Advances in Dynamical Systems and Applications.
ISSN 0973-5321, Volume 16, Number 2, (2021) pp. 1269-1281

https://www.ripublication.com/adsa.htm

Analysis of Homogeneous and Heterogeneous Ensemble Classifiers for Bank Marketing

M. Govindarajan

Associate Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar – 608002, Tamil Nadu, India.

Abstract

In the era of Internet and big data, the bank has gradually realized that the traditional data analysis cannot meet the demands of the existing marketing. So the bank direct marketing based on machine learning emerges. However, there are few references which are completely based on ensemble learning. As different banks have different structures of customer data, the existing model cannot be employed directly. Therefore, new ensemble classification methods are proposed with homogeneous ensemble classifier using bagging and heterogeneous ensemble classifier using arcing and their performances are analyzed in terms of accuracy. A Classifier ensemble is designed using Radial Basis Function (RBF) and Support Vector Machine (SVM) as base classifiers. The feasibility and the benefits of the proposed approaches are demonstrated by the means of dataset of direct marketing. The main originality of the proposed approach is based on three main parts: pre-processing phase, classification phase and combining phase. A wide range of comparative experiments are conducted for standard dataset of direct marketing. Furthermore, comparisons with prior work on standard dataset of direct marketing are also listed. The results illustrate that the proposed ensemble methods are competitive.

Keywords: Accuracy, Arcing, Bagging, Ensemble, Radial Basis Function, Support Vector Machine.

1. INTRODUCTION

With the trend of digital transformation, the bank has gradually realized that the traditional data analysis cannot meet the new marketing demands, and the bank

urgently needs direct marketing driven by big data. Under this background, bank direct marketing based on machine learning emerges. Direct marketing mode can insight the potential requirements and preferences of customers and help banks obtain target customer groups. On one hand, the application of machine learning in bank direct marketing can improve the accuracy of the bank marketing, on the other hand, it can also increase the number of customers. The goal of this work is to determine if it is possible to improve the performance of the data mining classification algorithm used for customer marketing response prediction through the use of homogeneous ensemble classifier using bagging and heterogeneous ensemble classifier using arcing. The work will rely on customers' bank transaction data aggregated electronically through the CRM process. The contributions of the paper are as follows:

(i) In this research work, new ensemble classification methods are proposed with homogeneous ensemble classifier using bagging and heterogeneous ensemble classifier using arcing and their performances are analyzed in terms of accuracy.

(ii) A Classifier ensemble is designed using Radial Basis Function (RBF) and Support Vector Machine (SVM) as base classifiers. The main originality of the proposed approach is based on three main parts: pre-processing phase, classification phase and combining phase.

(iii) The accuracy of base classifiers is compared with homogeneous and heterogeneous models for direct marketing.

(iv) The proposed ensemble methods provide significant improvement of accuracy compared to individual classifiers and heterogeneous models exhibit better results than homogeneous models.

(v) Furthermore, comparisons with prior work on standard dataset of direct marketing are also listed. The results illustrate that the proposed ensemble methods are competitive.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 presents proposed methodology and Section 4 explains the performance evaluation measures. Section 5 focuses on the experimental results and discussion. Finally, results are summarized and concluded in section 6.

2. RELATED WORK

In the field of direct marketing lot of research has been done in which many techniques are covered and still many remains to be covered.

Fatemeh Nemati Koutanaei et al. (2015) have developed a hybrid data mining model of feature selection and ensemble learning classification algorithms on the basis of three stages. The first stage, as expected, deals with the data gathering and pre-processing. In the second stage, four FS algorithms are employed, including principal component analysis (PCA), genetic algorithm (GA), information gain ratio, and relief attribute evaluation function. In here, parameters setting of FS methods is based on the classification accuracy resulted from the implementation of the support vector

machine (SVM) classification algorithm. After choosing the appropriate model for each selected feature, they are applied to the base and ensemble classification algorithms. In this stage, the best FS algorithm with its parameters setting is indicated for the modeling stage of the proposed model. In the third stage, the classification algorithms are employed for the dataset prepared from each FS algorithm. The results exhibited that in the second stage, PCA algorithm is the best FS algorithm. In the third stage, the classification results showed that the artificial neural network (ANN) adaptive boosting (AdaBoost) method has higher classification accuracy.

