Type 2 Diabetes Prediction using Gray Wolf Optimization Algorithm

Aliakbar Tajari Siahmarzkooh

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

Abstract

Background: Increasing the number of diabetic patients and the ignorance of most of these patients about the dangers arising from it is a challenge that threatens human lives.

Materials and Method: In this paper, a new solution based on the Gray Wolf Optimization (GWO) algorithm for predicting type 2 diabetes is presented. The main purpose of the proposed method is to increase the accuracy of prediction and also to reduce the probability of getting stuck in local optimal points. In more detail, the proposed method consists of two parts: 1- data preprocessing including data preparation and noise cancellation and 2- data classification using gray wolf algorithm. The Pima Indians Diabetes dataset in MATLAB simulation environment was used to analyze the data and compare the research results.

Results: The simulation results show that by adjusting the parameters of the gray wolf algorithm, about 6% better prediction accuracy is obtained than other researches.

Conclusion: Also, for a more accurate evaluation of the proposed method, two other datasets have been used for testing. The results of experiments show that the proposed model for health management in diabetes is effective.

Keywords: *Disease forecasting, Gray Wolf Optimization (GWO), Diabetes, Clustering.*

Introduction

Diabetes is a chronic disease that is diagnosed with high blood glucose levels. About half of all diabetics have inherited traits, which is one of the most important features of diabetes. Poor pancreatic insufficiency and insufficient use of insulin are among the causes of diabetes. The International Diabetes Association (IDF) has reported recent data on diabetes in the Diabetes Atlas (Seventh Edition).[1] Statistics show that in 2020, the number of diabetics worldwide was about 460 million, which given the growing number of diabetics is projected to reach 640 million.

It seems necessary to focus more on people at high risk of diabetes to reduce the prevalence and effects of diabetes. We need information-based methods to study high-risk groups for diabetes. In this regard, meta-heuristic algorithms are good tools that are used as computational processes to discover patterns in large datasets and include several solutions such as evolutionary clustering, machine learning and Gray Wolf Optimization (GWO) algorithm.[2-5]

In recent years, various data mining methods have been used to predict diseases. Patil presented a hybrid predictive model in which the k-means clustering algorithm was used to validate the data class label and C4.5 decision tree algorithm was used to create the final model.[6] The results of his proposed method have an accuracy of 92.38% in classification. Aliza compared the predictive accuracy of the MLP model in the neural network with the decision tree algorithms ID3 and J48.[7] 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%.[8] All of the studies used the Pima Indians diabetes database for

experiments. Also, the toolbox used by most researchers to perform the analyses was WEKA software.

Vijoyan examined the benefits of using different data processing methods to predict diabetes.[9] The preprocessing methods studied were PCA and discretization. Researches show that the preprocessing improves the accuracy of simple Bayesian classification and decision tree. This reduces the accuracy of the backup vector machine. In another paper, researcher analyzed the high-risk indicators of type 2 diabetes using association rules and the evaluation of false positive rates.[10] Zhou also suggested the area of the ROC curve, the values of sensitivity and specificity for validation and review of test results.[11]

Sojania presented an Android-based application solution for raising awareness about diabetes in his paper.[12] Wang proposed an improved k-means clustering algorithm by removing noise data.[13] Yanbui Sun proposed a solution to improve the selection of k-means initial clustering centers based on the Forubenius distance.[14] Wang proposed an improved k-means clustering algorithm with variance in which the primary clustering centers were selected using the Hoffmann tree structure.[15]

Omprakash and Saini presented the risk score for Indian overweight diabetes as a tool to show diabetes to solve the problem of diagnosis or late diagnosis of diabetes.[16] Longfei and Senlin proposed k-means clustering in pairs and limited to a certain size to represent the population at high risk of diabetes.[17] This solution 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 proposed 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 datasets. In this paper, the Pima Indians dataset is used to test the proposed model.

Materials and Methods

In this paper, a new solution based on the GWO algorithm for predicting type 2 diabetes is presented. In the proposed algorithm, stronger wolves are replaced by weaker wolves based on their fitness level. In each iteration of the algorithm, the degree of suitability is calculated and if it improves, the algorithm is repeated again, otherwise the algorithm terminates. The main purpose of the proposed method is to increase the accuracy of prediction and also to reduce the probability of getting stuck in local optimal points. To ensure the accuracy of the proposed model, the proposed method is tested on two datasets. In more details, the proposed method consists of two parts: 1- data preprocessing including data preparation and noise cancellation and 2- data classification using GWO algorithm.

Data Preprocessing

One of the most effective tasks in creating a model is data preprocessing, which plays an important role in the process modeling by increasing the quality of data in large quantities.[18] At this stage, the dataset optimization is done by using some appropriate methods. First, numerical properties that have a certain interval are transferred to the interval of zero and one and the normalization is performed on them. In the second stage of preprocessing, the output data is identified using k-means clustering and the mean value is recorded. At this point, some unknown values recorded in the dataset 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 the less effective properties are excluded from the feature set based on that. In this way, the complexity of the data is reduced.

