Indian Journal of Science and Technology, *Vol 8(8), 771–776, April 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Design and Implementation of Expert Clinical System for Diagnosing Diabetes using Data Mining Techniques

Srideivanai Nagarajan[1*] and R. M. Chandrasekaran[2]

[1]Annamalai University, Chidambaram, India; deivagayu@gmail.com
[2]Department of Computer Science and Engineering, Annamalai University, Chidambaram, India; aurmc@hotmail.com

Abstract

Objective: The aim of this paper is to design and implement an expert clinical system to diagnose the type of diabetes and the levels of risk among diabetic patients using the data mining techniques clustering and classification. **Methods:** The research design made use of primary and secondary data and the data were collected using data collection tools and techniques such as questionnaires, direct interview and survey of existing medical records from 650 diabetic patients. The study was based on purposive sampling type using a structured questionnaire that was pre tested with 25 respondents. After making necessary modifications from the feedback received from the pre-test, the final questionnaire was prepared. **Findings:** With six iterations the data could be successfully clustered into three clusters namely type-1, type-2 and gestational diabetes using Simple K-means algorithm. The classification algorithms - NaiveBayes, Random Tree, Simple Cart and Simple Logistic were used on the clustered data to classify the data into mild, moderate and severe types resulting into an expert clinical system. **Conclusion:** This paper demonstrates creation of expert clinical system for the diagnosis of the diabetic mellitus using clustering and classification techniques of data mining. However with suitable modification the same can be extended to evolve similar systems in other application areas in health care.

Keywords: Classification, Clustering, Data Mining Techniques, Diabetes Type, Diagnosis of Diabetes, Expert Clinical System, NaiveBayes, Random Tree, Simple Cart, Simple K-Means, Simple Logistic

1. Introduction

Type-1, type-2 and gestational diabetes are the three types of diabetes. Although there is a raise in the sugar level in the blood during pre-diabetes period, it is generally not very high above the normal value.

The type-1 diabetes mellitus is an auto immune disease where the pancreas secretes little quantity of insulin. Generally, type-1 diabetes affects people when they are young and below 20 years of age. At times, small children also get affected by type-1 diabetes. In type-1 diabetes, the pancreatic cells get affected and fail to function. Due to nil secretion of insulin, type-1 diabetic people suffer throughout their life and they depend on insulin injection. Besides being on insulin, they need to do regular exercises and follow healthy diet as suggested by dietitians[1].

Diabetes mellitus type-2 is formerly called Non-Insulin-Dependent Diabetes Mellitus (NIDDM) or adult-onset diabetes and is a metabolic disorder that is characterized by hyperglycemia (high blood sugar) caused by insulin resistance. The classic symptoms are excess thirst, frequent urination, and constant hunger. Obesity is one of the primary reasons for type-2 diabetes. At the initial stage, type-2 diabetes can be controlled by doing proper exercise and taking appropriate diet. If the glucose level is not reduced by the above methods then medicines can be administered. National Diabetes Statistics Report 2014 says that 29.1 million people or 9.3% of the U.S. population have diabetes[2].

**Author for correspondence*

According to the recent estimates by the International Diabetes Federation (IDF), the number of type-2 patients was 366 million in 2011 and by 2030 it may be increased to 552 million. Almost 80% of the affected people belong to middle- and low-income countries. Long-term complications from high blood sugar can include heart disease, kidney failure, strokes, and diabetic retinopathy[3].

The number of persons affected by type-2 diabetes will be more by 2025. Occurrences of diabetes are reduced by 2.7% in rural area in India compared to urban area[4]. Overweight, obesity and diabetes mellitus are strongly associated with prehypertension. Indian Diabetic Risk Score (IDRS) detects that a person who is having normal blood pressure but with high Indian diabetic risk score is likely to become hypertensive or diabetic in near future[5].

When a non diabetic woman has high blood sugar level during pregnancy, she is said to have gestational diabetes. According to recently announced diabetes criteria, it is found that around 18% of pregnant women have diabetes. Pregnancy during older age may have a risk of developing gestational diabetes.

Among all diabetes patients, nearly 90% of cases are of type-2 diabetes, with the other 10% as type-1 and gestational diabetes[6].

