Vol. 7(24), Apr. 2017, PP. 3397-3404

Contents list available at IJMEC

International Journal of Mechatronics, Electrical and Computer Technology (IJMEC)

Journal Homepage: www.aeuso.org

PISSN: 2411-6173 - EISSN: 2305-0543

A Hybrid Method of Feature Selection and Neural Network with Genetic Algorithm to Predict Diabetes

Seyyed Mohammad Hossein Dadgar[1] and Mostafa Kaardaan[2]

[1]*Young Researchers and Elite club, Central Tehran Branch, Islamic Azad University, Tehran, Iran*

[2]*Master Student of Software Engineering Department of Computer Engineering Islamic Azad University, Central Branch, Tehran, Iran.*

Phone Number: *+98-9176618775*

****Corresponding Author's E-mail:*** *sey.dadgar.eng@iauctb.ac.ir*

Abstract

Diabetes is one of the most serious challenges of health care in developing and developed countries. In the medical field, examination of the patient data using different classifications is used to derive a predictive model. A high number of hypotheses is necessary in order to analyze the system using thermodynamic method, without these hypotheses, thermodynamic analysis of the actual application requires a large number of non-linear equations, whose solutions are either impossible, or that too much time and effort is required computationally. Machine learning methods are often used as replacements for thermodynamic method to eliminate this barrier. The aim of this study was to facilitate the prediction of diabetes, which is increasing rapidly in the world. In this paper, a hybrid approach is proposed based on UTA algorithm and two-layer neural network, which are updated uses its weights genetic to enhance the prediction of diabetes. This method is developed in two stages: 1. Feature selection based on UTA algorithm, 2. Prediction using neural-genetic. The method was assessed using Pima Indians Diabetes which prediction accuracy of 87.46% was obtained. The result obtained in this study compared to other methods presented shows the high accuracy rate for predicting diabetes and also reducing the time consumed.

Keywords: *machine learning, UTA algorithm, neural network, genetic algorithms, diabetes*

1. Introduction

Diabetes is a metabolic disorder in the body. In this disease, patient's ability to produce insulin in the body disappears, or the body becomes resistant to insulin, so insulin produced can't perform its normal function. The main role of insulin is to reduce blood sugar in the body by different mechanisms. Diabetes is divided into two categories: I and Type II diabetes. In Type II diabetes, which will include more patients with diabetes, the body becomes resistant to insulin [3]. Prediction of a disease in the community can play an important role in increasing health care in the community. It is natural that doctors do not have enough accuracy to predict disease. Despite recent advances in the field of computers and its use in the medical field, this disease can be predicted based on population data. Different learning methods are used to predict and diagnose diseases. Features subset selection means identifying and selecting a useful subset of the features of primary data collection, as well as an important issue in the analysis of the correlation in the fields of classification and modeling, which is used to reduce the dimensions of the feature set. This work can be done by eliminating features that

Article History:

IJMEC DOI: 649123/10.225130

Received Date: Sep. 05, 2016

Accepted Date: Jan. 24, 2017

Available Online: Mar. 03, 2017

make noise or have low correlation with other features. Selection of feature used in the classification can have a significant impact on the accuracy of classifier function, the time required for classification, training data set required and the cost of implementation for the classification [11]. There have been numerous predictions for diabetes in the past, some of which are:

Hybrid intelligent system for classifying medical data [4], the decision tree for the diagnosis of type II diabetes [1], predicting diabetes using an artificial neural network [5], classification of diabetic patients using support vector machine [8], the use of Bayesian network for prediction of Type II diabetes [10].

In this study, a hybrid method is proposed to enhance the prediction of diabetes based on UTA algorithm and two-layer neural network, which its weights are updated using genetic algorithm. This method is composed of two phases: 1. The feature selection based on UTA algorithm, 2. prediction using genetics-neural network. In the following, in section two, the related works will be explained briefly, and in section three, the proposed method is presented. Section four evaluates the proposed method, and finally, section five expressed conclusions.

2. Related works

In recent years, much works have been done on predicting diabetes. Before introducing and examination of the tools presented, a brief look at the methods in this field will be provided in this section.

