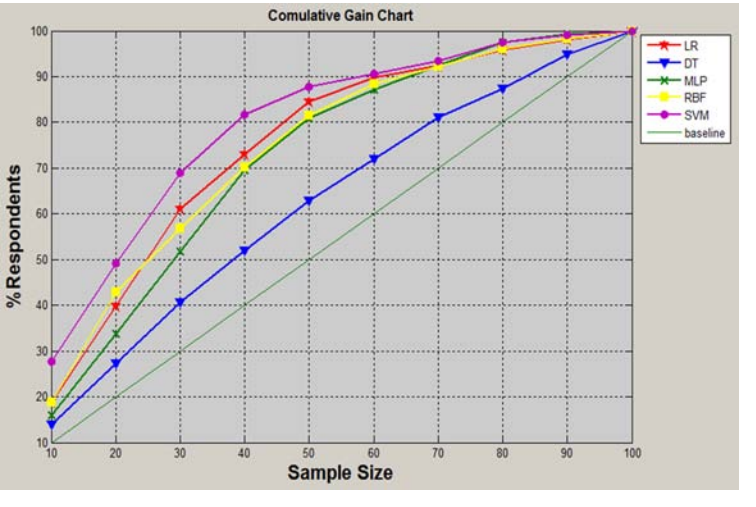

Fig. 7. Cumulative gain for five classifier ensembles.

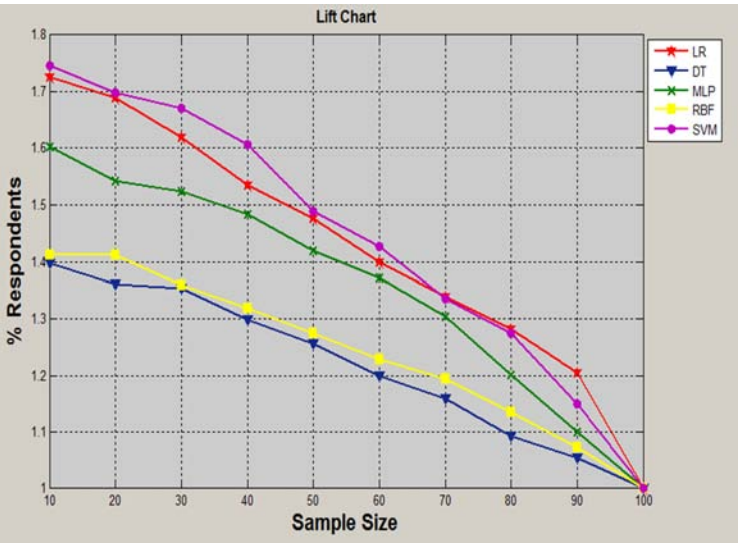

Fig. 8. Lift values for five classifier ensembles.

cluster-based data balancing sampling method for creating different samples for the base classifiers. In order to understand how effective this sampling method is, we compared the performance of our ensemble model with two other ensemble methods. One is the popular bagging method, in which bootstrap sampling (sampling with replacement) is done to create multiple samples with size n, when the size of the original sample is n.[5] It can be shown that each sample in this method will have 63.2% of the instances in the original dataset and as a result it will be clearly imbalanced. After training each base classifier, the final prediction for the ensemble is made by the majority voting scheme. Ha *et al.* used bagging of artificial neural networks for response modeling.[16] The other ensemble system is made by simple random under sampling for data balancing. N training samples are created for base classifiers using random under sampling. This way, we can evaluate the effect of clustering and sampling on data balancing method. In order to have a reasonable evaluation all three ensembles were created by the same base calssifiers, SVM.

Table 3 shows the measured performance of three mentioned ensemble methods: Ensemble with Cluster-based under-sampling method (CUS-Ensemble), ensemble

Table 3. Performance comparison of three ensemble methods using SVM base classifiers.

