Volume 5, Issue 2, February 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review of Softcomputing Methods for Diabetes

Shribangi Mukherjee[**] Mythili Thirugnanam[##] Mangayarkarasi R[*] Tamizharasi T[#]
[*]Programmer Analyst Trainee (PAT) , Cognizant Tecnology Solution, Chennai
[##]Associate Professor , School of Computing Science and Engineering , VIT University ,Vellore, Tamil Nadu -632014
[*]Assistant Professor (SG),School of Information Technology Engineering, VIT University, Vellore, Tamil Nadu -632014
[#] Assistant Professor, School of Computing Science and Engineering, VIT University, Vellore, Tamil Nadu -632014

Abstract — Diabetes mellitus, commonly known as diabetes is a silently killing disease which comprises of a group of metabolic diseases where a person has high blood sugar levels. The reason for the elevated sugar level is because the pancreas does not produce enough amount of insulin, or because the body cells do not respond to the insulin that is produced. It is one of the most common diseases found in the world today. An early prognosis of diabetes can help in making changes in the lifestyle of high risk patients and thus reduce complications. To achieve this, research has been carried out since a long time to find out various diagnosis techniques that give us proper prediction results. This work aims to compare the accuracies of diabetes diagnosis using techniques of support vector machines, decision trees, and logistic regression to find the method that produces a more efficient prediction rate of the disease.

Keywords— Data mining, Support vector machines, decision trees, logistic regression, machine learning, diabetes

I. INTRODUCTION

Diabetes is a major health problem in this world. It can be of three forms: Type 1 - which results from the body's failure to produce insulin, and thus requires to inject insulin or wear an insulin pump into their system; Type 2 - which is caused by insulin resistance or a failure to use insulin properly by the body cells, also known as non insulin-dependent diabetes mellitus (NIDDM); The third form is called gestational diabetes which occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level. Diabetes increases risks of developing diseases in kidney, blindness, nerve damage ,blood vessel damage and it also contributes to heart disease. The most common form of diabetes is the Type 2 diabetes. There are many factors to be analyzed in order to diagnose a person for diabetes. Identifying such parameters which help in producing an efficient prediction result has been disscussed by Mythili et al. (2013) in their proposed methodology where they considered a set of sixteen parameters and applied linear and non-linear regression techniques respectively. But the resulting sets of influential factors were different for the two individual methods. There is a need to find common factors that help in diabetes diagnosis [1]. Applying machine learning techniques like that of support vector machines, as used by Wei Yu et al. (2010) in their proposed work where they use an web based tool called diabetes classifier and applies two different diagnosis schemes, to produce two different results accordingly. A need of better diagnosis technique is required to conclude with a concrete diagnosis result [2].

II. MATERIALS AND METHODS

2.1 Data source

The data sets used in this study is the Pima-Indians-Diabetes dataset which has been taken from the UCI Machine Learning repository for performing the prediction analysis. The dataset has 8 attributes and 768 instances, and has the following attributes (all numeric values):

1. Number of times pregnant,
2. Plasma glucose concentration a 2 hours 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 [BMI] (weight in kg/(height in m)^2),
7. Diabetes pedigree function,
8. Age (years),
9. Class variable (0 or 1)

Where the class distribution is given as:
Class 1: normal (500 instances)
Class 2: Pima Indian diabetes (268 instances).

III. RELATED WORKS

Hasan Temurtas et al. (2009) presented a comparative study on pimadiabetes disease diagnosis which included the use of two methods - a multilayer neural network structure, trained by Levenberg–Marquardt (LM) algorithm and - a probabilistic neural network structure which was implemented on the datasets taken from the UCI machine learning

