Design of Classifier for Detection of Diabetes using Neural Network and Fuzzy k-Nearest Neighbor Algorithm

Mrs. Madhavi Pradhan[1], Ketki Kohale[2], Parag Naikade[3], Ajinkya Pachore[4], Eknath Palwe[5]

[1]Asst. Prof., Department of Computer Engineering, AISSMS's College of Engineering, Pune, Maharashtra, India-411001
[2]Department of Computer Engineering, AISSMS's College of Engineering, Pune, Maharashtra, India-411001

Abstract
Diabetes Mellitus is one of the growing vitally fatal diseases world-wide. A design of classifier for the detection of Diabetes Mellitus with optimal cost and precise performance is the need of the age. The current project implementation looks further to train self organizing neural networks and apply fuzzy logic to effectively classify a diabetic patient as such. Neural networks are so chosen due to their dynamic nature of learning and future application of knowledge. Fuzzy logic allows partial membership and rule base that allows direct mapping between human thinking and machine results. The proposed method here uses a neural network implementation of the fuzzy k-nearest neighbor algorithm for designing of classifier.

The system is to be run on small mobile devices to facilitate mobility to the user while the processing is to be done on a server machine.

Keywords: *classification, diabetes detection, fuzzy knn, hybrid soft computing, k nearest neighbor, neural networks, weka*

1. Introduction

Diabetes Mellitus is one of the fatal diseases growing at a rapid rate in developing countries like India. This rate is also critical in the developed countries, Diabetes Mellitus being one of the major contributors to the mortality rate. Detection and diagnosis of Diabetes at an early stage is the need of the day. It is required that a classifier be designed that is cost efficient, convenient and most importantly, accurate. Artificial Intelligence and Soft Computing Techniques mimic a great deal of human ideologies and are encouraged to involved in human related fields of application. These systems most fittingly find a place in the medical diagnosis. With the following literature survey, we propose a model for detection of diabetes in the coming sections.

As much as there does exist a need for exact classification with accuracy, it should be understood that detection of a pre-diabetic situation is highly beneficial to the community. With this model, we expect to find a methodology for detection of the pre-diabetic conditions so as to provide a sound warning before hand.

2. Related Literature

There has been a lot of encouragement in the soft computing field for the development of methods that ease the medical diagnosis. There has been a debate over whether to support the use of SVM in the current scenario or not. In the present literature survey, while ([1],[2],[6]) support the use of Support Vector Machine, the rest proposed models based on various types of neural networks ([7],[8]), neuro fuzzy systems ([3],[5]) as well as the genetic programming ([9],[11]).

One of the most accurate data mining techniques (such as [1],[2]) apply supportive methods like in [1] uses a Multiple Spline function to approximate the plus function for Smooth SVM. Generally, the integral sigmoid function is used for this purpose. Purnami, Zain and Embong prefer to use diabetes and heart dataset to verify their results. This method claims accuracy of 96.58%. On the other hand, [2] apply Principal Components for selection of optimal subset of the dataset and Mutual Information for weighing different features based on their importance. Finally, this model applies the Modified Cuckoo Search for the selection of the best parameters, with this getting an accuracy of 93.58%. Another method in this literature related to SVM is the GDA-LS-SVM, which uses the Generalized Discriminant Analysis for feature selection between the healthy persons and diabetic patients. Least Square SVM is used for classification scoring an accuracy of 82.05%.

Also, the use of the k-Nearest Neighbor is though not so prominent in the literature, but the results obtained are appreciable in [10]. This proposes a method of generalizing the kNN algorithm and is able to get an error-rate as low as 4%. While in [4], two variants of the k-Nearest Neighbor algorithm are seen. With use of the kNN for classification, this model manages to get an accuracy of 82.69%, which upsurges to 89.10% with the use of Fuzzy kNN instead of simply kNN.

The acceptance of neural networks for the same purpose is also visible. Diversified use of the Artificial Neural Networks can be seen spread all over the literature. A very complex structure of ARTMAP-IC, as modeled in [7] is capable of results are equal to or better than those of logistic regression, K nearest neighbor (KNN), the ADAP perceptron, multi surface pattern separation, CLASSIT, instance-based (IBL), and C4. Also in [8], a much simpler structure is drawn with general regression neural network (GRNN) and is examined on the Pima Indian Diabetes (PID) data set. These [7] and [8] claim an accuracy of 81% and 80.21% respectively.

