ACO-based Type 2 Diabetes Detection using Artificial Neural Networks

Aliakbar Tajari Siahmarzkooh

Assistant Professor, Department of Computer Science, Faculty of Sciences, Golestan University, Gorgan, Iran

Abstract

Background: Type 2 diabetes is one of the most common diseases among people. Early diagnosis and treatment can reduce mortality and morbidity. So far, various solutions have been proposed to predict this type of disease.

Materials and Method: In this paper, a method for diagnosing diabetes was proposed using the Ant Colony Optimization (ACO) algorithm. To this end, data set properties are first reduced using artificial neural network features and then prepared for classification purpose. Finally, some components of accuracy assessment on the proposed system were calculated.

Results: The simulation results show that by adjusting the parameters of ANN and ACO, about 3.2% better prediction accuracy is obtained than other researches.

Conclusion: The results of experiments represent that the proposed method is proper for health management in diabetes.

Keywords: *Diabetes detection, Artificial Neural Network (ANN), Ant Colony Optimization (ACO).*

Introduction

Diabetes is a chronic disease that is diagnosed with high blood glucose levels. About half of diabetics have hereditary characteristics, which is one of the most important features of diabetes. Poor pancreatic insufficiency and insufficient use of insulin are the causes of diabetes. There are two major types of diabetes. Type 1 diabetes (T1DM) is when pancreatic secretions damage β cells and prevent a timely drop in blood glucose levels. Insulin resistance and inefficient insulin secretion are the causes of type 2 diabetes.

We need information technology-based methods to study high-risk groups at risk for diabetes. In this regard, meta-heuristic algorithms are a good tool that is used as a computational process to discover patterns in large data sets and includes several solutions such as evolutionary clustering, machine learning and neural networks.

Meta heuristic algorithms have been successfully used to solve various problems in various fields of basic sciences, engineering and even humanities. Weather forecasting, stock market analysis, system suggesting better customer management in banks and shopping malls, disease forecasting and medical data analysis are examples of applications of this category of algorithm. In more detail, extracting logical patterns from patients' information in hospitals is essential for support as well as analysis, which requires the use of intelligent methods and data mining tools.

In recent years, various data mining methods have been used to predict diseases. Algorithms as well as various toolboxes have been developed and studied by researchers. Here are some examples of work done.

Patil proposed a hybrid predictive model that used the K-means clustering algorithm to validate the data class tag and the C4.5 decision tree algorithm to create the final model.[1] The results of his proposed method have an accuracy of 92.38% in classification. Aliza compared the predictive accuracy of the multilayer perceptron (MLP) model in the neural network with the decision tree algorithms ID3 and J48.[2] The comparisons showed the

superiority of the pruned J48 tree with 89.3% accuracy compared to the others with 81.9% accuracy. Codina proposed artificial flexibility on multilayer perceptron (AMMLP) as the final model for predicting diabetes with an accuracy of 89.93%.[3] All of the studies used the Pima Indians Diabetes Database for experiments. Also, the toolbox used by most researchers to perform analyzes was WEKA software.

Research shows that in order to obtain results with higher accuracy, data preprocessing operations must be performed before applying the proposed solution to clear the data and make it more meaningful and logical. Vijoyan examined the benefits of using different data processing methods to predict diabetes.[4] The preprocessing methods studied were principal component analysis (PCA) and discretization. Research has shown that preprocessing improves the accuracy of simple Bayesian classification and decision tree. This reduces the accuracy of the backup vector machine.[5] He analyzed the high-risk indicators of type 2 diabetes using association rules and the evaluation of false positive rates. Zhou also suggested the area of the ROC curve, the values of sensitivity and specificity for validation and review of test results.[6]

Sojania presented an Android-based application solution for raising awareness about diabetes in his article.[7] The application uses the decision tree classifier to predict users' blood sugar levels and provides information and suggestions about diabetes. The application uses data collected from a hospital in the Indian state of Chhattisgarh. Shi et al. have developed a model for assessing the risk of developing diabetes using a mobile device to prevent people from developing diabetes.[8]