Ashkan Zakaryazad and Ekrem Duman (2016) introduced an ANN model with a new penalty function which gives variable penalties to the misclassification of instances considering their individual importance (profit of correctly classification and/or cost of misclassification) and then considered maximizing the total net profit. The effectiveness of the proposed models is appraised on two real-life data sets from fraud detection and a University of California Irvine (UCI) repository data set about bank direct marketing.

Joaquín Abellán and Javier G. Castellano (2017) extended a previous work about the selection of the best base classifier used in ensembles on credit data sets. It is shown that a very simple base classifier, based on imprecise probabilities and uncertainty measures, attains a better trade-off among some aspects of interest for this type of studies such as accuracy and area under ROC curve (AUC). The results shown here present to this simple classifier as an interesting choice to be used as base classifier in ensembles for credit scoring and bankruptcy prediction, proving that not only the individual performance of a classifier is the key point to be selected for an ensemble scheme.

Arno De Caigny et al. (2018) proposed a new hybrid algorithm, the logit leaf model (LLM), to better classify data. This new hybrid approach is benchmarked against decision trees, logistic regression, random forests and logistic model trees with regards to the predictive performance and comprehensibility.

Stefan Lessmann et al. (2019) proposed profit-conscious ensemble selection, a modeling framework that integrates statistical learning principles and business objectives in the form of campaign profit maximization. Studying the interplay between data-driven learning methods and their business value in real-world application contexts, the paper contributes to the emerging field of profit analytics and provides original insights how to implement profit analytics in marketing. The paper also estimates the degree to which profit-concious modeling adds to the bottom line.

Ashalata Panigrahi et al. (2020) proposed neural network based techniques to construct a decision making model using six classifiers, namely, SMO, SVM, RBFN, MP, SOM, and HLVQ. Further, the most relevant attributes in the input data have been selected through a preprocessing stage using three attribute evaluator methods namely, filtered attribute evaluator, one-R attribute evaluator, Relief-F attribute evaluator.

Inspired by the above studies, a hybrid system is proposed using radial basis function and support vector machine and the effectiveness of the proposed bagged RBF,

bagged SVM and RBF-SVM hybrid system is evaluated by conducting several experiments on bank marketing dataset. The performance of the proposed bagged RBF, bagged SVM, and RBF-SVM hybrid classifiers are examined in comparison with standalone RBF and standalone SVM classifier and also heterogeneous models exhibits better results than homogeneous models and comparisons with prior work on standard dataset of direct marketing are also listed.

3. PROPOSED METHODOLOGY

3.1 Preprocessing

In preprocessing of dataset, cleaning and transformation are performed. Cleaning process means removing the redundant labels and filling missing value in the dataset. Transformation means translate full data set into the desired form (it means convert numeric value to the string type data).

3.2 Existing Classification Methods

3.2.1 Radial Basis Function Neural Network

This is an artificial neural network formulated by Broomhead and Lowe (1988). RBF uses radial basis functions for activation to change along the distance from a location. For functional approximation, it uses time-series prediction, classification, and system control. A multi-layer feed forward neural network, RBF is used to classify data in a non-linear mode and compare input data with training data. The production of the RBF neural network is weighted linear superposition of all basis functions. The frequently used basis function in the RBF model is the Gaussian basis function.

3.2.2 Support Vector Machine

This is widely used for training SVMs and was formulated by J. Platt (1998). SMO is one way to solve a quadratic programming (QP) issue that arises during SVM training. SMO divides the large QP problem into a series of very tiny sub-problems. These small sub-problems are solved analytically, preventing the use of time-consuming numerical QP optimization as an inner loop. It is the fastest for linear SVMs and sparse datasets and can be more than 1000 times faster than the chunking algorithm. The amount of memory needed for SMO is linear in the training dataset size, allowing SMO to handle very large training sets. It scales somewhere between linear and quadratic in the training set size for several test problems.

3.3 Homogeneous Ensemble Classifiers

3.3.1 Proposed Bagged RBF and SVM Classifiers

Given a set D, of d tuples, bagging (Breiman, L. 1996a) works as follows. For iteration i (i =1, 2,…..k), a training set, D_i, of d tuples is sampled with replacement from the original set of tuples, D. The bootstrap sample, D_i, created by

sampling D with replacement, from the given training data set D repeatedly. Each example in the given training set D may appear repeatedly or not at all in any particular replicate training data set D_i. A classifier model, M_i, is learned for each training set, D_i. To classify an unknown tuple, X, each classifier, M_i, returns its class prediction, which counts as one vote. The bagged RBF and SVM, M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: RBF and SVM ensemble classifiers using bagging