Seera & Lim have designed a smart hybrid system, which consists of fuzzy neural network, classification and regression tree and random forest model. They used hybrid system and could use it as a decision support tool for classification of medical data. Hybrid model consists of three important features to be able to manage medical decision support tasks:

1. Online training, whereby, model does incremental learning using sample data with a training procedure.
2. High Performance whereby an effective way is used to integrate models.
3. Extracting the rule whereby the model is used to show the extracted knowledge in the form of a decision tree.

The purpose of intelligent hybrid system is to exploit the benefits of forming models at the same time and reduce the restrictions. In this study, the maximum precision among the models used is 78.39% [4].

In an article published in the alJarullah, decision tree method is used to predict patients with type II diabetes. This study consisted of two stages. The first phase includes data pre-processing including feature identification and selection. The second phase is construction of predictive diabetes prediction models using decision tree. Pre-processing is done to increase the quality of data. These techniques include selecting and identifying features, missing values management and numerical differentiation. Next, the Weka J48 was used to categorize and make the decision tree model which the accuracy of this model is 78.1768% [1].

Pradhan & Sahu used the artificial neural network to classify patients with diabetes into two groups. They were used for achieving better results of genetic algorithm to select the features. Genetic algorithm is used to find the number of neurons in the hidden layer. Weights of this method have been taught using propagation algorithm and training genetic algorithm that the accuracy of this method is 73.438% [5].

In the proposed method by Kumari & Chitra, support vector machine is provided as a machine learning method for classification for detecting the disease. It was used to predict diabetes, and all data have been trained using Support Vector Machine. SVM with Radial Basis Function (RBF) is used for classification, and in this study, the accuracy is 78% [8].

Guo et al have suggested a way to predict patients with type II diabetes using Bayesian network, they could increase the data quality for classification through preprocessing of data. Techniques used in data pre-processing include the selection and identification of features, missing values management and numerical differentiation. Category is done at a later stage to make the NaIve Bayes model. At the end, Weka was used to simulate, which the accuracy of this model is 72.3% [10]. In the following, Table 1 shows summary of the works done on the Pima Indians Diabetes data set.

Table 1: Previous works.

Method	Accuracy	Advantages	Disadvantages
Sierra and Lim[4]	78.39%	The use of intelligent hybrid system consisting of fuzzy neural network, classification and regression tree and random forest model	No way to handle missing data in the data set
alJarullah [1]	78.1768%	Using decision tree method	The use of Weka ready software instead of writing Application needed
Pradhan & Sahu [5]	73.438%	Feature selection using genetic algorithm	Lack of comparison of the model's performance with similar works
Kumari & Chitra [8]	78%	Using Support Vector Machine	Not benefiting from the features selection in SMV classifier
Guo et al. [10]	72.3%	Using the Naive ***Bayes*** model	Using a Bayesian network to predict diabetes

3. The proposed Method

This section describes the proposed method for predicting diabetes. This method consists of two stages of feature selection based on UTA algorithm and genetic-neural network. The overall architecture proposed method is shown in Figure 1. Then, to explain each of these steps will be discussed.

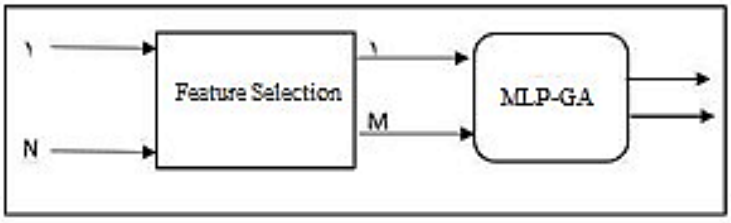

Figure 1: Magnetization as a function of applied field. Note that “Fig.” is abbreviated.

3.1. UTA algorithm

UTA optimization algorithm is applied to remove ineffective and destructive features to reduce the dimension of feature vectors, increase network efficiency, improve the accuracy and speed of recognition. The algorithm is based on the replacement value of a feature in the feature vector with the average of the feature on the entire test set [10]. In this algorithm, the average of every feature is calculated in the entire test set, after each test, the value of a deature is replaced with the average value of the feature. If the network uses this feature, the results will change, and if the results did not change, we can say that, that feature has been ineffective or destructive. In this paper, UTA algorithm was used on diabetes data set, which led to remove three features and increase the accuracy of the network.

