International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Predicting Diabetes Disease Using Effective Classification Techniques

Vachan O,Vishwanath Bhat,Pratheek M P,Sachin M S,NagaNandini D S

Eight Semester, Dept. of Cse, The National Institute of Engineering, Mysore

---

Abstract - *Diabetes mellitus[1], known as diabetes, is a group of metabolic disorders and has affected hundreds of millions of people. The detection of polygenic disorder is of nice importance, regarding its severe complications. There have been plenty of research studies about diabetes identification, many of which are based on the Pima Indian diabetes data set[2]. It's a knowledge set finding out ladies in Pima Indian population started from 1965, wherever the onset rate for polygenic disorder is relatively high. Most of the research studies done before mainly focused on one or two particular complex technique to test the data, while a comprehensive research over many common techniques is missing*

Key Words**:** *diabetes mellitus[1],pima Indian dataset[2]*

1.Introduction

1) Diabetes mellitus has a direct signal of high blood sugar, together with some symptoms including frequent urination, increased thirst, increased hunger and weight loss. Patient of diabetes usually need constant treatment, otherwise, it will possibly lead to many dangerous life-threatening complications. The diabetes is diagnosed with the 2-hour post-load plasma glucose being at least 200mg/dL [1], and the necessity of identifying diabetes timely calls in various studies about diabetes recognition.

2)Many previous research studies have been done about machine learning in diabetes identification. Research has been done focused on diabetes identification through GDA (Generalized Discriminant Analysis) and SVM (Support Vector Machine) [2] and they obtained some inspiring results. Another research was to do the same thing by GRNN (General Regression Neural Network)[3], which also had a very high accuracy. Comparing to the previous work, we make a more comprehensive study containing a number of common techniques used to diabetes identification, intending to compare their performance and find the best one among them.

3)Through this experiment, we compare several common and data pre processors for each of the classifiers we use, and find the best pre processor respectively. Then we compare these classifiers after we modify the parameters of them to reach their approximate maximum accuracy, and we particularly analyse how to modify the parameters in DNN (Deep Neural Network). At last, we also analyse the relevance of each feature with the classification result, and this will help to modify the data set in future studies.

2.SYSTEM ANALYSIS

2.1 Literature Survey

2.1.1 Research Papers on Literature Survey

1. Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases

Desc: This paper presents the summary of machine learning techniques in classification of polygenic disorder and vessel diseases (CVD) , Artificial Neural Networks (ANNs) and Bayesian Networks (BNs).The comparative analysis was performed on chosen papers that area unit printed within the amount from 2008 to 2017.The most usually used form of ANN in chosen papers is multilayer feedforward neural network with Levenberg-Marquardt learning rule. On the other hand, the most commonly used type of BN is Naive Bayesian network which shown the highest accuracy values for classification of diabetes and CVD, 99.51% and 97.92% respectively.More over, the calculation of mean accuracy of discovered networks has shown higher results ANN, that indicates that higher chance to get additional correct leads to polygenic disorder and/or CVD

Classification is when it is applied to ANN.

Drawbacks:

- ✧ System used for classification of diabetes and cardiovascular diseases.
- ✧ Less efficient

2. Performance Analysis of Classification Approaches for the Prediction of Type II Diabetes.

Desc: Medical decision making is characterized by an exponential evolution of knowledge. With the increasing trend of healthcare applications in medical domain, disease prediction has become the center of research. In an actual risk assessment process, the discovery of a disease prediction model is essential for patients and physicians. To estimate these risks, enormous classification and prediction algorithms have been developed in the field of data mining (DM). Recently, the World Health Organization has reported type II diabetes as the major cause of complications such as blindness, amputation and kidney failure. So, this paper has aimed to compare the performance of five classification approaches namely ant-miner, CN2, RBF network, boost and Bagging for the prediction of diabetes mellitus.

These classification approaches have been tested with three sets of type II diabetes datasets [PIMA, US, AIM'94] obtained from the UCI machine learning repository in terms of sensitivity, specificity, F-score, accuracy, chance agreement and kappa. The results indicate that ant-miner algorithm has achieved the highest kappa value of 0.982 which indicates a perfect level of agreement between the medical expert's opinion and the corresponding

classification approach.

Drawbacks:

- ✧ Prediction of only Type 2 diabetes.
- ✧ Time Consuming.

3. Association Rule Extraction from Medical Transcripts of Diabetic Patients.

Desc: Medical databases serve as rich knowledge sources for effective medical diagnosis. Recent advances in medical technology and intensive usage of electronic medical history systems, helps in massive production of medical text data in hospitals and other health institutions. Most of this text data that contain valuable information are just filed and not utilized to the full extent. Proper usage of medical info will create tremendous changes in medical field. This paper gift a brand new methodology of uncovering valid association rules from medical transcripts. The extracted rules describes Association of disease with other diseases, symptoms of a particular disease, medications used for treating diseases, the most prominent age group of patients for developing a particular disease. NLP (Natural Language Processing) tools were combined with data mining algorithms (algorithm and FP-Growth algorithm) for the extraction of rules. Interesting rules were elect the correlation live, lift.

