



International Journal of Engineering, Business and Enterprise Applications (IJEBA)

www.iasir.net

Design of an efficient data classification method for disease diagnosis

Veenu Mangat and Dimple Rani
University Institute of Engineering and Technology
Panjab University
Chandigarh, India

Abstract

Early diagnosis of any disease with less cost is always Preferable. Diabetes is one of such disease. It has become the 4th leading cause of death in developed countries and is also reaching epidemic proportions in many developing and newly industrialized nations. Due to its importance the need for detection of this disease is increasing. For this we present an efficient method for classification of patients for diabetes using optimized soft computing techniques.

Keywords: classification, diabetes detection, fuzzy knn, k nearest neighbor, neural networks, support vector machine.

I. Introduction

Diabetes is a major health problem not only in industrial but developing Countries as well and its incidence is inclining. It is a condition in which the body unable to produce or properly use the hormone called insulin that “unlocks” the cells of the body and allows glucose to enter and fuel them. Diabetes leads to increase in the risks of developing kidney disease, blindness, nerve damage, blood vessel damage and heart disease also. The exact cause of diabetes continues to be a mystery, although both genetics and environmental factors such as obesity and lack of exercise are considered to be key factors. There are two kinds of diabetes, type 1 diabetes and type 2. The most common form of diabetes is Type 2 diabetes. This type diabetes is a result of insulin resistance. However, in Type 2 diabetes, either the body does not produce sufficient insulin or the cells ignore the insulin. There are many factors which need to be analyzed to diagnose the diabetes of a patient, and this makes the physician’s job difficult. Without doubt, evaluation of data taken from patient and decisions of experts are the most important factors in diagnosis. But, this is not easy as number of factors has to be evaluated. To help the experts and minimize possible errors that can occur due to fatigued or inexperienced, classification systems provide medical data to be examined in shorter time and are more detailed. Expert systems and different artificial Intelligence techniques for classification systems in medical diagnosis are increasing slightly. As for other clinical diagnosis problems, classification Systems has been used for diabetes diagnosis problem.

II. Previous studies

Carpenter and Markuzon (1998)

This paper implemented an instance counting algorithm ARTMAP-IC (Adaptive resonance theory match tracking algorithm) and obtained an accuracy of 81% to test set. Conventional (one training and one test) validation method has been used by them. ARTMAP-IC modifies the ARTMAP search algorithm to allow the network to encode inconsistent cases, and it combines both instance counting during training with distributed category representation during testing to obtain probabilistic predictions, even with fast learning and only one training epoch. The ARTMAP algorithm, control predictive errors. A new version facilitates prediction with sparse or inconsistent data. When original match tracking algorithm is compared with the new algorithm, it provides better approximates the real-time network differential equations and also reduces memory without any loss of performance. Predictive accuracy is estimated by simulations on four medical databases: Pima Indian diabetes, breast cancer, heart disease, and gall bladder removal. ARTMAP-IC results are similar to or better than those of logistic regression, K nearest neighbor (KNN), the ADAP preceptron, multisurface pattern separation, CLASSIT, instance-based (IBL), and C4.

Polat and Gunes (2007)

This paper discussing using principal component analysis (PCA) and adaptive neuro-fuzzy inference system (ANFIS). The aim of this study is to bring an improvement in the diagnostic accuracy of diabetes disease combining

PCA and ANFIS. The proposed system has two stages. In the first stage, dimension of diabetes disease dataset that has 8 features, is minimized to 4 features with the usage of principal component analysis. In the second stage, diagnosis of diabetes disease is carried out via adaptive neuro-fuzzy inference system classifier. The dataset used in our study is taken from the UCI (from Department of Information and Computer Science, University of California) Machine Learning Database. The obtained classification accuracy of our system was 89.47% and it was very promising as compared to the other classification applications.