Some articles focus on improving the K-means clustering algorithm. For example, Wang proposed an improved k-means clustering algorithm by removing noise data.[9] Sun proposed a solution to improve the selection of primary k-means clustering centers based on the Forubenius norm distance.[10] Shoni Wang proposed an improved k-means clustering algorithm with variance in which the primary clustering centers were selected using the Hoffmann tree structure. Most articles improve the initial values of cluster centers.[11]

People at risk for developing diabetes need to develop a set of rating standards for prediction.[12] In this regard, Chandrakar and Saini presented the risk score of Indian overweight diabetes as a tool to show diabetes to solve the problem of diagnosis or late diagnosis of diabetes.[13] Hamm and Lou proposed k-means clustering in pairs and limited to a certain size to represent the population at high risk of diabetes.[14] This provided a tool for classifying the risk of the disease.

In summary, some of the research done to predict diabetes. However, the accuracy of the prediction and the validity of the data were not sufficient for real applications. In addition, most of the models presented by researchers work well only on specific datasets that do not have acceptable results on different datasets. Therefore, we need to create a new forecasting model with higher accuracy and compatibility with other data sets. In this paper, in addition to the Pima Indians dataset, two other datasets are used to test the proposed model.

Materials and Methods

In recent years, the use of data mining algorithms to predict diseases has increased. Some researchers have shown that it is possible to obtain predictive models from initial patient data. In particular, most of the published articles in the field of diabetes prediction have been aimed at improving the accuracy of the model. In this regard, some researchers have obtained good results using the WEKA toolbox on the Pima Indian dataset.

This section includes a review of the data set used for the experiment, the ant colony algorithm for data preprocessing, and an artificial neural network for data classification. All simulation and experimental processes have been performed using MATLAB 2018 software.

The Pima Indian Diabetes Database contains information on 768 patients living near Arizona. Tests performed with positive and negative results show whether the patient has diabetes or not. For all samples, 8 numerical properties are considered. This data includes data on a person's health as well as the results of tests performed. The features in the dataset are as follows:

• *Number of pregnancies (preg)*

• *Plasma glucose concentration in 2 hours in a glucose tolerance test (plas(*

• *Diastolic blood pressure (pres(*

• *Thickness of the skin*

• *Two-hour serum insulin (insu(*

• *Body mass index (bmi(*

• *Racial function of diabetes (pedi(*

• *Age*

• *Class label*

One of the most effective tasks in creating a model is data preprocessing, which plays an important role in the modeling process by increasing the quality of data in large quantities. At this stage, by using some appropriate methods, data set optimization is done. First, numerical properties that have a certain interval are transferred to the interval of zero and one and normalization is performed on them. In the second stage of preprocessing, outlier data is identified using ant colony optimization and the mean value is recorded instead. At this point, some unknown values recorded in the data set are also recorded with the mean value. Then in the next step, the degree of dependence of the properties on the class property is calculated and based on that, the less effective properties are excluded from the feature set. In this way, the complexity of the data is reduced.

• *Ant Colony Optimization Algorithm*

The ant colony optimization algorithm is a probabilistic technique for solving computational problems which can be reduced to find good paths through graphs. Artificial ants stand for multi-agent methods inspired by the behavior of real ones. The pheromone-based communication of ants is often the predominant paradigm used. Combinations of artificial ants and local search algorithms have become a proper method for numerous optimization tasks involving some sort of graph such as vehicle routing and internet routing.

In the real world, ants of some species wander randomly and upon finding food return to their colony while laying down pheromone trails. If other ants find such a path, they are likely not to keep travelling randomly, but instead to follow the trail, returning and reinforcing it if they eventually find food.[15]

In this algorithms, an artificial ant is a simple computational agent that searches for good solutions to a given optimization problem. To apply this optimization algorithm, the optimization problem needs to be converted into the problem of finding the shortest path on a weighted graph. Initially, each ant stochastically constructs a solution, i.e. the order in which the edges in the graph should be followed. Secondly, the paths found by the different ants are compared. The last step consists of updating the pheromone levels on each edge.