Drawbacks:

- ✧ Time Consuming.
- ✧ Less Efficient.

4. A Comprehensive Exploration to the Machine Learning Techniques for Diabetes Identification.

Desc: during this paper, we have a tendency to build a comprehensive exploration to the foremost in style techniques (e.g. DNN (Deep Neural Network), SVM (Support Vector Machine), etc.) used to identify diabetes and data preprocessing methods. Basically, we have a tendency to examine these techniques by the accuracy of cross-validation on the Pima Indian information set. We compare the accuracy of every classifier over many ways that of information preprocessors and that we modify the parameters to boost their accuracy. The best technique we discover has seventy seven.86% accuracy mistreatment 10-fold cross validation we have a tendency to conjointly analyze the relevancy between every feature with the classification result.

Drawbacks:

- ✧ Less Accuracy.

5. Clustering Medical Data to Predict the Likelihood of Diseases

Desc: Several studies show that information of a site will improve the results of bunch algorithms. In this paper, we have a tendency to illustrate the way to use the information of medical domain in bunch method to predict the chance of diseases. To find the likelihood of diseases, clustering has to

be done based on anticipated likelihood attributes with core attributes of disease in data point. To find the likelihood of diseases, we have proposed constraint k-Means-Mode clustering algorithm. Attributes of Medical data are both

continuous and categorical. The developed algorithmic program will handle each continuous and distinct

data and perform bunch supported anticipated chance attributes with core attributes of unwellness in datum. We have incontestable its effectiveness by testing it for a true world patient information set.

Drawbacks:

- ✧ Less Accuracy.

2.2 SYSTEM REQUIREMENTS

2.2.1 Hardware Requirements

- RAM: 4GB and Plus
- Processor: Intel Quad Core and Higher versions
- Processor Speed: 2.4ghz+
- Hard Disk: 40GB and more

2.2.2 Software requirements

- ✧ Frame work: DOTNET
- ✧ IDE: Visual Studio 2010 or higher
- ✧ Front end: ASP.NET 4.0
- ✧ Programming Language: C#.NET
- ✧ Back End: SQL Server
- ✧ OS: XP, Win7, Win8, Win10
- ✧ Browsers: IE, Firefox, Google Chrome etc..

3.EXISTING AND PROPOSED SYSTEM

3.1 EXISTING SYSTEM

The detection of diabetes is of great importance, concerning its severe complications. Current system is a manual process where the concerned doctor analyzes patients reports manually and it requires more time for the diabetes identification. Home glucose monitoring also done by the patients using the glucose monitoring device (mentioned in the below picture) to detect the diabetes. There have been plenty of research studies about diabetes identification, many of which are based on the Pima Indian diabetes data set. Most of the research studies done before mainly focused on one or two particular complex technique to test the data, while a comprehensive research over many common techniques is missing.

FIG 1:ONE TOUCH DEVICE

Limitations of Existing System

- ➢ Manual process (done by doctor)
- ➢ Manual analysis of test reports
- ➢ Glucose monitoring device such as one touch..
- ➢ Few techniques used for diabetes prediction.

3.2 PROPOSED SYSTEM

Diabetes mellitus has a direct signal of high blood sugar, together with some symptoms including frequent urination, increased thirst, increased hunger and weight loss. Patient of diabetes usually need constant treatment, otherwise, it will possibly lead to many dangerous life-threatening complications. Detection of diabetes in early stages and faster plays vital role in curing diabetes. Proposed system is an automation for diabetes identification using the old diabetes patients data. Proposed system is a medical sector application which is useful to Physicians (diabetic doctors) in identifying the disease. Proposed system uses machine learning techniques for diabetes identification.

Scope and Objectives

- Proposed system is a medical application used by diabetes doctors (physicians).
- Proposed system is a medical software for diabetes identification.
- Proposed system uses machine learning techniques for diabetes identification.
- Proposed system uses old diabetes patients data for diabetes identification of the new patient.
- Proposed aims at diabetes identification based on the attributes such as blood test reports, age, DOB, weight, BP, Height etc.
- Proposed system is a real time application which uses ASP.NET as front technology and SQL Server as backend technology.

4.METHODOLOGY AND RESULTS

4.1Classification Rules

Classification is a process of finding a model (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown.

Example: Suppose the sales manager of All Electronics want to classify a large set of items in the store, based on three kinds of responses to sales campaign: good response, mild response and no response. The model for each of these three classes is derived based on the descriptive features of the items, such as price, brand, place_ made, type and category. The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set.

4.1.1Naïve Bayes Algorithm Steps

Step 1: Scan the dataset (storage servers)retrieval of required data for mining from the servers such as database, cloud, excel sheet etc..

Step 2: Calculate the probability of each attribute value. [n, n _c, m, p]Here for each attribute we calculate the probability of occurrence using the following formula. (mentioned in the next step). For each class(disease) we should apply the formulae.

