Available online at sjuoz.uoz.edu.krd

Science Journal of University of Zakho

Vol. 5, No. 4, pp. 314 –318, Dec.-2017

p-ISSN: 2410-7549
e-ISSN: 2414-6943

NORMALIZATION METHODS FOR BACKPROPAGATION: A COMPARATIVE STUDY

Adel S. Eesa*, Wahab Kh. Arabo

Dept. of Computer Science, Faculty of Science, University of Zakho, Kurdistan Region - Iraq (adel.eesa@uoz.edu.krd).

Received: Jun. 2017 / Accepted: Dec., 2017 / Published: Dec., 2017 *https://doi.org/10.25271/2017.5.4.381*

ABSTRACT:
Neural Networks (NN) have been used by many researchers to solve problems in several domains including classification and pattern recognition, and Backpropagation (BP) which is one of the most well-known artificial neural network models. Constructing effective NN applications relies on some characteristics such as the network topology, learning parameter, and normalization approaches for the input and the output vectors. The Input and the output vectors for BP need to be normalized properly in order to achieve the best performance of the network. This paper applies several normalization methods on several UCI datasets and comparing between them to find the best normalization method that works better with BP. Norm, Decimal scaling, Mean-Man, Median-Mad, Min-Max, and Z-score normalization are considered in this study. The comparative study shows that the performance of Mean-Mad and Median-Mad is better than the all remaining methods. On the other hand, the worst result is produced with Norm method.

KEYWORDS: Normalization, Neural network, Back propagation.

1. INTRODUCTION

Artificial Neural Networks (ANNs) are one of the most successful learning models. They have the versatility to approximate a wide range of complex functions representing multi-dimensional input-output maps. NN also has inherent adaptability, and can perform strongly even in noisy environments. IT is used successfully for identification of complex, unclear, or incomplete patterns. The most successful applications of NN are classification such as in (B. Dębska, 2011; Cal, 1995; Diane M. Miller, 1995; Markku Siermala, 2008; Sajad JASHFAR, 2013; Tamer Ölmez, 2003). In addition, NN is widely used in pattern recognition such as in (Birendra Biswal, 2013; Bishop, 1995; Jiří Grim, 2008; Seref SAGIR OGLU, 2000; Teena MITTAL, 2016).
Normalizing data such as scaling data between (0, 1) may enhance the accuracy and the performance of the mining algorithms including NNs. Many researchers applied normalization methods on BP to improve learning process such as (Kim, 1999; Kyung Whan Kim, 2005; Norlida, 2004). In this paper several normalization methods are proposed for BP. The proposed normalization methods are tested using several UCI real world datasets.
The structure of this paper is organized as follows. The next sections (Section 2 and 3) provide a background and a brief overview of BP and normalization methods. While, Section 4 highlights the methodology of the current study. Section 5 presents the experimental setup, results and discussion. Finally, Section 6 introduces the conclusions and future work.

2. BACKPROPAGATION

Backpropagation algorithm (D. E. Rumelhart, 1986) is mostly known as the successful tool used as training for feed-forward neural networks. It uses the input vector and the corresponding target (class) to train a given feed-forward multilayer neural network. When each input sample is passed to the network, the network checks its output and then it compares its output with a desired output. The difference between the NN output and the desired output (error) is used to adjust the connection weights. BP algorithm uses widrow-Hoff delta learning rule to adjust the connection weights by calculating the mean square error of the NN output and the desired output. The set of the input samples are repeatedly passed to the network until the error value is minimized. The main steps of BP algorithm can be stated as follows:

(1) Initialization of weights: for each connection weight assign a small random value between (0, 1).
(2) Feed-forward computation: each neuron in the input layer receives an input value and propagates this input to each neuron in the hidden layer. At each hidden neuron the activation function is calculated and propagated to each neuron in the output layer. Then, the output neuron calculates the activation function to form the response of the given input pattern.
(3) Back propagation of errors: each output neuron calculates the difference between its output and the corresponding desired output to determine the associated error of that neuron. Then, these errors are distributed from output layer back to all the neurons in the previous layers.
(4) Update all weights and biases.

