The relationship between data skewness and accuracy of Artificial Neural Network predictive model

The relationship between data skewness and accuracy of Artificial Neural Network predictive model

A Larasati [1,*], A M Hajji[2], Anik Dwiastuti[1]

[1]Department of Industrial Engineering, Universitas Negeri Malang, Indonesia
[2]Department of Civil Engineering, Universitas Negeri Malang, Indonesia

E-mail: aisyah.larasati.ft@um.ac.id

Abstract. The purpose of this study is to investigate the relationship between data skewness in the output variable and the accuracy of artificial neural network predictive model. The artificial neural network predictive model is built using multilayer perceptron and consist of one output variable and six input variable, and the algorithm used is back propagation. Data used in this study is generated by conducting the simulations in 1000 cycles. Three categories of skewness used in the output variables are positive skewness, neutral, and negative skewness. The results show that data skewness does not have a significant effect on the accuracy of the artificial neural network predictive model. These results imply that artificial neural network predictive model has a higher capability to cope with skewed data due to its complexity in the hidden layer.

1. Introduction

Although data mining techniques have been widely adopted in various sectors, data mining models including predictive models built using data mining techniques, are generally still understood as black boxes. Therefore, understanding the arguments behind the predictive model built using data mining techniques is crucial because it is necessary to assess the level of confidence of the model [1]. Understanding model parameters also provide an insight into the conditions of the model that can be used or to convert unreliable models into reliable ones.

One important factor that influences the determination of model parameters is data skewness. Understanding the distribution and variability occurring in the data is generally accomplished by a skewness investigation that is often expressed as a symmetry level of data. Although skewness of data can be visually evaluated through a data plot, still many coefficients need to be considered to ensure data preprocessing performed produce an accurate analysis [2].

The application of machine algorithm, such as the artificial neural network model (ANN), becomes more popular when it links to the problem of skewed (imbalanced) data. Datasets obtained from survey studies are mostly skewed, in which majority of the cases were classified into a certain class and only small number were classified into the other classes (minor). Many existing classification system tend to build a model without considering the proportion of each class. When data is highly skewed, classification model tends to misclassify minority class. Thus, when skewed data is not handled properly, the model tends to have high misclassification rates (poor prediction capability) [3].

[*] Corresponding author: aisyah.larasati.ft@um.ac.id

Based on this background, this study aims to explore the relationship between data skewness and the accuracy of the artificial neural network as a predictive model.

2. Literature Review

2.1 Skewness

Skewness is intended to represent how close the data is from the symmetry (or sometimes expressed as how much data distribution to the distribution of normality), where one side of the data distribution is more "stretched" than the other. Although symmetrically distributed data is preferred in statistical processing with the distribution (-∞, ∞), the concept of asymmetry is also common with distribution (0, ∞). The study of data skewness has been discussed by many statisticians, e.g. Pearson proposed gamma distribution as an alternative to model non-symmetric and non-normal data distribution as well as pareto that is focused on skewed data during data pre-processing [2].

The problem of distribution imbalance in the data classes (skewness) is quite often found in the digital age today. Skewness is often associated with a marginal distribution of unbalanced data. The data collected massively in this era of information technology development is sometimes difficult to understand because of the problem of data skewness and the marginal distribution of unbalanced data [4]. This problem refers to the skewness in the underlying marginal data distribution, which in turn, poses many difficulties in the learning algorithm that was constructed when constructing a model. To solve emerging problems arising from skewness and marginal distribution of data, efforts can be made primarily in two categories: model-oriented solutions (finding appropriate models) and data-oriented techniques (looking for appropriate data processing techniques).