Method	TPR	Accuracy	BCR	AUC
CUS-Ensemble	83.23	73.61	71.52	0.82
RUS-Ensemble	65.75	67.14	67.17	0.74
Bagging	41.79	63.17	59.94	0.71

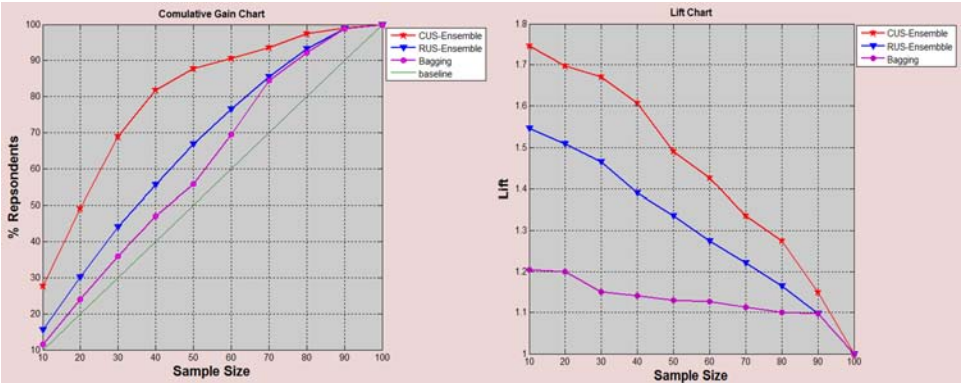

Fig. 9. Cumulative gain and Lift values for three ensemble methods.

with simple random under-sampling method (RUS-Ensemble) and the Bagging ensemble method.

Since SVM demonstrated the best performance for response modeling in this paper, we used this classifier to build these ensembles. In Table 3, the performance of our proposed ensemble methods is the best among the three methods.

The effect of clustering non-respondent customers and then sampling from the clusters will maintain the non-respondent distribution and improve the classification ability of the model. This makes our proposed method superior to RUS-ensemble. Bagging method has the worst prediction ability. Particularly, the TPR for bagging is considerably low. This is because the imbalanced data samples fed to the base classifiers will cause the base models to have bias towards non-respondent. As a result the TNR is increased and TPR will be reduced.

In Fig. 9, the cumulative gain and Lift charts of the three ensemble methods are presented. CUS-Ensemble is much more powerful than the other two ensemble method in predicting the actual respondents, especially for the sample size less than 60%. The cumulative gain and Lift values of the Bagging method is very low due to its inability to recognize true respondents since it has bias towards non-respondents.

5. Conclusion and Future Work

The efficiency and optimization of customer targeting is very necessary for direct marketing in banking industry since there is an increasing pressure to gain more profits and reduce costs. Accurate response models can help marketing managers detect potential customers and target them efficiently. Therefore, the accuracy of classification systems used for response modeling should be increased. On the other hand, data imbalance is a challenging problem in response modeling because it limits the performance of predicting models. In this paper, we proposed an ensemble method for response modeling to tackle the data imbalance problem and improve prediction performance. First, a data balancing method is used to create multiple

training samples using a clustering technique for dividing the non-respondents. Then, different training samples were given to base classifiers in an ensemble framework. In order to evaluate the performance of the proposed model, it was implemented on five different classification algorithms. The performance of five ensemble methods was compared against the corresponding single classifiers, using various performance measures applied in classification as well as direct marketing. The results analysis demonstrated that our proposed method improves the performance of response modeling in two ways: First, it increases the predictive accuracy by removing data imbalance in bank telemarketing data and keeping the customers characteristics simultaneously; second, it augments the profit obtained from direct marketing and reduces costs by predicting more actual respondents in the first deciles of customer sample size. Our ensemble method is also more powerful than the two alternative ensemble methods for response modeling.

New research directions can be followed after this study. It would be worthwhile to find new ways for combining base classifiers in order to maximize the ensemble prediction ability. Creating an ensemble using a combination of the different classifiers and devising innovative methods for their combination can be considered as an extension and future work for our research. Building response models for predicting the amount of money that each customer will deposit can be further investigated. Another future research direction is to make use of the proposed ensemble method for up-lift modeling. This means finding models which can predict customers who will subscribe a deposit only when they are targeted by the marketing campaign. In up-lift modeling we try to increase profit by diminishing the contacting costs through skipping the customers who will deposit anyway whether they are targeted or not.

