The Estimation of Ensemble Logistic Regression using Newton Raphson Parameter

Armin Lawi[1], Firman Aziz[2], Husna Gemasih[3] and Mursalin[4]
{armin@unhas.ac.id }

[1]Department of Computer Science, Hasanuddin University, Makassar, Indonesia
[2]Post-Graduate Program of Electrical Engineering, Hasanuddin University, Makassar, Indonesia
[3]Department of Informatics, Universitas Gajah Putih, Aceh, Indonesia
[4]Department of Mathematics Education, Universitas Malikussaleh, Aceh Utara, Indonesia

Abstract. The large volume of customer data in the credit industry makes the development of an effective credit scoring model extremely important. The use of an ensemble model on statistical methods to solve credit scoring problems managed to get the best predictive performance. ensemble performance can still be improved by estimating the parameters using nonlinear equations. This paper proposes the estimation of ensemble Logistic Regression using Newton Raphson parameter. The results showed that proposed method successfully achieved the best performance by improving the performance of a single classification with an increase of 2% accuracy

Keywords: Credit Scoring; Logistic Regression; Ensemble Bagging.

1 Introduction

Credit is a business activity that many done by the bank and become the largest source of income as well as the greatest risk. therefore the stability of banks is strongly influenced by the success in managing credit.

Credit Score is a technology that seeks to minimize the risk of credit borrowers who are unable to make timely payments. Credit Score is a classification problem that divides two potential borrowers into approved borrowers and unapproved borrowers based on several characteristics [1], such as age, economic conditions, social status, guarantees, etc. Objective.

Classification statistical method adopted in the credit scoring industry is Logistic Regression that is relatively easy to understand and implement. In research [2] estimating artificial neural networks and proposing a model in credit ratings. Logistic Regression models have better performance percentages than other methods of classifying good and bad borrowers. Research [3] learn about the latest classification algorithms on the application of credit cases. The conclusion shows that for the case of credit score, the Logistic Regression method gets better results.

On the ensemble, several sets of training are used to solve the same problem and the result of a single classification will be combined with the ensemble method into a single classifier to improve performance [4]. Research [5] divide the three data partitions to see the chi-square performance. The focus of this research lies in the selection of features and the research show

WMA-1 2018, January 20-21, Indonesia
Copyright © 2019 EAI
DOI 10.4108/eai.20-1-2018.2281940

that the accuracy of the classification performance was obtained on the data partition with 70% for training data and 30% test data.

Research [6] proposed to analyze the accuracy of ensemble method in classifying customers using three ensemble methods of AdaBoost, Bagging, Random Forest but in this research still using a single standard classification of the ensemble method of decision tree. In [7] applied the feature selection method on the german dataset and incorporated a single classification with a greedy stepwise search method but this study reduced the attributes from 20 to 14.

Research [8] performs ensemble bagging and boosting by using ANN as a single classification. ANN ensemble shows accurate results and low generalization errors but bagging has disadvantages when the number of attributes and data bit small. In [9] it examines the use of SVM and KNN as a basic classifier and uses ensemble bagging and boosting to improve the accuracy of basic classification. The results show advantage of ensemble model in terms of accuracy. Research [10] proposed a systematic homogeneous and heterogeneous ensemble model based on three classifications of LR, ANN and SVM.

The results show that the heterogeneous ensemble classifier provides high predictive accuracy with low error than homogeneous and singular clustering but in this study, no estimates were obtained on ensemble by bagging technique.

In this research is proposed ensemble Logistic Regression method with bagging technique for credit scoring. the bagging technique is used because it reduces the overfitting problem and does not depend on single classifier and bagging techniques more effectively on unstable learning algorithms. the focus of this study is to estimate the ensemble Logistic Regression using the Newton Raphson parameter. Newton Raphson is chosen because it can solve nonlinear data types by interpreting the initial value for its maximum function.

2 Methods

2.1 Classification

Single classification in this study using logistic regression. which is a statistical method used to solve classification and regression problems [11]. Logistic Regression is used to make example of binary result variable, usually, it is represented by 1 or 0. The scores of the models use binary numbers where 1 for good debtor and 2 for bad debtors and each independent variable gives dependence on each score [12]. The function of the classifier shown by equation:

$$log\left(\frac{p}{1-p}\right) = \sum_{i=1}^{i=n} \beta^{(i)} * x^{(i)} + e = \beta^T x + e \tag{1}$$

Where $\beta =(\beta^{(1)},\beta^{(2)},....,\beta^{(n)}$is coefficient vector of hyperplane. The probability of customer stop on equation (1) can be simply formulated as in the following eqution (2).

$$p = \frac{e^{\beta^T x+e}}{1 + e^{\beta^T x+e}} \tag{2}$$

To obtain an estimate of Logistic Regression parameters can be done in two ways namely by Maximum Likelihood Estimation (MLE) and Newton Raphson. However, this study uses the parameters of Newton Raphson.

