MATEMATIKA, 2018, Special Issue, 83–90

On the Comparison of Deep Learning Neural Network and Binary Logistic Regression for Classifying the Acceptance Status of Bidikmisi Scholarship Applicants in East Java

Nita Cahyani[*], **Kartika Fithriasari, Irhamah and Nur Iriawan**
Institut Teknologi Sepuluh Nopember
Kampus ITS Sukolilo, 60111 Surabaya, Indonesia
[*]Corresponding author: cahyani.nc@gmail.com

Article history
Received: 22 August 2018
Received in revised form: 6 December 2018
Accepted: 17 December 2018
Published on line: 31 December 2018

Abstract Neural Network and Binary Logistic Regression are modern and classical data mining analysis tools that can be used to classify data on Bidikmisi scholarship acceptance in East Java Province, Indonesia. One form of Neural Network model available for various applications is the Resilient Backpropagation Neural Network (Resilient BPNN). This study aims to compare the performance of the Resilient BPNN method as a Deep Learning Neural Network and Binary Logistic Regression method in determining the classification of Bidikmisi scholarship acceptance in East Java Province. After preprocessing data and dividing them into two parts, i.e. sets of testing and training data, with 10-foldcross-validation procedure, the Resilient BPNN and Binary Logistic Regression methods are implemented. The result shows that Resilient BPNN with two hidden layers is the best platform network model. The classification G-mean resulted by these both methods is that Resilient BPNN with two hidden layers is more representative with better performance than Binary Logistic Regression. The Resilient BPNN is recommended to be used to predict acceptance of Bidikmisi applicants yearly.

Keywords Resilient backpropagation neural network; Bidikmisi; deep learning neural network; binary logistic regression.

Mathematics Subject Classification 62M45, 62J12.

1 Introduction

Referring to the Laws and Government Regulations through the Directorate General of Higher Education, Ministry of National Education in Indonesia, starting in 2010 provides scholarships and tuition fees for prospective students from economically disadvantaged and outstanding families called Bidikmisi Scholarships [1].

Classification is a statistical method that is usually used to determine the accuracy or mapping acceptance of Bidikmisi scholarship applicants as well as to find out how much accuracy of the predicted results of actual observation. There are several classification methods that are usually used, that is classical methods and modern methods. The classical methods that are often used include the method of Logistic Regression and Discriminant Analysis while for the modern method Neural Network, Genetica Algorithm and Support Vector Machine. We unusually used Neural Network model as a technique of data minings has a better performance capability compared to a classical method of data mining technique.

An Artificial Neural Network (ANN) model, especially Multi-Layer Perceptron, is used to predict the probability of a prospective student's performance to be considered for admission to an Engineering Department of University of Ibadan, Nigeria. Various factors that are expected to affect the student's performance, among others are subject value, matriculation test score, parental background, type and location of secondary school and sex. The results show that the ANN model is able to predict the performance of prospective students correctly by 70% [2].

Some statistical models of classification developed in other educational fields examine the admission of college students by comparing the methods of regression analysis with Neural Network [3]. Teshnizi and Ayatollahi [4], examined by comparing logistic regression with Naural Network to predict student academic failure by yielding each classification 77.5% for result of classification logistic regression method and 84.3% for classification of Neural Network method. The resulting Neural Network method is better than the logistic regression method [4], using the Bayesian Bernoulli mixture regression model for Bidikmisi scholarship classification [5], andanother method that uses a binary logistic regression method [6].

Backpropagation has a weakness in taking a long time in the learning process. Resilient Backpropagation is a further development of the Backpropagation algorithm which can accelerate the level of learning. Resilient Backpropagation adapts directly from the weighted value based on information from the weighted value [7].

Based on the description above, this research will use Binary Logistic Regression and Resilient Backpropagation method by using two hidden layer or Deep Learning Neural Network (DLNN) whose purpose is to compare the two methods to get better performance level and expected result will be an accurate classification in determining the accuracy of the target of Bidikmisi scholarship acceptance in East Java or provide more appropriate decisions in the consideration of the adoption of accepted students.

2 Methods

This section describes the methodology used for the results of the classification data of the Bidikmisi scholarship.

2.1 Binary Logistic Regression

Binary logistic regression is one type of logistic regression. Logistic regression is a method that can be used to find the relationship between dichotomous response variable (nominal or ordinal scale with two categories) or polychotomous (nominal or ordinal scale with more than two categories) with one or more predictor variables on a continuous scale or continuous category (Hosmer, *et al.* [8]). Binary logistic regression analysis is used to describe the relationship between response variables that have only two categories, the category that states the success event ($Y = 1$) with the probability $\pi(x_i)$, and the category that states the failure event ($Y = 0$) with the chance of $1 - \pi(x_i)$. The model obtained from binary logistic regression analysis can be used as a model in classifying predictor variables into response variables in the form of categorical data [8].

