A Survey on Data-Mining Technologies for Prediction and Diagnosis of Diabetes

Dr.B.L.Shivakumar
Professor
Sri Ramakrishna EngineeringCollege
Coimbtaore – 22
blshiva@yahoo.com

S. Alby
Research Scholar
Research and Development Centre
Bharathiar University, Coimbatore - 44
alby.siby@gmail.com

Abstract **–** The recent report of WHO shows a remarkable hike in the number of diabetic patients and this will be in the same pattern in the coming decades also. Early identification of diabetes is an important challenge. Data mining has played an important role in diabetes research. Data mining would be a valuable asset for diabetes researchers because it can unearth hidden knowledge from a huge amount of diabetes-related data. Various data mining techniques help diabetes research and ultimately improve the quality of health care for diabetes patients. This paper provides a survey of data mining methods that have been commonly applied to Diabetes data analysis and prediction of the disease.

I. Introduction

Data Mining refers to extracting knowledge from large amounts of data [1]. It enables us to explore the large patterns and analyze the same by means of Statistical and Artificial Intelligence in large datasets [2]. The Data Mining technique is used to predict possible future trends or to discover hidden patterns in the behaviour of the data. Techniques such as Artificial Neural Networks, Decision Trees, Classification, Clustering, Association rule algorithms etc are widely utilized by experts.

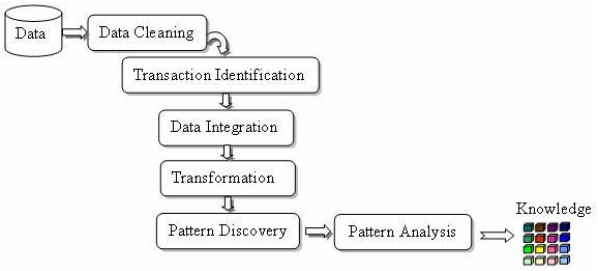

Fig. 1: Data Mining Process

Data mining techniques is widely being applied by researchers in Bioinformatics. Bioinformatics is the science of storing, extracting, organizing, interpreting and utilizing information from biological sequences and molecules [3]. In recent years, Knowledge Discovery and Data Mining techniques are widely used for extracting the patterns from the large biological databases. The amount of biological data is growing rapidly [4]. Analyzing these data sets requires making sense of the data by inferring structure or generalizations from the data. The interaction between data mining and bioinformatics plays a vital role in diagnosis many diseases.

The World Health Organization predicts that by 2030 there will be approximately 350 million people worldwide affected by diabetes [5]. Diabetes is a state or a condition in which the body is not able to produce or utilize the insulin properly which resulting in developing diseases affecting kidney,eyesight,nerve system, blood vessels and also cardiac related issues [6]. There are three types of diabetes. Type-1 diabetes used to be called juvenile-onset diabetes. It is usually caused by an auto-immune reaction where the body's defense system attacks the cells that produce insulin. Type-1 is also called insulin dependent. Almost 90% of the diabetes suffers from Type II diabetes which is alternatively called as non-insulin dependent diabetes or Adult-onset diabetes. This is characterized by resistance to insulin and deficiency of relative insulin which may either or both together be present at the time of which diabetes is diagnosed. In the case of Gestational Diabetes (GDM) which is often seen during pregnancy resulting from high glucose levels. This is often accomplished by renal complications, cardiac diseases and peripheral vascular diseases. Early identification of patients with

978-1-4799-3966-4/14 $31.00 © 2014 IEEE
DOI 10.1109/ICICA.2014.44

undiagnosed type-2 diabetes or those at an increased risk of developing type-2 diabetes is an important challenge in the field of medical. The availability of huge amounts of medical data leads to the need for powerful data analysis tools to extract useful knowledge. Researchers have long been concerned with applying statistical and data mining tools to improve data analysis on large data sets. Disease diagnosis such as diabetics is one of the applications where data mining tools are proving successful results in the recent years.

II. Association Rule

Association rules are more related when new rules are searched. Association rules are widely used in exploring the relationship between the symptom and syndrome type. B. M. Patil ,in literature [6] introduced a new approach to generate association rules on numeric data. They used pre-processing to improve the quality of data by handling the missing values and applied equal interval binning with approximate values based on medical expert's advice to Pima Indian diabetes data. Lastly apriori association rule algorithm is applied to generate the rules. Only type- 2 diabetic patients those who are pregnant woman below 21 years are included in their study. It proved that the results obtained are very promising. S.M.Nuwangi, in paper [7] used advanced and reliable data mining techniques to identify different risk factors behind the diabetes and the relationship between the diabetes and the other diseases. Using association rule generation, the relationship between edema and diabetes and wheezes and diabetes has been identified. The result shows, the females aged between 39 – 75 years with normal BMI range, systolic BP range and diastolic BP range and having wheezes will have a high risk towards developing high FBS (fasting blood sugar) level. Milan Zormana, in literature [8] addressed the problem of mining rules from the diabetes database using a combination of decision trees and association rules. About 1251 different cases from original database with selected attributes were considered and with the help of association rule approaches, different trees are built and converted them into different set of rules and these rules where further reduced and filtered. The main objective of this paper was to analyze the number of rules which are generated and how many rules will be balanced after performing filtering and reduction. It also analyze how many rules will be generated employing association rule approach on the same database and the conclusion is that, the sets of rules built by decision trees were much smaller than results of association rules. Recent research shows that the main cause of type -2 diabetes is the obesity and sedentary lifestyle which may affects the genetic elements. The paper [9] discussed the potential of applying the apriori – Gen algorithm to the association study for the type – 2 diabetes. The relative risk (RR), which is the risk of developing a disease relative to exposure and odds ratio (OR) , which is the ratio of the odds of an event occurring in control group, are used to prove that interaction of Multi SNPs is associated with the disease. The study [18] aims to propose a theory of discovering Association Rule of Diabetes Mellitus (DM) patients data with renal , Ophthalmic , Neurological and Peripheral circulatory complications . Diabetes Mellitus is a group of metabolic deseases in which a person has high Blood sugar , in etiher because the body not prodouce enough insulin or because cells donot respond to the insulin that is produced. Data Mining methords vis Association Rule discovery was processed by WEKA . The results of this study are

