International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 4038-4041
© Research India Publications. http://www.ripublication.com

An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach

Mustafa S. Kadhm[1] **Ikhlas Watan Ghindawi**[2] **Duaa Enteesha Mhawi**[3]

[1]*Computer Engineering Techniques, Imam Ja'afar Al-Sadiq University, Baghdad, Iraq.*

[2]*Computer science department, Collage of education, Al-Mustansiriya University, Baghdad, Iraq.*

[3]*Middle technical university, technical institute for administration, Baghdad, Iraq.*

Abstract

Diabetes prediction system is very useful system in the healthcare field. An accurate system for diabetes prediction is proposed in this paper. The proposed system used K-nearest neighbor algorithm for eliminating the undesired data, thus reducing the processing time. However, a proposed classification approach based on Decision Tree (DT) to assign each data sample to its appropriate class. By experiments, the proposed system achieved high classification result which is 98.7% comparing to the existing system using Pima Indians Diabetes (PID) dataset.

Keywords: Diabetes, PID, KNN, Decision Tree, Classification.

INTRODUCTION

Diabetes mellitus is a syndrome characterized by metabolic disorder and abnormal rise in the concentration of blood sugar caused by insulin deficiency, or low insulin sensitivity of tissues, or both of them. Diabetes leads to serious complications or even premature death. However, to diagnosing diabetes, several time consuming tests and analyzing critical factors are done. Now days, machine learning algorithms are used to classify and diagnosis the diseases, in order to eliminate the problem and reduce the required cost. Besides that, using the machine learning algorithm lead to meaningful and accurate decisions. [1]

Medical data sets that contain irrelevant data (noise) are used to train and evaluate the machine learning algorithms. The noise are affects the decision results of the used algorithm. [2]

This paper analyses and overcome the accompanied problems of the diabetes through removing the noise and predicting the diabetics using medical data set and machine learning algorithms.

RELATED WORKS

Several researcher have develop and design a systems for diabetes prediction based on various algorithms and methods. V. Anuja et al. [3] proposed a system for diabetes disease classification using Support Vector Machine (SVM). The authors used Pima Indian diabetes dataset for evaluation. The obtained accuracy was 78% base on using the Radial Basis Function (RBF) kernel of SVM as the classifier. Aiswarya et al. [4] used J48 Decision Tree and Naïve Bayes as a classifiers for classify the diagnosis of diabetes. Pima Indian diabetes dataset is used in the proposed system and the classification results was 74.8%, 79.5% for J48 Decision Tree and Naïve Bayes respectively. Rajesh et al. [5] proposed a system for diabetes classification based on using C4.5 algorithm for classification. The authors achieved classification rate of 91% by evaluating the training data through data feature relevance analysis. Harleen et al. [6] proposed a system based on a technique in data mining for diabetes disease prediction. The proposed system has three main steps which are: preprocessing, feature extraction and parameter evaluation. In preprocessing step, the empty and anomalies sets are removed from the used dataset. Besides that, the helpful hidden patterns and relationships of the dataset are explored in the feature extraction step in order to improve the decision making result. Furthermore, the proposed system evaluated based on using J48, Naive Bayes and the achieved rates are 73.8%, 76.3% respectively. In addition, Ravi et al. [7] fuzzy c means clustering and support vector machine for developing diabetes mellitus prediction. The authors used a dataset that consists of 768 cases and the obtained result was 59.5%. In [8] Krishnaveni et al. proposed a six various techniques to predict diabetic disease. The used techniques are Discriminant analysis, KNN Algorithm, Naïve Bayes, SVM with Linear Kernel function, and SVM with RBF Kernel function. The obtained results of the proposed system for the used techniques are 76.3% using discriminant analysis, 71.1% using KNN Algorithm, 76.1% using Naïve Bayes, 74.1% using SVM with Linear Kernel function, 74.1% using SVM with RBF Kernel function. However, several authors in [9] [10] [11] are used various methods in order to get the best prediction rate.

