An Integrated Approach towards the prediction of Likelihood of Diabetes

Subham Khanna
M.Tech, Software Engineering
Indian Institute of Information Technology, Allahabad
Allahabad, Uttar Pradesh, India
Subhamkhanna123@gmail.com

Sonali Agarwal
Assistant Professor, Information Technology
Indian Institute of Information Technology, Allahabad
Allahabad, Uttar Pradesh, India
sonali@iiita.ac.in

***Abstract*—With the growth of Information and communication technologies, the health care industry is also producing extensively large data. For managing such large amount of data, an efficient knowledge discovery process is required. This field is developing fast and there is a big scope of early planning towards the treatment of large number of diseases. The planning can be done by developing some strategic solutions based on Data Mining for the treatment of the disease. Classification based on supervised learning is a technique of Data Mining which helps in predicting the label of unknown samples as Class. This is extremely popular technique of Data Mining by which the treatment of a disease could be planned at an early stage. Diabetes is one of the chronic diseases produces metabolism disorder in human bodies. Metabolism refers a chemical process in human body responsible for energy conversion and utilization. The diabetes with type 1 and type 2 indicates excess glucose level in the blood could be cured if regular precautions have been taken persistently under certain clinical guidelines. This paper performs classification on diabetes dataset taken from SGPGI, Lucknow (A super specialty hospital in Lucknow, Uttar Pradesh, India).It predicts an unknown class label for given set of data and helpful to find out whether the class label for the dataset under consideration would be of low risk, medium risk or high risk. The classifier is further trained on the basis of weights assigned to different attributes which are generated by means of expert guidelines. The accuracy of classifier is verified by kappa statistics and accuracy, evolution criteria for classifiers**

Keywords*—Classification, Data Mining, Diabetes, Kappa Statistics

I. Introduction

There are huge amount of medical data available such as identification of cause and nature of diseases, patient details, resources available for hospitals, availability of doctors and patients. All these data are needed to be analyzed because everybody is seeking knowledge from such vast data with reduced costs. A classifier could be applied on such data for efficient decision making which in turns contribute significant knowledge about a system. Hidden patterns present inside the data, which is required by care givers, patients, and health sector experts could be automatically derived by the application of Data Mining. It enables patients, doctors and everybody in health care industry to make better decisions regarding healthcare and treatment for any disease at early stages. The foremost task of classification is predicting the upcoming behavior of a data sample for instance it can answer whether a person would be having any chances of suffering from diabetes or not, on the basis of hidden patterns on which the classifier is trained.

A recent survey of IDF (International Diabetes Federation) shows that more than 70.3 million people in South East Asia are suffering from Diabetes and the number would be increased to 120.9 million by 2030. The impacts of the disease on adults are increasing in such a way that one out of five people is having diabetes. The following table 1and 2 reveals the situation of diabetes in the countries of South East Asia region as well as worldwide [1].

TABLE I. THE TABLE SHOWS THE TOTAL NUMBER OF CASE OF DIABETES IN DIFFERENT COUNTRIES OF SEA REGION

Serial number	Countries by diabetes cases in SEA	
	Country	*Cases (in millions)*
1	India	63.0
2	Bangladesh	5.5
3	Sri Lanka	1.1
4	Nepal	0.506
5	Mauritius	0.141
6	Bhutan	0.022
7	Maldives	0.015

TABLE II. THE TABLE SHOWS THE GLOBAL PHENOMENON ABOUT THE DIABETES DISEASE

Serial number	Global figures for diabetes, 2012 (20-79 years)	
	Global Phenomenon	*Total Count*
1	Prevalence of diabetes in adults	8.3%
2	Number of People with Diabetes	371 million
3	Number of undiagnosed Cases	187 million
4	Deaths due to diabetes	4.8 million
5	Total healthcare expenditures in USD	471.6 billion

978-0-7695-5013-8/13 $31.00 © 2013 IEEE
DOI 10.1109/ICMIRA.2013.62

The above facts and figures very clearly indicate that the problem of diabetes disease is becoming severe day by day and requires efficient strategic planning specially in Indian scenario. It can be very well achieved with the application of Data Mining. The section II of this paper presents some related works done previously in the field of Data Mining especially in diabetes and section III presents proposed approach of classification followed by results in section IV and in section V concluding remarks with its future perspective has been described.

