Hindawi
International Journal of Mathematics and Mathematical Sciences
Volume 2022, Article ID 3584406, 9 pages
https://doi.org/10.1155/2022/3584406

Research Article

Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification

Saichon Sinsomboonthong

Department of Statistics, School of Science, King Mongkut's Institute of Technology Ladkrabang, Chalongkrung, Bangkok 10520, Thailand

Correspondence should be addressed to Saichon Sinsomboonthong; saichon.si@kmitl.ac.th

Received 7 January 2022; Accepted 7 March 2022; Published 22 April 2022

Academic Editor: Birendra Nath Mandal

In this research, the normalization performance of the proposed adjusted min-max methods was compared to the normalization performance of statistical column, decimal scaling, adjusted decimal scaling, and min-max methods, in terms of accuracy and mean square error of the final classification outcomes. The evaluation process employed an artificial neural network classification on a large variety of widely used datasets. The best method was min-max normalization, providing 84.0187% average ranking of accuracy and 0.1097 average ranking of mean square error across all six datasets. However, the proposed adjusted-2 min-max normalization achieved a higher accuracy and a lower mean square error than min-max normalization on each of the following datasets: white wine quality, Pima Indians diabetes, vertical column, and Indian liver disease datasets. For example, the proposed adjusted-2 min-max normalization on white wine quality dataset achieved 100% accuracy and 0.00000282 mean square error. To conclude, for some classification applications on one of these specific datasets, the proposed adjusted-2 min-max normalization should be used over the other tested normalization methods because it performed better.

1. Introduction

In a classification process, an attribute may have a small range of values or a very large range of values. The attribute with a large range of values may unreasonably dominate the outcome of the classification process solely due to its larger numerical values. Therefore, these variables must be normalized before classification. Normalization or transformation is a process of adjusting the measured values on different scales into a same scale [1] so that all attributes are parallel with each other. The following are some papers involving normalization methods based on different approaches. In the comparison of Z-score and min-max normalization in k-nearest neighbor classification, the min-max and Z-score had the accuracy 88% and 79%, respectively [1]. The normalization efficiency of decimal scaling for k-nearest neighbor was higher than that of Z-score, median and min-max for k-nearest neighbor, Naïve Bayes, and artificial neural network using accuracy in classification on six datasets [2]. The statistical column normalization for k-nearest neighbor performed better than decimal scaling for each of the following classification methods: decision tree, k-nearest neighbor, support vector machine, artificial neural network, Naïve Bayes, and binary logistic regression [3]. In the research studied on data perturbation procedure for privacy protection with min-max using NB tree classification, the result of analysis showed that min-max is capable to protect confidential data and keep the efficiency of data mining method [4]. In some paper, some widespread feature normalization methods were examined and estimated, and in studying the effect on performance of classifier with breast tumor classification for estimating the normalization methods, back-propagation artificial neural network (BPANN) and support vector machine (SVM) classifier were used in comparison. The result concluded that min-max and classification by BPANN is the best accuracy, followed by softmax scaling and classification using BPANN [5]. The efficiency of min-max in normalizing violence video dataset for multilayer perceptron classification was higher than the normalization efficiency of Z-score. In addition, the accuracy of min-max in the range of [0,1] was nearly 98% high, while the accuracy of min-max in the range of [−1,1] was lower, at 59%. Much lower was the accuracy of Z-score, at 50% [6]. Through the impact of the normalization technique used in various sample sizes on the artificial neural network by Z-score, median, min-max and adjusted-3 min-max normalization, it was showed that the adjusted-3 min-max was the highest amount of explained variance. Considering precise classification percentage, it was found no significant difference between research data that was not normally distributed. Nevertheless, a sample size over 1000 more regular results can be managed with artificial neural network [7]. Finally, the effect of data preprocessing on the prediction of 305-day milk yield by artificial neural networks was examined with the impact of normalizations: Z-score, median, min-max, sigmoid, decimal scaling, D-min-max, and median on mean absolute deviation and five back propagation algorithms. The results showed that the best efficiency was the decimal scaling using artificial neural network with the Bayesian regulation algorithm [8]. Currently, several normalization methods are widely used: median, Z-score, statistical column, decimal scaling, adjusted decimal scaling, min-max, and adjusted min-max [1–8]. Each normalization method uses a different approach. For example, the simplest approach is Z-score, which uses the mean and standard deviation of each attribute [1]. A widely used method at the time of this study, min-max uses the range and minimum value of each attribute [2]. The proposed method attempted to improve on the original min-max method by multiplying min-max equation with the range of a new maximum and a new minimum value. The background of min-max and previous adjusted min-max normalization methods can be found in these papers [1–8]. To be more useful, the performance of this proposed method was evaluated not only against the performance of the original min-max but also against the performances of every existing normalization method based on different approach. The exact methods to achieve these two objectives are reported in the Materials and Methods section below.