THE DATASET [12]

Pima Indians Diabetes (PID) dataset of National Institute of Diabetes and Digestive and Kidney Diseases. PID is composed of 768 instances as shown in Table 1. Eight numerical attributes are represent each patient in data set.

Table 1: PID Dataset.

No.	**Name**
1	Number of times pregnant
2	Plasma glucose concentration a two hours
3	Diastolic blood pressure
4	Triceps skin fold thickness
5	2-Hours Serum insulin
6	Body mass index
7	Diabetes pedigree function
8	Age

Table 2 illustrates all the statistical datum of each feature vector in the PID dataset.

Table 2: Statistical Datum of PID Dataset.

Attribute	**Min value**	**Max value**	**Mean value**	**Standard division**
1	0	17	3.845	3.37
2	0	199	120.859	31.973
3	0	122	69.105	19.356
4	0	99	20.536	15.952
5	0	846	79.799	115.244
6	0	67.1	31.993	7.884
7	0.078	2.42	0.427	0.331
8	21	81	33.241	11.76

THE PROPOSED SYSTEM

The proposed diabetes prediction system has two main stages that work together to achieve the desired results. The first stage of the proposed system is the data preparation, and the second one is the classification. However, the input into the system is the PID dataset and the output will be one class which represent the healthy or the diabetic. Figure 1 illustrates the diagram of the proposed system stages.

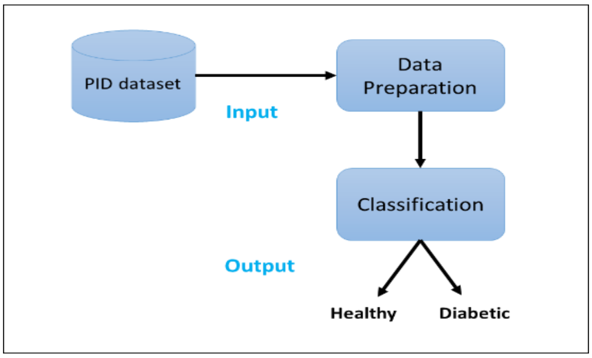

Figure 1: Main Stages of the Proposed System.

Data Preparation

In this stage of the proposed system the input data is processed through number of steps in order to improve the system performance. First of all, data reduction is applied on the input dataset to eliminate the noisy and inconsistent data. K-means clustering algorithm is performed on the input dataset in order to partition data to k clusters. In each cluster the most appropriate features will assigned based on its centroid. The process of data preparation stage is shown in Algorithm 1.

```
Algorithm 1: Data Preparation
Input: PID dataset
Output: Prepared dataset

Begin
Step1: Load the PID dataset
Step2: Remove the noisy and inconsistent data
Step3: Run K-means clustering
Step4: Count samples number of healthy and diabetes for each cluster
Step5: Remove the small size group in each cluster // Prepared dataset
Step6: Return (Prepared dataset)
End
```

Classification

In the classification stage, each input sample of the data that came from the previous stage will be assigned to whether healthy or diabetic. In the proposed system, a proposed classification approach based on using several Decision Trees (DTs) is used as in Figure 2.

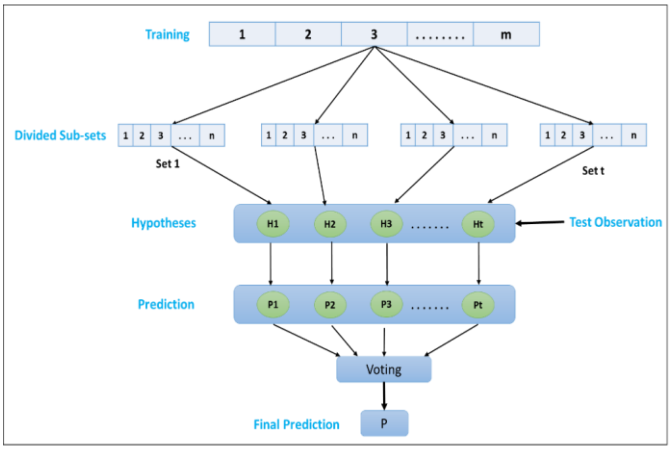

Figure 2: Proposed Classification Approach.