1.1 Clustering Algorithm

Clustering is a main task in data mining and a general technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bio-informatics[7].

1.2 Classification Algorithm

Classification is a supervised machine learning technique that assigns labels or classes to different objects or groups[8]. It is a two step process: First step is model construction, which is used to analyze the training dataset of a database. The second step is model usage where the constructed model is used for classification. The accuracy of the classification is estimated according to the percentage of test samples or test dataset that are correctly classified.

1.3 Significance of the Study

In this paper, data mining techniques namely clustering and classification are applied to diagnose the type of diabetes and its severity level for every patient from the data collected. The study of related works is presented in section II. The methods and materials which include description about dataset, system design supervised learning algorithms are discussed in section III. The section IV describes the experimental results and finally section V gives the conclusion of this paper.

2. Related Works

[9]Proposed an approach known as CoLe for detecting diabetes in the initial stage itself. CoLe is a multi-agent system with multiple data miner besides being a combination agent. The important aim of CoLe is to achieve a mixture of knowledge that describes data in different perspectives.

[10]Applied many classifications algorithms like C4.5, ID3, K-NN, LDA, NaiveBayes for diagnosing diabetes for the given dataset. They found that C4.5 is the best algorithm with less error rate of 0.0938 and more accuracy value of 91%.

[11]Used Artificial Intelligence to design an expert clinical system to diagnose diabetes. They used Extended Classifier System (XCS). This method has greater accuracy than the conventional data mining techniques.

[12]Applied Fuzzy Id3 algorithm for predicting diabetes. The author applied the clustering algorithm first and then used classification algorithm on the clustered data. The author suggested a combination of classification system which was developed using Em algorithm for clustering and fuzzy ID3 algorithm to attain decision tree for each cluster.

[13]Used Apriori association algorithms to classify type-2 diabetes. For the class value "yes", the author developed four association rules and for the class value "no", the author developed ten association rules. Preprocessing techniques were also applied by the author for improving the dataset quality.

[14]Used J48 algorithm to construct a decision tree for diagnosing type-2 diabetes. The accuracy of the model is 78.1768%.

[15]Developed a new frame work known as duo-mining tool for diagnosing diabetes. They also applied many classification algorithms like KNN, SVM, decision Tree for type-2 diabetes classification. They found that SVM algorithm gave the highest accuracy value of 95.49%.

[16]Used hierarchical clustering algorithm to identify the trends for controlling diabetes mellitus.

[17]Applied CART Method for monitoring Diabetes. The algorithm distinguishes between high risk and low risk patients. The system achieved an accuracy rate of 96.39%.

[18]Some of the classification algorithms like SimpleCart, J48, Simple Logistics, SMO, NaiveBayes and BayesNet

were used by the author for diagnosing neonatal jaundice and found out that the Simple Logistics algorithm to be the best.

[19]Applied the NaiveBayes and SVM classification algorithms for predicting diabetic retinopathy and found out that the NaiveBayes algorithm contributed to nearly 83% accuracy.

The clustering algorithm Simple K-Means was used to develop a model for this paper. This model groups the dataset into type-1, type-2 or gestational diabetes. After clustering the type of diabetes, the model also applies classification algorithms like RandomTree, NaiveBayes, SimpleCart and Simple Logistics for predicting the risk levels of diabetes as mild, moderate and severe.

3. Methodology

3.1 Algorithms Applied

The model developed for this paper uses Simple K-Means clustering algorithm for predicting the type of diabetes among the patients. After finding the type of diabetes the classification algorithms like RandomTree, NaiveBayes, SimpleCart and Simple Logistics are used to predict the risk levels of the disease.

3.2 Discussion

The model proposed in this paper has three stages.

- Data pre-processing.
- Application of Simple K-Means algorithm to the dataset for clustering the data into three clusters as cluster-0 (gestational diabetes), cluster-1 (type-1 diabetes), cluster-2 (type-2 diabetes).
- Application of Classification algorithms to classify the patient's risk of diabetes levels.
- During the preprocessing equal interval binning was used under a medical expert's guidance.

