The effect of Kurtosis on the accuracy of artificial neural network predictive model

Aisyah Larasati^{1*}, Anik Dwiastutik¹, Darin Ramadhanti¹ and Aal Mahardika¹

¹Industrial Engineering Department, Universitas Negeri Malang, 65145 Malang, Indonesia

Abstract. This study aims to explore the effect of kurtosis level of the data in the output layer on the accuracy of artificial neural network predictive models. The artificial neural network predictive models are comprised of one node in the output layer and six nodes in the input layer. The number of hidden layer is automatically built by the program. Data are generated using simulation approach. The results show that the kurtosis level of the node in the output layer is significantly affect the accuracy of the artificial neural network predictive model. Platycurtic and leptocurtic data has significantly higher misclassification rates than mesocurtic data. However, the misclassification rates between platycurtic and leptocurtic is not significantly different. Thus, data distribution with kurtosis nearly to zero results in a better ANN predictive model.

1 Introduction

Currently, data mining technique has become a popular approach to analyze abundant amount of data since its capability to discover knowledge based on data pattern by applying a specific algorithm [1]. The implementation of data mining techniques, such as a predictive models, requires data preparation process. Data characterization, such as mean, standard deviation, kurtosis, is necessary to be identified during data preparation process [2]. The aim of data characterization is to evaluate whether data is normally distributed or not.

Normality of data distribution has been examined previously by many researchers. However, in real world people are still less care about non-normal data although the potential problem of non-normal data have been recently studied in educational and psychological research [3]. One of the commonly measures of normality of data distribution is kurtosis.

Kurtosis is an important point for interpreting data variability. Kurtosis is an indication factor that reveals some characteristics of normality of data distribution. People can easily determine the normality of data distribution by interpreting the shape of the kurtosis curve of the data being examined. Kurtosis shows whether the data have a symmetrical distribution or not. The kurtosis value of normal distribution data is 3 (excess kurtosis is exactly 0). This normal kurtosis called mesocurtic. Platycurtic is a type of data distribution which has kurtosis value < 3 (excess kurtosis < 0). Platycurtic has a shorter tail and lower

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

center peak. The last type is leptokurtic (kurtosis value >3 , excess kurtosis >0). Leptokurtic has a longer tail and longer center peak [4].

Identifying non-normality in data distribution needs a specific measure of it, such as kurtosis. The problem is many result of kurtosis measure are hard to understand. Only few of researchers report normality measurement, such as kurtosis, in their data examination [3]. This fact implicitly shows that kurtosis is not a commonly used in the research since many people are still unfamiliar and feel this measure is too complicated to use. In addition, the use of kurtosis requires some basic statistic data such as means and standard deviation. Thus, some researchers feel too hard to measure them.

Furthermore, people also feel that kurtosis creates many difficulties in learning algorithm too. This makes people hardly create a predictive model. In many cases, people just know how to build a predictive model on a normal distribution data (symmetrical distribution). Therefore, this study aims to explore the effect of kurtosis level on the accuracy of artificial neural network predictive model.

2 Literature review

Data Mining is a collection of a number of computational approaches [5]. Data mining is a solution to organize the 'big data' (due to the large volume of data) in an efficient way. People can predict the result and take the conclusion with data mining. There are many techniques in data mining, one of them is a predictive model.

Artificial Neural Network (ANN) is a predictive model that is able to solve non-linear problem in modern way [6]. An artificial neural network is a network of a group of small processing units that are modeled for the behavior of human neural network. Sometimes, there is a problem that takes physical factor. ANN can solve the non-linear problem by analyzing node in the input layer and node (results) in the output layer. ANN predictive models have been used in wide application, such as market predictions, meteorological and network traffic forecasting [6].

ANN has some unique characteristics in its ways of working. Imitate how human brain work, ANN contain some neuron (nodes). In a neuron (node) contains one input data, indeed is raw data or data that has been processed from another neuron (node). Raw data is a type data that have not been processed yet. Therefore, for the output, a neuron (node) contains one final result or a data that have been processed by another neuron. A collection of some neuron is called neuron network.

Neuron network has some layers that has its own function. First, there is input layer for connecting data source to other neuron networks. Second, there is a hidden layer that must increasing network capability to solve the problem. Hidden layer imitates connector neuron in the human brain. Amount of hidden layer can be more than one, depend on the problem complication. The more amount of hidden layer, the more complicated to organize it. The last is output layer, this layer catches the last network result.

ANN has the same method to analyzing its program. The supervised learning algorithm is one of them. The difference between the output pattern generated and the desired output called error. Error is used to correct the weight of the ANN, so that the ANN will be able to produce output close to the desired output. Example of methods have used this algorithm are Back Propagation, Perceptron, Adaline, Hebian, etc. Back Propagation is the most popular method and is used widely in any kind of task [7].

Knowing kurtosis itself can help choosing the best method for this research. Kurtosis can analyze data normality by showing the result with the measurement. The best predictive model is a model that results in the highest accuracy result.