reposiratory. And then compared the results of the study using both methods and that of the results of previous studies, done with respect to validation methods, and concluded with the successful use of neural network structures in pima diabetes diagnosis and also that the results obtained in comparison of other methods had better classification accuracies. Wei Yu et al. (2010) used support vector machines modeling technique and applied on a set of data from the National Health and Nutrition Examination Survey (NHANES) and generated a SVM algorithm. They have used a web-based tool called diabetes classifier which provided efficient solutions to classify problems without any distribution and interdependency of data assumptions. And thus, proved to be a better technique in diagnosis. Sumathy et al. (2010) proposed a methodology that applies back propagation algorithm and takes inputs based on the possible symptoms at the early stage of diabetes ant also the physical conditions. The parameters are designed in a way such that it makes it easier for a person to predict by himself if one is having diabetes. Mythili T et al. (2012) proposed a novel two step approach called (FNC) of predicting diabetes where the initial stage uses fuzzy logic (F), neural network(N) and case based reasoning(C) techniques and then the final stage implements rule based algorithm to the data set values obtained from the previous stage. This approach results in more accurate prediction rates than using only the prediction stage in a system as is done in most cases. Mehdi Khashei et al. (2012) proposed a hybrid binary classification model for classifying diabetes type II based on soft computing concepts. The hybrid model worked better than other linear/non-linear and classic/intelligent, thus it can be used as an effective alternative model for medical classication and provides more accurate and improved medical diagnosis results. Asha Gowda Karegowda et al. (2011) proposed the use of neural networks as one of the data mining analytical tool that can be used to predict diabetes from the available medical data, where the application of hybrid model integrates genetic algorithm and back propagation networks (BPN) and the initialization is done by the genetic algorithm and optimize the connection weights of BPN. The hybrid of GA-BPN when provided with correct inputs gives us better categorization accuracy compared to results produced by GA-BPN individually. Pankaj Srivastava et al. (2012), proposed a diagnostic system using soft computing methods that detects the various phases of diabetes. This system is user friendly and guides a patient with strategies to maintain their blood sugar level so as to avoid acquiring diabetes. Madhavi Pradhan et al. (2012) proposed the implementation of self organizing neural networks and apply fuzzy logic which thusly gives an effective classification of a diabetic patient. Neural networks are chosen accordingly as they have a dynamic nature of learning and future application of knowledge and Fuzzy logic allows partial membership and rule base. Thus, these allow a direct mapping between human thinking and machine results hence, can be used for designing of classifier. K. Rajesh et al. (2012) applied data mining techniques on a dataset of 768 records with 8 attributes which was taken from the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases from UCI Machine Learning Repository. They applied the technique of feature selection and performed feature relevance analysis on the datasets and compared ten classification algorithms to conclude with using the C4.5 algorithm that gives 91% accuracy. A confusion matrix is then constructed and an accuracy analysis is done to conclude if a person is having diabetes or not. Mythili T et al. (2013) proposed an approach that considers a set of sixteen parameters and uses regression analysis which is a statistical technique that is used for modelling and analyzing variables, for the diagnosis of diabetes. They have applied linear and non-linear regression techniques respectively on the data to find the most influential parameters among the considered parameters and thus, help achieve better accuracy of diagnosing if a person has diabetes.

Table 1. Tabular comparison of earlier methodologies applied and their results

AUTHOR(Ref #)	**PARAMETERS/FACTORS**	**TECHNIQUES APPLIED/USED**	**ACCURACY RATE/ RESULTS**
Madhavi Pradhan et al. (2012) [8]	P, Pl, oral glucose tolerance test, Diastolic Blood Pressure, Triceps Skin fold thickness, Hour serum insulin,B, Diabetes Pedigree Function, A.	Neural Network and Fuzzy k-Nearest Neighbor Algorithm.	Performs better.
Wei Yu et al. (2010) [3]	A,G,F, race and ethnicity, weight, height, waist circumference, B, hypertension ,physical activity, smoking, alcohol use, education, household income.	Support Vector Machines.	classification scheme-I gives 83.5% classification scheme-II gives 73.2%
Mythili T et al. (2013) [2]	A, F, Medication for High blood pressure, High blood glucose during illness, Smoking or any tobacco products, Vegetable or fruit intake, Physical activity 30 minutes daily, Waist Hip ratio ,Increased urination, hunger and thirst, Poor wound healing, Lifestyle-Sedentary work , Frequent intake of non - vegetarian food ,Itching all over the body, Patients' blood glucose (GRBS) in Normal Range,B.	Linear and Non-Linear regression models.	For linear regression :- 1.Age 2.Medication for high blood pressure 3.Physical Activity 30 min daily 4. Increased Urination, Hunger, and thirst For non linear regression :- 1. Frequent intake of no vegetarian food (more than twice a week) 2.BMI-weight/height^2 3. Medication for High blood pressure 4. Age 5. Waist Hip ratio

