A Neural Network based Approach for the Diabetes Risk Estimation

Deepti Jain[1]
Department of Computer Science & Engineering
BUIT, Bhopal, India

Divakar singh[2]
Department of Computer Science & Engineering
BUIT, Bhopal, India

ABSTRACT
Diabetes is one of the most common anddramatically increasing metabolic diseases causes the increase in blood sugar. The patient having high blood sugar either caused by the bodyfailure to produce enough insulin (type 1) or the cells failure to respond to the produced insulin (type 2). Since the present medication cannot cure it hence the only way is to estimate the risk of diabetes for each person and take precautions according to the risk factor. This paper presents a Feed forward neural network based approach for the estimation of diabetes risk which estimates the risk factor for any person on the basis of body characteristics (like weight,Bloodpressure etc.).

Keywords: Feed forward Neural Network (FFNN), Diabetes Risk Estimation.

1. INTRODUCTION

Diabetes mellitus, or simply diabetes, is a group of metabolic diseases in which a person has high blood sugar, either because the pancreas does not produce enough insulin, or because cells do not respond to the insulin that is produced [1-2]. This high blood sugar produces the classical symptoms of polyuria (frequent urination), polydipsia (increased thirst) and polyphagia (increased hunger). The World Health Organization estimates that by 2030 there will be approximately 350 million people with type 2 diabetes. Associated with renal complications, heart disease, stroke and peripheral vascular disease, early identification of patients with undiagnosed type 2 diabetes or those at an increased risk of developing type 2 diabetes is an important challenge [1-3]. During the past two decades, many such predictionmodels have been developed.7 8 9 10 11 Threerecent reviews on this topic described existing prediction models and the predictive value of specific risk factors (such as metabolic syndrome) over a wide range of populations. Surprisingly, however, the performance of less than a quarter of the prediction models was externally validated. Because the performance of a prediction model is generally overestimated in the population in which it was developed, external validation of such models in an independent population, ideally by researchers not involved in the development of the models, is essential to broadly evaluate the performance and thus the potential utility of such models in different populations and settings.Consequently, certain prediction models to identify those at high risk of diabetes cannot be recommended when external validity of available models is unknown. Moreover, a direct comparison of the performance of the existing models in the same (external) validation cohort is essential to bridge the gap between the development of models and the conduct of studies for clinical utility [1-4].

2. LITERATURE REVIEW

Muhammad AkmalSapon et al [1] presents a study on the prediction of diabetes using different supervised learning algorithms of Artificial Neural Network. The network is trained using the data of 250 diabetes patients between 25 to 78 years old. The performance of each algorithm is further discussed through regression analysis. Akkarapol Sa-ngasoongsong et al [2] categorize their analysis into three different focuses based on the patients' healthcare costs, then examine whether more complex analytical models sing several data mining techniques in SAS® Enterprise Miner™ 7.1 can better predict and explain the causes of increasing diabetes in adult patients in each cost category. The preliminary analysis shows that high blood pressure, age, cholesterol, adult BMI, total income, sex , heart attack, marital status, dental checkup, and asthma diagnosis are among the key risk factors. Artificial Neural Networks to Detect Risk of Type 2 Diabetes is presented by B. Y. Baha et al [3] in their research, 7 risk factors and their strength of association to the development of Type 2 diabetes was used as relative weight of input variables. A multilayer feedforward architecture with backpropagation algorithm was designed using Neural Network Toolbox of Matlab. The network was trained using batch mode backpropagation with gradient descent and momentum. ManaswiniPradhan et al [5] experimented and suggested an Artificial Neural Network (ANN) based classification model as one of the powerful method in intelligent field for classifying diabetic patients into two classes. For achieving better results, genetic algorithm (GA) is used for feature selection. The GA is used for optimally finding out the number of neurons in the single hidden layered model. Further, the model is trained with Back Propagation (BP) algorithm and GA (Genetic Algorithm) and classification accuracies are compared. support vector machine modeling for prediction of common diseases is presented by Wei Yu et al [7] they used data from the 1999-2004 National Health and Nutrition Examination Survey (NHANES) to develop and validate SVM models for two classification schemes: Classification Scheme I (diagnosed or undiagnosed diabetes vs. pre-diabetes or no diabetes) and Classification Scheme II (undiagnosed diabetes or pre-diabetes vs. no diabetes). The SVM models were used to select sets of variables that would yield the best classification of individuals into these diabetes categories.