3.3 Dataset Used

The Dataset collected from known sources is a clinical dataset containing records of 650 patients of all age group. Table 1 shows all the attributes used for the research.

3.4 Data Preprocessing

As the collected data contain some inconsistencies data preprocessing was done to remove the inconsistencies. During this study, instances which had zero values for the attributes - Pregnant, Plasma Glucose, Diastolic BP, job type and food habit were removed. In this study for data preprocessing, supervised attribute filtering technique was used. Discretize filter was used for deriving good intervals of data. After pre-processing, only 599 valid instances remained out of 650.

Table 1. The attributes used in the experimentation

Attribute	Description	Type
Gender	Male or Female	Numeric
Insulin dependent	100% Insulin dependent	Numeric
Plasma	Plasma glucose concentration - oral GTT	Numeric
HbA1c	glycated haemoglobin	Numeric
Systolic	blood pressure (Systolic)	Numeric
Diastolic	blood pressure (Diastolic)	Numeric
Mass	BMI	Numeric
bg	Blood group	Nominal
Age	Patient's age	Numeric
Pedigree	Family history details	Numeric
Pregnancy	Number of pregnancies (only for the female patients who are pregnant)	Numeric
Living area	Living area(Urban or Rural)	Numeric
Job type	Type of job	Numeric
Food habit	Food habits of the patients (healthy, moderate, junk food)	Numeric

3.5 System Design

This section explains the steps involved in building the model. The frame work is shown in Figure 1.

3.6 Accuracy Measures

In this research RandomTree, NaiveBayes, SimpleCart and Simple Logistics algorithms were used. The tests were performed by means of internal cross validation 10-folds.

Each algorithm's accuracy indicates how far the datasets are being classified. Recall and precision are the accuracy measures used for this study.

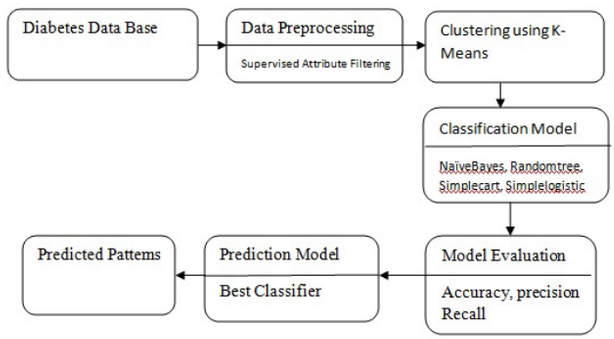

Figure 1. System Design.

Precision = TP/(TP + FP).
Recall = TP/(TP + FN).
Accuracy = (TP + TN)/(TP + TN + FP + FN).

TP - Positive tuples.
TN - Negative tuples.
FP - Incorrectly classified positive tuples.
FN - Incorrectly classified negative tuples.

The classifiers with corresponding precision and recall values are listed in Table 2.

4. Results and Discussion

The Proposed system was executed in three stages:
Stage 1 - The entire dataset was preprocessed.
Stage 2 - After preprocessing the clean dataset was clustered to find the type of diabetes for each patient as type-1, type-2 and gestational diabetes.
Stage 3 - The clustered dataset was classified into three classes as mild, moderate and severe. This was done in order to predict the risk levels of diabetes for each patient.

4.1 Performance of the Simple K-Means Algorithm

The Simple k-means algorithm clusters the entire dataset into 3 clusters as cluster-0 - for gestational diabetes, cluster-1 for type-1 diabetes (juvenile diabetes), cluster-2 for type-2 diabetes. The time taken to build the model was 0.03 seconds. Among the 599 instances of the data, 94 were in cluster-0 (i.e. gestational diabetes), 85 were in cluster-1 (i.e. type-1 diabetes) and 420 in cluster-2 (i.e. type-2 diabetes).

This clustered dataset was given as input to the model which classified each patient's risk levels of diabetes as mild, moderate and severe. This model uses all classification algorithms discussed in section III. Table 3 shows the results of the classification algorithms.

The Error rate and accuracy value of each classification algorithm are shown in Table 4.

4.2 Classifier Performances

From the Table 4, it is clear that SimpleCart algorithm has less error rate and more accuracy value. Figure 2 shows error rates of all classifiers.

