Applied Mathematical Sciences, Vol. 8, 2014, no. 114, 5667 - 5672
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2014.47222

Feature Selection with Data Balancing for Prediction of Bank Telemarketing

Chakarin Vajiramedhin

Department of Innovation Management, Faculty of Management Science
Ubon Ratchathani University, Ubon Ratchathani, 34190, Thailand

Anirut Suebsing

Department of Innovation Management, Faculty of Management Science
Ubon Ratchathani University, Ubon Ratchathani, 34190, Thailand

Abstract

Nowadays, Telemarketing is an interactive technique of direct marketing that many banks apply to present a long term deposit to bank customers via the phone. Although the offering like this manner is powerful, it may make the customers annoyed. The data prediction is a popular task in data mining because it can be applied to solve this problem. However, the predictive performance may be decreased in case of the input data have many features like the bank customer information. In this paper, we focus on how to reduce the feature of input data and balance the training set for the predictive model to help the bank to increase the prediction rate. In the system performance evaluation, all accuracy rates of each predictive model based on the proposed approach compared with the original predictive model based on the truth positive and receiver operating characteristic measurement show the high performance in which the smaller number of features.

Keywords: Data Mining, Feature Selection, Prediction

1 Introduction

Telemarketing is an interactive technique of direct marketing that a telemarketer solicits prospective customers via the phone to make a sale of merchandise or service. The direct marketing is the marketing discovered pinpoint prospects for additional services based on the customer data collected in the database known as the database marketing. A database of potential customers can benefit greatly from the direct marketing such as communication, advertisement and analysis.

The most successful telemarketing is to focus on the quality of prospect data, attempting to predict the expected customers that have a higher probability to use the service by using data mining technique. To understand customer behavior, many banks have adopted the predictive technique based on the data mining to predict the customer data for classifying the customers before offering special services. The prediction or classification is the most important task in the data mining that is usually applied to classify the group of data [2]. Thus, many predictive models ([3], [4], [6], [7], [9], [10]) have been proposed that each model has its own advantages and disadvantages vary.

One important factor affecting the performance of the prediction is the number of input feature. Especially, the information of bank customer is that normally has many features thus it makes forecasting performance of the prediction decrease. In this paper, we focus on how to reduce the feature of input data for predictive model to help banks to increase the prediction rate that the goal of scrutinization is to identify a group of customers who have a high probability to subscribe to a long term deposit. In data mining based on the feature selection, it helps the banks to predict customers' profiles whether they can trust on the customers' profiles or not if they will offer any services for customer.

The remainders of this paper are organized as follows: Section 1 describes the research objectives. Section 2 reviews of the feature selection, the predictive model algorithms and dataset. Section 3 describes the methodology in this paper. Section 4 illustrates the experimental results and discusses the performance evaluation. Finally, the conclusion is discussed in Section 5.

2 Related Work

2.1 Feature Selection

Feature selection ([4], [9]) is a method where only the relevant features will be selected, discarding the irrelevant or weak features in the dataset. Minimum set of features, which is close enough to represent the original dataset, will be selected. The selected features will form the smallest size of dataset to enable an efficient result. Feature selection algorithms typically fall into two categories; filter and wrapper approach. Filter approach filters irrelevant features out keeping a good feature set before learning process. On the other hand, wrapper approach searches for a good feature set using a learning algorithm.

Utilizing filter approach to generate a feature set is generally faster than wrapper approach because filter approach uses heuristics based on general characteristics of the data rather than wrapping a learning algorithm into the selection process to evaluate the merit of feature subsets.

2.2 Classification

Classification [5] is one of the most popular data mining techniques. Examples of classification applications based on classification include pattern recognition, medical diagnosis, detecting faults in industry application, and classifying financial market trends. Classification is a process of learning a function mapping a data item into one of some predefined classes. Every classification based on supervised learning is given as input a set of samples consisting of vectors of attribute values and a corresponding class.

The input of a classification is a training set which each record consists of attributes and a class label. The target of classification is to build a classification model or function, called as a classifier, which uses for predicting a class of objects whose class label that is unknown. In other words, classification is the process of finding a model which describes and distinguishes data classes in order to employ the model to detect class label. The built model may be represented in diverse forms such as IF-THEN rules, neural networks or decision trees.

Table 1. Detail for the Bank Telemarketing Dataset

Num.	Attribute Name	Description	Type
1	Age	It is age of client.	Numeric
2	Job	It is type of client’s job.	Categorical
3	Marital	It is client’s marital status.	Categorical
4	Education	What is the highest education of client?	Categorical
5	Default	Does client has credit?	Categorical
6	Housing	Does client has housing loan?	Categorical
7	Loan	Does client has personal loan?	Categorical
8	Contact	What is a contact communication type of client?	Categorical
9	Month	What is the last month of the year contracting to the client?	Categorical
10	Day of Week	What is the last day of the week contracting to the client?	Categorical
11	Duration	How long does it contact to the client?	Numeric
12	Campaign	Number of contacts performed during this campaign and for this client	Numeric
13	Pdays	Number of days that passed by after the client was last contacted from a previous campaign	Numeric
14	Previous	Number of contacts performed before this campaign and for this client	Numeric
15	Poutcome	Outcome of the previous marketing campaign	Categorical
16	Emp.var.rate	Employment variation rate	Numeric
17	Cos.price.idx	Consumer price index	Numeric
18	Cons.conf.idx	Consumer confidence index	Numeric
19	Euribor3m	Euribor 3 month rate	Numeric
20	Nr.employed	Number of employees	Numeric
21	Label	Does the client has subscribed a term deposit?	Categorical

2.3 Data Set

This study considers real data provided by The UCI Machine Learning Repository [1]. This dataset was collected from a Portuguese retail bank [8], from May 2008 to June 2013, in a total of 41,188 phone contacts. The dataset is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The dataset is composed of 21 attributes including a label attribute shown in Table 1.