2. Materials and Methods

2.1. Materials. Comparison of normalization methods was done with Microsoft Excel 2016 software, while performance comparison of neural network classification was done with RStudio version 4.1.0 software. The software programs were run on a notebook with i3, 2 GHz, Intel CPU, and 4 Gb of RAM under Windows 10.

2.2. Methods

2.2.1. Data Collection. Classification performance depends strongly on a training dataset; therefore, the author attempted to use a large variety of datasets to evaluate the classification performance of various normalization methods with an identical artificial neural network classifier in a testing dataset so that the reported performances would be useful to as many applications as possible. Six datasets were gathered from three websites: (1) the first was white wine quality dataset from UCI.com consisting of 1,500 total values with a high coefficient of variation (C.V.) (>30%) [9]. The dataset included Vinho Verde white wine quality data collected from Portugal on October 7, 2009. The dependent variable was the quality of the white wine (yes was good quality, and no was bad quality). The independent variables consisted of acidity (unit: pH), volatile acidity (unit: pH), citric acid (unit: pH), reduced sugar (unit: gram), chloride (unit: gram), sulfur dioxide (unit: gram), total sulfur dioxide (unit: gram), density (unit: kilograms per cubic meter), acidity (unit: pH), salt of sulfuric acid (unit: pH), alcohol (unit: gram), and quality score (unit: score). (2) The second was Pima Indian diabetes dataset from Kaggle.com consisting of 768 total values with a high C.V. [10]. The dataset included Pima Indian diabetes data collected from the National Institute of Diabetes and Digestive and Kidney Diseases of India in 1990. The dependent variable was whether a patient has contracted diabetes or not (yes represents contracted, and no represents not contracted). The independent variables consisted of number of times of pregnancy (unit: times), oral glucose concentration (unit: millimoles per liter), blood pressure (unit: mmHg), thickness of subcutaneous fat (unit: millimeter), concentration of insulin in the body 2 hours after serum injection (unit: microunit per milliliter), body mass index (unit: kilograms per square meter), risk of genetic diabetes (unit: percent), and age (unit: years). (3) Vertebral column dataset consisted of 310 total values with a high C.V. [11]. This dataset came from Kaggle.com. Data on human vertebral column were collected from Monte Klinikum Hospital, Brazil. The dependent variable represented vertebral column of the human body (normal represented normal bone, and abnormal represented abnormal bone). The independent variables consisted of pelvic floor (unit: millimeter), pelvic tilt (unit: millimeter), cervical spondylosis (unit: millimeter), sacral vertebrae (unit: millimeter), pelvic radius (unit: milliliter), and displacement of bone (unit: millimeter). (4) Indian liver patient dataset consisted of 575 total values with a low C.V. (<30%) [12]. This dataset came from Mldat.com containing Indian liver patient data from northeastern Andhra Pradesh, India, collected in 2012. The dependent variable represented whether or not liver disease was detected (yes represented liver disease detected, and no represented liver disease not detected). The independent variables consisted of patient's age (unit: years), total bilirubin (unit: milligrams per deciliter), bilirubin (unit: milligrams per deciliter), alkaline phosphatase (unit: unit per liter), enzyme alanine aminotransferase (unit: unit per liter), aspartate aminotransferase (unit: unit per liter), total protein in blood (unit: grams per deciliter), protein (unit: grams per deciliter), and ratio of albumin and globulin (unit: grams per deciliter). (5) Working hour dataset consisted of 956 total values with a low C.V. [13]. This dataset came from UCI.com containing data on working hours collected in the United States in 1995. The dependent variable represented a residential status (1 represented rented house and 2 represented owned house). The independent variables consisted of working time (unit: hour), income (unit: dollars), age (unit: years), year of study (unit: years), number of children (unit: person), and unemployment rate in the area (unit: percentage). (6) Avocado dataset consisted of 1,149 total values with a low C.V. [14]. This dataset came from Kaggle.com containing data on avocado collected from the National Institute of Diabetes and Digestive and Kidney Diseases, India, in 1990. The dependent variable represented the type of avocado (conventional represented conventionally grown avocados, and organic represented organically grown avocados). The independent variables consisted of total number of avocados sold (unit: piece), total number of avocados available for sale in PLU 4046 (unit: piece), total number of avocados available for sale in PLU 4225 (unit: piece), total number of avocados available for sale in PLU 4770 (unit: piece), avocado bag (unit: bag), small bag of avocado (unit: bag), medium size bag of avocado (unit: bag), and large bag of avocado (unit: bag). In this study, coefficient of variation (C.V.) is defined as $C.V. = (S\ D/\overline{X})$x100%, where $\overline{X}$ is the mean of the data, and $S\ D$ is the standard deviation. The author employed C.V. as the indicator of the level of difference of data points in a dataset and attempted to collect an extensive group of representative datasets with both low and high C.V.s to use in the comparison so that the evaluated performances would be applicable to as many real-world datasets as possible.