Yue Huang (2007)

This paper discussing feature selection technique, feature selection via supervised model construction (FSSMC), was used to identify the important attributes affecting diabetic control. After selection of suitable features, three complementary classification techniques (Naïve Bayes, IB1 and C4.5) were applied to the data in order to predict how well the condition of patient was controlled. FSSMC identified patients' 'age', 'diagnosis duration', the need for 'insulin treatment', 'random blood glucose' measurement and 'diet treatment' as the most important factors which influence blood glucose control. With the usage of this technique, the best predictive accuracy of 95% and sensitivity of 98% was achieved. The factors, like 'type of care' delivered, the use of 'home monitoring', and the importance of 'smoking' are not so much important in diabetes control. The more important factors that are identified include: 'age of patients', 'diagnosis duration' and 'family history', which are beyond the control of physicians.

Kemal Polat (2008)

This paper we discussing Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM) for diagnosis of diabetes disease. Also, we proposed a new cascade learning system based on Generalized Discriminant Analysis and Least Square Support Vector Machine. The proposed system includes two stages. The first stage, we have used Generalized Discriminant Analysis to discriminant feature variables between healthy and patient (diabetes) data as pre-processing step. The second stage, we have used LS-SVM for classification of diabetes dataset. While LS-SVM obtained 78.21% classification accuracy using 10-fold cross validation, the proposed system called GDA-LS-SVM obtained 82.05% classification accuracy using 10-fold cross validation and it is very promising compared to the previously reported classification techniques.

Hasan Temurtas (2009)

In this study, a multilayer neural network structure, trained by Levenberg–Marquardt (LM) algorithm and a probabilistic neural network structure were used. The results of the study were compared with the results of the previous studies that also focused on diabetes disease diagnosis and by using the same UCI machine learning database obtains 79.62% accuracy. The classification accuracy of MLNN with LM obtained by this study using correct training was comparatively bit better than those obtained by other studies except the classification accuracies by Polat and Gunes.

Santi Wulan (2009)

This paper implemented MKS-SSVM technique to improve accuracy of the result has been developed by many researchers. It is called Multiple Knot Spline SSVM (MKS-SSVM). Implement an experiment on Pima Indian diabetes dataset to evaluate the effectiveness of our method. The accuracy of previous results of this data is still below 80% using SSVM that is smooth support vector machine. Then, the proposed MKS-SSVM showed better performance in classifying diabetes disease diagnosis with accuracy of 93.2% which is better than previous reported results.

Santi Wulan Purnami (2010)

This paper presents a novel method for diabetes disease diagnosis using modified spline smooth support vector machine (MS-SSVM) to obtain optimal accuracy results, firstly uniform Design method was used for selection of most relevant features. The performance of this method is evaluated using 10-fold cross validation accuracy, confusion matrix. The obtained classification accuracy using 10-fold cross validation is 96.58% in comparison with other spline SSVM technique. The results of this study showed that the modified spline SSVM was effective to detect diabetes disease diagnosis and this is very promising result compared to the previously reported results.

Muhammad Waqar Aslam (2010)

This paper implemented genetic programming (GP) and a variation of genetic programming called GP with comparative partner selection (CPS) for detection of diabetes. The proposed system includes two stages. In first

stage, genetic programming is used to produce 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. In the second stage, test data is used for testing of that individual features. The proposed system was able to achieve $78.5 \pm 2.2\%$ accuracy. The results showed that GP based classifier perform better in the diagnosis of diabetes disease.

Nahla H. Barakat (2010)

In this paper, support vector machines (SVMs) are proposed for the diagnosis of diabetes. In this, we used an explanation module, which is called as "black box" model of an SVM which we used for diagnostic (classification) decision. Results obtained on diabetes dataset by using "black box" shows that it is a promising tool that is provided by intelligible SVM's for the prediction of diabetes, with prediction accuracy of 94%, sensitivity of 93%, and specificity of 94%. Future work is to conduct a prospective study to further refine the predictive results obtained by the proposed rules.