3.2. Multi Layer Perceptron

This method is used to classify data that are not separable linearly. The network consists of an input layer that is network interface with the outside world, and the number of neurons in this layer, based on the data set features, and hidden layers which are responsible for train process and adjust the weights of each neuron. Finally, the output layer which is the number of neurons in the output layer based on the number of classes. In this model, a two-layer neural network is used for data classification. The network consists of an input layer, and two hidden layers, and an output layer. And the connections between neurons are done on the basis of weights in MLP .

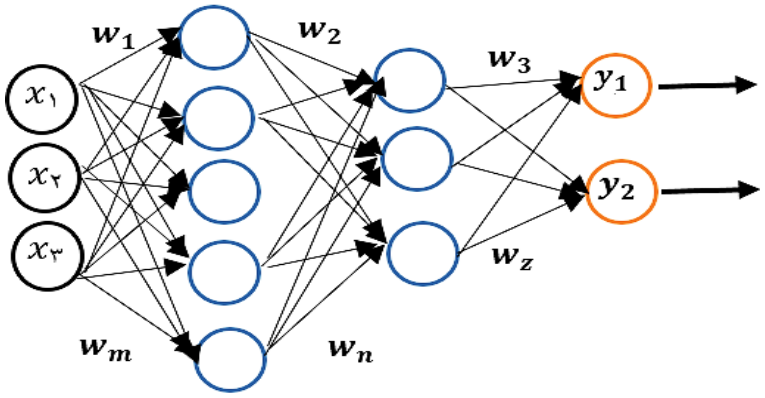

Figure 2: Two-layer perceptron.

In the two-layer perceptron that is designed, the number of neurons in the first hidden layer is twelve, and the number of neurons in the second hidden layer is seventeen. The weights entered the network are chosen randomly, which using a genetic algorithm, the network tries to learn the weight parameter, so that the mean square error in each iteration will be reduced. Error is calculated in each row, and consequently, leads to adjust the weights in each row. The difference between network output with target output is called the error. Using equation 1, the product weights in values in neurons are calculated.

$$v_j(n) = \sum_{i=1}^{m} w_{ij}(n)\, x_i(n) \qquad (1)$$

Where, w is weight parameter and x represents the features values. m represents a number of features in the data. Using the equation 2, network output is achieved.

$$y_i(n) = \varphi_j\left(v_j(n)\right) \qquad (2)$$

Where, Φ function used in this paper is Gradian Discent function. This function is obtained using the following equation.

$$\varphi_j = a\left(\frac{1-e^{-2by_j}}{1+e^{-2by_j}}\right) \quad (3)$$

Here, the values of a and b are considered on the basis of a proposal number in Haykin book a = 1.716, b = 0.66. Using the equation 4, the error rate is calculated for each example.

$$e_k = targ_k - out_k \quad (4)$$

Where out is parameter containing the network output and targ is actual output in the database. After calculating output, mean square error is achieved using equation 5.

$$E = \frac{1}{2}\sum_{k=1}^{n} e_k^2 \quad (5)$$

To update the weights in each row, weight derivative of the error was calculated. In this model, we have used a genetic algorithm and this algorithm is described in the next section.

3.3. Genetic Algorithm

Genetic algorithm is based on Darwinian Theory of evolution. This algorithm is one group of stochastic optimization algorithms to solve complex problems in the unknown search space. This approach represents a practical solution for certain problems that with the development and successive repetition of the generations through a competitive process to be controlled. First, the initial population chromosomes are entered with random values into the repetition ring, and in the first phase, the fitness of each chromosome is calculated, fitness function is based on solving problem. Then, one of the chromosomes is chosen based on the selection procedures for combination action, which the resulting chromosome is called child chromosome. Mutation is change of a random gene on a series of random chromosomes. The process of calculating the amount of fitness can be calculated for child chromosome from mutation, and the best chromosome is chosen for the next generation from the set of chromosomes. This process continues until that the best solution is reached. Here, we have used a genetic algorithm to generate random weights and choose the best weight for neural network layers. The network weights are the equivalent the chromosomes of genetic algorithm. Fitness function can be used to calculate the network output using chromosomes and its different from that objective function, the chromosomes that has less difference with network output, is more likely to survive. The advantage of using genetic algorithm in multi-layer neural perceptron is reducing the time needed to reach the desired weight. The use of this method on the data set used has achieved the desired result by reducing 1700 repetitions.