References

1. D. A. Aaker, V. Kumar and G. S. Day *Marketing Research* (John Wiley & Sons, 2008).
2. B. Baesens, S. Viaene, D. Van den Poel, J. Vanthienen and G. Dedene, Bayesian neural network learning for repeat purchase modelling in direct marketing, *Eur. J. Oper. Res.* **138** (2002) 191–211.
3. P. Berkhin, A survey of clustering data mining techniques, in *Grouping Multidimensional Data* (Springer, 2006), pp. 25–71.
4. I. Bose and X. Chen, Quantitative models for direct marketing: A review from systems perspective, *Eur. J. Oper. Res.* **195** (2009) 1–16.
5. L. Breiman, Bagging Predictors, *Machine Learning.* **24** (1996) 123–40.
6. G. Brown, Diversity in Neural Network Ensembles, Ph.D thesis, The University of Birmingham (2004).
7. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, arXiv:1106.1813 (2011).
8. F. Coenen, G. Swinnen, K. Vanhoof and G. Wets, The improvement of response modeling: Combining rule-induction and case-based reasoning, *Expert Syst. Appl.* **18** (2000) 307–313.
9. R. Colombo and W. Jiang, A stochastic Rfm model, *J. Interact. Mark.* **13** (1999) 2–12.
10. T. G. Dietterich, Ensemble methods in machine learning, in *Multiple Classifier Systems* (Springer, 2000), pp. 1–15.

11. G. Dreyfus, *Neural Networks: Methodology and Applications* (Springer Heidelberg, 2005).
12. R. Elsner, M. Krafft and A. Huchzermeier, Optimizing rhenania's direct marketing business through dynamic multilevel modeling (DMLM) in a multicatalog-brand environment, *Mark. Sci.* **23** (2004) 192–206.
13. F. F. Gönül and F. T. Hofstede, How to compute optimal catalog mailing decisions, *Mark. Sci.* **25** (2006) 65–74.
14. F. F. Gönül, B.-D. Kim and M. Shi, Mailing smarter to catalog customers, *J. Interact. Mark.* **14** (2000) 2–16.
15. G. Guido, M. I. Prete, S. Miraglia and I. De Mare, Targeting direct marketing campaigns by neural networks, *J. Mark. Manage.* 27 (2011) 992–1006.
16. K. Ha, S. Cho and D. MacLachlan, Response models based on bagging neural networks, *J. Interact. Mark.* **19** (2005) 17–30.
17. D. J. Hand, H. Mannila and P. Smyth, Principles of data mining in *Adaptive Computation and Machine Learning* (The MIT Press, 2001).
18. J. A. Hartigan and M. A. Wong, Algorithm as 136: A K-means clustering algorithm, *Appl. Stat.* **28** (1979) 100–108.
19. T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman and R. Tibshirani, *The Elements of Statistical Learning*, Vol. 2. (Springer, 2009).
20. D. Haughton and S. Oulabi, Direct marketing modeling with Cart and Chaid, *J. Dir. Mark.* **7** (1993) 16–26.
21. D. W. Hosmer Jr and S. Lemeshow, *Appl. Logistic Regression* (John Wiley & Sons, 2004).
22. Y. Kim, W. N. Street, G. J. Russell and F. Menczer, Customer targeting: A neural network approach guided by genetic algorithms, *Manage. Sci.* 51 (2005) 264–276.
23. J. Kittler, M. Hatef, R. P. Duin and J. Matas, On combining classifiers, *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (1998) 226–239.
24. A. Krogh and J. Vedelsby, Neural network ensembles, cross validation, and active learning, *Adv. Neural Inf. Process Syst.* **7** (1995) 231–238.
25. M. Kubat, R. Holte and S. Matwin, Learning when negative examples abound in *Machine Learning: Ecml-97* (Springer, 1997), pp. 146–153.
26. K.-N. Lau, H. Chow and C. Liu, A database approach to cross selling in the banking industry: Practices, strategies and challenges, *J. Database Mark. Customer Strategy Manage.* **11** (2004) 216–234.
27. A. Lemmens and C. Croux, Bagging and boosting classification trees to predict churn, *J. Mark. Res.* **43** (2006) 276–286.
28. C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan and Q. Zou, Libd3c: Ensemble classifiers with a clustering and dynamic selection strategy, *Neurocomputing* **123** (2014) 424–435.
29. C. X. Ling and C. Li, Data mining for direct marketing: Problems and solutions, in *KDD* (1998), pp. 73–79.
30. D. Martens and F. Provost, Pseudo-social network targeting from consumer transaction data (2011).
31. S. Moro, P. Cortez and P. Rita, A data-driven approach to predict the success of bank telemarketing, *Decis. Support Syst.* **62** (2014) 22–31.
32. S. Moro, R. Laureano and P. Cortez, Using data mining for bank direct marketing: An application of the Crisp-Dm methodology, in *European Simulation and Modelling Conf. - ESM'2011*, ed. P. Novais *et al.* (2011), pp. 117–121.
33. S. A. Neslin, S. Gupta, W. Kamakura, J. Lu and C. H. Mason, Defection detection: Measuring and understanding the predictive accuracy of customer churn models, *J. Mark. Res.* **43** (2006) 204–211.
34. N. C. Oza and K. Tumer, Classifier ensembles: Select real-world applications, *Inf. Fusion* **9** (2008) 4–20.