- **Gray Wolf Optimization**

The gray wolf algorithm is derived from the social life of gray wolves to determine the leader, which was actually proposed by Mirjalili et al. in 2014.[19] Four types of gray wolves, including alpha, beta, delta, and omega, are used to simulate the leadership hierarchy. Also, three main stages of hunting, namely bait search, bait siege and bait attack are performed to perform optimization. Gray wolves are considered agile predators, meaning that they are at the top of the food chain. They prefer to

live in groups. The average size of the group is 5 to 12 wolves. The remarkable thing is that they have a special and difficult hierarchy for social domination.

The leaders of the group are a man and a woman named alpha, who are mostly responsible for deciding on hunting, where to sleep, when to wake up, and so on. Alpha decisions are communicated to the group. However, there is also a kind of democratic behavior in which the alpha follows the other wolves in the group. At gatherings, the whole group acknowledges the alpha by holding their breath. Alpha is also called the dominant wolf because his commands must be carried out by the group. Alpha wolves are only allowed to mate. Alpha is not the strongest member of the group, but the best in terms of group management. This shows that the organization and discipline of a group is much more important than its strength.[20]

The second level is in the beta gray wolf hierarchy. Beta are sub-wolves that help alpha make decisions or other group activities. The beta wolf can be male or female, and will probably be the best candidate for alpha if one of the alpha wolves dies or gets old. The wolf beta must respect alpha while commanding other lower level wolves. In fact, it plays the role of an alpha consultant and regulator of the group. Beta boosts alpha commands throughout the group and gives feedback to alpha.

The lowest grade is the omega gray wolf. Omega plays the role of the victim. Omega wolves must always surrender to all dominant wolves. They are the last wolves allowed to eat. It seems that omega is not an important person in the group, but on the other hand, if omega is lost, the whole group will face civil war and problems. This is due to omega causing violence and frustration in all wolves. This helps to maintain the whole group and the hierarchical structure. In some cases, omegas support the group.

If the wolf is not alpha, beta, or omega, he is called delta. Delta wolves must submit to alpha and beta, but they dominate omega. Scouts, guards, elders, hunters and caretakers belong to this category. The scouts are responsible for watching the boundaries of the group's territory in the event of any danger. Guards ensure the safety of the group. The elders are experienced wolves that used to be alpha or beta. Hunters help with alpha and beta when hunting prey and providing food for the group. Finally, caregivers are responsible for caring for weak, sick, and injured wolves in this group.

Results

The proposed algorithm is simulated using MATLAB 2020 software on Pima Indians. Validation methods with 10 repetitions of k-fold cross validation and percentage split with different percentages were used to obtain the most accurate answer. In the first validation method, the dataset 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 evaluation parameters are true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), false negative rate (FNR), accuracy and f-measure.

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

The simulation results using each of the methods listed in Table 1. As can be seen, the application of k-fold method results in an accuracy of 98.94% and the use of percentage split method results in an accuracy of 97.59%. With this result, if the accuracy of the model is considered as the main criterion of the accuracy of the proposed model, the k-fold validation method with a value of k = 10 is superior to the other. However, a closer look at the Table 1 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 is a better option than the first method.

Table 1. Best test results on Pima Indian dataset

Validation Strategy	TPR	TNR	FPR	FNR	Accuracy	F-measure
10-fold cross validation	0.991	0.995	0.043	0.025	98.94	0.984
Percentage split	0.985	0.962	0.018	0.076	97.59	0.943

Also, the correct negative rates in the second method is superior to the first method (higher values), 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. The accuracy and f-measure values in the first method are better than the second method.

Figure 1 shows the ROC diagrams of the experiments performed on the dataset. The below area of the graph is large. This means that the ratio of the number of true predictions to false predictions on the dataset is higher. Therefore, considering the ROC diagram, the accuracy of the proposed model on the dataset is acceptable.

