Improvement of Customer Loans Prediction Accuracy in *Neural Networks*

Irwan Syah[1], Amron[2], Herry Subagyo[3]
{irwansyahcadangan2@gmail.com[1], amron@dsn.dinus.ac.id[2], herry.subagyo@dsn.dinus.ac.id}

Doctor of Management Program, Dian Nuswantoro University, Indonesia[123]

Abstract. Credit screening is divided into two categories of credit applicants based on their potential repayment capacity. A suitable applicants will be accepted high probability of default on installment obligations. Predicting creditworthiness is one of the considerations in granting credit to customers. Therefore the prediction accuracy results are needed. The neural network algorithm integrated use bagging and feature selection Support Vector Machine Recursive Feature Elimination (SVM-RFE). This study showed that this method improved the accuracy of predictive model performances. The lowest accuracy value was 86.52%. While 87.15% using the neural network model, joint bagging and optimizing the SVM-RFE selection. The experiment results revealed that the neural network model with bagging and optimizing the SVM-RFE selection could improve the performance of at customer credit prediction model.

Keywords: Neural Network; SVM-RFE; Bagging; Classification; Optimaze Selection

1 Introduction

A loan, or commonly known as credit, is the ability for an individual or business to borrow money to purchase a product and repay the customer within a certain period of time. Cooperative loans cannot be arbitrarily loaned and must go through various checks [1]. By implementing a credit score, loan applicants can be divided into two categories based on their potential ability to pay. Eligible loan applicants who tend to pay installment debts after the loan is obtained; Bad loan applicants who are likely not to repay the loan.[2]. In addition, loans are also related to life insurance. A customer is required to take out a life insurance policy in order to exchange an amount to cover the risks that affect someone's life. Insurance companies will cover the accident risk and pay out the insured when they pass away. Insurance is a mutual transfer of risk from the insured as, the owner of the risk to the insurer as the transfer of risk beneficiary. An insured person is a person who needs certainty about the risks to which he or she is exposed, such as financial loss due to illness, death or retirement, and other related financial risks to an individual's life, and the insurer undertakes to transfer risk from and promise financial compensation to the insured. It is a business insurance company that accepts. There is room for complaints. As a result of the transfer of risk, the insured pays the insurance company an amount called the premium [22]. To predict the risk of lending, a system is required and one of the following systems will be investigated.

Data mining is the process of finding patterns and information in selected data using specific techniques and methods. One of the data mining techniques is classification using neural network algorithms [2]. Neural network algorithms are commonly used for decision making and have proven to be highly accurate [3]. Backpropagation is a supervised learning

ICEMBA 2022, December 17, Tanjungpinang, Indonesia

DOI 10.4108/eai.17-12-2022.2333265

technique of a neural network wich is also popular for its learning ability [4]. The backpropagation algorithm is a simple and direct iterative algorithm that works well even with complex data. Backpropagation algorithms can analyze past year data and recognize the data patterns. Therefore, the backpropagation pattern can be used to examine and predict what will happen in the future. In contrast to other learning algorithms, the backpropagation algorithm has good computational properties, especially for a huge amount of data. The weights are iteratively ordered in the backpropagation training process to diminish the introduced error [5].

However, some researchers often find data with imbalanced class conditions. Those conditions can influence the performance of neural networks, which causes, overfitting in the algorithm and reduce accuracy [6]. An imbalanced data class occurs when the data class distribution is unbalanced in which one data class (instance) one or more classes compared to other data classes. The smallest data class is called the minority group, and the is called the majority group [15].

Data with unbalanced classes usually have classification problems in Machine Learning because the amount of data per class is not evenly distributed. Thoise conditions are usually found in credit, medicalhealth, and other data [7]. There are many approaches that have been The researchers employed many approaches to cope with the unbalanced class problemss. An effective way to deal with that problem is the Ensemble method. Two common techniques proved to increase the accuracy of predictive models or learning algorithms are Boosting and Bagging. A study by Yoon et al. [8] showed that the Bagging technique improved task performances and achieved higher classification accuracy than other techniques. The Pima Indian Diabetes dataset results [4] showed that the average neural network classification achieved 0.761% accuracy, while the neural network -integrated with Bbagging achieved an 0.768% accuracy. In the Wisconsin Breast Cancer dataset, the average neural network achieved 0.959% accuracy, and the neural network -integrated with Bagging achieved 0.965% accuracy.