• Artificial Neural Network

An Artificial Neural Network (ANN) is based on a collection of connected nodes called artificial neurons, which loosely model the neurons in a biological brain. All connection like the synapses in a biological brain can transmit a signal to other neurons. Artificial neurons that receive signals then process them and can signal neurons connected to them. The signal at a connection is a real number and the output of each neuron is computed by some nonlinear functions of the sum of its inputs.[16] These connections are called edges. Neurons and edges usually have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at connections. Neurons may have threshold points such that a signal is sent only if the aggregate signal crosses that threshold. Generally, neurons are aggregated into layers. Different layers may perform different transmissions on their inputs. Signals travel from the first layer to the last layer, likely after traversing the layers multiple times.

```
procedure ACO_MetaHeuristic is
  while not_termination do
    generateSolutions()
    daemonActions()
    pheromoneUpdate()
  repeat
end procedure
```

Figure 1. Pseudo-code for ACO. [15]

Results

To obtain the most accurate answer, 10-fold cross

validation and percentage split validation methods with different percentages were used. In the first validation method, the data set is divided into 10 subsets and in 10 consecutive periods, 9 subsets are used as training sets and another set is used for testing. In the second validation method, the data is used as a training set and the rest as a test set. Also found parameters are true positive rates (TPR), false positive (FPR), true negative (TNR), false negative (FNR), accuracy and f-measure.

In addition to the mentioned parameters, the ROC diagram related to the simulation is also calculated. This graph shows the ratio of positive rate to true positive false that the higher the level below the chart, the more accurate the model is.