From Figure 2, it was observed that SimpleCart algorithm has zero error rate when compared to other algorithms as RamdomTree, NaiveBayes, and Simple Logistics in predicting the risk levels of diabetes. The accuracy values of various classifiers are shown in Figure 3.

The diabetes data for about 599 tuples with 14 attributes were analyzed for their accuracy and error rate with various classification algorithms. From the above analysis it was found that SimpleCart algorithm was the best one

Table 2. Accuracy measures precision and recall

Classifier	Precision			Recall		
	Mild	Moderate	Severe	Mild	Moderate	Severe
NaiveBayes	0.93	0.897	0.957	0.955	0.789	1
RandomTree	0.992	0.953	0.943	0.973	0.917	0.99
SimpleCart	1	1	1	1	1	1
Simple Logistics	1	1	1	1	1	1

Table 3. The results of the classification algorithms

Diabetes Type	Number of Patients	Risk Level	Number of Patients
Type-1 Diabetes	85	Mild	30
		Moderate	23
		Severe	32
Type-2 Diabetes	420	Mild	185
		Moderate	93
		Severe	142
Gestational Diabetes	94	Mild	49
		Moderate	17
		Severe	28

Table 4. Classifiers error rate and accuracy values

Classifier	Error Rate	Accuracy Value
NaiveBayes	0.0793	0.989
Random Tree	0.022	0.975
SimpleCart	**0**	**1**
Simple Logistics	0.0367	1

Figure 2. Basic classifier error rate.

when compared to other classifiers for diagnosing diabetes dataset.

4.3 Random Tree

From Figure 4 RandomTree, it was discovered that among the 14 attributes used for the study, the attributes HbA1c, plasma, systolic, age, weight and food habit played a vital role in diagnosing diabetes. The tree reveals that for a diabetic patient with HbA1c value less than 5.55%, the diabetic level will be mild and if HbA1c value is greater than 5.85%, it may lead to severe level of diabetes. When the age of the patient is less than 50 with systolic pressure value less than 167.5 mm Hg, then the patient may have moderate level of type-2 diabetes. If the patient's age is greater than 50 with weight greater than 61 Kilograms and HbA1c value greater than 5.95%, it may lead to severe level of diabetes. The tree also shows that if the food habit of the patient is greater than 2.5 (i.e. if they take junk food frequently) and with age value greater than 50, it will lead to severe level of diabetes.

5. Conclusion

Diabetes is one of the commonly occurring diseases. Preventing or controlling diabetes is important as diabetes leads to other potential health problems. Type-1 and type-2 diabetes may lead to heart problems, kidney

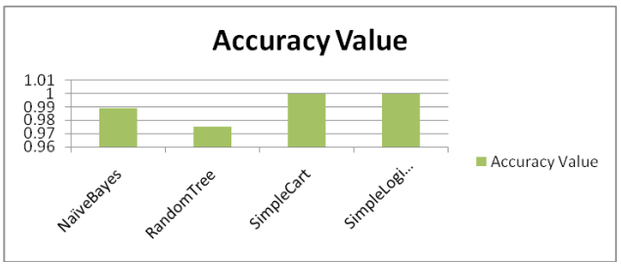

Figure 3. Basic classifiers accuracy value.

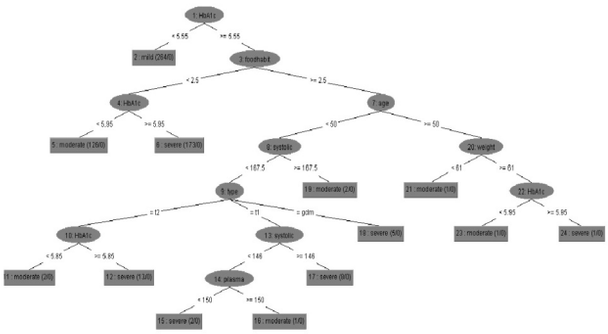

Figure 4. RandomTree.

diseases and eye related ailments. Gestational Diabetes Mellitus (GDM) may disappear post pregnancy, but GDM women are 7 times prone for type-2 diabetes than the non GDM women. The children of the GDM mother have the risk of obesity and type-2 diabetes[20]. All these complications can be handled by controlling the blood sugar levels[21]. From this study, it was found out that data mining techniques can be used for predicting the type and risk levels of diabetes. So from this research paper, the researcher recommends to use data mining techniques in the medical field which will improve the diagnosis of the diseases like diabetes.