Meanwhile, in the Indian Liver Patient dataset, the average neural network attained 0.707% accuracy, and the neural network -integrated with Bagging attained 0.713% accuracy. The experimental results show that Bagging technique performs better for class imbalance problems than the Boosting technique. Furthermore, Bagging performs well with the neural network classifier [6]. There are various problems in classification apart from unbalanced data classes, such as missing values, redundant data, correlations, and irrelevant features. It often leads to large misclassification [6]. Feature selection is a popular approach to solvinging classification in irrelevant data [8]. Feature selection is a selecting subset of originalrelevant features to optimize accuracy results. Feature selection is applied to complete data and , improve the classification of incomplete data (missing values). Feature selection increased the classification accuracy and produced a less complex model [9]. In addition, feature selection is performed before the classification process [6].

The process of feature selection is divided into two categories are:

independent and dependent classifiers. Independent classifiers are known as filter methods which. do not depend on any classification method. Filter methods only consider unique features of the data to determine which features to select, and do not use classification methods to assess feature relevance. Meanwhile, dependent classifiers are categorized into two methods: that are Wrapper and Embedded methods. The wrapper selects a subset of the feature by inputting all the dataset features. With the overall work of the classifier, theA wrapper method can discover possible interactions by measuring the accuracy of the classifier's predictions in two variables [10]. Feature selection has an algorithm method, the Support Vector Machine Recursive Feature Elimination (SVM-RFE). This method is a combination of SVM and RFE. Recursive Feature Elimination (RFE) is a technique that recursively selects data set features based on minimum feature values. Thus, by implementing RFE, the SVM-RFE eliminates irrelevant features, and features with the lowest weight will be removed. Therefore, weighting features from the highest to the lowest weight value are required [11]. The SVM-RFE has been shown to outperform the SVM methods without feature selection. SVM-RFE and 1-Dimensional Naive Bayes Classifier (1-DBC) algorithms have previously been used to classify prostate and breast cancer data to produce high-level classification scores. Those algorithm methods are also applied in this study. This method gives 95.61% precision, 100% precision, and 93.61% recall. In additional evaluation, the SVM-RFE has a faster run time than the 1-DBC [9]. A study from Luo et al. [12] showed that SVM-RFE outperforms existing SVM-based feature selection algorithms in terms of the positive sampling rate (rrp) and G mean (G). The

SVM-RFE can be further improved by adding optimization techniques such as Optimize Selection [13] to refine and improve the accuracy of the solution obtained by SVM RFE. A streamlined selection optimization technique has been shown to improve algorithm performance [14].

This study obtained data from customer credit data in the Rukun Abadi Savings and Loans Cooperative. The number of datasets used was 1276 data. The data consists of 164 smoothed bad credit data and 1,112 smoothed current credit data, which causes unbalanced class problems. The attributes or characteristics includes marital status, number of dependents, age, recent educational history, occupation, monthly income, house ownership, collateral, loan size, loan term, and facilitation. Based on a large amount of feature or attribute data, features the relevant features for classification weare analyzed using SVM-RFE feature selection . This with the Optimize Sselection and unbalanced class problems by applying ensemble bagging techniques. In addition, a neural network algorithm was used to classify the algorithm..

Based on the background explanation, there is a problem in predicting customer creditworthiness. Those problems are irrelevant traits and imbalanced classes, which can lead to bad performance results of a customer credit prediction. Therefore, there should be a way to overcome those problems and improve the accuracy of customer creditworthiness predictions, which credit analysts can use as a guide in assessing customer creditworthiness.

2 Research Methods

Research is an investigation activity carried out systematically in a field. The research aims to find or revise facts, theories, applications, and many more, as new knowledge is to be published. This study starts with a problem analysis, literature review and data collection. This study used secondary data, namely data that is not obtained directly but is collected by other parties. The data was collected from customer credit datasets inat the Rukun Abadi Savings and Loans Cooperative. This study applied customer credit attributes such as marital status, the number of dependents, age, educational history, work, income per month, house ownership, guarantee, loan amount, loan duration, and loan rate smoothing.

This study SVM-RFE feature selection to select the relevant attributes. After selecting the attributes, the dataset was divided into ten parts using 10fold cross-validation, where all parts of the dataset become training and test data. The next class balance process will be carried out with the bagging method, with artificial neural networks as the classification process. The results's performance will be measured by Precision and Area Under Curve (AUC).

3 Results and Discussion

Neural Network (NN) and NN with Bagging, NN and NN with SVM-RFE and Bagging. The model wasis tested using a customer credit dataset fromat Rukun Abadi Savings and Loan Cooperative. In this section, the researchers measured the model tested using a neural network with the proposed method, namely neural network, neural network and SVM RFE, Neural Network and Bagging (BG), and Neural Network with SVM- RFE and Bagging (SFB). Measurements are recorded based on the confusion matrix, accuracy, and AUC results.