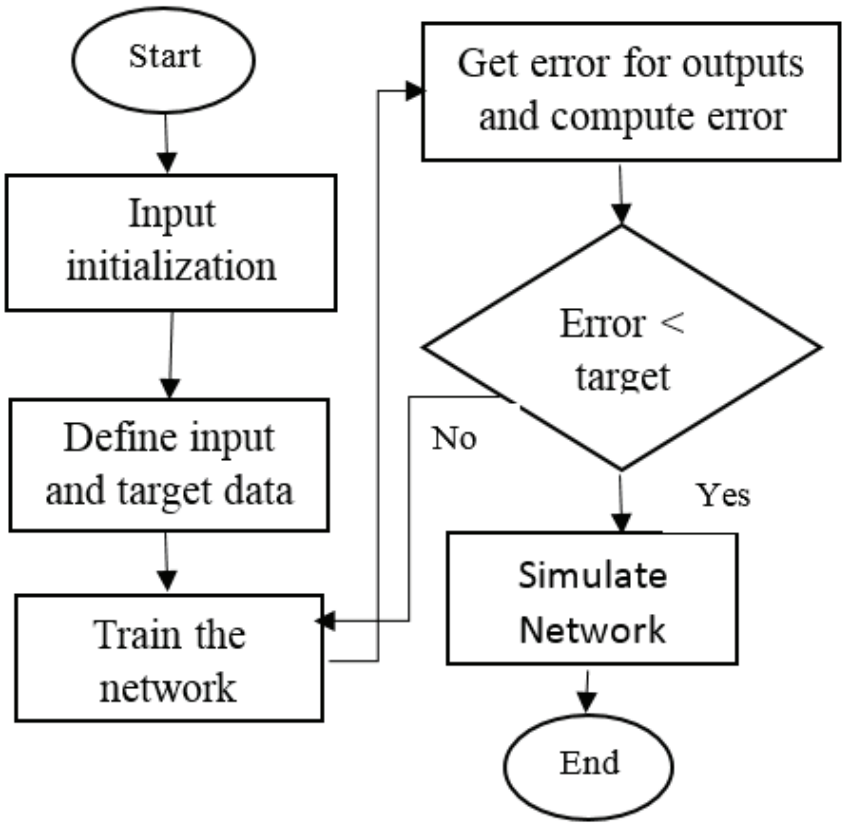

Figure 2. Flowchart for ANN

The simulation results are shown on the dataset using each of the methods listed in Table 1. As we see, the application of the 10-fold method results in an accuracy of 99.24% and the use of the percentage split method results in an accuracy of 98.73%. If the accuracy of the model is considered as the main criterion of the accuracy of the proposed model, then the 10-fold validation method is superior to another. However, a closer look at the table shows that the amount of FPR in the second method is lower than in the first method, which means that in this method a person without diabetes is less likely to be labeled diabetic and in certain circumstances can it is a better option than the first method. Also, the true negative rate in the second method is superior to the first method (higher value), and this is useful when we want to identify people who do not have diabetes more accurately, in which case the percentage split validation method is better than the other. In other cases, the first method is superior.

Examining the table, it can be seen that in all accuracy parameters, the k-fold method is superior to the percentage split method; both in the parameters that have a positive aspect (such as true positive and false positive) or in negative parameters (true negative and false negative). The accuracy and f-measure values in

the first method are better than the second method.

Table 1. Simulation results on the dataset

Method	TPR	TNR	FPR	FNR	Accuracy	F-measure
10-fold	0.995	0.984	0.052	0.076	99.24	0.989
Cross-validation	0.983	0.997	0.053	0.083	98.73	0.978

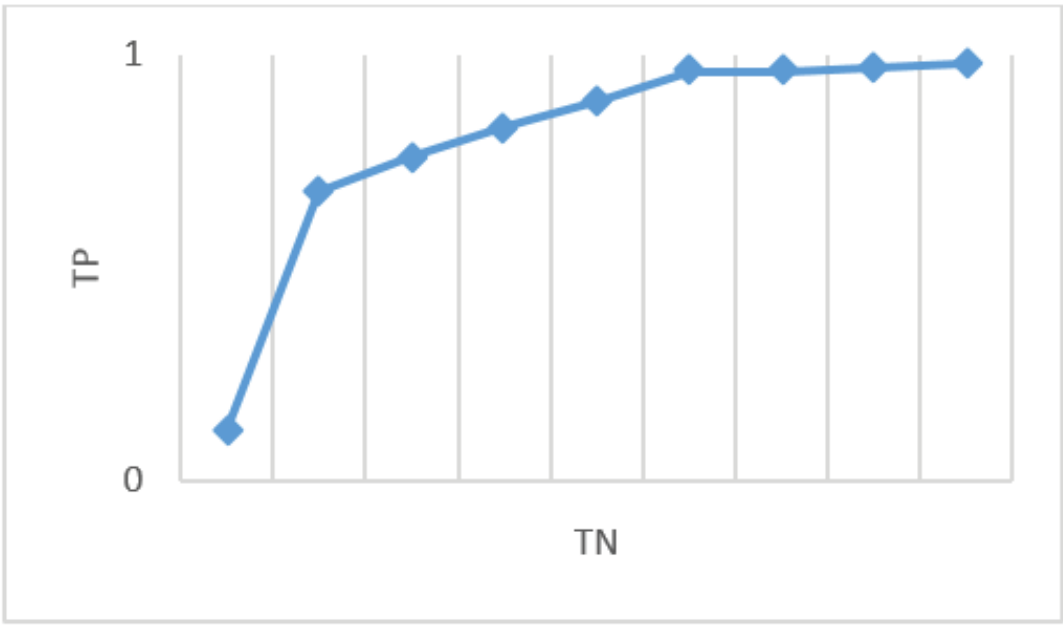

Figure 3. ROC on the dataset

Conclusion

As To prove that the proposed model improves the accuracy of the prediction, we compare the results with the experiments of other researchers in this field. Table 2 summarizes this comparison.

The accuracy obtained from the proposed method is 98.73% in the lowest case and 99.24% in the best case. As can be seen in Table 2, indicates the comparison of the accuracy of the proposed method with some recent approaches. Therefore, the proposed method is more appropriate than the other proposed methods.