Step 3: Apply the formulae

P(attribute value(ai)/subject value v j)=(n_ c + m p)/(n +m)

Where:

- n = the number of training examples for which v = v j
- n c = number of examples for which v = v j and a = ai
- p = a priori estimate for P
- m = the equivalent sample size

Step 4: Multiply the probabilities by p

for each class, here we multiple the results of each attribute with p and final results are used for classification.

Step 5: Compare the values and classify the attribute values to one of the predefined set of classes.

Sample Example

Attributes(Constraints) – S1,S2,S3 [m=3]

Subject (Disease) – Yes, No [p=1/2=0.5]

Formulae to calculate P=[n _c + (m*p)]/(n +m)]

Patient Name	S1(X,Y,Z)	S2 (A,B,C)	S3 (P,Q,R)	Disease (subject)
Anil	X	A	P	Yes
Ajay	X	B	Q	Yes
Arun	Y	B	P	No
Kumar	Z	A	R	Yes
Naveen	Z	C	R	No

New Patient data – Akash Constraints (S1 -X,S2-A,S3-R) Disease – Yes/ No

P=[n _c + (m*p)]/(n +m) WHERE P=PROBABILITY FACTOR

Yes	No
X P=[n_c + (m*p)]/(n+m) n=2, n_c=2,m=3,p=0.5 p=[2+(3*0.5)]/(2+3) p=0.7	X P=[n_c + (m*p)]/(n+m) n=2, n_c=0,m=3,p=0.5 p=[0+(3*0.5)]/(2+3) p=0.3
A P=[n_c + (m*p)]/(n+m) n=2, n_c=2,m=3,p=0.5 p=[2+(3*0.5)]/(2+3) p=0.7	A P=[n_c + (m*p)]/(n+m) n=2, n_c=2,m=3,p=0.5 p=[2+(3*0.5)]/(2+3) p=0.3
R P=[n_c + (m*p)]/(n+m) n=2, n_c=1,m=3,p=0.5 p=[1+(3*0.5)]/(2+3) p=0.5	R P=[n_c + (m*p)]/(n+m) n=2, n_c=1,m=3,p=0.5 p=[1+(3*0.5)]/(2+3) p=0.5

Results:

Yes– 0.7 * 0.7 * 0.5 * 0.5 (p) =0.1225

No-0.3 * 0.3 * 0.5 * 0.5 (p)= 0.0225

Since Yes> No So this new patient is classified to Yes

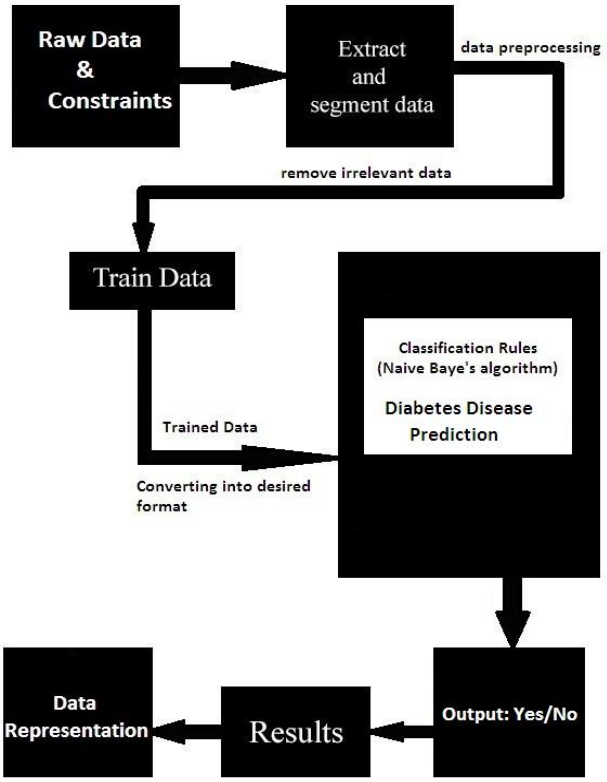

FIG : MAIN ARCHITETURAL DESIGN

5.CONCLUSIONS

One of the biggest causes of death worldwide are diabetes diseases. early identification of this disease can be achieved by developing machine learning models. System mainly concentrates to diabetes identification using some diabetes disease in early stages so that proper treatment may be given. System uses some machine learning techniques for prediction, so as to get more accurate results.

6.REFERENCES

[1] World Health Organization, "Report of a study group: Diabetes Mellitus," World Health Organization Technical Report Series, Geneva, 727, 1985.

[2] Kemal Polat, Salih Gunes and Ahmet Arslan, "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine," Expert Systems with Applications, vol. 34. 1, January. 2008, pp. 482-487.

[3 Yildirim T, "Medical diagnosis on Pima Indian diabetes using general regression neural networks," Proceedings of the international conference on artificial neural networks and neural information processing, 2003, pp. 181-184.

[4] Jack W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," Proc. Annu.Symp. Compute. Appl. Med. Care, November 9. 1988, pp. 261-265.

[5] Karegowda A. G., Manjunath A. S. and Jayaram M. A., "Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes," International Journal on Soft Computing, vol. 2. 2, 2011, pp. 15-23.