Advances and Applications in Mathematical Sciences
Volume 21, Issue 8, June 2022, Pages 4761-4770
© 2022 Mili Publications, India

EFFECT OF PRE-PROCESSING TECHNIQUES IN PREDICTING DIABETES MELLITUS WITH FOCUS ON ARTIFICIAL NEURAL NETWORK

OVASS SHAFI*, JAHANGIR S. SIDIQ and TAWSEEF AHMED TELI

[1,2]School of Computer Science Application
Lovely Professional University
Phagwara, Punjab -144411, India
E-mail: owaisfour03@gmail.com

[3]Department of Computer Applications
Amar Singh College
Cluster University Srinagar
Jammu and Kashmir - 190008, India

Abstract

Diabetes mellitus is a deadly disease that affects people all over the globe. An early prediction of diabetes is very beneficial as it can be controlled before the onset of the disease. The data collected for the study of any problem suffers from various anomalies and is not in a form to be directly used for research. The data to be used for the research study must fulfill various characteristics like accuracy, consistency, completeness, and interoperability. There are various reasons for the presence of inconsistencies. This paper discusses and compares various pre-processing techniques for the prediction of Diabetes Mellitus. Also, various data mining techniques have been compared for accuracy based on missing values and focus on ANN to deal with missing values using z-score and MinMax techniques are deliberated.

1. Introduction

Diabetes Mellitus is one of the chronic diseases that has affected people all over the globe irrespective of their age and gender. Diabetes mellitus leads to various diseases like heart problems, eye-related problems, liver problems, kidney problems, and many more. With the advancement of machine learning and due to the availability of a huge collection of medical datasets, it is

2020 Mathematics Subject Classification: 68T01.
Keywords: Diabetes, prediction, pre-processing, ANN, z-score, MinMax.
Received November 30, 2021; Accepted February 4, 2022

possible to predict the onset of diabetes mellitus well in advance. The data collected for the study of any problem suffers from various anomalies and is not in a form to be directly used for research. The data to be used for the research study must fulfil various characteristics like accuracy, consistency, completeness, and interoperability. There are various reasons for the presence of inconsistencies, inaccuracies, and other anomalies in data. Some of the reasons include computer errors while data entry, disguised information given by users, transmission errors in data, etc. to make the data suitable for research the data need to be pre-processed. The major phases that data processing includes are data cleaning, integration, reduction, and transformation. The data cleaning step is done to compensate for missing values, remove the noise from data by identifying outliers, and make the inconsistent data consistent. Data integration is the step that merges data from more than one source to reduce redundancies and inconsistencies. The data reduction process aims to improve the data processing and computational time by reducing the number of variables in the dataset without affecting efficiency. In the data transformation step, the data is transformed to make data mining results more efficient and productive.

2. Literature Review

A research study [1] states the importance of pre-processing of dataset due to the availability of an unbalanced range of values that ultimately degrades the quality of classification results. They used two pre-processing methods viz min-max normalization and z-score normalization using the equations given below:

$$\text{Normalized}\,(X') = \frac{X - \min}{X - \max - X - \min}$$

$$\text{New Value}\,(X') = \frac{\text{acturvalue} - \text{mean}}{\text{stdev}}$$