3. DATA NORMALIZATION

Normalization can be a primary process in the analysis to compare data having different domain values. It is important to ensure that the data being compared is really comparable. Normalization transfers data from its domain to a specific range such as between (0, 1).

3.1 Decimal scaling normalization method

The normalization process in this method is by moving the decimal point of values of attribute *X*. this movement of decimal points totally depends on the maximum absolute value of X. A

* Corresponding author

new value *nv* corresponding to *v* is produced using Equation (1):

$$nv = f(v) = \frac{v}{10^c} \qquad (1)$$

Where *c* is the minimum number of position that such that the maximum value drop to (0,1) (Luai Al Shalabi, 2006). To illustrate, suppose that the range of feature *X* is −600 to 35. The maximum absolute value of *X* is 600. The normalization by decimal scaling will divide each value by 1000 (*c* = 3). So, −600 becomes −0.6 while 35 will be 0.035

3.2 Min-Max normalization method

In this technique the attribute will be rescaled from its domain to a new range of values such as between (0, 1) (Luai Al Shalabi, 2006). The formulation of this method is as follows:

$$nv = f(v) = \frac{v - \min(v)}{\max(v) - \min(v)} \qquad (2)$$

3.3 Norm normalization methods

Norm or length of any vector is equal to the Euclidean distance.(HELM, 2004). For instance, suppose that we have the following vector: *y* = {35, 36, 46, 68, 70}. Then, norm of y is calculated as follows:

$$||y|| = \sqrt{34^2 + 37^2 + 42^2 + 69^2 + 71} = 118.71$$

Thus, any element in y vector can be normalized using Equation (3):

$$nv = f(v) = \frac{v}{||y||} \qquad (3)$$

3.4 Z-Score normalization method

In this method the mean and standard deviation are used to normalize the input attributes values (Chen, 2012). The transformation is given in the Equation (4):

$$nv = f(v) = \frac{v - \mu}{\sigma} \qquad (4)$$

Where μ represents the mean value while σ is represent the standard deviation of data.

3.5 Median-Mad normalization method

This technique is based on the calculation of the Median Absolute Deviation (Christophe Leys, 2013). The normalized scores *nv* are calculated as follows:

$$nv = f(v) = \frac{v - median(v)}{MAD(v)} \qquad (5)$$

The Median-MAD normalization is insensitive to the presence of aberrant scores, does not keep the input distribution and does not transform the scores in a common interval. Hence, *MAD* is calculated as in (6):

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - median(x)| \qquad (6)$$

3.6 Mean-Mad normalization method

Instead of using Median, mean here is used in the above normalization as in (6) (Pham-Gia, 2011).

$$nv = f(v) = \frac{v - mean(v)}{MAD(v)} \qquad (7)$$

Hence, *MAD* is calculated as in (7):

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - mean(x)| \qquad (8)$$

4. METHODOLOGY

As it is mentioned before, several normalization methods (Min-Max, Decimal scaling, Norm, Z-Score, Mean-Mad, and Median-Mad) will be tested on BP learning algorithm using eight UCI datasets: Balance scale, Breast tissue, Gesture phase A, Glass identification, Haber man, Iris, User knowledge modelling, and Wine. The number of dimensions, the number of instances and the number of classes are varied in each set. All these datasets are available in (Lichman, 2013) and they are described in Table 1. The architecture of the artificial NN contains three layers, input layer, hidden layer and output layer. Number of nodes in the input and the output layers is depends on the number of features and the number of the classes in the used dataset, respectively. While the number of nodes in the hidden layer is calculated as: (number of nodes in the input layer)*1.5. To make fair comparison, the same initialize weights, learning rate and momentum are considered for each normalization method. The value of learning rate is set to (0.1), momentum is set to (0.5), while number of epoch is fixed to 5000.

Table 1. Datasets information

Name	# instances	#Dimensions	# classes
Balance scale	625	4	3
Breast tissue	699	9	2
Gesture Phase Segmentation (A1)	1747	19	5
Glass identification	214	9	7
Haber man	307	3	2
Iris	150	4	3
User Knowledge Modeling	258	5	4
Wine	178	13	3

5. EXPERIMENTS

The experiments were carried out using C# on a Dual-Core CPU 2.10 GHz laptop with 2 GB RAM. The obtained result shown in Figure 1, 2, 3, 4, 5, 6, 7, and 8 are the average of 10 independent runs. The experiments on each dataset are described below.