3 Research Methodology

This study is conducted in three phase. The first phase aims to generate data set. The data is generated using simulation approach as many as 1000 times based on various kurtosis. Each data set consists of 200 pair data points (input and output). The input data represents six independent variables and the output data represents one dependent variable. The second phase is to build the artificial neural network (ANN) using the generated data. Before being used to build the ANN models, data set is split into three data sets: training, validating and testing data sets. Each of training, validating, and testing data set comprise of 20, 30, and 50 data points respectively. The last step is to measure the performance of the ANN model, which is represented by the misclassification rate. The lower the misclassification rate, the better the model is. The misclassification rate is calculated based on the proportion of disagreement between the predicted-output from the ANN model and the actual outcome from a testing data set. In brief, the procedure of this research is shown in Figure 1.

Fig. 1. The Procedure of the Research.

4 Results and Discussion

The results of the accuracy of the artificial neural network predictive model is shown in Table 1. The ANOVA test results are shown in Table 2. Since the ANOVA results show that there are significant differences among the means of misclassification rates resulted

from different level of kurtosis, then the post hoc test is conducted using LSD test. The result of post hoc test is shown in Table 3.

Table 1. Descriptive statistics of misclassification rates.

Type of Kurtosis	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
				Lower Bound	Upper Bound
platycurtic	.38390	.007310	.002312	.37867	.38913
mesocurtic	.37380	.009716	.003072	.36685	.38075
leptocurtic	.38900	.014000	.004427	.37899	.39901
Total	.38223	.012165	.002221	.37769	.38678

The ANOVA test results show p-value = 0.012. Assuming $\alpha = 0.05$, it implies that there is a significant differences in mean misclassification rate of the artificial neural network model due to various level of kurtosis in the output data distribution. Kurtosis in dependent variable means the mode of the frequency on the dependent variable. The higher the excess kurtosis from 0, the higher the noise in the model and at the end the higher the misclassification rate. As explain in [8], excess kurtosis in dependent variable causes higher noise in the predictive model than excess value in variance. A post hoc test using LSD test is conducted to determine which type of kurtosis leads to the differences.

Table 2. Anova of misclassification rates.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.001	2	.001	5.221	.012
Within Groups	.003	27	.000		
Total	.004	29			

Table 3. Post hoc test using LSD.

(I) KURTOSIS	(J) KURTOSIS	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
platycurtic	mesocurtic	.010100*	.004788	.044	.00028	.01992
	leptocurtic	-.005100	.004788	.296	-.01492	.00472
mesocurtic	platycurtic	-.010100*	.004788	.044	-.01992	-.00028
	leptocurtic	-.015200*	.004788	.004	-.02502	-.00538
leptocurtic	platycurtic	.005100	.004788	.296	-.00472	.01492
	mesocurtic	.015200*	.004788	.004	.00538	.02502
*. The mean difference is significant at the 0.05 level						

The post hoc test results, as shown in Table 3, indicate that platycurtic data has no significant differences in its mean misclassification rate compared to leptocurtic data since the p-value = 0.296 (p-value > α) for the mean differences between platycurtic and leptokurtic. Meanwhile, mesocurtic data has a significant difference in its mean misclassification rate compared to platycurtic and leptokurtic data since both p-value for these mean differences are less than 0.05. The results of post

hoc test using LSD also confirms that mesocurtic data has a lowest mean misclassification rate compares to other type of data distribution, platycurtic and leptokurtic data. As explain in [8], leptokurtic curve results in higher volatility model. A volatility model means the accuracy of a model is easily to change due to randomization in building the predictive model. Thus, it will be better to do a transformation on leptokurtic or platycurtic data in order to increase accuracy of a predictive model as well as its robustness.

5 Conclusions

Based on the results, this study concludes that show that the kurtosis level of the node in the output layer is significantly affect the accuracy of the artificial neural network predictive model. Platycurtic and leptocurtic data has significantly higher misclassification rates than mesocurtic data. However, the misclassification rates between platycurtic and leptocurtic is not significantly different. Thus, data distribution with kurtosis nearly to zero results in a better ANN predictive model. This results implies that a data transformation is needed if data distribution of a dependent/target/output variable show a platycurtic and leptocurtic distribution in order to increase the accuracy of the ANN predictive model.

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

References

1. T. L. Tseng, C. C. Huang, F. Jiang, and J. C. Ho, *Int. J. Prod. Res.*, **44** (2006)
2. C. Giraud-Carrier and O. Povel, *Intell. Data Anal.*, **7** (2003)
3. M. K. Cain, Z. Zhang, and K. H. Yuan, *Behav. Res. Methods*, **49** (2017)
4. A. D. Ho and C. C. Yu, *Educ. Psychol. Meas.*, **75** (2015)
5. N. S. Gill and P. Mittal, **87** (2016)
6. J. P. Donate, G. G. Sanchez, and A. S. De Miguel, *Int. J. Artif. Intell. Tools*, **21** (2012)
7. K. Soltanian, F. A. Tab, F. A. Zar, and I. Tsoulos, *2013 21st Iran. Conf. Electr. Eng.*, **1–5** (2013)
8. S. J. Koopman, A. Lucas, and M. Scharth, *Rev. Econ. Stat.*, **98** (2016)