Input:

- D, a set of d tuples.
- $k = 2$, the number of models in the ensemble.
- Base Classifiers (Radial Basis Function, Support Vector Machine)

Output: Bagged RBF and SVM, M^*

Method:

(1) for $i = 1$ to k do // create k models

(2) Create a bootstrap sample, D_i, by sampling D with replacement, from the given training data set D repeatedly. Each example in the given training set D may appear repeated times or not at all in any particular replicate training data set D_i

(3) Use D_i to derive a model, M_i;

(4) Classify each example d in training data D_i and initialized the weight, W_i for the model, M_i, based on the accuracies of percentage of correctly classified example in training data D_i.

(5) endfor

To use the bagged RBF and SVM models on a tuple, X:

1. if classification then
2. let each of the k models classify X and return the majority vote;
3. if prediction then
4. let each of the k models predict a value for X and return the average predicted value;

3.4 Heterogeneous Ensemble Classifiers

3.4.1 Proposed RBF-SVM Hybrid System

Given a set D, of d tuples, arcing (Breiman. L, 1996) works as follows; For iteration i (i =1, 2,…..k), a training set, D_i, of d tuples is sampled with replacement from the original set of tuples, D. some of the examples from the dataset *D* will occur more than once in the training dataset D_i. The examples that did not make it into the

training dataset end up forming the test dataset. Then a classifier model, M_i, is learned for each training examples *d* from training dataset D_i. A classifier model, M_i, is learned for each training set, D_i. To classify an unknown tuple, X, each classifier, M_i, returns its class prediction, which counts as one vote. The hybrid classifier (RBF-SVM), M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: Hybrid RBF-SVM using Arcing Classifier

Input:

- D, a set of d tuples.
- $k = 2$, the number of models in the ensemble.
- Base Classifiers (Radial Basis Function, Support Vector Machine)

Output: Hybrid RBF-SVM model, M^*.

Procedure:

1. For i = 1 to k do // Create k models
2. Create a new training dataset, D_i, by sampling D with replacement. Same example from given dataset *D* may occur more than once in the training dataset D_i.
3. Use D_i to derive a model, M_i
4. Classify each example d in training data D_i and initialized the weight, W_i for the model, M_i, based on the accuracies of percentage of correctly classified example in training data D_i.
5. endfor

To use the hybrid model on a tuple, X:

1. if classification then
2. let each of the k models classify X and return the majority vote;
3. if prediction then
4. let each of the k models predict a value for X and return the average predicted value;

The basic idea in Arcing is like bagging, but some of the original tuples of D may not be included in Di, where as others may occur more than once.

4. PERFORMANCE EVALUATION MEASURES

4.1 Cross Validation Technique

Cross-validation is a statistical technique which involves partitioning the data into subsets, training the data on a subset and use the other subset to evaluate the model's performance.

4.2 Criteria for Evaluation

Accuracy is one of the essential measures for describing the performance of any algorithm. It is the degree to which an algorithm can properly predict positive and negative instances, and it can be determined by the following formula: Accuracy = TP + TN/TP + FN + FP + TN.

5. EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Bank Marketing dataset Description

The data 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 (or not) subscribed. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Table 1: Properties of Direct Marketing Dataset

Datasets	Instances	Attributes
Bank Marketing	600	11

5.2 Experiments and Analysis

5.2.1 Performance comparison of the Homogeneous and Heterogeneous Ensemble Classifiers

In this section, new ensemble classification methods are proposed using classifiers in both homogeneous ensembles using bagging and heterogeneous ensembles using arcing classifier and their performances are analyzed in terms of accuracy.

Table 2: Performance Comparison of the Homogeneous and Heterogeneous Ensemble Classifiers for Direct Marketing dataset

Dataset	Classifiers	Classification Accuracy
Bank Marketing	RBF	71.16 %
	Proposed Bagged RBF	76.16 %
	SVM	69.00 %
	Proposed Bagged SVM	73.33 %
	Proposed Hybrid RBF-SVM	88.33 %