The following table 1and 2 reveals the situation of diabetes in the countries of South East Asia region as

II. Related WoRk

Diabetes prediction using Data Mining has been explored by various researchers from time to time and developed encouraging solution for medical expertise and researchers. As a result of all these research, diagnostic and prognostic models have been developed and influenced the existing clinical practices.

In a research work, characteristics of various diagnostic models have been analyzed and some abnormal factors have been identified which may be improved for further clarity in clinical decision making, highlighted by Wyatt and Altman[2].The authors also highlighted the benefits of the Glasgow Coma Scale and specified that confidence, accuracy, effectiveness and interoperability factors, responsible for different situations, are not available up to a certain level which indicates the prime reason of un-usefulness of the approach [3].Apart from the above methodology some authors have specified that the decision tree and C4.5 are also not applicable in every case and for using them, it has to be used in a specific order[4].

The technique of Bayesian approach was also analyzed by authors to show that some amount of reverse engineering was needed to calculate the training proportion of the classification [5].Some other accurate approaches of Data Mining like Neural Network and Support Vector Machines are supposed to classify appropriately but they fall into categories of "Black Box"[6][7]. These black box techniques are not suitable since the internal details cannot be understood properly by the researchers and by analyzing the classifier none can understand the core of the problem domain. Among the many statistical approaches available, logistic regression are currently popular used in many medical applications. Although they have solid theoretical foundation, Wyatt and Altman established that one in every five statistical models, the underlying assumptions were violated the integrity of the approach [2].In sequence, with the research works in the field of diabetes classification, an expert system for predicting the diabetes disease was proposed by working of an expert system that was mechanized on chaining inference depending on backward, forward and forward-backward chaining technique [8] [9]. It suggested an uncertainty principal which calculates the probability of illness and severity of the disease as well as the potential complications of the disease. The dataset of PIMA INDIA was studied by various researchers to develop a certain classification model by using Rapid Miner tool [10]. The author also analyzed how to handle missing values. The impact of preprocessing of diabetes data in artificial neural networking based technique is also examined [11].The research in the field of diabetes has also been studied based on Association Rules [12][13].

This research takes advantage from the support system which proposed the process of knowledge extraction with the help of Data Mining [14] [15]. The idea of weighted classifier is formed the basis of this newly proposed approach based on applied weights of different attributes [16].

III. Methodology

This proposed classification method is considering the impacts of different attributes, present in dataset, for the severity of the diabetics. The method intended to find out the total score for a patient indicates various categories such as low, medium and high risk patients.

At the beginning the Data Preprocessing is performed by using binning and substitution for example non numeric attributes like male and females are converted into 0 and 1 and same pattern is followed for rest of the attributes. Some other continuous attributes are also broken into ranges with some assigned values. Removal of duplicate records also performed in preprocessing stage. The missing values in the dataset have been calculated by taking the average of rest of the values. The division of range for any particular attribute is totally influenced by expert advice.