Newton Rhapson is a method for solving nonlinear equations such as solving likelihood equations in Logistic Regression models [13]. The newton rhapson method requires an initial estimate of its maximum function value, which is an estimate using a polynomial approach of degree two. In this case to determine the $\hat{\beta}$ value of β which is the maximum function of $g(\beta)$ Suppose $q' = \left(\frac{\partial g}{\delta\beta_1}, \frac{\partial g}{\delta\beta_2}, \dots\right)$, and suppose H is denoted as a matrix with members $h_{ab} = \frac{\partial g}{\delta\beta_{1,\delta\beta_2}}$ suppose $q^{(t)}$and $\boldsymbol{H}^{(t)}$ is a form of evaluation of $\beta^{(t)}$ estimate to t at $\hat{\beta}$. In step t in the iteration process (t = 0, 1, 2, ...).

2.2 Classification Ensembles

The focus of this classification is to solve similar problems by combining a set of classifications to obtain a more accurate classification [14]. The ensemble method can reduce classification errors effectively, and is believed to perform well compared to the use of a single classifier. Compared to an individual classifier the learn and train one data only. in contrast to the ensemble classification. from the original data, classification ensemble doing the learning and training of various data and the training results will build the hypothesis and produce better accuracy [15].

Some ensemble classifiers techniques have been developed such as random forest, rotation forest, boosting, and bagging. because the focus of this research is using ensemble bagging, then the ensemble bagging classification will be presented.

Bagging is an algorithm that provides good performance and is very easy to apply. Number of the different bag (n-bag) obtained through training data and the results of the training dataset on every bag will be stuffed with randomly generated data from the training dataset [10].

Bagging (Bootstrap Aggregating) algorithm creates M bootstrap samples T_1, T_2, . . . ,T_M randomly drawn from the original training set T of size n[10]. Each bootstrap sample T_i of size n is then used to train a base classifier Ci. Predictions on new observations are made by taking the majority vote of the ensemble C^* built from C_1, C_2, . . . ,C_M [16].

Ensemble bagging algorithm

a single classification algorithm $C_t(\mathrm{x})$ will be given a training set of size n.

- Input sequence of training samples $(x_1 : y_1), \dots (x_n : y_n)$ with labels $y \in Y = (-1,1)$
- Initialize probability for each example in learning set $D_1(i) = \frac{1}{n}$ and set $t = 1$.
- Loop while $t < B = 100$ ensemble members
- Form training set of size n by sampling with replacement from distribution D_t
- Get hypothesis $ht : X \to Y$
- Set $t = t + 1$

  End of loop
- Output the final ensemble hypothesis

$$C^*(x_i) = h_{final}(x_1) = argmax \sum_{t=1}^{B} I(C_t(x) = y). \tag{3}$$

2.3 Performance Evaluation

The performance evaluation of classification method can be seen from the level of classification error. To count mark of classification error can use confusion matrix, the confusion matrix is usually called with contingency table like on Table 1.

Table 1.Confusion Matrix/ Contingency Table

Actual/Prediction	Good debtors	Bad debtors
Good debtors	True Positive (TP)	False Negative (FN)
Bad debtors	False Positive (FP)	True Negative (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 (3).

$$accuracy = \frac{TP + TN}{TN + FP + FN + TP} \tag{4}$$

3 Experimental

3.1 Dataset

To evaluation the accuracy of the proposed model uses two data ie German and Australian datasets. in the credit research literature, these two datasets are used to test the model and is available in UCI machine learning repository. Generally, the dataset can be seen in Table 2.

Table 2. Description Of Dataset

Dataset	Number of debtor	Good debtors	Bad debtors	Number of Attributes
German dataset	1000	700	300	20
Australian dataset	690	307	383	14

Both datasets will be divided into 70% data training and 30% data testing to see the classification results of the proposed model.

3.2 Normalization Data

The data must be transformed from different value scales to eliminate redundancy and build data models at the same interval. The dataset attribute is normalized with values 0 for 'bad' customer and 1 for 'good' customer.