Mythili et al. (2012) [4]	A, G, F, taking medication for high blood pressure, found to have high blood glucose in a health examination during illness, smoking or using tobacco products, amount of vegetable and fruit intake, physical activity (30 minutes daily), B, waist hip ratio, increased urination, hunger, thirst, poor wound healing, life style (labor class, sedentary work, retired persons and house wife's), gestational diabetes, frequent intake of non Vegetarian food and itching all over the body.	Neural network, fuzzy approach, case based reasoning.	Higher prediction rates than before.
Hasan Temurtas et al. (2009) [10]	P, Pl, Diastolic blood pressure, Triceps skin fold thickness, 2-h serum insulin, B, pedigree function,2-h serum insulin, A.	Multilayer neural network (trained by Levenberg–Marquardt [LM] algorithm), Probabilistic neural network.	79.62% 78.05%

*

A –Age
P – Number of times Pregnant
B- BMI
F- Family history of diabetes
G- Gender
Pl - Plasma glucose concentration a 2 hours in an oral glucose tolerance test

From the above table, it can be concluded that support vector machines and decision trees gives a better result as compared to neural Network and Fuzzy k-Nearest Neighbor Algorithm. And thus, can be used for comparison to find the best technique among them, to yield better results of diabetes diagnosis.

IV. PROPOSED FRAMEWORK

The proposed framework consists of two phases. The first phase is the testing phase where the patient data is collected first and then data pre-processing is done to screen the data and remove missing values and other anomalies to get proper set of data to test upon. WEKA tool is implemented using java to carry out the first phase. The second phase is to apply the three techniques of support vector machines, decision trees and logistic regression respectively on the testing data which is implemented using java by integrating WEKA tool.

4.1 First Phase

Data Pre-processing: Data Pre-processing is a data mining technique that includes the methods of cleaning, normalization, transformation, feature extraction and selection. It is an important step in Knowledge discovery process. In this phase, the incomplete, inconsistent and noisy data are applied with cleaning, integration, transformation, reduction, and discretization such that the missing values are filled, noisy data is smoothened, outliers are removed, and inconsistencies are resolved. The resultant product of this step gives us the training data set which is then used for the next step in the comparisons of the diagnosis of diabetes. This step is carried out using the WEKA tool implementation using java. WEKA stands for Waikato Environment for Knowledge Learning which has been developed by the University of Waikato, New Zealand. It is a software for implementing machine learning algorithms which contains preprocessing, classification, regression, clustering,association rules, and visualization. And thus, is an efficient tool for the first phase of the proposed work.

Training data: In case of Support Vector Machines, Sequential minimal optimization (SMO) algorithm is applied to train data sets. It is an an iterative algorithm for solving optimization problems where the problem is broken into a series of smallest possible sub-problems, which are then solved analytically. It involves Lagrange multipliers , because of the inequality constraint. The constraints are reduced to as below, in case of two multipliers:

$$y_1\alpha_1 + y_2\alpha_2 = k,$$

where, k is the sum over the rest of terms in the equality constraint, and is fixed for every iteration.

In case of Decision trees, the K-fold cross-validation method is used, where k equals 10 is used. This method divides the data into 10 blocks and averages the results of the blocks. And uses all the tuples for the training the data and uses any one block for testing the data. This is an efficient criterion when applying on data with multiple attributes.

In case of Logistic Regression, training data includes finding the regression coefficients after normalizing the datasets. And then applies 10 fold cross validation to split the datasets into testing data and training data for classification.

4.2 Second Phase

Applying SVM technique: Support vector machines (SVM) are supervised learning models that analyze data and recognize patterns. This method was invented by Vladimir N. Vapnik. It is a widely used classifier because of its high accuracy. A SVM classifies data by creating a hyper plane in an n-dimensional space that separates data points of one class from another. A large marginal model gives better results. It is based on mathematical functions and thus, gives good results when used on data sets with multiple attributes. The training data is mapped into kernel function which can be linear, quadratic, polynomial, radial basis function kernel. Selecting the best kernel function for the considered data set is very important. It is a relatively new approach.

4.3 Applying Decision Tree

Decision Tree is a decision making tool that is similar to that of a graph or decision model. It depicts the possible outcomes for an event and the resource costs and utility for a particular algorithm. It is one of the basic data mining techniques. Every decision tree has a root node followed by branches that lead to leaf nodes. The branching into leaf nodes is based on the splitting criterion as per the algorithm used, which also indicates the importance of each attribute. The k-fold cross validation method can be used for the split criterion in this case.