Logical rendering can be used in many predictor variables known as multivariable models. Hosmer *et al.* [8] assume that a set of free variable p is represented as a vector $x^{'} = (x_1, x_2, \ldots, x_3)$. The logit form of multivariable logistic regression is given in the equation

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{1}$$

The logistic regression model with independent variable p, namely the number of predictor variables shown in the equation

$$\pi\,(x) = \frac{e^{g(x)}}{1+e^{g(x)}} = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1+\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}. \tag{2}$$

2.2 Resilient Backpropagation Neural Network (Resilient BPNN)

Artificial neural network learning algorithm Backpropagation was first formulated by Werbos [9] and popularized by Rummelhart and Mc. Clelland [9]. Neural Network Backpropagation is a guided learning algorithm that has many layers. Backpropagation uses an error output to use the weight values in the backward direction. To get this error, the stage of forwarding propagation must be done first [9].

Backpropagation network architecture in Figure 1 shows that the network consists of 3 units of neurons in the input layer namely X_1, X_2, and X_3; one hidden layer with two neurons, namely Z_1 and Z_2 and one unit in the output layer, namely Y. The weights that connect X_1, X_2, and X_3 with the first neurons in the hidden layer are V_{11}, V_{21}, and V_{31} or V_{ij} is the weight that connects the i^{th} input neurons to the jth neuron in the hidden layer. The terms $b1_1$ and $b1_2$ are the bias weights leading to the first and second neurons in the hidden layer. The weight that connects Z_1 and Z_2 with neurons in the output layer, is W_1 and W_2. The bias weight b_2 connects the hidden layer with the output layer. In this research, the sigmoid activation function is used between the input layer and the hidden layer.

Backpropagation has a weakness in taking a long time in the learning process. Resilient Backpropagation is a further development of the Backpropagation algorithm which can accelerate the level of learning.Resilient Backpropagation is an efficient new learning scheme, that performs a direct adaptation of the weight step based on local gradient information. So to achieve this, each weight its individual update value Δ_{ij}. This adaptive update value evolves during the learning process based on its local sight on the error function E as follows [5]:

$$\Delta_{ij}^{(t)} = \begin{cases} \eta^{+} * \Delta_{ij}^{(t-1)}, \text{if } \dfrac{\partial E^{(t-1)}}{\partial w_{ij}} * \dfrac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ \eta^{-} * \Delta_{ij}^{(t-1)}, \text{if } \dfrac{\partial E^{(t-1)}}{\partial w_{ij}} * \dfrac{\partial E^{(t)}}{\partial w_{ij}} < 0 \\ \Delta_{ij}^{(t-1)}, \qquad\qquad \text{else,} \end{cases} \tag{3}$$

where $0 < \eta^{-1} < 1 < \eta^{+}$.

Once the update value for each weight is adapted, furthermore if the derivative is positive (increasing error), the weight is decreased by its update-value, if the derivative is negative the

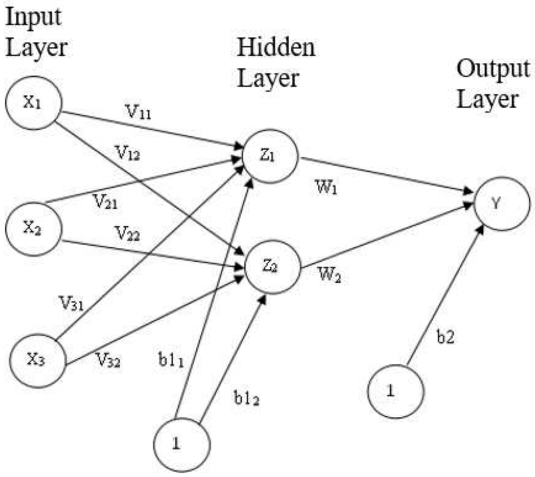

Figure 1: Backpropagation Architecture

update-value is added, according to the following learning rule:

$$\Delta w_{ij}^{(t)} = \begin{cases} -\Delta_{ij}^{(t)}, \text{if } \dfrac{\partial E^{(t)}}{\partial w_{ij}} > 0 \\ \Delta_{ij}^{(t)}, \quad \text{if } \dfrac{\partial E^{(t)}}{\partial w_{ij}} < 0 \\ 0, \qquad \text{else.} \end{cases} \tag{4}$$

2.3 Crosss Validation

Cross-validation is an approach used to estimate the classification accuracy of a model and is based on the separation of available samples between learning data and testing data [10]. The basic form of cross-validation is k-fold cross-validation. k-fold cross-validation, first the data is partitioned into the same k sections or folds. Furthermore, the iteration of learning and validation is done in such a way that in each iteration a different fold is used for validation while the remaining k-1 fold is used for learning. This study uses 10-fold cross-validation because it tends to provide accuracy and less bias [11].