Furthermore, the precision of Artificial Neural Networks is enhanced with supportive methods as in SVMs, though not with the same methods. [3] Suggests a method to improve the diagnosis accuracy of diabetes disease by combining PCA and ANFIS. The proposed system reduces the features of the diabetes dataset from 8 features to 4 features using principal component analysis and performs diagnosis of diabetes disease thorough adaptive neuro-fuzzy inference system classifier. The classification accuracy of this system was 89.47%. [5] applies a new method based on FCM and ANFIS to diagnose the diabetes diseases. The proposed approach FCM-ANFIS gets high accuracy with fewer rules. FCM-ANFIS approach has given the best results with CC = 83.85%, Se = 82.05% and Sp = 84.62% comparing to the other cases.

[9] uses genetic programming (GP) and a variation of genetic programming called GP with comparative partner selection (CPS) for diabetes detection. The system produces an individual from training data, that converts the available features to a single feature such that it has different values for healthy and patient (diabetes) data and in the next stage use test data for testing of that individual. The proposed system was able to achieve 78.5±2.2% accuracy. Similarly, [11] scores an accuracy more than 89%.

The next focus is on selection of a implementation technique to develop the system with the current problem statement. The highest bidders in the current literature survey, SVM and ANNs, are compared in the coming section.

3. Support Vector Machines Versus Neural Networks

The selection of method for implementation of a system depends on getting optimal solution along with a good precision for classification and cost effective in terms of modeling with performance in time and resources usage as well.

Support vector machines are supervised learning models that are associated learning algorithms which analyze data, recognize patterns and are used for classification as well as regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input, making it a non-probabilistic binary linear classifier, as in [14].

[15] depicts that more complicated structure of the data require more support vectors to be created in order to achieve reasonable performance. Since most real-world data is noisy and complex and many support vectors will be created. All these vectors have to be stored leading to model-sizes much higher than those required for neural networks. More support vector mean less precision and speed. In many cases, the number of support vectors is correlated to the size of the training data which often leads to prediction-speeds magnitudes lower than neural networks. If training time is not an issue but fast prediction is, a scenario very common in real-world-applications, use of neural networks is preferred.

NNs are heuristic, while SVMs are theoretically founded. A SVM is guaranteed to converge towards the best solution in the probably approximately correct sense. For example, for two linearly separable classes SVM will draw the separating hyperplane directly halfway between the nearest points of the two classes. A neural network would draw any line which separates the samples, which is correct for the training set, but might not have the best generalization properties. For SVMs, a kernel function is used whose property is that the computational complexity doesn't rise with the number of dimensions, while for NNs it obviously rises with the number of neurons.

4. Propose System:

A classifier is a detection function which classifies between two or more classes by assigning labels to the individuals. Specific to our project, classifier can be defined as -

$$F: D^8 \rightarrow R$$

Where,

1. F is the function that matches the domain inputs to the range outputs.
2. D^8 is the 8 dimensional input-domain that has following the 8 attributes that are present in Pima Indian Database -
   i. No. of times pregnant
   ii. Plasma Glucose concentration a 2 hrs oral glucose tolerance test

iii. Diastolic Blood Pressure (mmHg)
iv. Triceps Skin fold thickness (mm)
v. Hour serum insulin (mu U/ml)
vi. Body Mass Index
vii. Diabetes Pedigree Function
viii. Age

3. R is the range i.e. {Diabetic, Non-diabetic}

The proposed system will use a Neural Networks for design of a classifier for detection of diabetes. Neural Networks are popular for their dynamic nature in terms of learning, which is why they are preferably chosen for medical diagnosis. In spite of having such a great quality, the neural networks cannot predict with a remarkable accuracy. The fuzzy systems, though not as dynamic as neural networks, can work accurately owing the fact that they have rule bases that mimic that human thinking. When neural networks are made to control the rules generated by the fuzzy systems, the neuro-fuzzy systems are created. Instead of modifying the fuzzy systems as per the predictions of the neural networks, the presented approach focuses on implementing the neural networks as a fuzzy system.

This method of detection of diabetes proposes a system that will be implemented in client-server architecture. Here, the training dataset will be kept on the server, which will be used to train the neural network classifier on the mobile device. The mobile device is a feature add-on for convenience of the doctor. The relevant processing of the input data will be done on the server to spare the adverse effect on the performance of the device. The processing will include the search for k nearest neighbors using k-NN algorithm and fuzzy allotment of class for the input. The neural networks will be made to do this implementation.

The accuracy of this propose method is calculated to be 72.8281% for 10 fold CV in WEKA classifier. The accuracy surges to 100% when all the attributes are known and the training set is used as the test set. Owing to erroneous dataset of the UCI Pima Indians Diabetic Dataset, removal of records with missing values is considered. With removal of various attributes that are less significant in the context, the accuracy of classification ranges between 75% and 100%.