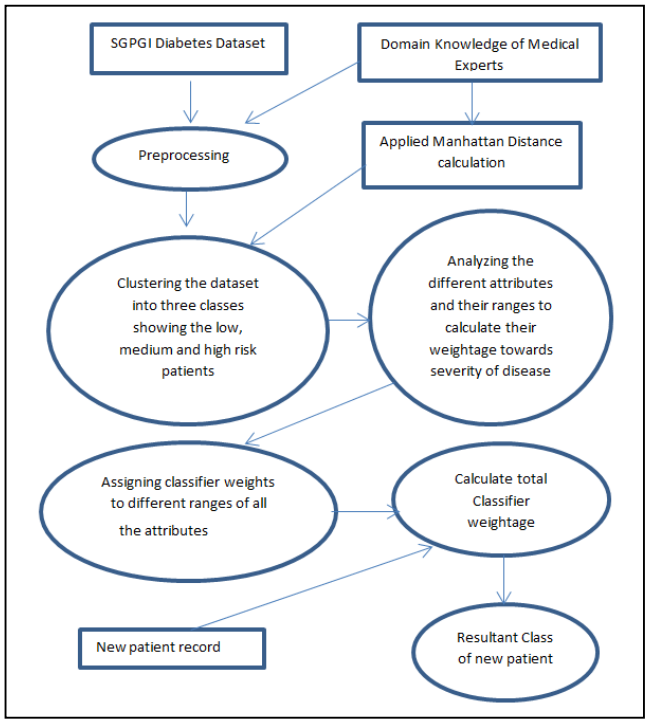

Fig. 1. Steps of Proposed Model

The next step is clustering in which the enhancement of K means algorithm is newly introduced to show the impact of expert advice in medical domain which is not very important in commercial domain. There exists a significant difference between standard algorithm for a typical commercial domain and specific algorithm of a medical domain. In commercial dataset each attribute has been considered with an equal weightage while in medical field the impact of attributes towards the formation of a disease are different from each other. So, different weightage and priorities are assigned to different attributes.

The newly proposed Clustering algorithm is applied on the dataset to divide it into three classes and the modification is achieved by distance calculation between two clusters in a novel way. Here the distance is calculated on the basis of influences, for example more influencing attribute plays more crucial role in finding the distance.

Here the modified distance is based on the Applied Manhattan Distance given by:

$$D = \sum_{i=0}^{i=n-1} Abs\,(X_{1,i} - X_{2,i})/(i+1)$$

i.e.

$$D = \frac{Abs(X_{1,1} - X_{2,1})}{(1)} + \frac{Abs(X_{1,2} - X_{2,2})}{(2)} + \frac{Abs(X_{1,3} - X_{1,3})}{(3)} + \cdots,$$

(1)

Where $(x_{1,i})$ and $(x_{2,i})$ are two patient records where (i = 1,2……n) denotes attributes (in this case n=8). Since first attribute is having more impact on disease as per expert advice, which is glucose level. In distance formula it is assigned more weight than other attribute. For example $(x_{1,1}-x_{2,1})/1$ is having more impact and$(x_{1,2}-x_{2,2})/2$ is having comparatively less impact and so on.

As a result of clustering, three classes of the patients, Low, Medium and High have been obtained.

Then in every class, the impact of attributes is analyzed to find their corresponding values responsible for the formation of the particular cluster. The analysis also calculates classifier weights for all the attributes.

The calculation of weights is done by making all other attributes constant and analyzing only one attribute at a time to calculate its effect in formation of a particular class of disease.

The steps of proposed algorithm for generating clusters and calculating the classification weights of the attributes to classify the new patients are shown below:

Step 1: Read Diabetes Dataset D.

Step 2: Prioritize the Dataset according to the expert knowledge of medical domain.

Step 3: consider the most influential attribute and identify the mean values of every range, here three ranges has been chosen..

Step 4: establish the mean values as the center of each class.

Step5: for each record in data set D

a: Calculate the distance according to the equation 1 with all the three centers.

b: Add it to the cluster with which it has Minimum distance

c:Update the center of the cluster by Averaging the values.

Step6: After the scanning of whole dataset three clusters will be generated.

Step7: for each cluster:

a: for each record:

b: Analyze the lowest priority attribute and, Fix the rest of the attributes at their Minimum range in the cluster.

c: Calculate the classifier weight of the range Of the attribute and upgrade the range.

d:If all the ranges have shown classifier Weight then repeat a-c for next higher Priority attribute.

Step8: Take the new patient record and apply the classifier weight to the ranges of value and the resultant class will be obtained.

Now the attributes are assigned classification weights to classify the unknown data sample.