2.4 Evaluation of Performance of Classification Method

Actual data and prediction data from the classification model are presented using a confusion matrix, which contains information about the actual data class represented in the matrix row and prediction data class in the matrix column [12]. The confusion matrix can be seen in Table 1.

The accuracy of classification can be seen from the accuracy of classification. Classification accuracy shows the performance of the overall classification model, where the higher the classification accuracy means the better the performance of the classification model is.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \times 100\%. \tag{5}$$

Table 1: Confusion Matrix

Actual	**Prediction**	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

To get an optimal and more specific classification, sensitivity and specificity can be tested. Sensitivity is the percentage of positive data that is predicted to be positive while specificity is the percentage of negative data predicted as negative.

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \times 100\%, \tag{6}$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \times 100\%. \tag{7}$$

Performance evaluation of the classification model can be done using G-mean, is the average of geometric sensitivity and specificity. G-mean will be zero if the positive class is unpredictable.

$$\text{G-Mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}}. \tag{8}$$

3 Results

In this section, we firstly present the data description, then the result of comparison study.

3.1 Data

The Bidikmisi 2017 scholarship is applied throughout the territory of Indonesia which covers 33 provinces, 497 districts, and cities. East Java is the largestor most registered registrant will be discussed in this study. Thus, the classification of Bidikmisi Scholarship acceptance status in East Java

The proportion accepted by Bidikmisi is 93,74% while the proportion not accepted by Bidikmisi is 6,26%, this indicates that the status is accepted more (majority) than the status of less accepted (minority) or what is called an imbalanced class. The imbalance of this class is also known by looking at the value of IR (Imbalance Ratio), which is the comparison between the percentage of the majority class and the percentage of minority classes, which is 14.97%. The higher the IR value, the more unbalanced the dataset.

The research variables used in this study are Bidikmisi Student Enrollment Year 2017. The data is sourced from Database Kemenristek DIKTI Bidikmisi channel. The research variables used in this study are based on the Bidikmisi Program registration form. The response variable (Y) is the status of the Bidikmisi scholarship throughout the districts and cities in East Java Province. Predictor variables (X) are characteristics of prospective students registering Bidikmisi scholarships in all Regencies and Cities in East Java Province, among authors Father's work (X_1), Mother's work (X_2), Father Education (X_3), Mother's Income (X_4), Home

Ownership (X_5), Power Source used by the Family (X_6), Land Size House Family House (X_7), Family Household Building Area (X_8), Public bathing (X_9), washing (X_{10}), and toilet facilities (X_{11}), Number of Dependents (X_{12}).

3.2 On The Comparison of Regression Logistic Binary and Resilient Backpropagation

This section describes on the comparison of the 3 methods that have been done, namely classification analysis with Binary Logistic Regression, Resilient Backpropagation with one hidden layer and Resilient Backpropagation with two hidden layers or Deep Learning Neural Networks. These methods areperformed by using R Software.

Table 2: Classification Performance Level on Binary Logistic Regression, Resilient Backpropagation 1 Hidden Layer, and Resilient Backpropagation 2 Hidden Layer

Classification Performance	**Regresi Logistic Binary**		**Resilient BPNN 1-Hidden Layer**		**Resilient BPNN 2-Hidden Layer**	
	All variables	Significant variables	All variables	Significant variables	All variables	Significant variables
G-mean at Training Data	0.00%	1.22%	19.02%	6.11%	13.38%	12.37%
G-mean at Testing Data	2.43%	1.22%	9.02%	10.38%	13.99%	15.42%
Sensitivity at Training Data	99.97%	99.99%	99.90%	98.27%	86.73%	95.95%
Sensitivity at Testing Data	99.79%	99.81%	99.90%	97.63%	86.47%	94.08%
Specificity at Training Data	0.00%	0.49%	3.79%	1.73%	13.64%	3.87%
Specificity at Testing Data	0.31%	0.15%	1.19%	2.38%	13.42%	5.35%
Accuracy at Training Data	93.73%	93.70%	93.88%	92.23%	82.16%	90.18%
Accuracy at Testing Data	93.56%	93.58%	93.72%	91.67%	81.89%	88.52%