2.2.2. Research Methods. In this study, the following research methods were employed in the evaluation of the selected normalization methods: (1) normalization calculation by Excel; (2) training and testing dataset formation method; (3) concepts of tested normalization methods; (4) classification; and (5) comparison of normalization performance.

(1) Normalization Calculation by Excel. Normalization is a process for adjusting measured values in different scales into the same scale and can even be more complex to bring the probability distributions of the adjusted values into alignment [1]. The authors of [2–8] showed that good methods for any normalization processes were statistical column, decimal scaling, adjusted decimal scaling, and min-max. In addition to the four normalization methods mentioned above, this work evaluated four adjusted min-max methods: adjusted-1 min-max and adjusted-3 min-max were two min-max methods adjusted with a different adjusting equation reported in [6, 7], and adjusted-2 min-max and adjusted-4 min-max were min-max adjusted with our proposed adjusting equations. Normalization outcome of each method was calculated by Microsoft Excel 2016 on a notebook under Windows 10.

(2) Training and Testing Dataset Formation Method. Each collected dataset was divided into a training dataset and a testing dataset at 70 : 30 ratio. Five rounds of random-seed generation were performed with five different assigned initial seeds: 10, 20, 30, 40, and 50 in a 70 : 30 ratio. The training dataset was used to create the classification model, and the testing dataset was used to test the prediction accuracy of the classification model [15–19]. Table 1 shows the number of training and testing data points from each collected dataset.

(3) Concepts of Tested Normalization Methods. The concept of each of the eight tested normalization methods, i.e., its governing equation, is described below.

(i) Statistical column normalization: this technique normalizes each column as a transformed column value with length of one, $n(c_a)$. The value in each column (X) is subtracted by $n(c_a)$, divided by $n(c_a)$, and then multiplied by a small biased value (0.1) [20].

$$X^* = \{[X - n(c_a)]/n(c_a)\} \times 0.1. \tag{1}$$

(ii) Decimal scaling normalization: this method is a data transformation method, like the conventional Z-score normalization. The number of decimal points of each value of every attribute is changed according to the highest number of placeholders for the values in all columns [21].

$$X^* = X/10^j, \tag{2}$$

where the integer j is equal to the highest number of placeholders for the values in all columns.

(iii) Adjusted decimal scaling normalization: this procedure is decimal scaling normalization adjusted by replacing j with c+1 as the power of 10 in (2) [20].

$$X^* = X/10^{(c+1)}, \tag{3}$$

where

$$c = \log_{10}\max(X_i). \tag{4}$$

(iv) Min-max normalization: this technique [1, 2, 4, 5, 8, 22] subtracts the data values with the minimum and divides it by the range, i.e., the difference between maximum and minimum.

$$\begin{aligned} X^* &= [X - \min(X)]/\text{range}(X) \\ &= [X - \min(X)]/[\max(X) - \min(X)], \end{aligned} \tag{5}$$

where $\min(X)$ is the minimum; $\max(X)$ is the maximum; and range(X) is the difference between maximum and minimum.

The range is in the interval of [0, 1], and the length of the interval is 1.

(v) Adjusted-1 min-max normalization: this way is min-max normalization adjusted by a power of c over the range of $\max(X)$ and $\min(X)$ [20] as follows:

$$X^* = [X - \min(X)]/[\max(X) - \min(X)]^c, \tag{6}$$

where $c = 2$ is a constant added to the power of the denominator.

This range of (6) is in the interval of [0, 1], and the length of the interval is 1.

Table 1: Numbers of total, training and testing data points in six datasets.

Dataset	Total	Training dataset (70%)	Testing dataset (30%)
White wine quality	1,500	1,050	450
Pima Indian diabetes	768	537	231
Vertebral	310	217	93
Indian liver patient	575	402	173
Working hour	956	669	287
Avocado	1,149	804	345

(vi) Proposed adjusted-2 min-max normalization: the proposed method is min-max normalization adjusted by multiplying with a new range term, $\text{new}\max(X) - \text{new}\min(X)$.

$$X^* = \{[X - \min(X)]/[\max(X) - \min(X)]\}[\text{new}\max(X) - \text{new}\min(X)], \tag{7}$$

where $\text{new}\min(X) = 0.5$ and $\text{new}\max(X) = 2$.