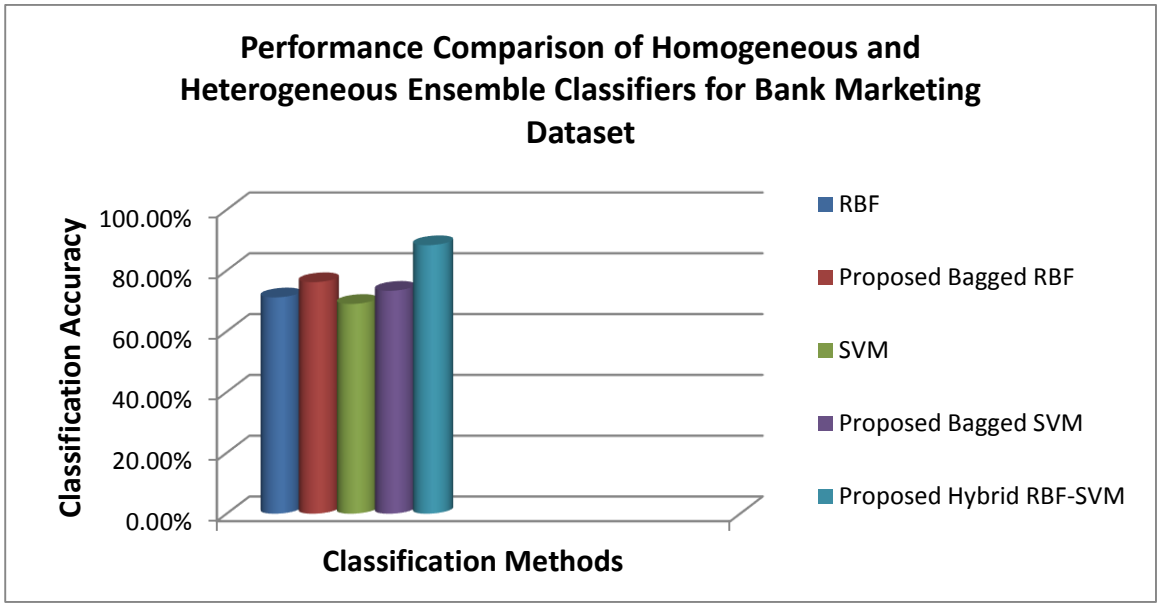

Figure 1: Accuracy for Homogeneous and Heterogeneous Ensemble Classifiers in Bank Marketing dataset

A wide range of comparative experiments are conducted for bank marketing dataset. The accuracy of base classifiers is compared with homogeneous and heterogeneous models for bank marketing dataset as given in table 2. According to Figure 1, the proposed hybrid methods provide significant improvement of accuracy compared to individual classifiers and also heterogeneous models exhibits better results than homogeneous models.

5.2.2 Performance comparison with prior research work

Table 3: Experimental Results for Homogeneous and Heterogeneous Ensemble Classifiers for Bank Marketing Dataset

Techniques	**Accuracy Claimed**
RBF	71.16%
SVM	69.00%
Homogeneous Ensemble Classifiers	
Proposed Bagged RBF	**76.16 %**
A.I. Marqués et al., 2012	71.25%
Ali AghaeiRad et al., 2017	71.48%
Rogelio A. Mancisidor et al., 2018	73.66%
Debjyoti Das Adhikary et al., 2019	71.24%
Khairul Nizam Abd Halim et al., 2020	75.45%

Proposed Bagged SVM	**73.33%**
A.I. Marqués et al., 2012	72.92%
Mohammad Amini et al.,2015	63.17%
Stamatis Karlos et al., 2016	66.67%
Bing Zhua et al., 2017	72.79%
Uma R. Salunkhe et al., 2018	72.00%
Debjyoti Das Adhikary et al., 2019	71.37%
Maisa Cardoso Aniceto et al., 2020	63.50%
Heterogeneous Ensemble Classifiers	
Proposed Hybrid RBF-SVM	**88.33%**
V. Ravi, et al, 2008	83.25%
Fatemeh Nemati Koutanaei et al., 2015	85.90%
Ali AghaeiRad et al., 2017	71.79%
Yajiao Tang at al., 2018	85.27%
Nazeeh Ghatasheh et al., 2020	84.18%
Nam N. Nguyen et al., 2021	87.03%

Finally, Table 3 shows a performance comparison of the proposed method with some recent hybrid methods using the bank marketing dataset in terms of accuracy. It can be seen from the table that higher accuracy is achieved with the homogeneous and heterogeneous models in comparison with prior work on the direct marketing. In addition to the performance analysis, the statistical significance tests prove that the better performance of the proposed classifier is statistically significant when compared to state of the art techniques.