3. DATA DESCRIPTION

The data source that is used to perform data mining analysis in this study is provided by SAS in the national 2010 SAS Data Mining Shootout competition. With 50,788 records, the dataset consists of 43 variables in which 35variables are discrete variables and the other eight variables are continuous variables. This dataset is assumed to be representative of the population and used for analysis as a snapshot of the country and its health care costs at a point in time. Our first task in this study is to get a sense of the dataset for any inconsistencies, errors, or extreme values in the data. Frequency distribution, descriptive statistics, and cross-tab analysis are used in this section. Table 1 presents the list of the variables in the dataset.

Table 1. List of the variables in the dataset

S. No.	Sub Category	Variable Name
1	Physical Characteristics	Age
2		Sex
3		Rel.
4		R/U
5		Occ.
6	Past History	DM
7		HT
8		IHD
9		PTB
10		CVA
11		Other
12	Family History	HT
13		IHD
14		DM
15		CVA
16	Addiction History	Dur. (Years)
17		Amt (gm.)
18		Dur. (Years)
19		Amt (No.)
20		Dur. (Years)
21		Amt (gm.)
22	General Examination	Pulse
23		SBP
24		DBP
25		RR
26		BMI
27		W/H Ratio
28	Systematic Examination	CVS
29		RS
30		PA
31		CNS
32	Investigation	B.S. Fasting
33		Chol.
34		S.T.
35	Tread Mill Test Report	Protocol
36		Ex. T.
37		MHR
38		SBP
39		DBP
40		METS
41		Ter. Criteria
42		THR
43		Ex. Toler
44		Hr and BP
45		Angina
46		Arryth.
47		ST T.C
48		Recovery

4. FEEDFORWARD NEURAL NETWORK

The feedforward neural network begins with an input layer. The input layer may be connected to a hidden layer or directly to the output layer. There can be any number of hidden layers, as long as there is at least one hidden layer or output layer provided. In common use, most neural networks will have one hidden layer, and it is very rare for a neural network to have more than two hidden layers. The term "feedforward" indicates that the network has links that extend in only one direction. Except during training, there are no backward links in a feedforward network; all links proceed from input nodes toward output nodes [11].

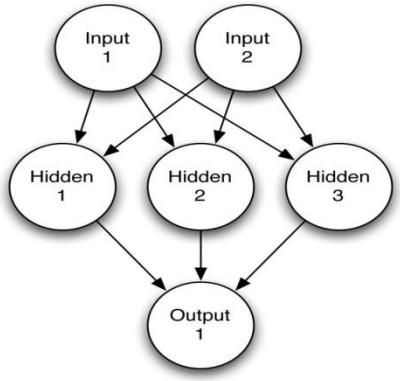

Figure 1: Feedforward Neural Network.

5. PROPOSED ALGORITHM

The algorithm can be described in detail by following steps:

Step 1: Read the dataset.

Step 2: Preprocess the dataset and remove unwanted symbols.

Step 3: Filter the dataset and take only selected features.

Step 4: Normalize the each parameter by detecting its maximum and minimum values according to the following formula

$$V_{norm} = \frac{V - V_{min}}{V_{max} - V_{min}}$$

$Where$:
$V = Designates\ the\ actual\ value\ of\ parameter.$
$V_{min} = Designates\ the\ minimum\ value\ of\ parameter\ from\ all\ senarios.$
$V_{max} = Designates\ the\ maxiimum\ value\ of\ parameter\ from\ all\ senarios.$
$V_{norm} = Designates\ the\ normalized\ value\ of\ parameter\ from\ all\ senarios.$

Step 5: The normalized values set are arranged in an array to represent system condition by a vector this vector can be represented by

$$Trn_{vect} = [V_{norm\,1}, V_{norm\,2}, V_{norm\,3}, \ldots . V_{normn}]$$

Hence the system states can be treated as n dimension vector.

Step 6: Group the dataset vectors according to the attack they belong to.

Step 7: Configure the neural network by selecting proper number of layers and neurons.

Step 8: Now the vectors with their classification group are used to train the Feedforward Neural Network (FFNN).

Step 9: Ones FFNN got trained it can now be used as anRisk detector.

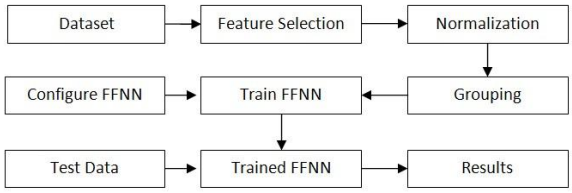

Figure 2: Block Diagram of the proposed system.