The range of (7) is in the interval of [0, 1.5], and the length of the interval is 1.5.

(vii) Adjusted-3 min-max normalization: this way is min-max normalization adjusted by multiplying with a new range term, $\text{new}\max(X) - \text{new}\min(X)$, and then adding a new term, $\text{new}\min(X)$.

$$X^* = \{[X - \min(X)]/[\max(X) - \min(X)]\}[\text{new}\max(X) - \text{new}\min(X)] + \text{new}\min(X). \tag{8}$$

The range of (8) is in the interval of [0.5, 2], and the length of interval is 1.5 [6, 7, 23].

(viii) Proposed adjusted-4 min-max normalization: the proposed method is similar to adjusted-3 min-max normalization, adjusted by adding a new term, $\text{new}\max(X)$.

$$X^* = \{[X - \min(X)]/[\max(X) - \min(X)]\}[\text{new}\max(X) - \text{new}\min(X)] + \text{new}\max(X). \tag{9}$$

The range of (9) is in the interval of [2, 3.5], and the length of interval is 1.5.

(4) Classification. Classification is manipulation of parts in a collection into target classes. The objective of classification is precisely to predict the target class for any dataset. Classification models are trained by a training dataset and evaluated by running them on a testing dataset: comparing the predicted values with the target values of the testing dataset. Classification usually starts from partitioning the interested dataset into two smaller datasets: a training dataset for creating the model and a testing dataset for evaluating the model [1]. Since previous research studies [5–8] showed that the best classification technique was artificial neural network, the author was interested in using it for this comparative study of various normalization methods.

(i) Artificial neural network: this method is an artificial intelligence technique for calculating summation and activation functions from a group of data. Artificial neural network is a machine learning method that learns from data points in a dataset and predicts the outcome of new data points in another dataset. It consists of three layers or more than three layers: an input layer with several input nodes, a hidden layer with several hidden nodes, and an output layer with several output nodes. A hidden layer may consist of one hidden sublayer or multiple sublayers. Artificial neural network uses backpropagation to adjust the weights of connections between every two nodes. Initially, the weights for every connection between nodes are randomly generated between 0 and 1, and as the run progresses, these weights will be adjusted according to the nature of the data points. Essential components in each neuron are summation function and activation function. The summation function computes the sum of weight and input node from the input layer, while the activation function normalizes the values of the data from the summation function to the desired range. The activation function that we adopted in this study was a sigmoidal function [24].

(ii) Command: the code for artificial neural network in R programming language is as follows (some variables are specific to white wine dataset).

```
command > getwd()
>data = read.csv(file.choose(),header = T)
>str(data)
>gp = runif(nrow(data))
>gp
>data = data[order(gp), ]
>ind = sample(2,nrow(data),                    replace = T,
prob = c(0.7,0.3))
>train = data[ind = = 1,]
>test = data[ind = = 2,]
>library(neuralnet)
>set.seed(10)
>n = neuralnet(class ~ X1+X2+X3+X4+X5+
X6+X7+X8+X9+X10+X11+ X12,
data = train, hidden = 1,linear.output = FALSE)
>plot(n)
>output = compute(n,test[-13])
```

Table 2: Confusion matrix.

		Predicted class	
		Positive	Negative
Actual class	Positive	True positive (TP)	False negative (FN) type II error
	Negative	False positive (FP) type I error	True negative (TN)

```
>predict = output$net.result
>actual = data[ind = = 2,13]
>MSE = sum((predict-actual)^2)/nrow(test)
>MSE
>table(actual, round(predict)).
```

(5) Comparison of Normalization Performance. The performance of every tested normalization method was evaluated in terms of accuracy and mean square error of predictions achieved by the artificial neural network classifier, common to every method. Accuracy and mean square error were computed by a computer program coded by the author with RStudio version 4.1.0.

(i) Accuracy: the accuracy is an actual prediction percentage divided by total number of dataset [25]. In this study, accuracy was defined by

$$\text{Accuracy} = \{(TP + TN)/(TP + TN + FP + FN)\} \times 100\%, \tag{10}$$

where TP is true positive; TN is true negative; FP is false positive; and FN is false negative, defined as elements of confusion matrix (Table 2) described in [26].

(ii) Mean square error (MSE): the mean square error is a measure of prediction accuracy of a model. Smaller is better. The lower the value of MSE is, the closer the predicted value is to the actual value.

$$\text{MSE} = \sum_{i=1}^{n} \frac{e_i^2}{n} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}, \tag{11}$$

where e_i is an error; n is the sample size; y_i is the actual value; and $\hat{y}_i$ is the predicted value [27].