In the proposed approach, the input data is divided into number of subsets. The divided subsets must cover all the input set and any input data sample could be used in more than one divided subsets. After that, each divided subsets is classified by one decision tree, then the results of all the used decision trees are collected. In addition, the voting process is taking place be selectin the most frequent result of the decision trees. The main steps of the proposed classification approach is sowed in Algorithm 2.

```
Algorithm 2: Classification
Input: Prepared dataset
Output: Classification result

Begin
Step1: Load the input dataset
Step2: Divide the input data into number of subsets
Step3: Run a decision tree classifier on each subsets
Step4: Collect the all results of decision tree classifiers
Step5: Apply voting majority on the collected results //select the most frequent result
Step6: Return (Classification result)
End
```

EXPERIMENTAL AND DISCUSSION

In this research paper, the proposed system code written by Matlab 2016a, and run in Microsoft Windows 10 environment. In another hand, the performance of any classification system could negatively affected by the duplicate, noisy, uncertain, and inconsistent data. The data preparation stage of the proposed system address this problem by using k-mean clustering algorithm. It leads to reduce the time consumption rate of the system. An example of removing the inconsistent and noisy data based on applying K-mean clustering is shown in Figure 3.

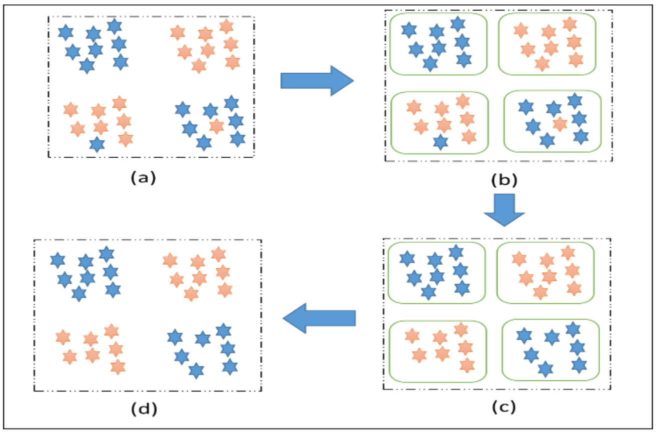

Figure 3: Noisy data removal. (a) Input data, (b) Apply K-means clustering, (c) Noise removal, (d) Reduced set.

In the proposed system ten clusters are used in order to get the best separation points and most accurate results. Each of the separated clusters contains number of samples that may belong to either first class which is (diabetic) or to the second one which is (healthy) or for the both classes. The proposed system keeps the group that have more samples than others which then remove it.

An example in such a situation, cluster 1 in Table 2 there are 129 healthy samples and 15 diabetic samples. To obtain an accurate decision the inconsistent samples of diabetic should be removed from the cluster.

Table 2: Number of Healthy and Diabetes Samples in Each Custer.

Clusters	1	2	3	4	5	6	7	8	9	10
Healthy	129	22	8	48	67	3	35	70	114	4
Diabetic	15	48	39	45	36	2	21	4	41	17

Totally, in this experiment, 570 samples out of 768 samples are taken as input data for the proposed classifier approach and 198 samples as noisy and inconsistent data is removed. Moreover, all the kept samples pass into DTs for classification. The experiment results that achieved by the proposed system are shown in Table 3.

Table 3: Experiment Results of the Proposed System.

Criteria	**Result**
Sensitivity	98%
Specificity	98.2%
Accuracy	98.7%
Time (Second)	1.28

In another hand, the proposed system achieved an efficient results comparing to the existing systems that work for diabetic classification as illustrated in Table 4.

Table 4: Comparison Results of the Existing Systems and the Proposed System.