5.1 Balance Data set

Converge carves, shown in Figure 1, describes the obtained result for all normalization methods. It can be seen that the performance of Min-Max is better when compared with the performance of other normalization methods. With Mean-Mad, Median-Mad, Z-Score and Decimal-Scaling, the obtained results seem to be the same, while, the worst result is obtained when Norm normalization method is used.

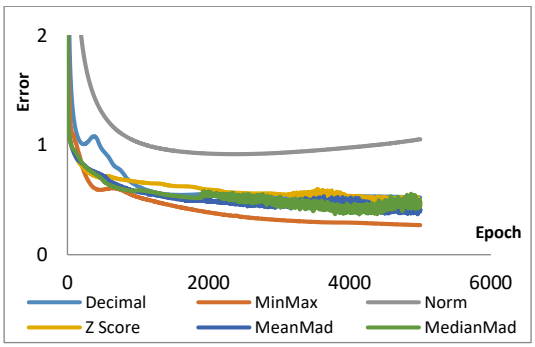

Fig. 1. Converge carves using Balance-scale dataset

5.2 Breast tissue dataset

From Figure 2 it is clearly seen that the performances of Median-Mad, Mean-Mad and Z-Score are better than the performances of Min-Max, Decimal-Scaling and Norm methods. The best result is obtained with Median-Mad, while the worst results are obtained with both Decimal-Scaling and Norm methods.

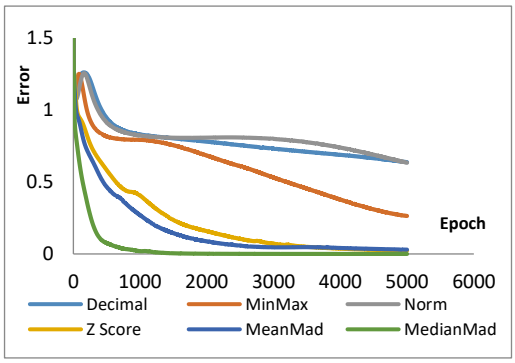

Fig. 2. Converge carves using Breast-tissue dataset

5.3 Gesture Phase Segmentation (A1) dataset

For this dataset, the obtained performances, shown in Figure 3, purport that the performance of Z-Score is better than the performances of all other normalization methods. Again the worst result is obtained, when Norm methods is used. The next best performance is obtained when the Mean-Man is applied on this dataset.

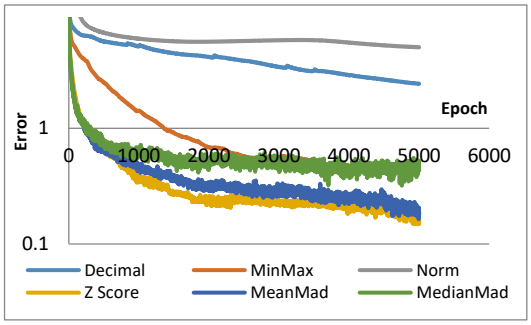

Fig. 3. Converge carves using Gesture phase A1 dataset

5.4 Glass identification dataset

As it is shown in Figure 4, it can be seen that the best performance is obtained with Mean-Mad method, while the next best is obtained with Z-Score method. On the other hand, the worst result is produced with both Norm and Decimal scaling methods.

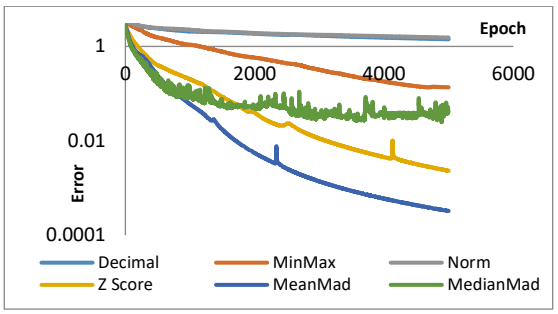

Fig. 4. Converge carves using Glass identification dataset

5.5 Haberman dataset

With this data, Mean-Mad method is performed better than other normalization methods as it is shown in Figure 5. The next best result is obtained with both Z-Score and Median-Mad which produced the same performance. While the worst result is produced with remained methods, but the worst one was Min-Max method.