The total weight for a unknown sample is calculated as

$$\propto_{1,i} + \propto_{2,i} + \propto_{3,i} + \cdots = \text{----------(2)}$$

Where $\propto_{1,1}, \propto_{1,2}, \propto_{1,3} \ldots$ are the weights of n ranges of attribute x_1 where the value of x_1 is divided into n ranges. Similarly attributes $x_2, x_3, \ldots x_n$ has been assigned different weightage and in this manner total weight has been calculated from equation 2.

It is further clarified that there are three ranges of W which are denoting low risk , medium risk and high risk patients. Thus the value of w_i shows in which class the patient will belong.

IV. Result and Discussion

This research work is based on SGPGI Diabetic Dataset. The data description with all its attributes as well as their assigned priorities is shown with the help of figure 2:

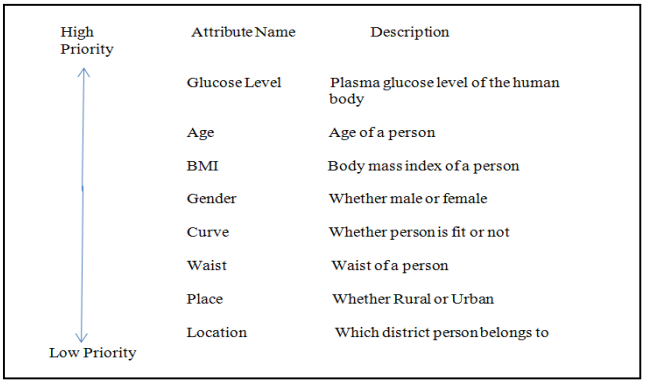

Fig. 2. Attribute details of SGPGI diabetes Dataset

The figure 2 indicates that attributes are considered in specific order based on their priority values and clustering is performed by using newly proposed applied Manhattan Distance in which distance is calculated between the cluster points to reflect the impact of different attributes as well as to increase intra class similarities and also reduce interclass similarities.

The table 3 presented below is showing the performance of Clustering where three clusters have been identified as Low Risk, Medium Risk and High Risk patients. It is further clearly noted that out of 403 records 120 low risk patients, 114 medium risk patients and 44 high risk patients are belonging to three different distinct clusters. There were duplicate records which were removed

TABLE III. PERFORMANCE OF CLUSTERING

Countries by diabetes cases in SEA			
Approach	***Low Risk Patient***	***Medium Risk Patients***	***High Risk Patients***
Applied Mean	120	114	44

The next step after generating three clusters is analyzing the clusters according to the change in the values of any attributes, which in turns calculates the classifier weights for attributes and represent particular class of disease.

After calculating the individual weights of different attributes the formula is applied to calculate the total weight for any new unknown sample. Finally this successfully classifies the patient records into low, medium or high risk categories according to equation2.

The performance evaluation of the proposed classifier has been performed on the basis of four important characteristic: accuracy, sensitivity, specificity and kappa values. These are defined in table 4:

TABLE IV. PERFORMANCE ANALYSIS OF THE CLASSIFIER

Serial No.	**Accuracy Measure of Different Algorithms**			
	Accuracy Measure / Algorithms	***Applied Weight Classifier (Proposed)***	***Decision Tree***	***Naïve Bayesian***
1	Accuracy $Accuracy = ((TP + TN)/(TP + TN +$	0.832	0.711	0.64
2	Specificity $Specificity = (TP)/(TP + FN))$	0.897	0.765	0.6375
3	Sensitivity $Sensitivity = (TN)/(TN + FP))$	0.709	0.607	1.0
4	Kappa Value $K = ((P(A) - P(E))/(1 - P(E)))$	1.003	1.026	0.5311

Here in the table TP indicated True positive value for example it includes all diabetic patients correctly diagnosed as Diabetic similarly FP represents False Positive which contains the group of healthy people those are incorrectly identified as diabetic patient. The TN represents True Negative which consist the group of Healthy People correctly recognized as Healthy and Lastly the FN stands for False Negative which holds the group of Diabetic Patients incorrectly identified as healthy.