3. Results

The comparative prediction efficiency based on accuracy and mean squared error by artificial neural network in combination with each of the eight normalization methods on the six datasets (described in Research Methods section above) is shown in Tables 3–9. Each table shows those values on a particular testing dataset.

3.1. Comparative Results on White Wine Quality Dataset. The comparative results yielded by every normalization method on white wine quality dataset are tabulated in Table 3.

All normalization methods except statistical column normalization had the same full accuracy of 100%. In terms of MSE, adjusted-3 min-max normalization method provided the minimum MSE of 0.0000024, followed by proposed adjusted-2 min-max and adjusted-1 min-max, with MSE of 0.00000282 and 0.0000033, respectively.

Table 3: Comparative results of prediction efficiency yielded by every tested normalization method.

Normalization method	Accuracy	Mean square error
Statistical column	68.0586	0.2076
Decimal scaling	**100**	0.00000528
Adjusted decimal scaling	**100**	0.000085
Min-max	**100**	0.000019
Adjusted-1 min-max	**100**	0.0000033
Proposed adjusted-2 min-max	**100**	0.00000282
Adjusted-3 min-max	**100**	**0.0000024**
Proposed adjusted-4 min-max	**100**	0.0000214

Note. The best values are in bold text.

3.2. Comparative Results on Pima Indian Diabetes Dataset. The comparative results yielded by every normalization method on Pima Indian diabetes dataset are tabulated in Table 4.

The normalization method that provided a maximum accuracy of 75.6624 was adjusted-3 min-max, followed by adjusted-1 min-max at 75.1759 and adjusted decimal scaling that provided the accuracy of 74.9129. In terms of MSE, adjusted-3 min-max normalization method provided the minimum MSE of 0.1689, followed by adjusted decimal scaling and proposed adjusted-4 min-max, with the same MSE of 0.1709.

3.3. Comparative Results on Vertebral Column Dataset. The comparative results yielded by every normalization method on vertebral column dataset are tabulated in Table 5.

The normalization method that provided a maximum accuracy of 83.7289 was adjusted-1 min-max, followed by proposed adjusted-2 min-max at 83.5132 and adjusted decimal scaling that provided the accuracy of 83.2869. In terms of MSE, proposed adjusted-2 min-max provided a minimum MSE of 0.1033, followed by adjusted-1 min-max and adjusted decimal scaling methods with MSE of 0.1159 and 0.1183, respectively.

3.4. Comparative Results on Indian Liver Disease Dataset. The comparative results yielded by every normalization method on Indian liver disease dataset are tabulated in Table 6.

The normalization method that provided a maximum accuracy of 73.7657 was the statistical column method,

Table 4: Comparative results of prediction efficiency on Pima Indian diabetes dataset yielded by every tested normalization method.

Normalization method	Accuracy	Mean square error
Statistical column	68.0586	0.2076
Decimal scaling	72.9108	0.1774
Adjusted decimal scaling	74.9129	0.1709
Min-max	74.6519	0.1711
Adjusted-1 min-max	75.1759	0.1712
Proposed adjusted-2 min-max	74.6524	0.1711
Adjusted-3 min-max	**75.6624**	**0.1689**
Proposed adjusted-4 min-max	74.3727	0.1709

Table 5: Comparative results of prediction efficiency on vertebral column dataset yielded by every tested normalization method.

Normalization method	Accuracy	Mean square error
Statistical column	82.1972	0.1399
Decimal scaling	72.4307	0.1788
Adjusted decimal scaling	83.2869	0.1183
Min-max	82.4666	0.1236
Adjusted-1 min-max	**83.7289**	0.1159
Proposed adjusted-2 min-max	83.5132	**0.1033**
Adjusted-3 min-max	80.2123	0.1413
Proposed adjusted-4 min-max	79.9925	0.1408

Table 6: Comparative results of prediction efficiency on Indian liver disease dataset yielded by every tested normalization method.

Normalization method	Accuracy	Mean square error
Statistical column	**73.7657**	0.1882
Decimal scaling	70.2721	**0.1854**
Adjusted decimal scaling	69.3007	0.3070
Min-max	69.5860	0.1887
Adjusted-1 min-max	70.5003	0.2950
Proposed adjusted-2 min-max	70.5003	0.2950
Adjusted-3 min-max	70.9425	0.2985
Proposed adjusted-4 min-max	67.9579	0.3204

followed by adjusted-3 min-max at 70.9425 and adjusted-1 min-max and proposed adjusted-2 min-max that provided the same accuracy of 70.5003. In terms of MSE, decimal scaling provided the minimum MSE of 0.1854, followed by statistical column and min-max with MSE of 0.1882 and 0.1887, respectively.