3.3 Implementation

The focus of this study is to see the use of Newton Raphson parameters from ensemble Logistic Regression classification because the estimation of Newton Raphson parameter can improve the performance of ensemble Logistic Regression classification. Logistic Regression will be used as a single classification and ensemble technique used is bagging.

4 Results

To see the performance of the proposed method uses test processes with python programming language. The level of accuracy based on an estimation of Newton Raphson parameter of ensemble Logistic Regression.

Table 3. Classification Results

Classification	German Dataset	Australian Dataset
Logistic Regression	77 %	85.9 %
EnsembleLogistic Regression	79.6 %	86.9 %

Table III shows the classification results on the Australian dataset and German dataset. The Australian dataset, single logistic regression model generates classification accuracy of 85.9% and German dataset, generates classification accuracy of 77.0%.

Shows the results of the proposed model that is an estimation of ensemble Logistic Regression using Newton Raphson parameter. The proposed method successfully achieved the best performance by improving the performance of single classification with 79.6% accuracy for German dataset and 86.9% of Australia dataset of the results obtained, the increase is not very significant due to the australian dataset only increased by 1% and the german dataset only increased by 2.6%. but in this study obtained when the number of iterations 10 to 50 affects the accuracy of the accuracy of the ensemble classification Logistic Regression but if iteration> 50 accuracy value becomes constant. the proposed method is sensitive to small amounts of data.

5 Conclusions

In this research is proposed ensemble Logistic Regression method with Bagging technique for credit scoring. the focus of this study is to estimate the ensemble Logistic Regression using the Newton Raphson parameter. The proposed method successfully achieved the best performance of a single classification. to credit data accuracy german yield 79.6% and 86.9% Australian dataset. The number of iterations 10 to 50 affects the accuracy of the classification but when the iteration > 50 constant accuracy value and in this study the proposed method is sensitive to the small amount of data.

References

[1] Kim, Yoon Seong, and So Young Sohn. "Managing loan customers using misclassification patterns of credit scoring model." Expert Systems with Applications 26.4 (2004): 567-573.

[2] Desai, Vijay S., Jonathan N. Crook, and George A. Overstreet. "A comparison of neural networks and linear scoring models in the credit union environment." European Journal of Operational Research 95.1 (1996): 24-37.

[3] Baesens, Bart, et al. "Benchmarking state-of-the-art classification algorithms for credit scoring." Journal of the operational research society 54.6 (2003): 627-635.

[4] Tsai, Chih-Fong. "Combining cluster analysis with classifier ensembles to predict financial distress." Information Fusion 16 (2014): 46-58.

[5] Dahiya, Shashi, S. S. Handa, and Netra Pal Singh. "Credit scoring using ensemble of various classifiers on reduced feature set." Industrija 43.4 (2015): 163-174.

[6] Devi, CR Durga, and R. Manicka Chezian. "A relative evaluation of the performance of ensemble learning in credit scoring." Advances in Computer Applications (ICACA), IEEE International Conference on. IEEE, 2016.

[7] Onik, Abdur Rahman, et al. "An analytical comparison on filter feature extraction method in data mining using J48 classifier." International Journal of Computer Applications 124.13 (2015).

[8] West, David, Scott Dellana, and Jingxia Qian. "Neural network ensemble strategies for financial decision applications." Computers & operations research 32.10 (2005): 2543-2559.

[9]Nanni, Loris, and Alessandra Lumini. "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring." Expert systems with applications 36.2 (2009): 3028-3033.

[10]Ala'raj, Maher, and Maysam Abbod. "A systematic credit scoring model based on heterogeneous classifier ensembles." Innovations in Intelligent SysTems and Applications (INISTA), 2015 International Symposium on. IEEE, 2015.

[11]Akkoç, Soner. "An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data." European Journal of Operational Research 222.1 (2012): 168-178.

[12] Thomas, Lyn C. "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers." International journal of forecasting 16.2 (2000): 149-172.

[13]Agresti, Alan. Analysis of ordinal categorical data. Vol. 656. John Wiley & Sons, 2010.

[14]Zhou, Ligang, Kin Keung Lai, and Lean Yu. "Least squares support vector machines ensemble models for credit scoring." Expert Systems with Applications 37.1 (2010): 127-133.

[15]Kuncheva, Ludmila I. Combining pattern classifiers: methods and algorithms. John Wiley & Sons, 2004.

[16]Marqués, A. I., Vicente García, and Javier Salvador Sánchez. "Exploring the behaviour of base classifiers in credit scoring ensembles." Expert Systems with Applications 39.11 (2012): 10244-10250.