Used Method	**Accuracy**
Naive Bayes [13]	79.56%
SVM [14]	78%
SVM + Modified k-means [15]	96.71%
LDA- MWSVM [16]	89.74%
SVM + Fuzzy C-means clustering [17]	94.30%
Proposed	**98.7%**

CONCLUSION

A fast and accurate diabetes prediction system is proposed in this paper. The proposed system used 768 instances within 8 attributes for each one of PID dataset. The used data is preprocessed in order to remove the unwanted data, and lead to faster processing time. Moreover, the dividing technique of the dataset into subset, made an optimal classification result. The proposed system focused on the features analysis and classification parts. The propositions of these parts lead to an optimal achievements. The results of experiments illustrated the effects of using the algorithms of the proposed system through achieving a higher classification rate that the other systems.

REFERENCES

[1] Yilmaz N., Inan O., Uzer M.S., " A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases," J Med Syst, vol. 38, no. 5 2014.

[2] Lowongtrakool C., Hiransakolwong N., "Noise filtering in unsupervised clustering using computation intelligence," International Journal of Math, vol. 6, no. 59, pp. 2911–2920, 2012.

[3] V. Anuja and R.Chitra., "Classification Of Diabetes Disease Using Support Vector Machine", International Journal of Engineering Research and Applications (IJERA), vol.3,Issue 2, pp. 1797-1801, 2013.

[4] Aiswarya I., S. Jeyalatha and Ronak S., "Diagnosis Of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), vol.5, ,No. 1, pp. 1-14, 2015.

[5] K.Rajesh and V.Sangeetha,"Application of Data Mining Methods and Techniques for Diabetes Diagnosis," in proceedings of International journal of Engineering and Innovative Technology, vol.2, Issue 3, pp. 43-46, 2012.

[6] Harleen and Dr. Pankaj B.,"A Prediction Technique in Data Mining for Diabetes Mellitus," Journal of Management Sciences and Technology, vol. 4, Issue 1, pp. 1-12, 2016.

[7] Ravi S. and Smt T., "Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine," International Journal of Computer Trends and Technology (IJCTT), vol. 11, No. 2, pp. 94-98, 2014.

[8] G. Krishnaveni*, T. Sudha," A Novel Technique To Predict Diabetic Disease Using Data Mining Classification Techniques" in International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT-2017), vol. 3, Issue 1, pp. 5-11, 2017.

[9] Raj A., Vishnu P., and Kavita B.,"K-Fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data Set Using Higher Order Neural Network and PCA", International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-6, pp. 436-438, January 2013.

[10] Vrushali B., and Rakhi W., "Review on Prediction of Diabetes using Data Mining Technique", International Journal of Research and Scientific Innovation (IJRSI), Volume IV, Issue IA, pp. 43-46, January 2017.

[11] Thirumal P., and Nagarajan N.," Utilization of Data Mining Techniques for Diagnosis of Diabetes Mellitus - A Case Study", ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 1, pp. 8-13, January 2015.

[12] Jain A.K., Murty M.N., Flynn P.J., "Data clustering: A review," ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

[13] Iyer A., Jeyalatha S., Sumbaly R., "Diagnosis of diabetes using classification mining techniques," International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. 5, no. 1, 2015.

[14] Kumari A.V., Chitra R., "Classification of diabetes disease using support vector machine," Int J Eng Res Appl, vol. 3, no. 2, pp. 1797-1801, 2013.

[15] Yilmaz N., Inan O., Uzer M.S., " A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases," J Med Syst, vol. 38, no. 5 2014.

[16] Çalişir D., Doğantekin E., "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine classifier," Expert Syst Appl, vol. 38, no. 7, pp. 8311–8315, 2011.

[17] Sanakal S., Jayakumari S.T., "Prognosis of diabetes using data mining approach-Fuzzy C means clustering and support vector machine," International Journal of Computer Trends and Technology (IJCTT), vol. 11, no. 2, 2014.