6. References

1. Type-1 diabetes. Available from: http://www.diabetes.org/diabetes-basics/type-1
2. National Diabetes Statistics Report. 2014. Available from: http://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf
3. Type-2 diabetes in India: Challenges and possible solutions. Available from: http://www.apiindia.org/medicine_update_2013/chap40.pdf

4. Jali MV, Hiremath MB. Diabetes. Indian Journal of Science and Technology. 2010 Oct; 3(10).
5. Lanord M, Stanley J, Elantamilan D, Kumaravel TS. Prevalence of Prehypertension and its Correlation with Indian Diabetic Risk Score in Rural Population. Indian Journal of Science and Technology. 2014 Oct; 7(10):1498–503.
6. Diseases and conditions with subheading women's health. Available from: http://www.thehealthsite.com
7. Han J, Kamber M. Data mining concepts and techniques. 2nd ed. Amsterdam, Netherlands: Elsevier Publisher; 2006. p. 383–5.
8. Han J, Kamber M. Data mining concepts and techniques. 2nd ed. Burlington, Massachusetts: Morgan Kaufmann; 2006. p. 285–8.
9. Gao J, Denzinger J, James RC. CoLe: A cooperative data mining approach and its application to early diabetes detection. Proceedings of the 5th International Conference on Data Mining (ICDM'05); 2005.
10. Rajesh K, Sangeetha V. Application of data mining methods and techniques for diabetes diagnosis. International Journal of Engineering and Innovative Technology (IJEIT). 2012; 2(3):224–9.
11. Afrand P, Yazdani NM, Moetamedzadeh H, Naderi F, Panahi MS. Design and implementation of an expert clinical system for diabetes diagnosis. Global Journal of Science, Engineering and Technology; 2012. p. 23–31. ISSN:2322-2441.
12. Adidela DR, Lavanya DG, Jaya SG, Allam AR. Application of fuzzy ID3 to predict diabetes. Int J Adv Comput Math Sci. 2012; 3(4):541–5.
13. Patil BM, Joshi RC, Toshniwal D. Association rule for classification of type-2 diabetic patients. 2nd International Conference of IEEE on Machine Learning and Computing; 2010. p. 67. DOI 10.1109/ICMLC.
14. Aljarullah AA. Decision tree discovery for the diagnosis of type II diabetes. International Conference on Innovative in Information Technology; 2011. p. 303–7.
15. Jaya Rama Krishnaiah VV, Chandra Shekar DV, Satya Prasad R, Rao KRH. An empirical study about type-2 diabetes suing duo mining approach. International Journal of Computational Engineering Research. 2012; 2(6):33–42.
16. Mandal S, Dubey V. Implementation and evaluation of diabetes management system using clustering technique. Special Issue of International Journal of Computer Science and Informatics. 2(2):33–6.
17. Kavitha K, Sarojamma RM. Monitoring of diabetes with data mining via CART Method. International Journal of Emerging Technology and Advanced Engineering. 2012; 2(11):157–62.
18. Ferreira D, Oliveira A, Freitas A. Applying data mining techniques to improve diagnoses in neonatal jaundice. BMC Med Informat Decis Making. 2012; 12:143. DOI: 10.1186/1472-6947-12-143.
19. Ananthapadmanaban KR, Parthiban G. Prediction of chances - diabetic retinopathy using data mining classification techniques. Indian Journal of Science and Technology. 2014 Oct; 7(10):1498–503.
20. National Center for Chronic Disease Prevention and Health Promotion. Gestational Diabetes. Centers for Disease Control and Prevention. U.S. Department of Health and Human Services; 2011. Available from: http://www.cdc.gov/
21. Type-2 diabetes complications. Available from: http://www.mayoclinic.org/diseases-conditions/type-2 diabetes/basics/complications/con-20031902