The authors [2] in a research study for early prediction of diabetes disease using machine learning propose the use of pre-processing techniques on the dataset for further improvement of experimental results. The authors propose the use of pre-processing techniques to replace values that are missing by application of various missing value imputation methods like mean, median, K-nearest neighbor, fuzzy K-means, expectation-maximization algorithm, and singular value decomposition. A research study [3] uses the pre-processing techniques for the dataset to make the research study more efficient and more accurate. The dataset collected for the study contains irrelevant and noisy data for the application of data mining algorithms. So, they pre-process the dataset by application of data cleaning, data reduction, and data transformation. With the application of data cleaning the missing qualities were filled and exceptional data were excluded. The authors in [4] use Kalman filtering and Kalman smoothing as a pre-processing technique for the dataset before using the data from the dataset for the research study. The Kalman Smoothing method generates an interpolated time series of the glucose level with mean and variance as output. The method can automatically fix errors in the Continuous Glucose Monitoring reading by utilizing estimated variance for the determination of intervals when the data is reliable. The authors in [5] for early diagnosis of diabetic retinopathy uses medical images of the patient's retina. As the medical images are subjected to the presence of a nose, the use of such images for research study without any pre-processing may produce inefficient results. Therefore, the author applies pre-processing techniques for the elimination of the noise from the images. The authors in [6] emphasized the application of pre-processing techniques as a key step for data cleaning. To achieve better results by optimizing image data and by eliminating unnecessary information like distortion and improving image characteristics, the application of an appropriate pre-processing technique is an essential step. The author applies pre-processing techniques like Cropping, Resizing, Image Enhancement, and Noise Removal, and Down Sampling. In a research study [7], the authors use a medical image dataset to predict diabetes mellitus at an early stage. A research study [2] proposed the use of pre-processing techniques in the future for improving the efficiency of classification results. The authors in [8] use the pre-processing techniques on PIMA dataset on the attributes of Blood Sugar Level and Body Mass Index for cleaning of the data. The research study in [9] for the prediction of diabetes using machine learning uses the PIMA India data set and applied pre-processing techniques for cleansing of data. The data cleaning is done using normalization and transformation applied to some of the attributes. The authors in [10] collect data from Kaggle's website for early diabetes diagnosis using ANN. The dataset contains some missing attribute values that the authors handle using pre-processing carried by some statistical techniques. Due to missing attribute values, the results would not be accurate. So, they apply the Numpy Package of python for handling missing attribute values. This results in better accuracy of the results. A research study [11] gives the importance of data handling before using it for the classification process for the prediction of diabetes mellitus. The authors collect the dataset and apply pre-processing techniques using normalization. The pre-processing technique used by authors utilizes mean and standard deviation of every feature of the training dataset using the equation:

$$Z = \frac{x_i - \mu}{\sigma}$$

The authors also used the Synthetic Minority Oversampling Technique to balance the data as the imbalanced data in the dataset may develop various biases for the majority class. The researchers in [12] focused on the classification techniques used for the onset of diabetes mellitus. They state that the classification techniques used for diabetes diagnosis after the dataset is pre-processed do not produce efficient results as the results of the pre-processing technique in attributes that are computed from original attributes and the classification is not done on original attributes. A research study [13] uses retinal images for the diagnosis of various diseases like diabetes mellitus. Before using the images for the classification process, the authors pre-process the images in DR screening to separate the fundus image and decline along the edges.

3. Pre-processing Techniques

Data means a huge collection of rows and columns; however, the data can be in the forms of images, videos, tables, audio, etc. To make the data suitable for decision making the computers can’t use the data directly in the forms of free text, images, audio, video but has to be transformed into a form that can be used to train the machine. This process of transforming data into a form that can easily be parsed by a machine is known as data pre-processing. After data pre-processing, the data can be easily interpreted by the machine learning algorithm. To pre-process the dataset, multiple steps are performed however it is not necessary to apply all steps for each problem. The number of steps applicable for a particular situation is highly dependent on the data we are working with. Among various pre-processing aspects, normalization is the most commonly used one. Data Normalization involves changing values measured on multiple scales to some common scale. The normalization allows the values to be modified in columns to some common scale in the case of data frames. The normalization is applicable for numeric columns only. Two such methods of normalization are:

The Min Max technique transforms each number to a value within the range of 0 and 1 as in single feature scaling. The modified value is a series of arithmetic operations that involve the subtraction of minimum value from the current value and then division by the whole range of column values. If we consider column *x* the Min max function can be used as:

$$df[x] = \frac{(df[x] - df[x], \min(\ ))}{(df[x], \max(\ ) - df[x], \min(\ ))}$$