Table 1. Model Measurement Results

	accuracy	**AUC**
NN	86.52	0.638
NN+Bagging	86.6	0.644
NN+SVM RFE+Bagging	87.15	0.602

Different tests were carried out using statistical methods to test hypotheses on the Neural Network (NN) model with Neural Network and Support Vector Machine Recursive Feature Elimination and Bagging (NN+ SVM RFE+ BG)
H_0 : There is no difference in the average accuracy of NN and NN+BG+SVM RFE.
H_1 : There is a difference between the average accuracy of NN and NN+BG+SVM RFE.

In table 8 it can be seen the difference in accuracy values between the NN model and the NN+BG+SVM RFE model.

Table 2 Comparison of NN and NN+BG+SVM RFE Accuracy

	NN	**NN+BG+SVM RFE**
Accuracy	86.52	87.15
AUC	0.638	0.602

After comparing the results of the accuracy values of NN and NN+BG+SVM RFE, an analysis was then performed using the Paired Two Sample for Means t-Test with the results shown in Table 3.

Table 3. Statistical Difference Test Results for NN and NN+BG+SVM RFE Accuracy

	NN	*NN+BG+SVM RFE*
Means	43,579	43,876
Variances	3687.858962	3745.278152
Observations	2	2
Pearson Correlation	1	
Hypothesized Mean Difference	0	
Df	1	
t Stats	-0.891891892	
P(T<=t) one-tailed	0.268169377	
t Critical one-tail	6.313751515	
P(T<=t) two-tailed	**0.536338755**	
t Critical two-tail	12.70620474	

Table 3 showed that the NN and NN+BG+SVM RFE models have a higher average value than the NN model, which was 43,876. For the statistical difference test, the alpha value was set to 0.05, if the p-value is less than the alpha value (p<0>0.05), H0 is accepted, and H1 is rejected, so there is no significant difference between the models. The result showed that there was no significant difference between the compared models, but the NN+BG+SVM RFE model was able to improve accuracy by 0.63%.

4 Conclusions

Based on the results of research experiments, the feature used is the owner with a weight of 1, and 10 iterative bagging methods can improve the performance accuracy of the predictive model by 0.63%. This study's lowestt accuracy value using the neural network model was 86.52%. Optimize selection achieved the highest accuracy value of 87.15% using a neural network model with bagging and SVM RFE. This research contributes is to overcoming the noise in the data and addressing the problem of class imbalance. Improvements for the further studies include how to use bagging integration and SVM RFE in neural networks. This improves the performance accuracy of the predictive model. change. The metrics used in this study were limited to Precision and AUC. The additional assessment measures may be added in future studies. This study used limited comparisons of method performance by t-tests only. The further researchers are expected to add other methods of performance comparison and more definitve methods.

References

[1] F. Husaini, "Naive Bayes Classification Algorithm for Assessing Creditworthiness (Case Study: Bank Mandiri Micro Credit)," *Progr. Studs. Tech. inform.* , vol. 1, no. 3, pp. 2–12, 2016.

[2] Ilayani, J. Nangi, and Yuwanda Purmasari Pasrun, "Data Mining Application for Credit Assessment Using Decision Tree Algorithm Id3 Case Study Pt. Mandala Multi Finance Branch Kendari Ilayani*1," *semanTIK* , vol. 4, no. 1, pp. 65–76, 2018.

[3] A. Ilham, "Comparison of Classification Algorithms with Data Level Approaches to Handle Class Unbalanced Data," *J. Ilm. Computing Science.* , vol. 3, no. 1, pp. 1–6, 2017, doi: 10.35329/jiik.v3i1.60.

[4] S. Setti and A. Wanto, "Analysis of Backpropagation Algorithm in Predicting the Most Number of Internet Users in the World," *J. Online Inform.* , vol. 3, no. 2, p. 110, 2019, doi: 10.15575/join.v3i2.205.

[5] S. P. Siregar and A. Wanto, "Analysis of Artificial Neural Network Accuracy Using Backpropagation Algorithm In Predicting Process (Forecasting)," *IJISTECH (International J. Inf. Syst. Technol.* , vol. 1, no. 1, p. 34, 2017, doi: 10.30645/ijistech.v1i1.4.

[6] I. Fakhruzi, "An artificial neural network with bagging to address imbalance datasets on clinical prediction," *2018 Int. Conf. inf. commun. Technol. ICOIACT 2018* , vol. 2018-January, no. 1, pp. 895–898, 2018, doi: 10.1109/ICOIACT.2018.8350824.