Another Parameter Kappa statistic is also used to define the classification performance which holds following terms:

P(A) reflects the agreement percentage in between classifier and underlying truth

P(E) is the chance of agreement

The proposed methodology is alsowith already existing decision that ion tree and Naïve Bayesian Classifier which clearly indicat.es that the performance of Applied Weighted classifier is much better than the rest

v. Conclusion & Future Perspective

In this research work an efficient and accurate measure of classification for diabetes patients has been presented. The proposed applied weighted classifier utilizes a newly developed Applied Manhattan Distance formula for cluster formation which considers different attributes on the basis of their priority level. Three clusters has been identified which further represent three different class levels for the diabetic patients. Here it is also explored that different attributes have different impact in classification process and playing a crucial role for determining the performance of the classifier.

Thus the research work suggests an efficient way of diagnosis of the disease as per its severity.

The Efficiency of the result has been clearly established by calculating the accuracy, sensitivity, specificity and kappa values.

There are numerous opportunities in this domain where the diagnostic model can be applied and various chronic disease may be treated in better way.

References

[1] http://www.idf.org/sites/default/files/WP_5E_Update_Country.pdf

[2] Altman, Douglas G., et al. "Prognosis and prognostic research: validating a prognostic model." *BMJ: British Medical Journal* 338.7708 (2009): 1432-1435.

[3] Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern classification*. Wiley-interscience, 2012

[4] Wang, Lipo. *Data mining with computational intelligence*. Springer-Verlag, 2009.

[5] Gil-Jiménez, P., et al. "Shape classification algorithm using support vector machines for traffic sign recognition." Computational Intelligence and Bioinspired Systems (2005): 494-497.

[6] Thirugnanam, Mythili, et al. "Improving the Prediction Rate of Diabetes Diagnosis Using Fuzzy, Neural Network, Case Based (FNC) Approach."*Procedia Engineering* 38 (2012): 1709-1718.

[7] Guo, Yang, Guohua Bai, and Yan Hu. "Using Bayes Network for Prediction of Type-2 Diabetes." Internet Technology And Secured Transactions, 2012 International Conferece For. IEEE, 2012.

[8] Al Jarullah, Asma A. "Decision tree discovery for the diagnosis of type II diabetes." *Innovations in Information Technology (IIT), 2011 International Conference on*. IEEE, 2011.

[9] Zhou, Xuezhong, et al. "Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support." *Artificial Intelligence in Medicine* 48.2-3 (2010): 139-152.

[10] Jianchao Han; Rodriguez, J.C.; Beheshti, M.; , "Diabetes DataAnalysis and Prediction Model Discovery Using RapidMiner," Future Generation Communication and Networking, 2008. FGCN '08. Second International Conference on , vol.3, no., pp.96-99, 13-15 Dec. 2008

[11] Jayalakshmi, T.; Santhakumaran, A.; , "A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial NeuralNetworks," Data Storage and Data Engineering (DSDE), 2010International Conference on , vol., no., pp.159-163, 9-10 Feb. 2010

[12] Patil, B.M.; Joshi, R.C.; Toshniwal, D.; , "Association Rule for Classification of Type-2 Diabetic Patients," Machine Learning and Computing (ICMLC), 2010 Second International Conference on , vol., no., pp.330-334, 9-11 Feb. 2010

[13] Nuwangi, S. M., et al. "Usage of association rules and classification techniques in knowledge extraction of diabetes." Advanced Information Management and Service (IMS), 2010 6th International Conference on. IEEE, 2010.

[14] Chen, Jian-xun, Shih-Li Su, and Che-Ha Chang. "Diabetes care decision support system." Industrial and Information Systems (IIS), 2010 2nd International Conference on. Vol. 1. IEEE, 2010.

[15] Nuwangi, S. M., et al. "Utilization of Data Mining Techniques in Knowledge Extraction for Diminution of Diabetes." Information Technology for Real World Problems (VCON), 2010 Second Vaagdevi International Conference on. IEEE, 2010.

[16] Quinn, Anthony, et al. "AWSum-applying Data Mining in a health care scenario." Intelligent Sensors, Sensor Networks and Information Processing, 2008. ISSNIP 2008. International Conference on. IEEE, 2008.