The aim of this article was to create a suitable predictive model for diagnosing high-risk diabetes. In this paper, a new model for forecasting is proposed which includes two stages: data preprocessing phase and categorization phase. In the preprocessing phase, the ant data are identified using the ant colony algorithm and replaced with the mean value. The second phase is based on a neural network algorithm that uses which data is categorized. The results obtained from the simulation were compared with the results of other research works in this field and it was found that the accuracy of the proposed model is higher than other researches.

Table 2. Comparison of the proposed method with some other approaches

Method	Accuracy
Proposed Approach	99.24%
Decision Tree	93.75%
Fuzzy	96.23%
Fuzzy + Neural Network	98.19%
ACO	95.83%
Bayesian Network	93.74%
Firefly + PSO	97.44%

Ethical Clearance: This article has been routed through the anti-plagiarism cell of Institutional Review Board.

Conflict of Interest: The author declares that they have no conflict of interests.

Source of Findings: Golestan University is the source finding of this paper.

References

1. Patil B. Hybrid prediction model for Type-2 diabetic patients. Expert Systems and Applications. 2010; 37: 8102–8108.
2. Aliza A, and Aida M. Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus. International Conference on Digital Information Processing and Communications, Springer. 2011; 537-545.
3. Cedeno M, Joaquín T, and Diego A. A prediction model to diabetes using artificial meta-plasticity. International Work-Conference on the Interplay Between Natural and Artificial Computation. 2011; 418-425.
4. Vijayan V, and Anjali C. Decision support systems for predicting diabetes mellitus– A review. Proceedings of 2015 global conference on communication technologies (GCCT 2015), Thuckalay. 2015; 98-103.
5. Zhe W, Guangjian Y, and Nengcai W. Analysis for risk factors of type 2 diabetes mellitus based on FP-growth algorithm. China Medical Equipment. 2016; 13: 45–58.
6. Guo Y. Application of artificial neural network to predict individual risk of type 2 diabetes mellitus. Journal of Zhengzhou Univiversity. 2014; 49: 180-183.
7. Sowjanya M.K. MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices. IEEE International Advance Computing Conference (IACC), Bangalore. 2015; 397-402.
8. Shi G, Liu S, and Ye D. Design and Implementation of Diabetes Risk Assessment Model Based On Mobile Things. 7th International Conference on Information Technology in Medicine and Education, Huangshan. 2015; 425-428.
9. Wang J, and Su X. An improved K-Means clustering algorithm. 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN), Xi'an. 2011; 44-46.
10. Sun Y, Fang L, and Wang P. Improved k-means clustering based on Efros distance for longitudinal data. 2016 Chinese Control and Decision Conference (CCDC), Yinchuan. 2016; 3853-3856.
11. Wang S. Improved K-means clustering algorithm based on the optimized initial centroids. 3rd International Conference on Computer Science and Network Technology (ICCSNT), Tiruchengode.2013; 450-453.
12. Songthung P, and Sripanidkulchai K. Improving

Type 2 Diabetes Mellitus Risk Prediction Using Classification. 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen. 2016; 1-6.

13. Omprakash C, and Saini J.R. Development of Indian weighted diabetic risk score (IWDRS) using machine learning techniques for Type-2 diabetes. ACM COMPUTE, Gandhinagar. 2016; 125-128.
14. Longfei H, and Senlin L. An intelligible risk stratification model based on pairwise and size constrained K-means. IEEE Journal of Biomedicine Health Information. 2016; 21:1288–96.
15. Ojha V.K, Abraham A, and Snasel V. ACO for Continuous Function Optimization: A Performance Analysis. 14th International Conference on Intelligent Systems Design and Applications (ISDA). 2017; 145 – 150.
16. Ojha V.K, Abraham A, and Snasel V. Meta heuristic design of feed forward neural networks: A review of two decades of research. Engineering Applications of Artificial Intelligence. 2017; 60: 97–116.