[7] L. H. F. Giovanini, EF Manffra, and JC Nievola, "Evolutionary ensemble approach for behavioral credit scoring," *Springer Nat. 2018* , no. June, pp. 350–357, 2018, doi: 10.1007/978-3-319-93713-7.

[8] H. J. Yoon *et al.* , "Accelerated training of bootstrap aggregation-based deep information extraction systems from cancer pathology reports," *J. Biomed. inform.* , vol. 110, no. september, p. 103564, 2020, doi: 10.1016/j.jbi.2020.103564.

[9] CT Tran, M. Zhang, P. Andreae, and B. Xue, "Bagging and feature selection for classification with incomplete data," *Lect. Computer Notes. sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* , vol. 10199 LNCS, pp. 471–486, 2017, doi: 10.1007/978-3-319-55849-3_31.

[10] Z. Rustam and SAA Kharis, "Comparison of Support Vector Machine Recursive Feature Elimination and Kernel Function as feature selection using Support Vector Machine for lung cancer classification," *J. Phys. Conf. Ser.* , vol. 1442, no. 1, 2020, doi: 10.1088/1742-

6596/1442/1/012027.

[11] A. Bustamam, A. Bachtiar, and D. Sarwinda, “Selecting features subsets based on support vector machine-recursive features elimination and one-dimensional-naïve bayes classifier using support vector machines for classification of prostate and breast cancer,” *Procedia Comput. sci.* , vol. 157, pp. 450–458, 2019, doi: 10.1016/j.procs.2019.08.238.

[12] K. Luo, G. Wang, Q. Li, and J. Tao, “An Improved SVM-RFE Based on F-Statistics and mPDC for Gene Selection in Cancer Classification,” *IEEE Access* , vol. 7, pp. 147617–147628, 2019, doi: 10.1109/ACCESS.2019.2946653.

[13] A. Fauzi, T. Informatics, U. Pamulang, RP No, and K. Pamulang, "Bank Direct Marketing Data Analysis with Comparison of Data Mining Classification Based on Optimize Selection (Evolutionary)," vol. 6, no. 1, pp. 102–111, 2021.

[14] H. Amalia, A. Puspitasari, and AF Lestari, “Student Performance Analysis Using C4 . 5 Algorithm TO,” pp. 149–154, 2018.

[15] H. Sanz, C. Valim, E. Vegas, JM Oller, and F. Reverter, “SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels,” *BMC Bioinformatics* , vol. 19, no. 1, pp. 1–18, 2018, doi: 10.1186/s12859-018-2451-4.

[16] X. Lin, C. Li, Y. Zhang, B. Su, M. Fan, and H. Wei, “Selecting feature subsets based on SVM-RFE and the overlapping ratio with applications in bioinformatics,” *Molecules* , vol . 23, no. 1, 2018, doi: 10.3390/molecules23010052.

[17] K. R. Kavitha, UN Harishankar, and MC Akhil, “PSO based feature selection of genes for cancer classification using SVM-RFE,” *2018 Int. Conf. Adv. Comput. commun. Informatics, ICACCI 2018* , pp. 1012–1016, 2018, doi: 10.1109/ICACCI.2018.8554429.

[18] T. M. Dantas and FL Cyrino Oliveira, “Improving time series forecasting: An approach combining bootstrap aggregation, clusters and exponential smoothing,” *Int. J. Forecasts.* , vol. 34, no. 4, pp. 748–761, 2018, doi: 10.1016/j.ijforecast.2018.05.006.

[19] M. A. Rahman and RC Muniyandi, "Feature selection from colon cancer dataset for cancer classification using Artificial Neural Network," *Int. J. Adv. sci. Eng. inf. Technol.* , vol. 8, no. 4–2, pp. 1387–1393, 2018, doi: 10.18517/ijaseit.8.4-2.6790.

[20] S. Karthik, R. Srinivasa Perumal, and PVSSR Chandra Mouli, “Breast cancer classification using deep neural networks,” *Knowl. Comput. Its Appl. Knowl. Manip. Process. Tech. Vol. 1* , pp. 227–241, 2018, doi: 10.1007/978-981-10-6680-1_12.

[21] G. Athanasopoulos, H. Song, and JA Sun, “Bagging in Tourism Demand Modeling and Forecasting,” *J. Travel Res.* , vol. 57, no. 1, pp. 52–68, 2018, doi: 10.1177/0047287516682871.

[22] A. Amron,”Electronic and traditional word of mouth as trust antecedents in life insurance buying decision,” *International Journal of e-Business Research*, Vol. 14, No. 4, pp. 91-103. 2018.