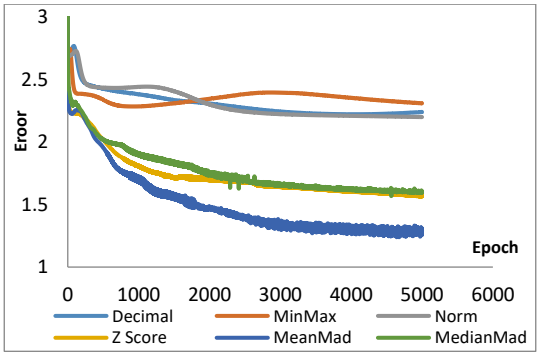

Fig. 5. Converge carves using Haber man dataset

5.6 Iris dataset

From Figure 6, the best and the worst results are obtained with Medina-Mad and Norm, respectively. On the other hand, there is no significant difference between the performances of the other methods.

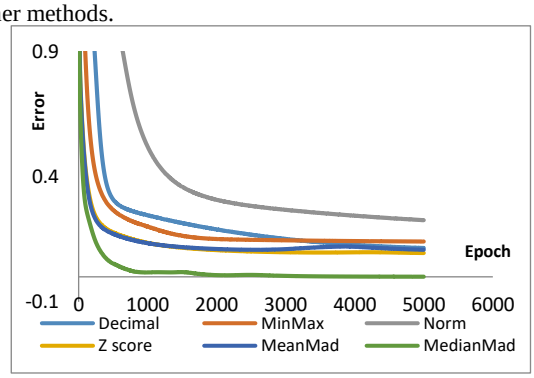

Fig. 6. Converge carves using Iris dataset

5.7 User Knowledge Modelling dataset

The best performance for this dataset is obtained with Z-Score, Mean-Mad and Median-Mad and there is no significant difference between them. But, they are also given unstable converges as shown in Figure 7. The worst result is produced when Min-Max and Decimal scaling is used with no significant difference between their performances.

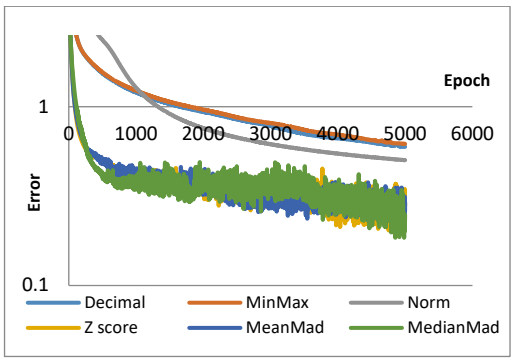

Fig. 7. Converge carves using User knowledge model dataset

5.8 Wine dataset

From Figure 8, the best result is produced with Median-Mad, Mean-Mad and Z-Score with small difference between their performances. The next best result is obtained with Min-Max while the worst result is obtained with Decimal scaling and Norm, respectively.

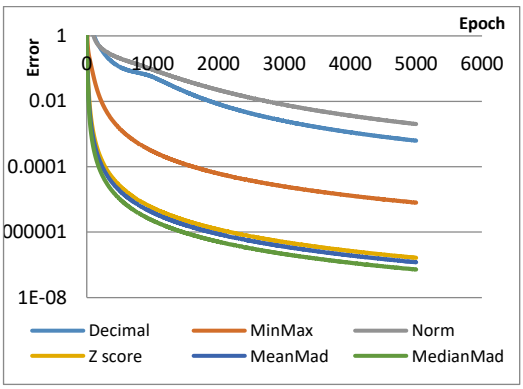

Fig. 8. Converge carves using Wine dataset

6. DISCUSSION

Table 2 describes the order of the best method for each dataset and it shows that the Median-Mad normalization method is performed in 4 datasets as the 1st best and for remaining 4 datasets it is performed as the 3rd best. Mean-Mad method is performed in 2 datasets as the 1st best and 1 dataset as 3rd best. But it is come as the 2nd best in 5 datasets. While Z-Score method is performed the 1st best in 1 dataset, 2nd best in 3 datasets, 3rd best in 3 datasets and 5th best in 1 dataset. The worst result is produced with Norm normalization method, where it comes the 4th best in 2 dataset and 6th best in 6 datasets.