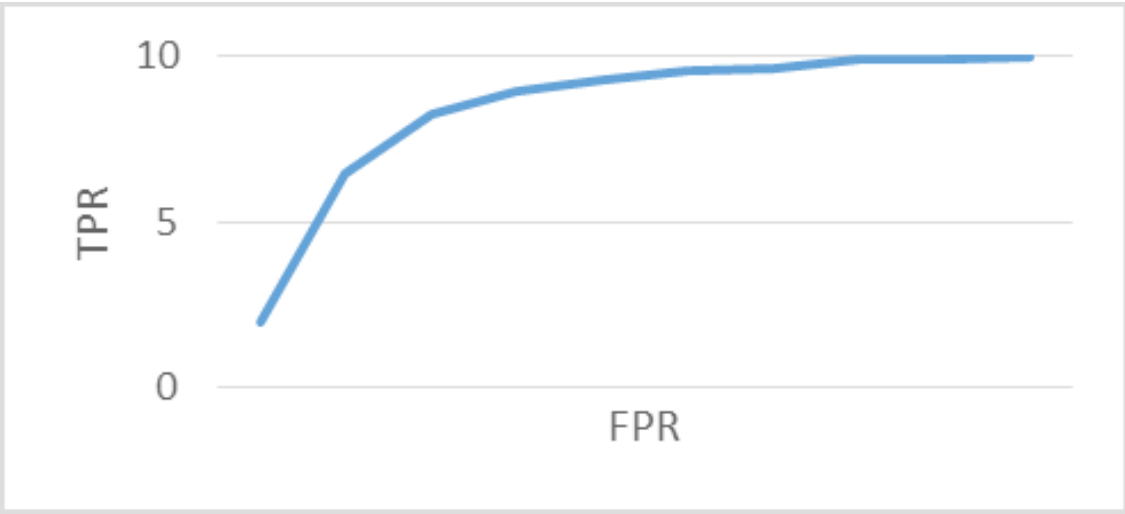

Figure 1. ROC diagram of the proposed model on the Pima Indian dataset

· **Model Validation**

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

The obtained accuracy of the proposed method is 97.59% in the lowest case and 98.94% in the best case. As can be seen in Table 2, the accuracy of the proposed Patil method is 92.38%, which is the closest to the accuracy of the proposed method in this paper. The accuracy of the Cedeno method, which is a model based

on a neural network, is 89.93%. Other methods based on decision trees are also less than expressed values. Therefore, the proposed method is more appropriate than the other proposed methods.

Table 2. Comparison of the proposed model with other methods

Method	Accuracy (%)
Our model	98.94
Patil [6]	92.38
SVM+ k-means	89.93
Decision Tree	89.30
PCA	84.50
Logistic	78.22
Meta-Heuristic	75.75
Simple Bayesian	74.94

Discussion and Conclusion

The aim of this article was to create an appropriate predictive model for the diagnosis of high-risk diabetes. In this paper, a new model for forecasting is proposed, which includes two stages: the data preprocessing phase and the categorization phase. In the preprocessing phase, the k-means clustering algorithm identifies the outgoing data and replaces it with the mean value. The second phase is based on the gray wolf algorithm, which uses the categorized data. 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.

Data collection from the country's hospitals and examining patients with internal diabetes and creating an application on a mobile phone to test for diabetes can be examples of prospects and achievements in future works.

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: No.

References

1. Exarchos, K. P. Prediction of coronary atherosclerosis progression using dynamic Bayesian networks. IEEE EMBC 2013; 192-203.
2. Jordan, M and Mitchell, T. Machine learning: Trends, perspectives, and prospects. Science 2015; 349: 255-260.
3. Kandhasamy, P and Balamurali, S. Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science 2015; 47: 45-51.
4. Olaniyi, E.O and Adnan, K. Onset diabetes diagnosis using artificial neural network. International Journal of Science and Engineering Research 2014; 5: 754-759.
5. Sarkar, R.P and Maiti, A. Investigation of Dataset from Diabetic Retinopathy Through Discernibility-Based k-NN Algorithm. Cont Adv in Innov and App Inf Tech 2018; 5: 93-100.
6. Patil, BM. Hybrid prediction model for Type-2 diabetic patients. Expert Syst Appl 2010; 37: 8102–8108.
7. 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 2011; 34-43.
8. 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; 121-133.
9. 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 114-125.
10. Zhe, W, Guangjian, Y and Nengcai, W. Analysis for risk factors of type 2 diabetes mellitus based on

FP-growth algorithm. China Med Equip 2016; 13: 45–48.

11. Guo, Y. Application of artificial neural network to predict individual risk of type 2 diabetes mellitus. J Zhengzhou Univ 2014; 49: 180-183.
12. Sowjanya, M.K. MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices. IEEE International Advance Computing Conference (IACC), Bangalore 2013; 108-121.
13. Wang, J and Su, X. An improved K-Means clustering algorithm. 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN) 2011; 125-132.
14. Sun, Y, Fang, L, and Wang, P. Improved k-means clustering based on Efros distance for longitudinal data. Chinese Control and Decision Conference (CCDC), Yinchuan 2016; 29-41.
15. Wang, S. Improved K-means clustering algorithm based on the optimized initial centroids. 3rd International Conference on Computer Science and Network Technology (ICCSNT), Tiruchengode, Tamil Nadu 2013; 117-133.
16. 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; 44-62.
17. Longfei, H, and Senlin, L. An intelligible risk stratification model based on pairwise and size constrained K-means. IEEE J Biomed Health Inf 2016; 25: 1288–96.
18. Kamalesh, M.D. Predicting the Risk of Diabetes Mellitus to Subpopulations Using Association Rule Mining. The International Conference on Soft Computing Systems 2016; 143-161.
19. Mirjalili, S, Gandomi, A.H, Mirjalili, S.Z, Saremi, S, Faris, H, and Mirjalili, S.M. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. Adv in Eng Sof 2017; 114: 163-191.
20. Mirjalili, S, and Lewis, A. Grey wolf optimizer. Advances in engineering software 2014; 69: 46-61, 2014.