35. R. Polikar, Ensemble learning, *Ensemble Machine Learning* (Springer, 2012), pp. 1–34.
36. A. Rahman and B. Verma, Novel layered clustering-based approach for generating ensemble of classifiers, *IEEE Trans. Neural Netw.* **22** (2011) 781–792.
37. C. Rao and J. Ali, Neural network model for database marketing in the new global economy, *Mark. Intell. Planning.* **20** (2002) 35–43.
38. R. E. Schapire, The strength of weak learnability, *Mach. Learn.* **5** (1990) 197–227.
39. H. Shin and S. Cho, Response modeling with support vector machines, *Expert Syst. Appl.* 30 (2006) 746–760.
40. M. Sokolova and G. Lapalme, A systematic analysis of performance measures for classification tasks, *Inf. Process. Manage.* **45** (2009) 427–437.
41. B. Sun, S. Li and C. Zhou, "Adaptive" learning and "Proactive" customer relationship management, *J. Inter. Mark.* 20 (2006) 82–96.
42. B. Thomas and M. Housden, *Dir. Mark. Pract.* (Routledge, 2002).
43. C.-F. Tsai, Combining cluster analysis with classifier ensembles to predict financial distress, *Inf. Fusion* **16** (2014) 46–58.
44. P. van der Putten, M. de Ruiter and M. van Someren, Coil challenge 2000 tasks and results: Predicting and explaining caravan policy ownership, *Coil Challenge.* 2000 (2000).
45. B. Verma and A. Rahman, Cluster-oriented ensemble classifier: Impact of multicluster characterization on ensemble classifier learning, *IEEE Trans. Knowl Data Eng.* **24** (2012) 605–618.
46. S. Viaene, B. Baesens, D. V. d. Poel, G. Dedene and J. Vanthienen, Wrapped input selection using multilayer perceptrons for repeat-purchase modeling in direct marketing, *Intelligent Systems in Accounting, Finance Manage.* **10** (2001) 115–126.
47. S. Viaene, B. Baesens, T. Van Gestel, J. A. K. Suykens, D. Van den Poel, J. Vanthienen, B. De Moor and G. Dedene, Knowledge discovery in a direct marketing case using least squares support vector machines, *Int. Intell. Syst.* **16** (2001) 1023–1036.
48. G. Wang, J. Hao, J. Mab and L. Huang, A new approach to intrusion detection using artificial neural networks and fuzzy clustering, *Expert Syst. Appl.* **37** (2010) 6225–6232.
49. J. Zahavi and N. Levin, Applying neural computing to target marketing, *J. Interact. Mark.* **11** (1997) 5–22.