4. Results and evaluation

This section explains the results obtained using the proposed method and compares this method with previous methods for predicting diabetes.

4.1. Data set

Pima Indians Diabetes Data Set extracted from the site UCI [9], refers to a medical problem which diagnoses that a patient may have diabetes or not. This data set includes 768 samples taken from women with at least 21 years of age. This data set consists of 8 features and a binary value to the class, that if this amount is equal to 1, means that the patient test for diabetes is positive, and if it is zero, means that, tests for disease is negative. Features of this data set are given below:

- Class A: normal (500 samples)
- Class II: diabetic patients (268 samples)
- Feature 1: The number of times a female is or has been pregnant
- Feature 2: Blood sugar levels two hours after taking glucose
- Feature 3: blood pressure
- Feature 4: triceps Size
- Feature 5: Two hours after the insulin injection
- Feature 6: Body Size (weight / (height ^ 2))
- Features 7: effective coefficient of hereditary diabetes
- Features 8: age

A brief statistical analysis of the database is shown in Table 2 [6].

Table 2: Statistical analysis on the desired data.

Feature	Average	Standard Deviation
1	3.8	3.4
2	120.9	32.0
3	69.1	19.4
4	20.5	16.0
5	79.8	115.2
6	32.0	7.9
7	0.5	0.3
8	33.2	11.8

4.2. Evaluation Criteria

Accuracy is used in order to determine the performance of supervised learning algorithms. This metric shows the accuracy of classification or prediction of each method, and is considered as an index to compare the algorithms. In this section, we will explain how to calculate it. First, effective parameters on the calculation of it, and then how to calculate it will be explained.

True Negative (TN): including the number of samples that haven't been belonged to the desired class and the network is detected correctly.

True Positive (TP): including the number of samples that have been belonged to the desired class and the network is detected correctly.

False Negative (FN): including the number of samples that haven't been belonged to the desired class and the network is detected correctly.

False Positive (FP): including the number of samples that have been belonged to the desired class and the network isn't detected correctly.

According to the above parameters, the criteria for the diagnosis and accuracy is calculated, which is achieved from the equation 6.

$$Accuracy = \frac{TN+TP}{TN+FN+TP+FP} \quad (6)$$

4.3. Results

In the proposed method, using the feature selection which was done at the first step, it has led to remove the features 4, 6 and 7. Based on the five remaining features, first back-propagation algorithm was used for perceptron learning, in the 1976 repetitions, the accuracy was 86.5%, then, genetic algorithm was used to learn perceptron, in the 276 repetitions, the accuracy of 87.46% was obtained; results can be seen in table 3.

Table 3: The results of applying the proposed method.

Method	The number of repetition	Accuracy
Mlp+back-propagation	1976	86.5
Mlp+GA	276	87.46

Reduced error chart due to the results are shown in Table 3 and Figure 3. According to studies conducted in the train section, by reducing excessive error, too much learning happens to the machine which doesn't lead to the desired result or with several repetitions with various errors stop, the best error rate using this method has been 0.1954.

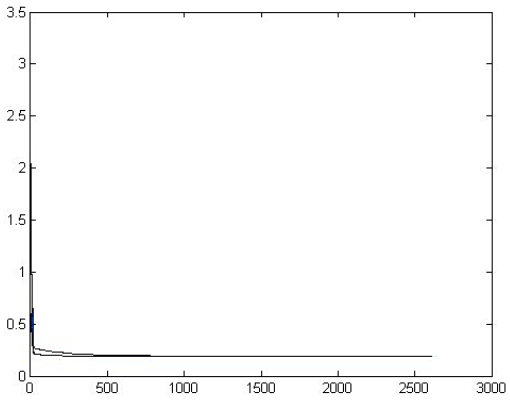

Figure 3: Reduced errors diagram.