As a final result, we can decide that the first best normalization methods that can be used with BP learning algorithm is Mean-Mad method and the second best comes beside Median-Mad method. While the third best performance is produced with Z-Score method. However, these three methods produced unstable coverage in 5 datasets, this is because of using a fixed learning rate and a fixed momentum for all experiments.

In addition, and to gain more valid and authentic results, these normalization methods were tested on intrusion detection dataset the (KDD Cup 99 dataset) available in (California, 1999). In this experiment, the number of train data is equal to 3000 records selected randomly from KDD Cup 99 dataset. While for test dataset the complete KDD Cup 99 test dataset is used which is equal to 494021 records. The number of the input neuron in neural network is equal to 39. The number of neurons in the hidden layer is set to 77, while the number of the output neurons is set to 2 corresponding to the number of classes in KDD Cup 99 dataset (Normal or Attack). The parameters of Backpropagation learning algorithm are set as follows: learning rate is set to 0.1, momentum is set to 0.5 and the number of epoch is set to 5000. The obtained result shown in Figure 9 describes the performance in the term accuracy rate of Mean-Mad, Z-Score, Decimal and Norm normalization methods. From Figure 9, as expected, it can be seen that the best performance is obtained with Mean-Mad method while the second best is obtained with Z-Score method. The worst result in this experiment is obtained when both Decimal and Norm normalization method are used.

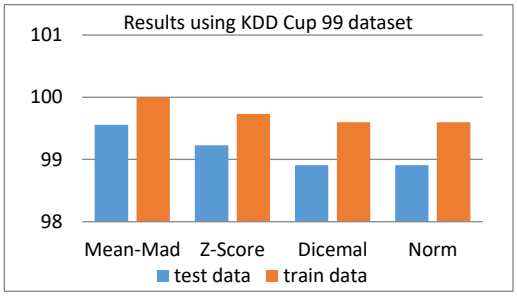

Fig. 9. Accuracy rate using KDD Cup 99 dataset

Table 2. The order of used normalization method for each dataset

Dataset	1st Best	2nd Best	3rd Best	4th Best	5th Best	6th Best
Balance scale	Min-Max	Mean-Mad	Median-Mad	Decimal	Z-Score	Norm
Breast tissue	Median-Mad	Mean-Mad	Z-Score	Min-Max	Decimal	Norm
Gesture phase A1	Z-Score	Mean-Mad	Median-Mad	Min-Max	Decimal	Norm
Glass identification	Mean-Mad	Z-Score	Median-Mad	Min-Max	Decimal	Norm
Haber man	Mean-Mad	Z-Score	Median-Mad	Norm	Decimal	Min-Max
Iris	Median-Mad	Z-Score	Mean-Mad	Decimal	Min-Max	Norm
User knowledge modeling	Median-Mad	Mean-Mad	Z-Score	Norm	Decimal	Min-Max
Wine	Median-Mad	Mean-Mad	Z-Score	Min-Max	Decimal	Norm

7. CONCLUSION AND FUTURE WORK

In this study, we investigated the use of six normalization methods to find the best method that could work better with Backpropagation learning algorithm. The performance of each method was evaluated using eight datasets. Empirical results revealed that the performance of Mean-Mad method was better than all the other methods. While the second best performance was produced with Median-Mad method. However, the performances of these two methods were unstable converge. On the other hand, the worst result was produced with Norm normalization method when it comes as a sixth best in six datasets. As a future work, one can use the best produced method with Backpropagation algorithm for classification purposes.

REFERENCES

B. Dębska, B. G.-Ś. (2011). *Application of artificial neural network in food classification*. Paper presented at the 12th International Conference on Chemometrics in Analytical Chemistry.