3.5. Comparative Results on Working Hour Dataset. The comparative results yielded by every normalization method on working hour dataset are tabulated in Table 7.

The normalization method that provided a maximum accuracy of 78.2005 was min-max, followed by decimal scaling at 78.0755 and statistical column that provided the accuracy of 54.1547. In terms of MSE, min-max provided a minimum MSE of 0.1674, followed by decimal scaling and statistical column methods with MSE of 0.1755 and 0.6439, respectively.

Table 7: Comparative results of prediction efficiency on working hour dataset yielded by every tested normalization method.

Normalization method	Accuracy	Mean square error
Statistical column	54.1547	0.6439
Decimal scaling	78.0755	0.1755
Adjusted decimal scaling	34.5326	2.8364
Min-max	**78.2005**	**0.1674**
Adjusted-1 min-max	34.7777	0.9007
Proposed adjusted-2 min-max	34.4172	2.8233
Adjusted-3 min-max	34.5326	2.8364
Proposed adjusted-4 min-max	34.8154	2.8281

Table 8: Comparative results of prediction efficiency on avocado dataset yielded by every tested normalization method.

Normalization method	Accuracy	Mean square error
Statistical column	**99.4265**	**0.0036**
Decimal scaling	99.3303	0.0065
Adjusted decimal scaling	89.2215	0.0962
Min-max	99.2074	0.0073
Adjusted-1 min-max	90.0255	0.0898
Proposed adjusted-2 min-max	99.1513	0.0080
Adjusted-3 min-max	90.3030	0.0897
Proposed adjusted-4 min-max	89.6828	0.0925

3.6. Comparative Results on Avocado Dataset. The comparative results yielded by every normalization method on avocado dataset are tabulated in Table 8.

The normalization method that provided a maximum accuracy of 99.4265 was statistical column, followed by decimal scaling at 99.3303 and min-max that provided the accuracy of 99.2074. In terms of MSE, statistical column provided the minimum MSE of 0.0036, followed by decimal scaling and min-max with MSE of 0.0065 and 0.0073, respectively.

3.7. Summary. All normalizations achieved the same maximum ranking of accuracy method in white wine quality, except statistical column normalization. A minimum ranking of MSE method was adjusted-3 min-max, the second was proposed adjusted-2 min-max, and the third was adjusted-1 min-max normalization. For Pima Indian diabetes dataset, a maximum ranking of accuracy method was adjusted-3 min-max, followed by adjusted-1 min-max and adjusted decimal scaling normalization. A minimum ranking of MSE method was adjusted-3 min-max, the second were proposed adjusted-4 min-max and adjusted decimal scaling normalizations. For vertebral column dataset, a maximum ranking of accuracy method was adjusted-1 min-max normalization, the second was proposed adjusted-2 min-max, and the third was adjusted decimal scaling normalization. A minimum ranking of MSE method was proposed adjusted-2 min-max, the second was adjusted-1 min-max and the third was adjusted decimal scaling normalization. Indian liver disease dataset, a maximum ranking of accuracy method was statistical column

Table 9: Comparative results of prediction efficiency of all datasets: statistical column, decimal scaling, adjusted decimal scaling, min-max, adjusted-1 min-max, proposed adjusted-2 min-max, adjusted-3 min-max, and proposed adjusted-4 min-max normalizations using the artificial neural network classification.

Datasets	Efficiency comparison	Normalization							
		Statistical column	Decimal scaling	Adjusted decimal scaling	Min-max	Adjusted-1 min-max	Proposed adjusted-2 min-max	Adjusted-3 min-max	Proposed adjusted-4 min-max
White wine quality	Ranking of accuracy	2	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	Ranking of MSE	8	4	7	5	3	2	**1**	6
Pima Indian diabetes	Ranking of accuracy	8	7	3	5	2	4	**1**	6
	Ranking of MSE	8	7	2	4	6	4	**1**	2
Vertebral column	Ranking of accuracy	5	8	3	4	**1**	2	6	7
	Ranking of MSE	5	8	3	4	2	**1**	7	6
Indian liver disease	Ranking of accuracy	**1**	5	7	6	3	3	2	8
	Ranking of MSE	2	**1**	7	3	4	4	6	8
Working hour	Ranking of accuracy	3	2	6	**1**	5	8	6	4
	Ranking of MSE	3	2	6	**1**	4	5	6	8
Avocado	Ranking of accuracy	**1**	2	8	3	6	4	5	7
	Ranking of MSE	**1**	2	8	3	6	4	5	7
Total	Average ranking of accuracy	74.27698	82.16992	75.20916	**84.01871**	75.70144	77.03913	75.27555	74.47027
	Average ranking of MSE	0.23183	0.12062	0.58816	**0.10971**	0.26214	0.56685	0.58917	0.59218