Table 2 shows the results of the performance level of binary logistic regression classification, Resilient BPNN one hidden layer, and Resilient BPNN two hidden layer by using a 10-fold cross-validation procedure. Trial and error were carried out to get the best BPNN Resilient structure. Resilient BPNN of a hidden layer using all significant input variables at 10-fold cross-validation obtained the best classification accuracy with the number of neurons in the hidden layer as many as 4 neurons and for significant inputs using 2 neurons, while Resilient BPNN two hidden layers by using all significant input variables obtained the best classification accuracy with the number of neurons in the first hidden layer as many as 1 neuron and for the second hidden layer as many as 10 neurons and for significant input variables using 1 neuron in the first hidden layer and as many as 10 neurons for the second hidden layer.

Based on the results of the analysis it is known that the data used in this study imbalance in the category so that the G-mean value is used to see which method is the best. Table 2 shows that the two hidden layer resilent back propagation method with a significant variable produces the highest G-mean value with the other method which is 15.42% of the testing data and the results of a high sensitivity value of 94.08% and a low sensitivity value of 5.35% and an accuracy value of 88.52%.

4 Conclusion

Based on the research that has been done, the average results of the sensitivity values of the three methods produce a high percentage of value. It means that almost all positive classes can be correctly predicted. While the average yield of the three methods produce a small percentage value. It means that not all negative classes (not accepted by Bidikmisi) are correctly predicted so that the classifier is not good at producing negative class predictions, this is because the amount of data between positive and negative classes is not balanced. Although the classifier is not good at producing negative class predictions but, it can be said the resilient backpropagation method with two hidden layers can be used to classify Bidikmisi scholarship acceptance status in East Java because this method produces a classification performance value that is better than other methods.

Acknowledgement

This research was supported by Directorate for Research and Community Service (DRPM) Ministry of Research, Technology and Higher Education Indonesia which support this research under PUPT research grant no. 1049/PKS/ITS/2018.

References

[1] Ministry of Education. *Bidikmisi Scholarship Program: Educational Scholarships for Prospective Students with Achievement from Underprivileged Families*. Jakarta: Directorate General of Higher Education. 2010.

[2] Oladokun, V. O., Adebanjo, A. T., and Charles-Owaba, O. E. Predicting students' academic performance using artificial neural network: a case study of an engineering course. *The Pacific Journal of Science and Technology*. 2008. 9(1): 72-79.

[3] Walczak, S., and Sincich, T. Comparative analysis of regression and neural networks for university admission. *Information Sciences*. 1999. 119(1-2): 1-20.

[4] Teshnizi, S. H., Taghi, S. M., and Ayatollahi. A comparison of logistic regression modeland artificial neural networks in predicting of students' academic fFailure. *Acta Inform Med*. 2015. 23(5): 296-300.

[5] Iriawan, N., Fihriasari, K., Ulama, B, S, S., Suryaningtyas, W., Susanto, I., and Pravitasari, A. A. Bayesian Bernoulli mixture regression model for Bidikmisi scholarship classification. *Jurnal Ilmu Komputer dan Informasi*. 2018. 11(2): 67 -76.

[6] Oktaviana, P., and Fihriasari, K. Bayesian network inference in binary logistic regression: a case study of Salmonella sp bacterial contamination on vannamei shrimp. *Journal of Mathematics and Statistics*. 2018. 13: 306-311.

[7] Riedmiller, M. and Braun, H. A fast adaptive learning algorithm. *International Symposium on Computer and Information Sciences*. 1992. 7: 279-285.

[8] Hosmer, D. W., Lemeshow, S., and Sturdivant, X. R. *Applied Logistic Regression* (3$^{\text{rd}}$ed.). New York : John Wiley and Sons. 2013.

[9] Kusumadewi, S. *Building Neural Neural Networks Using MATLAB & EXCEL LINK*. Yogyakarta: Graha Ilmu. 2004.

[10] Last, M. The uncertainty principle of cross-validation. *IEEE Conference Publications*. 2006. 275-280.

[11] Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *The Fourth International Joint Conference on Artificial Intelligence*. American Association for Artificial Intelligence. 1995. 14(2): 1137-1145.

[12] Han, J., and Kamber, M. *Data mining: Concepts and Techniques*. California: Second Edition, Morgan Kaufmann. 2006.