Increased accuracy in the classification method of backpropagation neural network using principal component analysis

Kristin Lorensi Sitompul, Muhammad Zarlis, Poltak Sihombing

Email: sriwijaya11121987@gmail.com, maniadu@yahoo.com

Abstrac. Water and air in life are needed in every human and living creature on earth, especially with the status of water quality and air quality status that must be known to humans. Water and air quality status has 120 records with 8 attributes consisting of 4 classes and 1096 records with 5 attributes consisting of 6 classes. Water and air quality classification can affect performance in data grouping. So from that the author tries to increase accuracy in classification by using the Neural Network Backpropagation algorithm with PCA. In this study, it is expected that the Backpropagation Neural Network algorithm using PCA is able to increase accuracy in the classification method.

1. Introduction

The method of classification neural network backpropagation is a very good method in the classification process given its ability to adapt network conditions to the data provided by the learning process. Behind the advantages of the classification method has weaknesses, this classification method can affect performance in grouping data. The backpropagation Neural Network Method has several weaknesses including in the accuracy of the data and the selection of attributes that are suitable for getting the best results in accuracy. So this study was proposed with the intention of improving the performance of neural network methods through distance weighing (similarity) using Principal Component Analysis (PCA). It is expected that this method is able to overcome weaknesses in backpropagation Neural Network and produce performance enhancements in classifying the data used.

2. Riview Of Literature

2.1. Backpropagation Neural Network Algorithm

Back propagation or back-propagation is one of the most used learning / supervised learning techniques. This method is one method that is very good in dealing with problems of complex patterns. Inside the back propagation network, each different unit in the input layer is connected to each unit in the hidden layer. Each unit in the hidden layer is connected to each unit with an output layer. This network consists of many layers (multilayer). When the network is given an input pattern as a training pattern, then the pattern leads to hidden layer units to then be forwarded to the units in the output layer [1].

Then the output layer units will give a response as output of the neural network. When the output is not as expected, the output will be spread backward to the hidden layer then from the hidden layer to the input layer This training step is a step to train an artificial neural network, namely by making changes in weight. While resolving the problem will be done if the training process has been completed, this phase is called the testing phase [2].

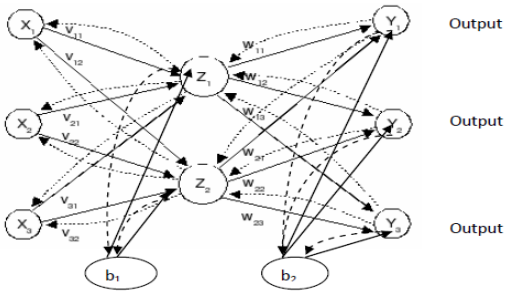


Figure 1. Backpropagation algorithm

Backpropagation is a model of artificial neural networks with multiple layers. As with other artificial neural network models, backpropagation trains the network to get a balance between the ability of the network to recognize patterns used during training and the ability of the network to respond correctly to input patterns that are similar (but not the same) to the patterns used during training [3].

The Backpropagation architecture consists of n entries (plus a bias), a hidden layaer consisting of p units (plus a bias), and m units of output units. W_{0j} andI_{0k} respectively are biases for the jth hidden unit and for kth output. The bias of $I_{0J}$and 0_{0k} behaves like weights where the bias output is always the same as 1. $W_{ij}$is the weighing weight between the i unit of the input layer with the hidden unit j layer, while $W_{jk}$is the connection weight between the hidden layer i unit and the unit to -j output layer [4]

2.2 Principal Component Analysis

PCA is used to explain the structure of the covariance variance matrix from a set of variables through linear combinations of these variables. In general, the principal component (PC) can be useful for feature selection and interpretation of variables. The conceptual scheme illustrated is how PCA can help to simplify the dimensions of the data through hypotheses, the dataset amounts to m the variables can be shown in Figure 2 below.