6. SIMULATION RESULTS

The proposed model is simulated using matlab neural network toolbox for different size of dataset. Finally the performance of the proposed algorithm is measured in terms of following parameters:

- True positive (TP): The individual has the condition and tests positive for the condition.
- True negative (TN): The individual does not have the condition and tests negative for the condition.
- False positive (FP): The individual does not have the condition but tests positive for the condition.
- False negative (FN): The individual has the condition but tests negative for the condition.
- Accuracy: the accuracy is the proportion of true results (both true positives and true negatives) in the population. It is a parameter of the test.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: precision or positive predictive value is defined as the proportion of the true positives against all the positive results (both true positives and false positives).

$$Precision = \frac{TP}{TP + FP}$$

- Recall: Recall in this context is also referred to as the True Positive Rate or Sensitivity.

$$Precision = \frac{TP}{TP + FN}$$

- F-Measure: A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Table 2: Performance of Proposed based System for type 1.

Dataset	TPR	TNR	FPR	FNR	Acc.	Prec.	Recall	F-meas
1000	0.9021	0.7934	0.2066	0.0979	0.8050	0.3417	0.9021	0.4956
2000	0.7832	0.9061	0.0939	0.2168	0.8102	0.9674	0.7832	0.8656
3000	0.6837	0.9885	0.0115	0.3163	0.9540	0.8836	0.6837	0.7709
4000	0.7846	0.9035	0.0965	0.2154	0.8259	0.8914	0.7846	0.8156
5000	0.9021	0.7934	0.2066	0.0979	0.9167	0.7167	0.0833	0.0833

Table 3: Performance of Proposed based System for type 2.

Dataset	TPR	TNR	FPR	FNR	Acc.	Prec.	Recall	F-meas
1000	0.9680	0.9926	0.0074	0.032	0.9900	0.9397	0.9680	0.9536
2000	0.9931	0.9754	0.0246	0.0069	0.9892	0.9931	0.9931	0.9931
3000	0.9594	0.9984	0.0016	0.0406	0.9940	0.9873	0.9594	0.9731
4000	0.9866	0.9798	0.0202	0.0134	0.9898	0.9867	0.9866	0.9866
5000	0.9735	0.9888	0.0112	0.0265	0.9911	0.9734	0.9735	0.9733

7. CONCLUSION

The model of the diabetes risk estimator is presented in this paper. The Detection accuracy of the system is up to 90% which is excellent also the algorithm have very low FPR (max 8.3%) hence reduces the chances of false alarming. Further it could achieve much better performance by increasing the number of samples taken and increasing the number of characteristics parameter selected.

8. REFERENCES

[1] Muhammad AkmalSapon, Khadijah Ismail and SuehazlynZainudin "Prediction of Diabetes by using Artificial Neural Network",2011 International Conference on Circuits, System and Simulation IPCSIT vol.7 (2011)

[2] Akkarapol Sa-ngasoongsong and JongsawasChongwatpol "An Analysis of Diabetes Risk Factors Using Data Mining Approach", Paper PH10-2012.

[3] B. Y. Baha, Bank, Yola and G. M. Wajiga" Artificial Neural Networks to Detect Risk Of Type 2 Diabetes", JORIND 10 (2), June, 2012.

[4] ZaritaZainuddin, Ong Pauline and CemalArdil "A Neural Network Approach in Predicting the Blood Glucose Level for Diabetic Patients", International Journal of Information and Mathematical Sciences 5:1 2009.

[5]ManaswiniPradhan, Dr. Ranjit Kumar Sahu "Predict the onset of diabetes disease using Artificial Neural Network (ANN)", International Journal of Computer Science & Emerging Technologies Volume 2, Issue 2, April 2011.

[6] DavarGiveki, Hamid Salimi, GholamRezaBahmanyar, YounesKhademian "Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search".

[7] Wei Yu, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, Muin J Khoury" Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes", BMC Medical Informatics and Decision Making 2010.

[8] Shoback, edited by David G. Gardner, Dolores (2011). Greenspan's basic & clinical endocrinology (9th ed.). New York: McGraw-Hill Medical. pp. Chapter 17. ISBN 0-07-162243-8

[9] Gary S Collins*, Susan Mallett, Omar Omar and Ly-Mee Yu "Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting", BMC Medicine 2011, 9:103.

[10] "Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study", BMJ 2012; 345 doi: http://dx.doi.org/10.1136/bmj.e5900