2.2 Artificial Neural Network

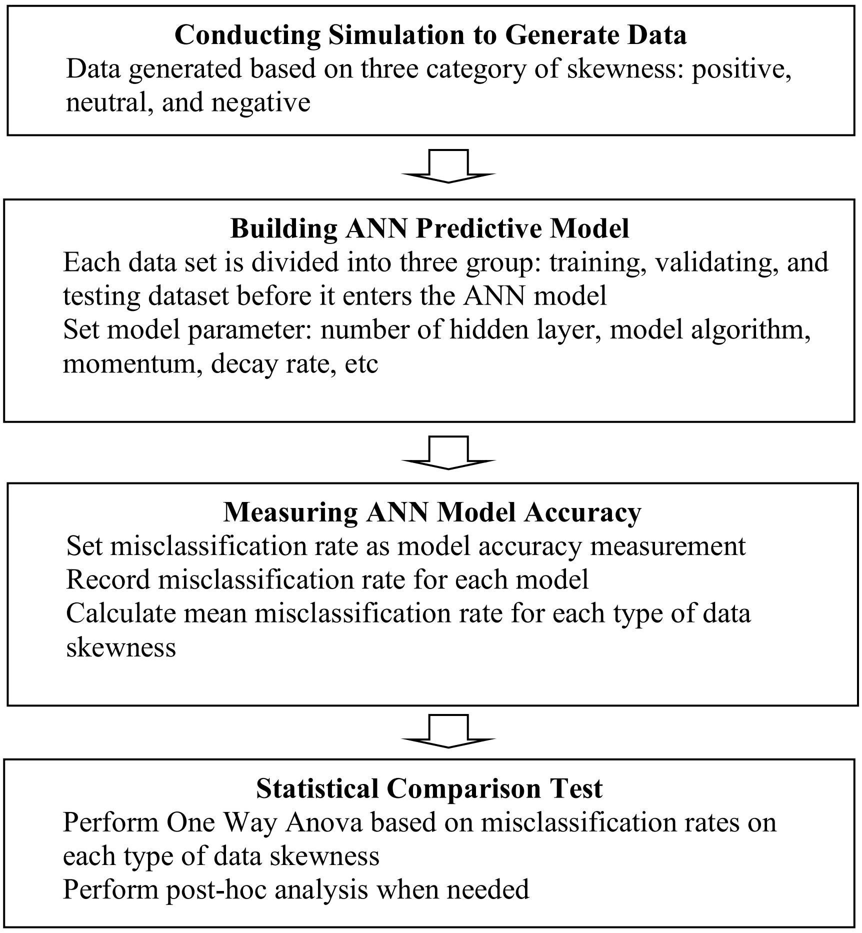
Artificial Neural Network (ANN), also known as Neural Network model (NN), is a mathematical or computational model based on neural networks (neurons). Because ANN is designed based on human biological systems, this model is built on the relationship seen in the training data set by adapting synaptic connections that exist between neurons. ANN consists of artificially linked artificial neurons and processes information using a linking approach for calculation. An ANN model consists of input, hidden, and output layer. The hidden layer in ANN contains some nodes that calculate the weights of input based on external or internal information that gets into the network during the learning process. ANN is able to derive the desired information from complicated or incorrect data, to determine patterns and to recognize trends that are actually very complex or almost impossible identified by humans or other statistical techniques. ANN is the best example of adaptive learning that can be designed to perform real time operations with a high level of fault tolerance [5].

In the simplest ANN framework, there is only one hidden layer, a layer containing functions to connect the input layer and output layer. Each layer in ANN can contain more than one variable (node). Each variable (node) is connected by using a one-to-one relationship so that there is a relationship pattern between each node in the input layer and each node in the output layer.

3. Research Methodology

This study is a type of simulation study. Data is generated in 1000 cycle in which each cycle contain 200 data set. The study calculates mean of misclassification rates in every 100 cycles. Resulted data from simulation has negative, neutral, and positive skewness. Generated data in this study is classified to be negatively skewed if it has skewness level < - 0.05, and it is classified as positively skewed if it has skewness level > 0.05. Otherwise, data is classified to be neutral.

Each data set is comprised of one dependent and six independent variables. After generating data, this study continued with building Artificial Neural Network Model (ANN) based on that various data skewness. The ANN model is built by splitting each data set into three categories, training (50%), validating (30%), and testing (20%) data set. The accuracy of ANN model is measured by the average value of mean classification rates resulted from each cycle. The misclassification rates mean the proportion of disagreement between predicted target variables compared to the actual target variable. Briefly, the research procedures applied in the study is shown in Figure 1.



Conducting Simulation to Generate Data
Data generated based on three category of skewness: positive, neutral, and negative

Building ANN Predictive Model
Each data set is divided into three group: training, validating, and testing dataset before it enters the ANN model
Set model parameter: number of hidden layer, model algorithm, momentum, decay rate, etc

Measuring ANN Model Accuracy
Set misclassification rate as model accuracy measurement
Record misclassification rate for each model
Calculate mean misclassification rate for each type of data skewness

Statistical Comparison Test
Perform One Way Anova based on misclassification rates on each type of data skewness
Perform post-hoc analysis when needed

Figure 1. Research Procedures

4. Results and Discussions

Statistic descriptive of the mean misclassification rates results show that negative skewness data has the highest mean and the lowest standard deviation of misclassification rates, while neutral skewness results in the highest standard deviation and the lowest mean of misclassification rates. Detail results of statistics descriptive of mean of misclassification rates are shown in Table 1.