The Z-score technique of normalization converts changes each value in a column around zero. Most common values obtained by the application of Z-Score ranges between-3 and 3. The subtraction of average from current value and then division by the standard deviation gives the new value. For example, consider the column *x*, the Z-Score can be calculated as:

$$df[x] = \frac{(df[x] - df[x], mean)}{df[x], std(\ )}$$

Following this, the min and max value can be calculated by the z-score transformation as:

$$df[x], \min(\ )$$

$$df[x], \max(\ )$$

4. Artificial Neural Network

ANN is another sub-field of artificial intelligence inspired by the Human Nervous System [14-15]. It is a computational network that simulates the construction of the human brain and has interconnected neurons within a layer as well as interconnected neurons between various layers [16]. The idea behind ANN is to mimic the behavior of the human brain so that the computers will be able to understand situations and make human-like decisions with being programmed explicitly. The ANN is composed of a huge number of artificial Neurons arranged in multiple layers. The concept of layers is shown in Figure 1:

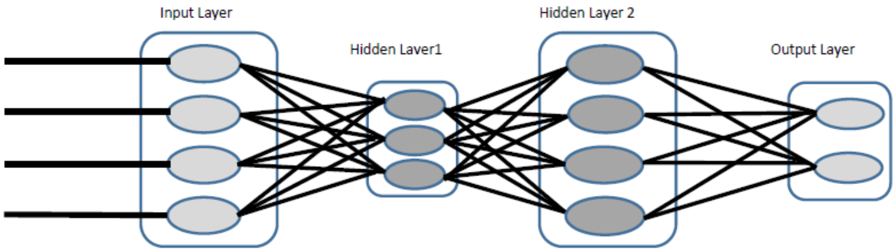

Figure 1. Architecture of ANN.

The input layer accepts data known as input from the external environment. The layers that are sandwiched between input and output players are referred to as hidden layers. The basic functioning of hidden layers is to perform calculations to find the hidden features and patterns in the input data. After going through multiple transformations in the hidden layers, the output layer finally produces the output that is presented to the external environment. The input layer takes the input and computes the weighted sum of all the inputs with the help of bias. The computation is performed using multiple transform functions as:

$$\sum_{i=1}^{n} W_i * x_i + b$$

The above equation computes the weighted total and is feed as input to another function known as activation function to produce the output. The purpose of activation function is to choose whether a node should fire or not. The nodes that are fired contributes to the output layer. Different activation functions are used for different types of tasks to be performed. The input applied to the ANN is multiplied by its weight.

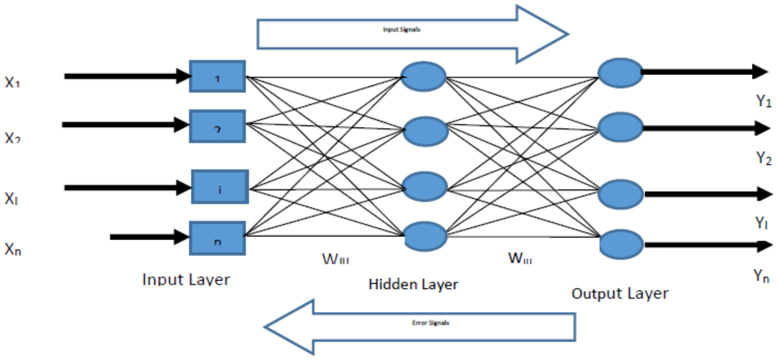

Figure 2. Working of Artificial Neural Network.

The weights represent the details used by ANN for solving a particular problem. The weights act as the strength of the interconnection that exists between neurons. The weights are aggregated using the summation function. If the sum of the weights is equal to zero, then bias is added to the output of ANN non-zero to scale up the system's response. The input and the weight of bias are equal to 1. The total weighted inputs are in the range of 0 to some positive infinity, Figure 2. To keep the response within some bounds, a threshold is used. The output of Binary function is either 0 or 1. To accomplish this function, the threshold value is set. If the final output of Binary activation functions 1 if the net weighted input of neurons is more than 1, otherwise the output returned is 0 [17]. This Sigmoidal Hyperbolic activation function is in the form of an "S" shaped curve. The tan hyperbolic function is applied to approximate the output of actual net input [18-28].

