On Identifying Potential Direct Marketing Consumers using Adaptive Boosted Support Vector Machine

Armin Lawi Department of Computer Science Universitas Hasanuddin
Makassar, Indonesia
armin@unhas.ac.id

Ali Akbar Velayaty*)
Post-graduate Program of Electrical Engineering, Universitas Hasanuddin
Makassar, Indonesia
arvasit@gmail.com

Zahir Zainuddin
Dept. Electrical Engineering
Universitas Hasanuddin
Makassar, Indonesia
zainuddinzahir@gmail.com

***Abstract*— Identifying potential consumers for direct marketing to very large data is a difficult and impossible task to do manually. Therefore, the machine learning approach needs to help analyze the data to contribute in determining the marketing strategy policy. In this paper, vector machine support methods using the Adaboost algorithm are investigated to classify potential customers for direct marketing of large bank data marketing. The adaboost algorithm aims to build a better model than the model generated from a single classifier. Data obtained from UCI machine learning repository. Total data is processed as many as 9280 data with the number of attributes of 20 classes and 1 target class. Training data and test data are divided into 70% and 30%. This classification predicts the prospects for a deposit subscription. The results show that the SVM method using Adaboost algorithm obtained accuracy is 95.07% and the sensitivity is 91.65% higher than the ordinary SVM approach.**

Keywords—Bank Direct Marketing; Support Vector Machine; Adaboost.

I. INTRODUCTION

Business Intelligence is a term related to the use of information space and Intelligence mechanisms to support a decision [1]. One sector that implements this is the banking sector that receives and processes information from customers every day. One area of banking that collects information from customers is the marketing sector. There are two ways of banking in introduce the goods or services that is with the general advertising either through television, radio, newspapers, etc. or by targeting customers in particular or called the bank direct marketing [2].

Direct marketing is a marketing technique applied to the bank to introduce a product either in the form of goods or services to customers by phone, email and others [3]–[5]. However, the disadvantages of direct marketing are those customers who feel disturbed that can cause negative ratings for the bank and the cost and time spent by telemarketers is not proportional to the number of customers contacted [2]

Over time, the amount of data that contains information from customers will increase. The increasing amount of data can be used to find information that serves to assist in making a decision. One of the computational methods that discuss the above problem is Data Mining [1].

Data mining is a field in the computational process that allows finding a pattern of a data that is often ignored by using a particular method. The pattern can be used to help take in a decision. One method of data mining is a classification method that is a technique of grouping data obtained based on the pattern produced previously. There are several models of classification among them is Support Vector Machine (SVM) method. Support Vector Machine is the method that forms the best separator in hyperplane field on the data that separates the two classes by maximizing the margin value of the two classes.

According to [6] Although ordinary SVM shows good performance in classification, but is sensitive to sample and parameter settings. Some research indicates that the accuracy and performance of ordinary SVM can be improved by using the Ensemble method [7]. The Ensemble method is a method that combines the predictions generated by some classifier. One of the Ensemble Methods of a classification is boosting algorithm. which is more popular used compared to bagging algorithm [7]. So, in this research will compare the performance of ordinary SVM with SVM using adaboost algorithm.

II. LITERATURE

As for some related research that discusses the predictions of potential customers to subscribe deposits are as follows. Moro (2011), et al. [2] proposed Cross Industry Standard Process for Data Mining (CRISP-DM) method indicates that the support vector machine gets better AUC and ALIFT results than the Decision Tree and Naive Bayes methods. Feng, et al. [7] proposed a Bayesian networks method and their application shows that a single support vector machine method only produces an average level of accuracy that is not much different than other classification methods, and thus they conducted ensemble method to improve the accuracy of the classification method. Moro (2014), et al. [8] proposed a data-driven approach to predict the success of bank telemarketing by comparing four methods of data mining classification, i.e., Decision Tree, Neural Networks, Support Vector Machine and Logistic Regression. Their result showed that neural network is the best

*) Corresponding Author: arvasit@gmail.com

performance, and support vector machines and neural networks are more flexible and have better learning skills among other classification methods. Alhakbani and Rifaie [9] proposed a handling class imbalance in direct marketing dataset using a hybrid data. An algorithmic level solution using the Syntethic Minority oversampling technique is implemented to balance the dataset and then use grid search to optimize the *c*, gamma, and kernel parameters. The models used the support vector machine to build the HybridDA model. The results obtained indicate that the technique is better than the techniques used by previous research.