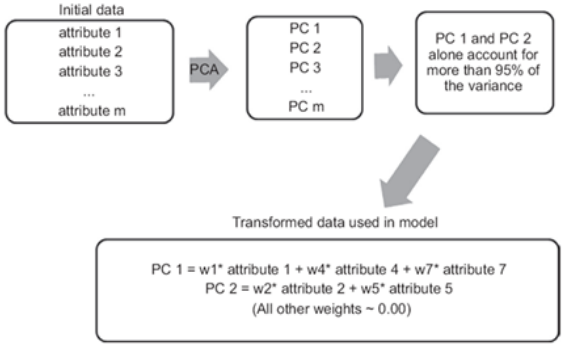


Figure 2. PCA conceptual model for the feature selection stage [5]

[6] Principal Component Analysis work procedures aim to simplify and eliminate factors or indicators that are less dominant and less relevant without reducing the intent and purpose of the original data from x random variables (matrix size nxn, where rows containing observations as much as n of variables random x)

3. Research methods

In this study, to find out the performance of the method used, 2 data sets will be used. The data used consists of a set of Water Quality Status data from the results of the Denades et al. (2016) [8] and a data set of Kota Pekan Baru Air Quality during 2014, 2015 and 2016 originating from City Government Air Laboratory Data Processing

Tabel 1. Data *Water Quality Status*

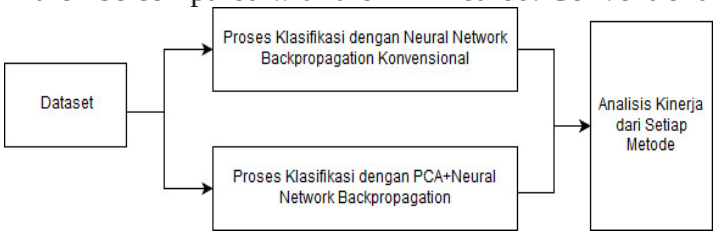
No	TSS (mg/L)	DO (mg/L)	COD (mg/L)	…	Quality Status
1	2	4	8	…	*Good Condition*
2	3	4.5	19.2	…	*Good Condition*
3	3	4.4	16	…	*Good Condition*
4	4	4.1	4.793	…	*Good Condition*
⋮	⋮	⋮	⋮	⋮	⋮
120	97	6.3	55.4	…	*Heavily Polluted*

Tabel 2. Data of Pekan Baru Air Quality

No	PM10	SO_2	CO	O_3	NO_2	Kategori
1	47	51	8	67	2	Sedang
2	48	51	9	37	2	Sedang
3	37	51	9	26	2	Sedang
4	24	50	2	51	1	Sedang
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1096	37	11	7	23	7	Baik

4. Stages of Method Performance Analysis

To see whether the proposed research model is able to improve the performance of classifying Neural Network Backpropagation (NN-BP), PCA + NN-BP performance analysis will be carried out based on the results of the Confusion Matrix (two-class prediction) tabulation which will then be compared with the NN method. Conventional BP.



5. Analysis and Discussion

This study seeks to improve the accuracy of the classification of Neural Network Backpropagation (NN-BP) using the Principal Component Analysis (PCA) method so that the two methods are compared with the conventional Backpropagation Neural Network classification method to measure method performance based on the level of accuracy produced.

6. PCA Analysis Results

Table 3. Dominant factors of w*ater quality status*

No	Factor	Variabel	Label	Eigenvalue	Loading value	Variance %
1	PC 1	X_3	COD (mg/L)	216.80	0.477	27.10%
2	PC 1	X_4	BOD (mg/L)	216.80	0.493	27.10%
3	PC 1	X_8	Pij	216.80	0.487	27.10%
4	PC 2	X_6	Fecal Coliform (mg/L)	191.20	0.366	23.90%
5	PC 2	X_7	Total Coliform (mg/L)	191.20	0.464	23.90%

Table 4. Dominant factors of Pekanbaru air quality

No	Factor	Atribute	Eigenvalue	Loading Value	Variance %
1	PC 1	PM10	284.5	0.489	56.90%
2	PC 1	CO	284.5	0.510	56.90%
3	PC 1	O_3	284.5	0.470	56.90%