5. Experiment and Results

We Various classifiers that include; Naïve Bayes, Random Forest, KNN, SVM, and Decision Tree [29-37] were implemented using Python. The dataset used for this experiment was taken from the National Institute of Diabetes and Digestive and Kidney Diseases popularly known as the PIMA dataset. From the dataset consisting of 768 records, a total of 70% samples were used for training and a total of 30% samples were taken randomly selected for testing.

Table 1. Accuracy (Different Classifiers) without pre-processing Techniques.

Model	Accuracy		
	Mean	Median	Most Frequent
Naïve Bayes	75.58	69.05	75.57
Random Forest	77.36	75.57	75.41
KNN	72.31	73.61	72.96
SVM	77.04	76.21	77.04
Decision Tree	70.36	67.43	75.57
ANN	71.66	58.50	62.89

The neural network use in the experiment is fully connected and a feed-forward with three hidden layers implemented in Python using Keras.

Table 2. ANN after pre-processing Techniques.

MISSING VALUE STRATEGY	Z-SCORE	MINMAX SCALER
MEAN	75.75%	84.77%
MEDIAN	60.89%	82.14%
MOST FREQUENT	65.19%	82.79%

6. Conclusion

Machine learning techniques have been used by researchers in medical diagnosis to assist in the proper treatment at the right time. The effect of pre-processing techniques on medical datasets like Diabetes Mellitus using these machine learning techniques is significant in improving the accuracy. Various machine learning models were compared using different missing value strategies in the dataset and the ANN was used to predict Diabetes Mellitus in missing values dataset using the pre-processing techniques that include z-score and Min Max. The results showed that the ANN accuracy is significantly improved using the pre-processing techniques.

References

[1] D. Westari, A. Halim and M. Eng, Performa Comparison of the K-Means Method for Classification in Diabetes Patients Using Two Normalization Methods 04(01) (2021), 18-23. doi:10.47191/ijmra/v4-i1-03

[2] M. Maniruzzaman, M. J. Rahman, B. Ahammed and M. M. Abedin, Classification and prediction of diabetes disease using machine learning paradigm, Heal. Inf. Sci. Syst. 8(1) (2020), 1-14, doi:10.1007/s13755-019-0095-z.

[3] K. Hirnak et al., Early prediction model for type-2 diabetes based on lifestyle 2 Review of Literature, International Conference on Automation, Computing and Communication (ICACC-2020) Art. 03053, 32 (2020), 1-5.

[4] F. Rabby, Y. Tu, I. Hossen, I. Lee, A. S. Maida and X. Hei, Stacked LSTM based deep recurrent neural network with kalman smoothing for blood glucose prediction (2021), 1-11.

[5] R. Rameshkumar, N. Rao, K. Dist and K. Dist, An efficient early diagnosis for diabetic retinopathy using quick convolutional 79 (2020), 2923-2940.

[6] R. Islam, Severity Grading of Diabetic Retinopathy using Deep Convolutional Neural Network 6(1) (2021), 1395-1401.

[7] N. Shivsharanr and S. Ganorkar, Predicting severity of diabetic retinopathy using deep learning models International Research Journal on Advanced Science Hub 03(01) (2021), 67-72.

[8] B. Pranto, S. M. Mehnaz, E. B. Mahid, I. M. Sadman, A. Rahman and S. Momen, Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh, Inf. 11(8) 2020. doi:10.3390/INFO11080374

[9] H. F. Ahmad, H. Mukhtar, H. Alaqail, M. Seliaman and A. Alhumam, Investigating health-related features and their impact on the prediction of diabetes using machine learning, Applied Sciences, 2021.

[10] S. Srivastava, L. Sharma, V. Sharma, A. Kumar and H. Darbari, Prediction of Diabetes Using Artificial Neural Network Approach, Springer Singapore, Engineering Vibration, Communication and Information Processing 478 (2018), 679-687.

[11] S Malik, S. Harous and H. El-Sayed, Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women, Proceedings of the 6th International Symposium, MISC 2020, Batna, Algeria, October 24-26, 2020.

[12] Jeffrey O. Agushaka and Absalom E. Ezugwu, Diabetes classification techniques: A brief state-of-the-art literature review, Third International Conference, ICAI 2020 Ota, Nigeria, October 29-31, 2020.