Table 1. Statistic descriptive of misclassification rates

Type of Skewness	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
				Lower Bound	Upper Bound
negative skewness	.3854	.00728	.00230	.3802	.3906
neutral	.3812	.01704	.00539	.3690	.3934
positive skewness	.3813	.01055	.00334	.3738	.3888
Total	.3826	.01204	.00220	.3781	.3871

Table 2. ANOVA of Misclassification Rates

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.000	2	.000	.379	.688
Within Groups	.004	27	.000		
Total	.004	29			

ANOVA test is conducted to test whether there are significant differences in the misclassification rates due to various skewness level of the dependent variable. The results of ANOVA as shown in Table 2 indicate that p-value = 0.688. Assuming $\alpha = 0.05$, the Anova test results indicate there is no significant difference in the mean misclassification rates caused by various skewness.

Skewness refers to a relative distribution of each category in polynomial data or data distribution toward its mean. Generally, skewed data affect the accuracy of the predictive model including artificial neural network since the performance of an artificial neural network model depend on how well the training data set to teach the model. As explained in [6], skewed data, also known as imbalanced data occurs when certain classes/categories have higher occurrence compared to other classes/category. Since artificial neural network model learns from the pattern of training data set, fewer occurrence data tend to be ignored than the higher one. As a result, the misclassification rates tend to be higher in a skewed data than the uniform distributed data. Thus, the solution to overcome this problem is performing data resampling or less probability sampling for the higher occurrence data until the number of occurrence between each category is equal. However, since in this study we define negative skewed data is data set in which the skewness value < -0.05, neutral data is data with skewness between -0.05 and 0.05, and positively skewed is data with skewness > 0.05, thus there is possibility this study generated simulated data with skewness not really differences although these data are negative, positive or neutral. As a result, the mean misclassification rates are not significantly different. Another possibility is the fact this study generates the same total number of positive, neutral and negative skewed data and applies continued learning, the ANN predictive model has learnt the pattern for positive, neutral and negative skewed data in the same amount thus the resulted mean misclassification rates are not significantly different.

Accuracy based on misclassification rate is not suitable to measure ANN predictive model performance with skewed data [7]. The more suitable accuracy measurement for skewed data is the geometric mean of accuracies per class (*g-mean*), which can be measured as the number of correct classification in a certain category divided by the number of total occurrence in that related category. This study also proposed the use of random over-sampling to cope with skewed data. The accuracy measurement could be not really suitable with the generated data set, thus the mean misclassification rates from positive, neutral, and negative skewed data set are not significantly different.

5. Conclusion

Based on the results, this study concludes that skewness in the dependent variable has no significant effect on the accuracy of the artificial neural network model (ANN). Data that have positive, neutral and negative skewness on the dependent variable have mean misclassification rates that do not differ significantly. One possible reason for this finding is the use of continued learning in the model building step tend to make the pattern of lower occurrence data have learned by the model better. Thus, the ANN predictive model has learned the pattern data at the same level even the data is positively or negatively skewed. Another possible reason for this finding is the mean of misclassification rates is not suitable to measure the predictive model accuracy when the data set is skewed. Thus, the study suggests further investigation to compare the accuracy when the performance measurement is based on mean misclassification rates and the geometric mean of accuracy.

6. Acknowledgment

The authors acknowledge to the Institute of Research and Community Services (LP2M) *Universitas Negeri Malang* for managing the university research fund (PNBP) 2018 as the financial source and support for this study.

7. References

[1] Ribeiro MT, Singh S, Guestrin C 2016 *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* ACM 1135–44.

[2] Belzunce F, Mulero J, Ruiz JM, Suárez-Llorens A 2016 *Environ Ecol Stat.* **23** (4) 491–512.

[3] Liu Y, An A, Huang X. 2006. Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles. 107–8.

[4] Abdi L, Hashemi S 2016 *IEEE Trans Knowl Data Eng.* **1** (1).

[5] Gill NS, Mittal P. A 2016 *J Theor Appl Inf Technol*. **87** (1) 1–10.

[6] Barandelaa R, Sanchezb JS, Garcia V 2003 *Pattern Recognit*. **63** (3) 849–51.

[7] Serrano-Guerrero J, Olivas JA, Romero FP, Herrera-Viedma E 2015 *Inf Sci* **311**, 18–38.