III. PRELIMINARIES

The datasets used in this paper is taken from the University of California at Irvine (UCI) Machine Learning Repository. It is a bank marketing data collected from March 2008 to November 2010. The number of attributes is 20 attributes with the addition of 1 target class, and the total observation is 41,188 data [8].

Table I. Attributes Description

Name	**Type**
Age	Numeric
Job	Categorical : "Admin.", "Blue-Collar", "Entrepreneur", "Housemaid", "Management", "Retired", "Self-Employed", "Services", "Student", "Technician", "Unemployed", "Unknown"
Marital	Categorical : "Divorced", "Married", "Single", "Unknown";
Education	Categorical : "Basic.4y", "Basic.6y", "Basic.9y", "High.School", "Illiterate", "Professional.Course", "University.Degree", "Unknown"
Default	Categorical: "No","Yes","Unknown"
Housing	Categorical: "No","Yes","Unknown"
Loan	Categorical: "No","Yes","Unknown"
Contact	Categorical: "Cellular","Telephone"
Month	Categorical: "Jan", "Feb", "Mar", ..., "Nov", "Dec"
Day_Of_Week	Categorical: "Mon", "Tue", "Wed", "Thu", "Fri"
Duration	Numeric
Campaign	Numeric
Pdays	Numeric
Previous	Numeric
Poutcome	Categorical : "Failure", "Nonexistent", "Success"
Emp.Var.Rate	Numeric
Cons.Price.Idx	Numeric
Cons.Conf.Idx	Numeric
Euribor3m	Numeric
Nr.Employed	Numeric
Y	Binary: "Yes","No"

A. Support Vector Machine

Support Vector Machine (SVM) is one of the methods used in the problem of pattern classification (pattern classification). The basic idea of the SVM method is to maximize the boundary-field separator (hyperplane) that separates the data into two classes in a feature space. Support vector machine method is suitable for use on linearly separable data [10].

To classifying the data cannot be separated linearly and it should be transformed into feature space dimension such that the data can be separated linearly from the feature space. Feature space in practice usually has a higher dimension of input space. Suppose there is a dataset whose data has two attributes and two classes of positive and negative classes when depicted in two-dimensional space of data cannot be separated linearly. Therefore, the data is transformed to a higher dimension so that it can be separated linearly. This method is called kernel methods, which transform a linear SVM into non-linear SVM

B. Adaboost

Boosting is a common and effective method for building an accurate classifier by combining weak classifiers [7] The use of boosting is preferred because it focuses on misclassified issues and has a tendency to increase higher accuracy compared to bagging. The commonly used boosting algorithm is the adaboost algorithm

Adaboost trains the basic classifier sequentially (iteratively) each iteration. The basic classification is trained by using data training with weight coefficients that depend on the classifier performance. On the previous iteration to give greater weight to the wrongly classified data (misclassified). If the classifier has been trained as much as desired, then the classifier is combined to form a final decision on the model that shows the best performance

The steps of the Adaboost algorithm are [11]

1. Input : Training data along with its label $\{(\boldsymbol{x_1}, \square_1), \ldots, (\boldsymbol{x_N}, \square_N)\}$, Component Learner, And the number of iterations of T.
2. Initialization of training data weight
$$D_{\square}^{1} = \frac{1}{N}, i = 1, \ldots, N. \quad (1)$$
3. For iteration $\square = 1, \ldots, T$
   a. Use the Component Learner algorithm to train a classification component, h_t, On training weights.
   b. Calculate the weight of classification error on h_t
$$\epsilon_t = \sum_{\square=1}(D_{\square} \quad \square_{\square} \neq h_t(\square))). \quad (2)$$
$$N \qquad t($$
   The learning trust index is calculated as:
$$c_t = \frac{1}{2} ln\,(\frac{(1-\epsilon_t)}{\epsilon_t}). \quad (3)$$
   c. Renew the sample training weight
$$D_{\square}^{t+1} = \frac{D_i{}^{t} \exp(c_t \times y_i \neq y_i^*)}{\sum_{i=1}^{N} D_i^{t}}. \quad (4)$$
   d. Testing model with testing data

4. The last learning output
   Combination of all classifications

$$h_{□} = □ign(\sum_{□=1}^{T} c_t\, h_t(□)). \quad (5)$$

C. Performance Evaluation

The performance result of the proposed method is evaluated based on the degree of accuracy and sensitivity. The measurement evaluation is obtained from the confusion matrix which consists of 4 parts of the classification results, i.e., True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The confusion matrix is given in Table II.