Meryem SAIDI (2011)

The use of expert systems and artificial intelligence techniques in disease diagnosis is increasing slightly. Artificial Immune Recognition System (AIRS) is one of the methods that are used in medical classification problems. AIRS2 is a modified version of the AIRS algorithm. In this paper, a modified AIRS2 called MAIRS2 was used where we replaced the K- nearest neighbor's algorithm with the fuzzy K-nearest neighbors to improve the diagnostic accuracy of diabetes diseases. The diabetes disease dataset which we are using are taken from UCI machine learning repository. The highest classification accuracy was obtained by applying the AIRS2 and MAIRS2 using 10-fold cross-validation, which was 82.69% and 89.10% respectively. MAIRS2 was applied to recognize diabetes disease. From this we observe that **an increase in the number of memory cells led to a better recognition rate** but side by side, it diminishes the degree of data reduction. In future we hope that the use of others classification techniques and affinity measure methods such as Manhattan distance increase the recognition rate of the diabetes disease.

Davar Giveki (2012)

This paper presents a novel automatic approach to diagnose Diabetes disease based on Feature Weighted Support Vector Machines (FW-SVMs) and Modified Cuckoo Search (MCS). The model consists of three stages: Firstly, there is application of PCA to select an optimal subset of features out of set of all the features. Secondly, Best feature weights are estimated using Mutual Information (MI) based on their degree of importance. Finally, MCS is applied to select the best parameter values from all parameters. An accuracy of 93.58% is obtained by the proposed MI-MCS-FWSVM method on UCI dataset. Moreover, Modified Cuckoo Search is used for speeds up the convergence of the algorithm and also to find the optimal values for parameters of SVM. This method can be combined with medical software's to assist physicians so that they can take make more accurate decisions about Diabetes disease.

Nesma Settouti (2012)

This paper discussing FCM and ANFIS to diagnose the diabetes diseases by using a reduced number of fuzzy rules with relatively small number of linguistic labels, removing the similarity of the membership functions, preserving the meaning of the linguistic labels, and in same time improving the classification performances. Experimental results show that the proposed approach FCM-ANFIS can get high accuracy with fewer rules. On the other hand, by using ANFIS more rules are needed to get a lower accuracy. Moreover, the features projected partition in ANFIS is ambiguous and cannot preserve the meaning of the linguistic labels. The best number of the rules is a trade-off between the accuracy and the rules number, also with a minimum of clusters ($c=2$) and just two fuzzy 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.

III. Proposed Data Classification Method

- We want to apply Bacterial Forging Optimization to classify Diabetes Clinical data and predict the likelihood of a patient being affected with Diabetes or not.
- To improve the accuracy, for lower classification error rate and to make the classification technique more efficient.

A. Algorithm

Bacterial forging optimization: Bacterial foraging optimization algorithm (BFOA) has been widely accepted as a global optimization algorithm of current interest for distributed optimization and control. BFOA is encouraged by

the social foraging behavior of *Escherichia coli*. BFOA has already drawn the attention of researchers because of its efficiency in solving real-world optimization problems rising in various application domains. The foraging strategy of *E. coli* bacteria can be explained by four processes namely Chemotaxis, Swarming, Reproduction, Elimination and Dispersal.

Chemotaxis: This process simulates the movement of an *E.coli* cell through swimming and tumbling via flagella. Biologically an *E.coli* bacterium can move in two different ways. It can swim for a period of time in the same direction or it may tumble, and alternate between these two modes of operation for the entire lifetime.

Swarming: For the bacteria to reach at the richest food location (i.e. for the algorithm to converge at the solution point), it is desired that the optimum bacterium till a point of time in the search period should try to attract other bacteria so that together they converge at the solution point more rapidly. To achieve this, a penalty function based upon the relative distances of each bacterium from the fittest bacterium till that search duration, is added to the original cost function. Finally, when all the bacteria have merged into the solution point this penalty function becomes zero. The effect of Swarming is to make the bacteria congregate into groups and move as concentric patterns with high bacterial density.

Reproduction: The original set of bacteria, after getting evolved through several chemotactic stages reach the reproduction stage. Here, the best set of bacteria (chosen out of all the chemotactic stages) gets divided into two groups. The healthier half replaces the other half of bacteria, which gets eliminated, owing to their poorer foraging abilities. This makes the population of bacteria constant in the evolution process. The survival and elimination behavior of any bacterium is better known as its ‘motile behavior’.