Note. The value of 1 represents a maximum (average) ranking of accuracy and a minimum (average) ranking of mean square error for each dataset, followed by 2 and 3, respectively.

normalization, the second was adjusted-3 min-max, and the third were adjusted-1 min-max and proposed adjusted-2 min-max normalizations. A minimum ranking of MSE method was decimal scaling, the second was statistical column and the third was min-max normalization. Working hour dataset, a maximum ranking of accuracy and a minimum ranking of MSE methods were min-max, followed by decimal scaling and statistical column normalizations respectively. Avocado dataset, a maximum ranking of accuracy and a minimum ranking of MSE methods were statistical column, followed by decimal scaling and min-max normalizations respectively.

4. Discussion

If the ranking of accuracy and the ranking of mean square error were the same, we give more importance to the accuracy. The most effective method for normalization of white wine quality was adjusted-3 min-max, followed by proposed adjusted-2 min-max and adjusted-1 min-max normalization. For Pima Indian diabetes dataset, the most effective method for normalization was adjusted-3 min-max, followed by adjusted decimal scaling and adjusted-1 min-max normalization. For vertebral column dataset, the best normalization method in this study was adjusted-1 min-max, the second was proposed adjusted-2 min-max, and the third was adjusted decimal scaling normalization. The best method found in this study was nearly identical to the best method found from a study in [7]. They said that adjusted-3 min-max normalization had the highest amount of explained variance. Nevertheless, the best normalization method for Indian liver disease dataset was statistical column, followed by decimal scaling, adjusted-1 min-max, and proposed adjusted-2 min-max normalization. For working hour dataset, the most effective method was min-max, followed by decimal scaling and statistical column normalization. Min-max method achieved the same effectiveness as the methods reported in [5, 6]. For the final avocado dataset, the best normalization method to use with artificial neural network was statistical column, the second was decimal scaling, and the third was min-max, when working with artificial neural network. The decimal scaling method achieved the same effectiveness as the methods reported in [8].

5. Conclusions

Considering an average ranking of accuracy, it was found that min-max had a maximum accuracy, followed by decimal scaling and proposed adjusted-2 min-max normalizations, respectively. For an average ranking of MSE, it was found that min-max had a minimum MSE, followed by decimal scaling and statistical column normalizations, respectively. In summary, the most effective method was min-max, followed by decimal scaling normalization using the artificial neural network classification. The proposed adjusted-2 min-max normalization was a fairly effective method when considering an average ranking of accuracy and an average ranking of MSE on white wine quality, Pima Indian diabetes, vertebral column, and Indian liver disease datasets. The proposed adjusted-4 min-max normalization was the lowest effective method. The proposed adjusted-4 min-max normalization is a good method for some datasets, for example, for white wine quality dataset, this method had the same maximum accuracy as the other six methods. For Pima Indians diabetes dataset, this method had the second minimum mean square error.

Data Availability

The six datasets used to support the findings of this study can be obtained from https://archive.ics.uci.edu/ml/datasets/Wine+Quality, https://www.kaggle.com/uciml/pima-indians-diabetes-database, https://www.kaggle.com/caesarlupum/vertebralcolumndataset, https://www.mldata.io/datase-details/indian_liver_patient, https://rdrr.io/rforge/Ecdat/man/Workinghours.html, and https://www.kaggle.com/neuromusic/avocado-prices.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The author thanks the committees of Department of Statistics, School of Science, King Mongkut's Institute of Technology Ladkrabang (KMITL), for consideration of funding the research project on performance comparison of new adjusted min-max with decimal scaling and statistical column normalization methods for artificial neural network classification (grant no. 2564-02-05-011). The author would also like to thank the 4th year students who helped in preparing the data during the research.

References

[1] A. Pandey, A. Jain, and A. Jain, "Comparative analysis of KNN algorithm using various normalization techniques," *International Journal of Computer Network and Information Security*, vol. 9, no. 11, pp. 36–42, 2017.

[2] T. Malai, P. Ninthanom, and S. Sinsomboonthong, "Performance comparison of transformation methods in data mining classification technique," *Thai Journal of Science and Technology*, vol. 10, no. 1, pp. 510–522, 2017.

[3] S. Sinsomboonthong, "Efficiency comparison in prediction of normalization with data mining classification," *Advances in Science, Technology and Engineering Systems Journal*, vol. 6, no. 4, pp. 130–137, 2021.