6. CONCLUSION

In this research work, new combined classification methods are proposed using classifiers in homogeneous ensembles using bagging and the performance comparisons have been demonstrated using bank marketing dataset in terms of accuracy. Here, the proposed bagged radial basis function and bagged support vector machine combines the complementary features of the base classifiers. Similarly, new hybrid RBF-SVM models are designed in heterogeneous ensembles involving RBF and SVM models as base classifiers and their performances are analyzed in terms of accuracy. The results indicate that higher accuracy is achieved with the homogeneous and heterogeneous models in comparison with prior work on the direct marketing.

The experiment results lead to the following observations.

- ❖ RBF exhibits better performance than SVM in the important respects of accuracy.
- ❖ The proposed bagged methods are shown to be significantly higher improvement of classification accuracy than the base classifiers.
- ❖ The hybrid RBF-SVM shows higher percentage of classification accuracy than the base classifiers.
- ❖ The proposed ensemble methods provide significant improvement of accuracy compared to individual classifiers.
- ❖ The heterogeneous models exhibit better results than homogeneous models for bank marketing dataset.
- ❖ Assessment of performance is based on the calculation of the χ^2 statistic for all the approaches and their critical values are found to be less than 0.455. Hence their corresponding probability is $p < 0.5$. This is smaller than the conventionally accepted significance level of 0.05 or 5%. Thus examining a χ^2 significance table, it is found that this value is significant with a degree of freedom of 1. In general, the result of χ^2 statistic analysis shows that the proposed classifiers are significant at $p < 0.05$ than the existing classifiers.
- ❖ The results indicate that higher accuracy is achieved with the homogeneous and heterogeneous models in comparison with prior work on the direct marketing.
- ❖ The future research will be directed towards developing more accurate base classifiers particularly for the bank marketing dataset.

ACKNOWLEDGEMENT

Author gratefully acknowledges the authorities of Annamalai University for the facilities offered and encouragement to carry out this work.

REFERENCES

[1] Ali AghaeiRad, Ning Chen, Bernardete Ribeiro. (2017). Improve credit scoring using transfer of learned knowledge from self-organizing map, Neural Computing and Applications, 28(6):1329–1342.

[2] Arno De Caigny, Kristof Coussement, Koen W. De Bock. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, European Journal of Operational Research, 269(2): 760-772.

[3] Ashalata Panigrahi, Manik Chand Patnaik. (2020). Customer Deposit Prediction Using Neural Network Techniques, International Journal of Applied Engineering Research, 15(3): 253-258.

[4] Ashkan Zakaryazad and Ekrem Duman. (2016). A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing, Neurocomputing, 175, Part A, 121-131.

[5] Bing Zhua, Bart Baesens, Seppe K.L.M. vanden Broucke. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction, Information Sciences, 408:84-99.

[6] Breiman. L, (1996). Bias, Variance, and Arcing Classifiers, Technical Report 460, Department of Statistics, University of California, Berkeley, CA.

[7] Breiman, L. (1996a). Bagging predictors, Machine Learning, 24(2):123–140.

[8] Broomhead, D.S.; Lowe, D. (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks. Complex Syst., 2: 321–355.

[9] Debjyoti Das Adhikary, Deepak Gupta. (2019). Ensemble Learning Models for Churn Prediction, International Journal of Innovative Technology and Exploring Engineering, 9(2S): 159-164.

[10] Dustin Baumgartner, Gursel Serpen. (2013). Performance of global–local hybrid ensemble versus boosting and bagging ensembles, International Journal of Machine Learning and Cybernetics, 4(4):301–317.

[11] Fatemeh Nemati Koutanaei, Hedieh Sajedi, Mohammad Khanbabaei. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring, Journal of Retailing and Consumer Services,27:11-23.

[12] Gend Lal Prajapati, Arti Patle. (2015). Formulation of Feature Selection with Support Vector Machine, International Journal of Computer Applications, 123(5):20-25.

[13] Gonzalo Martinez-Munoz, Alberto Suarez. (2010). Out-of-bag estimation of the optimal sample size in bagging, Pattern Recognition, 43(1):143-152.

[14] Guido Bologna and Yoichi Hayashi. (2018). A Comparison Study on Rule Extraction from Neural Network Ensembles, Boosted Shallow Trees, and SVMs, Applied Computational Intelligence and Soft Computing, Volume 2018:1-20.

[15] Han Liu, Mihaela Cocea. (2017). Granular computing-based approach for classification towards reduction of bias in ensemble learning, Granular Computing, 2(3):131–139.