3 Proposed Approach

In this study, the proposed approach used to enhance the predictive rate of the bank telemarketing dataset is a correlation-based feature subset selection algorithm and a dataset balancing technique. The study uses the dataset balancing techniques to make the label of dataset equivalent before using correlation-based feature subset selection algorithm to select the robust feature.

In this paper, the dataset balancing technique is used for make the label of dataset equivalent by randomly selecting any data of each label out of a dataset equally. In meanwhile, the correlation-based feature subset selection algorithm is used to extract the robust features. The correlation-based feature subset selection algorithm is a heuristic for evaluating the worth or merit of a subset of features. This heuristic consider of individual features for predicting the class label including the level of inter correlation among them. Its hypothesis is "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other"[6]. In equation (1), it shows the equation of heuristic.

$$Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)*\overline{r_{ff}}}} \qquad (1)$$

Where $Merit_S$ is the heuristic merit of a feature subset S containing k features, $\overline{r_{cf}}$ is the average feature-class correlation, and $\overline{r_{ff}}$ is the average feature-feature inter correlation [10]. In fact, equation (1) is Pearson's correlation, where all features have been standardized.

4 Performance Evaluation

In order to evaluate our proposed approach, the data sets from the UCI repository [1] the bank telemarketing dataset and C4.5 algorithm are used. Moreover, 10-fold cross-validation which is Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model is used also.

Note that the measurement in this paper of the experimental results is based on the standard metrics for evaluations of accuracy rate for truth positive (TP) and receiver operating characteristic (ROC) known as a relative operating characteristic curve that is used for comparing two operating characteristics (TPR and FPR) as the criterion changes.

This section shows the accuracy rate and ROC rate of the proposed approach compared with 2 methods as follows:

1. Method 1 - entire features without using the dataset balancing technique and the feature selection algorithm.
2. Method 2 - the correlation-based feature subset selection algorithm without employing the dataset balancing technique.

Table 2. The TP Rate and ROC Rate of each method

Technique	No. of Feature	TP Rate (%)	ROC Rate (%)
Method 1	20	91.19	88.40
Method 2	3	91.26	91.00
Proposed method	8	92.14	95.60

From Table 2, it shows all accuracy rates each method that is quite high; however, the proposed method can provide the highest rate than the others. Moreover, the method 2 that used the same algorithm as the proposed method but it did not utilize the dataset balancing technique can give higher rate than the method 1. In the meanwhile, when their ROC rate of each method was considered, the proposed method can demonstrate it has more effective than any other methods because its ROC rate is highest while the method 1 gives the worst ROC rate. Therefore, the proposed method can enhance the performance of the predictive model effectively as the proposed method can provide the highest rates of TP rate and ROC rate.

5 Conclusion

This paper presents an improved predictive model by utilizing the feature selection and the dataset balancing technique for the bank telemarketing prediction. The experimental results showed that our proposed approach can improve the performance of the predictive model with the number of smaller features. Note that our proposed method can enhance the predictive model performance both of the TP rate and the ROC rate while it employs the smaller storage space, reduces the computation time and gains the higher predictive performance.

Acknowledgements. We would like to thank the Faculty of Management Science, Ubon Ratchathani University for the financial support. We are also grateful to the owner of the datasets, the UCI repository, for providing the available datasets.

References

[1] A. Asuncion, D. Newman, CA: University of California, School of Information and Computer Science, UCI machine learning repository. Irvine, 2012.

[2] A. Suebsing, C. Vajiramedhin, Accuracy rate of predictive models in credit screening, Applied Mathematical Sciences, vol. 7 no.112, 2013, 5591-5597.

[3] A. Suebsing, N. Hiransakolwong, A novel technique for feature subset selection based on cosine similarity, Applied Mathematical Sciences, vol.6, no. 133, 2012, 6627-6655.

[4] G. John, R. Kohavi, and Pfleger. Irrelevant features and the subset selection problem. Int. Conf. on Machine Learning, Morgan Kaufman, San Francisco. 1994, 121-129.

[5] I.H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2005.

[6] Mark A. Hall, Lloyd A. Smith, Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper, American Association for Artificial Intelligence, 1998.

[7] R. Singh, R. R. Aggarwal, Comparative Evaluation of Predictive Modeling Techniques on Credit Card Data. International Journal of Computer Theory and Engineering, 3 (2011), 598-603.

[8] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 2014, 62:22-31.

[9] W. Duch, T. Winiarski, J. Biesiada, and A. Kachel, Feature Ranking Selection and Discretization. Int. Conf. on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP), Istanbul. 2003, 251-254.

[10] Xinguo Lo and et al., A Novel Feature Selection Method Based on CFS in Cancer Recognition, IEEE 6th International Conference on System Biology IISB), 2012.

Received: August 1, 2014