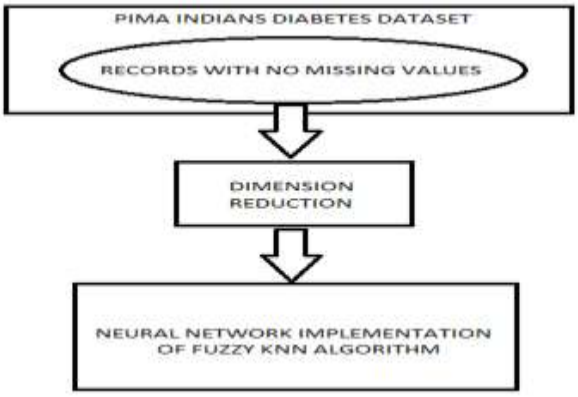

Fig. 1 Block Diagram of the Proposed Model

First of all, the proposed model eliminates the records containing the missing values from the Pima Indian Diabetes Dataset. Based from the literature survey [10], it is dimension of the dataset are reduced to half of the previous. The Fuzzy k-Nearest Neighbor algorithm is used to train the Neural Networks. Finally, the entire training set is used as test set to calculate the classification accuracy.

5. Analysis

WEKA (Waikato Environment for Knowledge Analysis) is freeware made available by the University of Waikato [16]. A newer remodeled version of WEKA with Fuzzy and Genetic Programming Classifiers [12] is used to pre-analyze the results of part of the proposed model. The following results are tabulated from various runs of the FuzzyNN algorithm for different tests.

Table 1. Use of 768-records of PID

TEST SET	ACCURACY
Training set	100%
10 fold CV	73.0469%

Table 2. Use of 768-records of PID and removal of four lesser significant attributes*

TEST SET	ACCURACY
Training set	97.3958%
10 fold CV	73.8281%

Table 3. Use of 392-records of PID (on removal of records with missing values)

TEST SET	ACCURACY
Training set	100%
10 fold CV	74.2347%

Table 4. Use of 392-records of PID and removal of four lesser significant attributes*

TEST SET	ACCURACY
Training set	100%
10 fold CV	72.9592%

*As reviewed from [13]

5. Conclusion

From these experiments, it can be concluded that, results of the proposed system are expected to perform better than those in the current literature survey. Though, the expected results shall be comparable with those of the SVMs, but with lesser complexity of the propose system, and minimum compromise with the quality, such a trade-off is acceptable.

6. References

[1] Santi Wulan Purnami, Jasni Mohamad Zain and Abdullah Embong. Data mining techniques for medical diagnosis using a new smooth SVM. Communications in Computer and Information Science, Volume 88, Part 1, 15-27. 2010.

[2] Davar Giveki, Hamid Salimi, GholamReza Bahmanyar, Younes Khademian. Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search. 2012.

[3] Kemal Polat, Salih Gunes. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Journal of Digital Signal Processing, Volume 17 Issue 4, 702-710. 2007.

[4] Mohamed Amine Chikh, Meryem Saidi, Nesma Settouti. Diagnosis of Diabetes Diseases Using an Artificial Immune Recognition System2 (AIRS2) with Fuzzy K-nearest Neighbor. Springer Journal of Medical Systems. 2011.

[5] Nesma Settouti, Meryem Saidi, and Mohamed Amine Chikh. Interpretable Classifier of Diabetes Disease. International Journal of Computer Theory and Engineering vol. 4, no. 3, pp. 438-442. 2012.

[6] Kemal Polat, Salih Gunes, Ahmet Arslan. A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine. Journal of Expert Systems with Applications: An International Journal, Volume 34, Issue 1, 482-487. 2008.

[7] Carpenter, G.A., Markuzon, N. ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases. 1998.

[8] K. Kayaer, T. Yildirim. Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks. International conference on artificial neural networks and neural information processing,181-184.

[9] M.W. Aslam, A.K. Nandi. Detection of diabetes using genetic programming. 18th European Signal Processing Conference. 2010.

[10] James C. Bezdek, Siew K. Chuah. Generalized *k*-nearest neighbor rules. Fuzzy Sets and Systems. Elsevier Volume 18, Issue 3, 237–256. 1986.

[11] M. A. Pradhan, Abdul Rahman, PushkarAcharya,RavindraGawade,AshishPateria. Design of Classifier for Detection of Diabetes using Genetic Programming. International Conference on Computer Science and Information Technology (ICCSIT'2011) Pattaya. 2011.

[12] Richard Jenson, Marcin Szczuka. Tutorial. Thirteenth International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. 2011

[13] Taufik Djatna, Yasuhiko Morimoto. Attribute Selection for Numerical Databases that Contain Correlations. International Journal of Software and Informatics, Vol.2, No.2, 125-139. 2008.

[14] Wikipedia- http://en.wikipedia.org/wiki/Support_vector_machine.

[15] Indiji Software Projects- http://indiji.com/svm-vs-nn.html

[16] WEKA- www.cs.waikato.ac.nz/ml/weka/