4.4. Assessment

Table 4 shows a comparison of the proposed method with other pervious methods for predicting diabetes. The result produced by this research is highlighted in Table 4 is. Seera & Lim have designed a smart hybrid system, which consists of fuzzy neural network, classification and regression tree and random forest model which they obtained the accuracy of 78.39%, respectively. alJarullah has used decision tree method predict diabetes and has reported the accuracy of 78.17%. Pradhan & Sahu have used the artificial neural network to classify patients with diabetes into two groups that the accuracy of this method is 73.40%. In the proposed method by Kumari & Chitra, support vector machine is provided as a machine learning method for classification for detecting the disease that its accuracy is 78% [8]. Guo et al have suggested a way to predict patients with type II diabetes using Bayesian network, that its accuracy is 72.30%. The result obtained using the proposed method is 87.46%. According to other methods presented, the proposed method has high accuracy rate for predicting diabetes.

Table4: Comparison of the results obtained in previous methods with the proposed method.

Accuracy	Method
78.39	The hybrid model of fuzzy system and regression tree [4]
78.17	Decision tree[1]
73.40	Artificial neural network [5]
78	SVM [8]
72.30	Bayesian network [10]
86.50	**Our approach with back-propagation algorithm**
87.46	**Our approach with genetic algorithm**

Conclusion

In recent years, the many investigations have been conducted to predict diabetes, which show that the use of machine learning techniques to predict the person with diabetes has been better than other methods such as regression, decision trees, etc. In this study, a hybrid approach based on UTA algorithm and two-layer perceptron that its weights are updated using genetic algorithm, it was proposed to increase diabetes prediction. The advantage of using genetic algorithm in multilayer perceptron is reducing the time consumed to reach the desired weight. The use of this method on the data set used has achieved the desired result by reducing 1700 repetitions. This method was assessed using Pima Indians Diabetes which the prediction accuracy was 87.46%. The result obtained in this study compared to other methods presented shows higher accuracy rate for predicting diabetes and also less the time needed.

References

[1] A. AlJarullah, "Decision Tree Discovery for the Diagnosis of Type II Diabetes", International Conference on Innovations in Information Technology, 2011.

[2] G. Chandrashekar and F. Sahin, "A survey on feature selection methods", Computers and Electrical Engineering, pp. 16–28, 2014.

[3] M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis ", Expert Systems with Applications, vol. 38,pp. 14650–14659, 2011.

[4] M. Seera and Ch. P. Lim, "A hybrid intelligent system for medical data classification", Expert Systems with Applications, vol.41, pp. 2239–2249, 2014.

[5] M. Pradhan and R. K. Sahu, "Predict the onset of diabetes disease using Artificial Neural Network (ANN)", International Journal of Computer Science & Emerging Technologies, vol. 2, pp 2044-6004, 2011.

[6] N. Dogantekin, A. Dogantekin, D. Avci and Avci, L.", An intelligent diagnosis system for diabetes on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System: LDA-ANFIS", Digital Signal Processing, vol.20, pp. 1248–1255,2010.

[7] T. Santhanam and M. S. Padmavath, "Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes", Procedia Computer Science, pp. 76 – 83, 2015.

[8] V. A. Kumari and R. Chitra, "Classification Of Diabetes Disease Using Support Vector Machine", International Journal of Engineering Research and Applications, vol. 3, pp. 2248-9622, 2013.

[9] V. Sigillito, Data set Pima Indians Diabetes collected (1990), https://archive.ics.uci.edu/ml/datasets/Pima+Indians +Diabetes, Accessed dec 8, 2015.

[10] Y. Guo, G. Bai and Y. Hu, "Using Bayes Network for Prediction of Type-2 Diabetes", The seventh International Conference for Internet Technology and Secured Transactions,2012.

[11] Y. Sheng, G. Jun, "Feature selection based on mutual information and redundancy-synergy coefficient", Journal of Zhejiang University SCIENCE ISSN 1009-3095, 2004.