Table II. *Confusion Matrix/ Contingency Table*

Actual \ Prediction	**True**	**False**
True	TP	FN
False	FP	TN

The accuracy level is defined as the degree of proximity between the predicted data with the actual data or the ratio of the amount of data that is correctly classified as shown by equation (6).

$$\text{Accuracy rate} = \frac{TP+TN}{TN+FP+FN+TP}. \quad (6)$$

The sensitivity level is defined as the ratio of data classified positively [12]given by the equation (7).

$$\text{Sensitivity} = \frac{TP}{FN+TP}. \quad (7)$$

IV. Experimental Design

A. Dataset

The total of 41,188 data will gets better accuracy results; however, since the amount of data on the target deposit subscribed customers is only 4,640 data which is less than compared with customers who do not subscribed of 36,548 data. This sitation will tend to predict the new data more incline to deposit unsubscribe costumer, and therefore, in this experiment, the amount of data is balanced to obtain the same probability outcome. Hence, in this experiment use 9,280 total data with the amount of comparison of training data is 70% and testing data is 30%.

B. Normalization data

For each attribute of type "categorical" will be changed to "numerical". This is because data other than the "numerical" type cannot be processed. The next type of target is changed to 1 for "yes" class and -1 for "No" class

C. Implementation

In this research will compare the performance results from a single vector support method with support vector machine using the adaboost algorithm.

V. Results

The performance comparison result of performance of ordinary SVM and adaboost SVM is given in Table III.

Table III. *Classification results of SVM Adaboost*

Method	Iteration	Accuracy	Sensitivity
Ordinary SVM	-	91,67 %	83,80 %
Adaboost SVM	10	92,53 %	85,93 %
	20	94,30 %	89,68 %
	30	**95,07 %**	**91,65 %**
	40	94,81 %	91,05 %
	50	94,86 %	91,05 %

Table III shows that the adaboost SVM method gives higher accuracy rate since the algorithm updates the weights of incorrect classified data to the specified iteration limit. The table indicates that the increasing number of iterations is not guaranteed the performance obtained will be better. In this result, the iteration 30th gives the best performance.

VI. Conclusion

In this paper proposed adaboost SVM method to improve prediction results of direct marketing of bank dataset compare to the ordinary SVM. The Adaboost algorithm successfully improves the performance of the SVM method. The results obtained shows that the accuracy of the ordinary SVM only get 91.67% and sensitivity of 83.80%, whereas our proposed adaboost SVM gives accuracy to 95,07% and sensitivity 91,65% in the 30th iteration.

References

[1] S. Abbas, "Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset," *Int. J. Comput. Appl.*, vol. 110, no. 3, pp. 975–8887, 2015.

[2] S. Moro and R. M. S. Laureano, "Using Data Mining for Bank Direct Marketing: An application of the CRISP-DM methodology," *Eur. Simul. Model. Conf.*, no. Figure 1, pp. 117–121, 2011.

[3] C. Vajiramedhin and A. Suebsing, "Feature Selection with Data Balancing for Prediction of Bank Telemarketing," *Appl. Math. Sci.*, vol. 8, no. 114, pp. 5667–5672, 2014.

[4] H. A. Elsalamony, "Bank Direct Marketing Analysis of Data Mining Techniques," *Int. J. Comput. Appl.*, vol. 85, no. 7, pp. 12–22, 2014.

[5] H. Elsalamony and A. Elsayad, "Bank Direct Marketing Based on Neural Network," *Int. J. Eng. Adv. Technol.*, vol. 2, no. 6, pp. 392–400, 2013.

[6] L. Zhou, K. K. Lai, and L. Yu, "Least squares support vector machines ensemble models for credit scoring," *Expert Syst. Appl.*, vol. 37, no. 1, pp. 127–133, 2010.

[7] G. Feng, J.-D. Zhang, and S. Shaoyi Liao, "A novel method for combining Bayesian networks, theoretical analysis, and its applications," *Pattern Recognit.*, vol. 47, no. 5, pp. 2057–2069, 2014.

[8] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decis. Support Syst.*, vol. 62, pp. 22–31, 2014.

[9] H. A. Alhakbani and M. M. Al-Rifaie, "Handling Class Imbalance in Direct Marketing Dataset using A Hybrid Data and Algorithmic Level Solutions," *SAI Comput. Conf. 2016*, pp. 1–6, 2016.

[10] C. J. C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[11] Y. Freund, “A more robust boosting algorithm,” vol. arXiv:0905, 2009.

[12] G. Nisbet, R., Elder, J., Miner, *Handbook of statistical analysis and data mining applications*. 2009.