4.4 Applying Logistic Regression

Logistic Regression is a statistical classification model that predicts a response from binary predictor variables. It maps a relationship between two dependent variables and the probability scores give the value of the dependent variable. It is mainly used for prediction purposes and yields significantly good results. It involves fitting into the following form of equation:

$\gamma = \beta 0 + \beta 1\ x1 + \beta 2\ x2 + \ldots + \beta n\ xn$ -eq. 1

4.5 Comparing the results

In this step, the classification accuracies of the above mentioned techniques, for the given datasets are compared to conclude with the one which produces the best result of diabetes diagnosis.

V. RESULT

The number of times pregnant, plasma glucose concentration a 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), a 2-Hour serum insulin (mu U/ml), Body Mass Index[BMI] (weight in kg/(height in m)^2), diabetes pedigree function, age (years), Class variable (0 or 1) are the dataset with 768 instances, used to apply the mentioned techiniques upon. The classification accuracies of the above mentioned techniques are given as in Table-2, to determine the method which produces the best result.

Table 2. Comparison of accuracies of the tested methods

Techniques applied	**Accuracy rates (in %)**
Support Vector Machine	50
Decision Tree	74.87
Logistic Regression	77.99

As per the table, the diagnosis result of diabetes reported on Pima Indian diabetes disease dataset using support vector machines gives 50% classification accuracy, and using decision tree method gives us 74.87% classification accuracy, and further using logistic regression the classification accuracy is 77.99%. Thus, it was seen that logistic regression could be successfully used as the better method in the diagnosis of pima-diabetes disease, as it gives the highest classification accuracy than the other tested methods.

VI. CONCLUSION AND FUTURE ENHANCEMENT

In this proposed work, the application of support vector machines, decision tree and logistic regression has been compared for the classification of pima indian diabetes dataset, to conclude that when compared their classification accuracies, the best result is given by the method of logistic regression. This approach of comparison can be implemented for other diseases as well, applying the same methods as above. Furthermore, the same methods can applied and comparison can be done for number attributes that are not included in the present case dataset, for a larger number of instances and the result can be noted.

REFERENCES

[1] Mythili T ,Abhiram Naidu B,Nikita Padalia,Sophia Jerald (2013). Identifying Influential Parameters for Diagnosis Diabetes. *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, Issue. 2, pp.449-455.

[2] Wei Yu, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, Muin J Khoury (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics & Decision Making* , pp.253-259. .

[3] Sumathy, Mythili Thirugnanam, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar (2010). Diagnosis of Diabetes Mellitus based on Risk Factors. *International Journal of Computer Applications,* Vol. 10, No.4, pp.1-4.

[4] Mythili Thirugnanam, Dr.Praveen Kumar, S Vignesh Srivatsan, Nerlesh C R (2012). Improving the prediction rate of diabetes diagnosis using Fuzzy, Neural Network,Case Based (FNC) approach. *International Conference on Modeling Optimization and Computing, Procedia Engineering* , Vol. 38, pp- 1709 – 1718.

[5] Mehdi Khashei , Saeede Eftekhari , Jamshid Parvizian (2012). Diagnosing Diabetes Type II Using a Soft Intelligent Binary Classification Model. *Review of Bioinformatics and Biometrics (RBB)* Vol.1 Issue. 1, pp. 9-21.

[6] Asha Gowda Karegowda , A.S. Manjunath , M.A. Jayaram,(2011). Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima indians diabetes. *International Journal on Soft Computing (IJSC*), Vol.2, No.2, pp.15 -23.

[7] Pankaj Srivastava,Neeraja Sharma,Richa Singh (2012). Soft Computing Diagnostic System for Diabetes. *International Journal of Computer Applications,* Vol.47,No.18,pp.22-27.

[8] Madhavi Pradhan, Ketki Kohale, Parag Naikade, Ajinkya Pachore, Eknath Palwe (2012) . Design of Classifier for Detection of Diabetes using Neural Network and Fuzzy k-Nearest Neighbor Algorithm. *International Journal of Computational Engineering Research*) Vol. 2 Issue. 5, pp.1384-1387.

[9] K. Rajesh, V. Sangeetha, (2012). Application of Data Mining Methods and Techniques for Diabetes Diagnosis. *International Journal of Engineering and Innovative Technology (IJEIT),* Vol. 2, Issue. 3,pp.224-229.

[10] Hasan Temurtas , Nejat Yumusak , Feyzullah Temurtas.(2009). A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with Applications , pp.* 8610–8615.