[4] Y. K. Jain and S. K. Bhandare, "Min max normalization based data perturbation method for privacy protection," *International Journal of Computer and Communication Technology*, vol. 4, no. 4, pp. 233–238, 2013.

[5] B. KumarSingh, K. Verma, and A. S. Thoke, "Investigations on impact of feature normalization techniques on classifier's performance in breast tumor classification," *International Journal of Computer Application*, vol. 116, no. 19, pp. 11–15, 2015.

[6] A. Ali and N. Senan, “The effect of normalization in violence video classification performance,” *IOP Conference Series: Materials Science and Engineering*, vol. 226, pp. 012082–012088, 2017.

[7] G. Aksu, C. O. Güzeller, and M. T. Eser, “The effect of the normalization method used in different sample sizes on the success of artificial neural network model,” *International Journal of Assessment Tools in Education*, vol. 6, no. 2, pp. 170–192, 2019.

[8] A. Akilli and H. Atil, “Evaluation of normalization techniques on neural networks for the prediction of 305-day milk yield,” *Turkish Journal of Agricultural Engineering Research*, vol. 1, no. 2, pp. 354–367, 2020.

[9] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Wine quality data set,” 2009, https://archive.ics.uci.edu/ml/datasets/Wine+Quality.

[10] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, “Pima Indians Diabetes Database,” 1988, https://www.kaggle.com/uciml/pima-indians-diabetes-database.

[11] H. D. Mota, “Vertebral Column Data Set,” 2011, https://www.kaggle.com/caesarlupum/vertebralcolumndataset.

[12] B. V. Ramana, “Indian Liver Patient,” 2012, https://www.mldata.io/dataset-%20details/indian_liver_patient.

[13] L. Myoung, “Working Hours,” 1995, https://rdrr.io/rforge/Ecdat/man/Workinghours.html.

[14] J. Kiggins, “Avocado Prices,” 2018, https://www.kaggle.com/neuromusic/avocado-prices.

[15] R. Shams, “Creating training, validation and test sets (data preprocessing),” 2014, https://www.youtube.com/watch?v=uiDFa7iY9yo.

[16] P. Thongpool, P. Jamrueng, R. Boonrit, and S. Sinsomboonthong, “Performance comparison in prediction of imbalanced data in data mining classification,” *Thai Journal of Science and Technology*, vol. 8, no. 6, pp. 565–584, 2019.

[17] S. Sinsomboonthong, “An efficiency comparison in prediction of imbalanced data classification with data mining techniques,” *Thai Journal of Science and Technology*, vol. 8, no. 3, pp. 383–393, 2019.

[18] N. Phonchan, P. Jaimeetham, and S. Sinsomboonthong, “Clustering efficiency comparison of outliers data in data mining,” *Thai Journal of Science and Technology*, vol. 9, no. 5, pp. 589–602, 2020.

[19] S. Sinsomboonthong, “An efficiency comparison in prediction of outlier six classifications,” *Thai Journal of Science and Technology*, vol. 9, no. 3, pp. 255–268, 2020.

[20] P. I. Dalatu and H. Midi, “New approaches to normalization techniques to enhance K-means clustering algorithm,” *Malaysian Journal of Mathematical Sciences*, vol. 14, no. 1, pp. 41–62, 2020.

[21] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, Burlington, Massachusetts, 2006.

[22] D. T. Larose, *Discovering Knowledge in Data : An Introduction to Data Mining*, John Wiley & Sons, NJ, USA, 2005.

[23] T. Jayalakshmi and A. Santhakumaran, “Statistical normalization and back propagationfor classification,” *International Journal of Computer Theory and Engineering*, vol. 3, no. 1, pp. 89–93, 2011.

[24] K. Waiyamai, C. Songsiri, and T. Rakthammanon, “Using data mining techniques to improve the quality of education for students of the faculty of engineering,” *NEC Technical Journal*, vol. 11, no. 3, pp. 134–142, 2011.

[25] O. G. Troyanskaya, M. Cantor, G. Sherlock, O. Patrick, and P. O. Brown, “Missing value estimation methods for DNA microarrays,” *Bioinformatrics*, vol. 17, no. 6, pp. 520–525, 2011.

[26] S. Sripaaraya and S. Sinsomboonthong, “Efficiency comparison of classifications for chronic kidney disease: a case study hospital in India,” *Journal of Science and Technology*, vol. 25, no. 5, pp. 839–853, 2017.

[27] S. Sinsomboonthong, *Data Mining: Discovering Knowledge in Data*, Jamjuree Product, Bangkok, 2017.