[13] R. Murugan, The retinal blood vessel segmentation using expected maximization algorithm, Advances in Intelligent Systems and Computing International Symposium, ISCMM 992 (2019).

[14] McCulloch Warren and Walter Pitts, A logical calculus of ideas immanent in nervous activity, Bulletin of Mathematical Biophysics 5(4) (1943), 115-133. doi:10.1007/BF02478259

[15] S. C. Kleene, Representation of events in nerve nets and finite automata, Annals of Mathematics Studies (34), Princeton University Press (1956), 3-41, Retrieved 17 June 2017.

[16] Hebb Donald, The organization of behavior, New York: Wiley, ISBN 978-1-135-63190-1. (1949).

[17] A. G. Ivakhnenko, Grigorevich Lapa, Valentin, Cybernetics and forecasting techniques, American Elsevier Pub. Co., (1967).

[18] Schmidhuber Jürgen, Deep Learning, Scholarpedia 10(11) (2015), 85-117. Bibcode:2015SchpJ..1032832S. doi:10.4249/scholarpedia.32832.

[19] Dreyfus E. Stuart Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure, Journal of Guidance, Control, and Dynamics 13 (5) (1990), 926-928. Bibcode:1990JGCD...13..926D. doi:10.2514/3.25422. ISSN 0731-5090.

[20] E. Mizutani, S. E. Dreyfus and K. Nishio, On derivation of MLP backpropagation from the Kelley-Bryson optimal-control gradient formula and its application, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, 2000.

[21] Ahmed Teli Tawseef and Masoodi Faheem, Blockchain in Healthcare: Challenges and Opportunities (July 8, 2021), Proceedings of the International Conference on IoT Based Control Networks and Intelligent Systems - ICICNIS 2021, Available at SSRN: http://dx.doi.org/10.2139/ssrn.3882 44.

[22] S. J. Sidiq, M. Zaman and M. Butt, An empirical comparison of classifiers for multi-class imbalance learning, International Journal of Data Mining and Emerging Technologies 8(1) (2018), 115-122.

[23] M. Ashraf, M. Zaman, M. Ahmed and S. J. Sidiq, Knowledge discovery in academia: a survey on related literature. Int. J. Adv. Res. Comput. Sci. 8(1) (2017).

[24] S. J. Sidiq, M. Zaman M. Ashraf and M. Ahmed, An empirical comparison of supervised classifiers for diabetic diagnosis, Int. J. Adv. Res. Comput. Sci., 8(1) (2017), 311-315.

[25] S. J. Sidiq, M. Zaman and M. Butt, A framework for class imbalance problem using hybrid sampling, Artificial Intelligent Systems and Machine Learning 10(4) (2018), 83-89.

[26] S. J. Sidiq, M. Zaman and M. Butt, A comprehensive review on class imbalance problem, Artificial Intelligent Systems and Machine Learning 10(3) (2018), 59-65.

[27] Abhishek Sharma, Prateek Agrawal, Vishu Madaan and Shubham Goyal, Prediction on diabetes patient's hospital readmission rates, 3rd International Conference on Advances Informatics on Computing Research (ICAICR'19), 1-5, Jul 2019, ACM-ICPS.

[28] A. Jain and C. Gupta, A genetic algorithm based approach in predicting and optimizing sickle Cell Anaemia, Global Journal of Enterprise Information System 8(4) (2016), 92-97.

[29] Charu Gupta, Prateek Agrawal, Rohan Ahuja, Kunal Vats, Chirag Pahuja and Tanuj Ahuja, Pragmatic analysis of classification techniques based on hyperparameter tuning for sentiment analysis, International Semantic Intelligence Conference (ISIC'21), Delhi (2021), 453-459.

[30] Anatoliy Zabrovskiy, Prateek Agrawal, Roland Matha, Christian Timmerer and Radu Prodan, ComplexCTTP: Complexity class based transcoding time prediction for video sequences using artificial neural network, 6th IEEE Conference on Big data in Multimedia (BigMM'20), New Delhi, Sep 20, IEEE Explore, pp. 316-325.