Birendra Biswal, A. J. P. (2013). *A New Approach to Time–Time Transform and Pattern Recognition of Non-stationary Signal Using Fuzzy Wavelet Neural Network.* Paper presented at the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA).

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*: Oxford University Press, Inc. New York, NY, USA.

Cal, Y. (1995). Soil classification by neural network. *Advances in Engineering Software, Elsevier, 22*(2), 95-97. doi: 10.1016/0965-9978(94)00035-H.

California, U. o. (1999). KDD Cup 99 Retrieved 10-2-2017, 2017, from http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

Chen, Y. W. a. H.-J. (2012). Handbook of Anthropometry V. R. Preedy (Ed.) doi:10.1007/978-1-4419-1788-1.

Christophe Leys, C. L. U., Olivier Klein, Philippe Bernard and Laurent Licata. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 40*(4), 764-766. doi: 10.1016/j.jesp.2013.03.013.

D. E. Rumelhart, G. E. H., R. J. Williams. (1986). *Learning internal representations by error propagation.* . Paper presented at the In D. E. Rumelhart, & J. L. McClelland (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition, Cambridge, MA: MIT Press.

Diane M. Miller, E. J. K., Soraya Rana. (1995). Neural network classification of remote-sensing data. *Computers & Geosciences, Elsevier, 21*(3), 377-386. doi: 10.1016/0098-3004(94)00082-6.

HELM. (2004). *Workbook Level 1 30.4: Introduction to Numerical Methods, VERSION 1*.

Jiří Grim, J. H. (2008). *Iterative principles of recognition in probabilistic neural networks*. Paper presented at the 17th International Conference on Artificial Neural Networks (ICANN).

Kim, D. (1999). Normalization methods for input and output vectors in backpropagation neural networks. *International Journal of Computer Mathematics, 71*(2), 161-171. doi: 10.1080/00207169908804800.

Kyung Whan Kim, D. K., Hun Young Jung. (2005). Normalization methods on backpropagation for the estimation of driver's route choice. *KSCE Journal of Civil Engineering, 9*(5). doi: 10.1007/BF02830631.

Lichman, M. (2013). UCI Machine Learning Repository, from http://archive.ics.uci.edu/ml.

Luai Al Shalabi, Z. S. (2006). *Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix.* Paper presented at the International Conference on Dependability of Computer Systems (DEPCOS-RELCOMEX '06), Washington, DC, USA.

Markku Siermala, M. J., Erna Kentala. (2008). Neural network classification of otoneurological data and its visualization. *Computers in Biology and Medicine, Elsevier 38*(8), 858-866. doi: http://dx.doi.org/10.1016/j.compbiomed.2008.05.002.

Norlida, H. (2004). *The Impact of Normalization Techniques on Performance Backpropagation Networks.* Masters thesis, Universiti Utara Malaysia. Retrieved from http://etd.uum.edu.my/id/eprint/1394.

Pham-Gia, T. L. H. (2011). The mean and median absolute deviations *Mathematical and Computer Modelling, 34*(7-8), 921-936. doi: 10.1016/S0895-7177(01)00109-1.

Sajad JASHFAR, S. E., Mehdi ZAREIAN-JAHROMI, Mohsen RAHMANIAN. (2013). Classification of power quality disturbances using S-transform and TT-transform based on the artificial neural network. *Turk J Elec Eng & Comp Sci, 21*, 1528-1538. doi: doi:10.3906/elk-1112-51.

Seref SAGIR OGLU, E. B., Mehmet ERLER. (2000). Control Chart Pattern Recognition Using Artificial Neural Networks. *Turk J Elec Eng & Comp Sci, 8*(2).

Tamer Ölmez, Z. D. (2003). Classification of heart sounds using an artificial neural network. *Pattern Recognition Letters, Elsevier, 24*(1-3), 617-629. doi: http://dx.doi.org/10.1016/S0167-8655(02)00281-7.

Teena MITTAL, R. K. S. (2016). Speech recognition using ANN and predator-influenced civilized swarm optimization algorithm. *Turk J Elec Eng & Comp Sci 24*, 4790 – 4803. doi: doi:10.3906/elk-1412-193