Elimination and Dispersal: In the evolution process a sudden unforeseen event can occur, which may drastically alter the smooth process of evolution and cause the elimination of the set of bacteria and/or disperse them to a new environment. Most ironically, instead of disturbing the usual chemotactic growth of the set of bacteria, this unknown event may place a newer set of bacteria nearer to the food location. From a broad perspective, elimination and dispersal are parts of the population-level long-distance motile behavior. In its application to optimization it helps in reducing the behavior of stagnation.

Diabetes Disease Dataset

In this study, we want to use the UCI Diabetes diseases dataset. This dataset contains 768 samples, where each sample has 8 features which are eight clinical findings:

- 1) *Number of times pregnant*
- 2) *Plasma glucose concentration a 2 h in an oral glucose tolerance test*
- 3) *Diastolic blood pressure (mm Hg)*
- 4) *Triceps skin fold thickness (mm)*
- 5) *2-hour serum insulin (mu U/ml)*
- 6) *Body mass index (kg/m²)*
- 7) *Diabetes pedigree function*
- 8) *Age (years)*

IV. Conclusion

We have presented an extensive review of data classification methods over the diabetes dataset using different data mining classification technique such as neural network structures, support vector machine, fuzzy set theory genetic programming, Artificial immune recognition system, PCA, MS-SSVM. Like by using AIRS2 an increase in the number of memory cells led to a better recognition rate but side by side, it diminishes the degree of data reduction using Modified Cuckoo Search is used for speeds up the convergence of the algorithm and also to find the optimal values for parameters of SVM. This method can be combined with medical software’s to assist physicians so that they can take make more accurate decisions about Diabetes disease. All these methods have some shortcomings. So we have proposed to use a novel Bacterial Forging Optimization applied to data classification to overcome these shortcomings.

References

- [1]. Carpenter, G. A., & Markuzon, N. ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases. *Neural Networks*, 11, 323–336.1998.

- [2] Polat, K., & Gunes, S. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing*, 17(4), 702–710.2007.
- [3] Yue Huang, Paul McCullagh, Norman Black, Roy Harper. “Feature selection and classification model construction on type 2 diabetic patients” data Volume 41 Issue 3, Pages 251-262.november 2007.
- [4] Polat, K., Gunes, S., & Aslan, A. A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. *Expert Systems with Applications*, 34(1), 214–221.2008.
- [5] T. Hasan, Y. Nejat, T. Feyzullah, A comparative study on diabetes disease diagnosis using neural networks, *Expert Systems with Applications*, vol. 36, pp. 8610- 8615,May 2009.
- [6] Santi Wulan Purnami, 1Abdullah Embong, Jasni Mohd Zain and S.P. Rahayu, A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis, *Journal of Computer Science* 5 (12): 1006-1011, 2009 ISSN 1549-3636,2009..
- [7] 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.
- [8] M.W. Aslam, A.K. Nandi. Detection of diabetes using genetic programming. 18th European Signal Processing Conference. 2010.
- [9] Nahla H. Barakat, Andrew P. Bradley, Mohamed Nabil H. Barakat, *Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus*, IEEE transaction on Information technology in bioinformatics, VOL. 14, NO. 4, JULY 2010.
- [10] 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.
- [11] 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
- [12] 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.
- [13] Nikhil Kushwaha, Vimal Singh Bisht and Gautam Shah. Genetic Algorithm based Bacterial Foraging Approach for Optimization. *IJCA Proceedings on National Conference on Future Aspects of Artificial intelligence in Industrial Automation NCFEAAIIA* (2):11-14.Published by Foundation of Computer Science, New York, USA, May 2012.
- [14] <http://archive.ics.uci.edu/ml/datasets/Diabetes>
- [15] http://en.wikipedia.org/wiki/Support_vector_machine
- [16] Kayaer, K., & Yıldırım, T. Medical diagnosis on Pima Indian diabetes using general regression neural networks. In *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)* (pp. 181–184). 2003.