[16] Joaquín Abellán and Javier G. Castellano. (2017). A comparative study on base classifiers in ensemble methods for credit scoring, Expert Systems with Applications, 73:1-10.

[17] Juan J. Rodriguez, Jesus M. Maudes, Carlos J. Alonso. (2006). Rotation-based Ensembles of RBF Networks, proceedings of the European Symposium on Artificial Neural Networks, Bruges (Belgium), 605-610.

[18] Khairul Nizam Abd Halim, Abdul Syukor Mohamad Jaya, Ahmad Firdaus Ahmad Fadzil. (2020). Data Pre-Processing Algorithm for Neural Network Binary Classification Model in Bank Tele-Marketing, International Journal of Innovative Technology and Exploring Engineering, 9(3): 272-277.

[19] Kudipudi Srinivas, R.Kiran Kumar, L.L.Surya Prasanthi. (2017). An Empirical Comparative study of classification algorithms for class imbalance data, International Journal of Computer Science and Information Security (IJCSIS), 15(7):36-41.

[20] Kuo-Wei Hsu. (2013). Weight-Adjusted Bagging of Classification Algorithms Sensitive to Missing Values, International Journal of Information and Education Technology, 3(5):560-566.

[21] Maisa Cardoso Aniceto, Flavio Barboza and Herbert Kimura. (2020). Machine learning predictivity applied to consumer creditworthiness, Futur Bus J, 6:37-51.

[22] Martínez-Muñoz, G., Sánchez-Martínez, A., Hernández-Lobato D. and Suárez, A. (2006). Building Ensembles of Neural Networks with Class-switching**,** Proceedings of the International Conference on Artificial Neural Networks (ICANN), Athens, Greece, Part I, LNCS 4131:178-187.

[23] Marqués, A.I., García, V., Sánchez, J.S. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles, Expert Systems with Applications, 39:10244–10250.

[24] Mohammad Amini, Jalal Rezaeenour, Esmaeil Hadavandi. (2015). A Cluster-Based Data Balancing Ensemble Classifier for Response Modeling in Bank Direct Marketing, International Journal of Computational Intelligence and Applications, 14(4):1550022-1-23.

[25] Nam N. Nguyen and Anh T. Duong. (2021). Comparison of Two Main Approaches for Handling Imbalanced Data in Churn Prediction Problem, Journal of Advances in Information Technology, 12(1): 29-35.

[26] Nazeeh Ghatasheh, Hossam Faris, Ismail AlTaharwa and Yousra Harb and Ayman Harb. (2020). Business Analytics in Telemarketing: Cost-Sensitive Analysis of Bank Campaigns Using Artificial Neural Networks, Appl. Sci., 10: 2581- 2596.

[27] Platt, J. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines; Technical Report; MSR-TR-98-14; Microsoft Research: Redmond,WA, USA.

[28] Ravi, V., Kurniawan, H., Peter Nwee Kok Thai, Ravi Kumar, P. (2008). Soft computing system for bank performance prediction, Applied Soft Computing, 8(1):305–315.

[29] Rogelio A. Mancisidor, Michael Kampffmeyer, Kjersti Aas, Robert Jenssen. (2018). Segment-Based Credit Scoring Using Latent Clusters in the Variational Autoencoder, arXiv:1806.02538v1 [cs.CE], 1-24.

[30] Stamatis Karlos, Sotiris Kotsiantis, Nikos Fazakis, Kyriakos N. Sgarbas. (2016). Effectiveness of semi-supervised learning in bankruptcy prediction", Proceedings of the 7th International Conference on Information, Intelligence, Systems & Applications (IISA), Chalkidiki, Greece.

[31] Stefan Lessmann, Johannes Haupt, Kristof Coussement, Koen W. De Bock. (2019). Targeting customers for profit: An ensemble learning framework to support marketing decision-making, Information Sciences, In press.

[32] Uma R. Salunkhe and Suresh N. Mali. (2017). Security Enrichment in Intrusion Detection System Using Classifier Ensemble, Journal of Electrical and Computer Engineering, Volume 2017:1-6.

[33] Yajiao Tang, Junkai Ji, Shangce Gao, Hongwei Dai, Yang Yu, and Yuki Todo. (2018). A Pruning Neural Network Model in Credit Classification Analysis, Computational Intelligence and Neuroscience, Volume 2018: 1-22.