7. Results of Accuracy of the Classification Model

Propagation Results Weight of LayerPCA + NN-BP Output Dataset Water Quality Status

Node	**Bobot*Output Layer (Class)***			
	Good Condition	**Lightly Polluted**	**Medium Polluted**	**Heavily Polluted**
Node 1	-4.221	-4.337	4.452	1.162
Node 2	-2.103	-1.658	-4.700	4.384
Node 3	-5.665	5.979	2.386	0.767
Node 4	-2.364	-1.849	-3.297	3.901
Node 5	-4.567	-3.878	5.214	0.958
Node 6	-2.068	-1.683	-4.444	4.163
Threshold	2.439	-2.490	-6.596	-7.545

Propagation Results for LayerPCA + NN-BP Output Weight Pekanbaru City Air Quality Data

Node	**Bobot*Output Layer (Class)***					
	Medium	**Good**	**No Healthy**	**Poor**	**Dangerous**	**No Data**
Node 1	3.373	0.036	1.064	-0.987	-2.807	-1.244
Node 2	-1.758	3.383	-2.246	-1.034	-2.090	-0.975
Node 3	-3.469	4.177	-2.619	-0.973	-1.986	-0.957
Node 4	-0.685	2.793	-2.039	-1.041	-2.103	-1.023
Node 5	-3.588	4.278	-2.671	-0.977	-2.020	-0.984
Node 6	3.932	0.811	0.239	-0.972	-2.867	-1.093
Threshold	-3.517	-6.719	-0.440	-1.606	1.092	-2.541

Performance metrics NN-BP classification with PCA + NN-BP (dataset for Water Quality Status)

Classification Model	Accuracy	Classification error
NN-BP	94.29 %	5.71%
PCA + NN-BP	97.14 %	2.86 %

8. Conclusion

1. In the research conducted on the classification model of Backpropagation Neural Network (PCA + NN-BP), using DatasetWater Quality Status which has been simplified into 5 attributes, 4 classes and 117 instances with an accuracy rate of 97.14% with a classification error rate of 2.86% . Meanwhile, the classification model of Conventional Backpropagation Neural Network (NN-BP) using 8 attributes with 4 classes from the dataset Water Quality Status has an accuracy rate of 94.29% with a classification error rate of 5.71%.
2. Based on the results of testing of the four classification models, it can be concluded that the success of PCA can be used as a reference to improve the performance accuracy of the Neural Network Backpropagation classification model.

References

[1] Solikhun, A. P. Windarto, Handrizal, and M.Fauzan, "Jaringan Saraf Tiruan Dalam Memprediksi Sukuk Negara Ritel Berdasarkan Kelompok Profesi Dengan Backpropogation Dalam Mendorong Laju Pertumbuhan

[2] A. Gupta and M. Shreevastava, "Medical Diagnosis using Back propagation Algorithm," Int. J. Emerg. Technol. Adv. Eng., vol. 1, no. 1, pp. 55–58, 2011.

[3] D. O. (Faculty of I. E.-G. U. Maru'ao, "Neural Network Implementation in Foreign Exchange Kurs Prediction," 2010.

[4] Sumijan, A. P. Windarto, A. Muhammad, and Budiharjo, "Implementation of Neural Networks in Predicting the Understanding Level of Students Subject," Int. J. Softw. Eng. Its Appl., vol. 10, no. 10, pp. 189–204, 2016

[5] Kotu, V., & Desphande, B. 2015. Predictive Analytics and Data Mining. Waltham, USA : Morgan Kaufmann Publishers.

[6] Jolliffe, I.T. 2002.*Principal Component Analysis*. 2nd Edition. Springer-Verlag: New York.

[7] Johnson, W.A. &Wichern, D.W. 2007.*Applied Multivariate Statistical Analysis*.6thEdition. Pearson Prentice Hall: New Jersey.

[8] Danades, A., Pratama, D., Anggraini, D., Anggriani, D. 2016. Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status. *International Conference on System Engineering and Technology*, pp. 137-141.