- Both Male & Female there has a chance DM Patients type II with Complications
- DM patients type II females aging 50 – 79 are prone to have Neurological, Ophthalmic and renal complications
- DM patients type II House work , Trade and no Occupation are prone to have Ophthalmic, Neurological & peripheral Circulatory complications .

The statistical data from Taiwan reveals that over the past 30 years , one of the main causes for death in the country is hypertension and Diabete Mellitus in metabolic syndrome [19]. So Mr Chien – Lung Chan , Mr Chien – Weichen and Mr Ban- Jhiune Liu [19] tried to study metabolic syndrome related desease by using Data Mining techniques and to understand the strength of association between Diabetes Mellitus , Hypertension and Hyperlipidemia. Metabolic syndrome is a combination of the medical disorders that when occurring together increase the risk of developing cardiovascular diseases and diabetes [27]. The data used for this research is from the National health Insurance research data base.

Association rules are used to find disease patterns of metabolic syndrome related disease. As a result of this study among 10,192,166 records it includes 163,045 diabetes patients , the prevalence of Diabetes Mellitus will increase with age and the aging effect on female is more apparent. Through the mining of Apriori, a strong relation between the diabetes and a lot of oral diseases like Dental caries, Pulpitis, Acute Gingivitis etc is found. From this research the authors proved that using data mining techniques can find the relation of diseases that are similar with the results which are experimented from clinics.

III. Clustering and Classification

Clustering is the process of analyzing and grouping the data into different clusters or classes such that objects within the same cluster will be having more similarity to each other, but will be having different properties in other clusters. The principle objective of clustering in this area is to find different groups of diabetes patients with similar symptoms within a group but different symptoms of other groups. The supervised learning of algorithm in contrast with clustering is called Classification [1]. It classifies or maps a data item into any one of many predefined classes.

Palivela Hemant at [10] combines K-means clustering with various different classification algorithms like SMO, Naive Bayes, Bagging, AdaBoost, J48, Rotation Forest and Random Forest to predict the positive and negative of disease. The data consists of 768 different entries in accordance with attributes like Skin , Mass , Age , Insulin , pregnant etc are used . SMO implements the sequential minimal optimisation algorithms for training a support vector classifier. Missing values are replaced globally, nominal attributes are transformed into binary ones and attributes are normalised. Bagging bags a classifier to reduce variance. By using various classifiers, the authors propose a hybrid model for the prediction of the positivity and negativity of the diabetes. The research [11] uses association rule generation and classification techniques to support decision making by considering a data set of diabetes type 1 and type 2 patients. Identified gender female is a major decision factor of high fasting blood sugar level. Through the data mining techniques, a direct and strong relationship between edema and diabetes as well as wheezing and diabetes was identified. According to the doctor's view points, the probability of developing diabetes for male and female are the same [11]. But various research studies state that females are having a higher risk for having diabets.

In [12] the authors presented the development of a hybrid model for classifying Pima Indian diabetic database. Firstly, the datasets were identified and the incorrectly classified instances were eliminated using K-means clustering. Then using decision tree classifier (C4.5) the classification is done on these correctly clustered instances. The resultant dataset is used to train and test the diabetic data set using two method (a) dividing training data and test data using 60-40 ratio (b) 10 fold cross validation method. Experimental results show the improvement in accuracy Diabetic data set using proposed cascaded method: k-means with Decision tree (with categorical data) by an order of 19.50% of classification compared to Decision tree C4.5 alone with unprocessed data. The rules generated by the proposed cascaded model are given below.

1. *If Plasma=low then class=> Tested Negative*

2. *If Plasma =medium & Age=low & Pedigree =low then Class => Tested Negative*

3. *If Plasma =medium & Age=low & Pedigree =medium & Diastolic BP=medium then Class=> Tested Negative*

4. *If Plasma=medium & Age=low & Pedigree =medium & Diastolic BP =low then Class=> Tested Negative*

5. *If Plasma=medium & Age=low & Pedigree =medium &Diastolic BP =high then Class=> Tested Positive*

6. *If Plasma =medium & age=high then Class => Tested Positive*

7. *If Plasma =medium & Age=low & Pedigree =high then Class=> Tested Positive*

8. *If Plasma =medium & Age=medium then Class=> Tested Positive*

9. *If Plasma =high then Class=> Tested Positive*

The research [20] was based on three techniques of Expectations – Maximisation (EM) algorithm, H – means clustering and Genetic algorithm (GA). Pima Indian Diabetes Datasets were used on WEKA software tool. About 35 % of the total of 768 test samples was found with diabetes presence.

Xiao Fang ,at [21] tried to identify diabetics patients of a large health care enterprise using Data Mining Techniques includes clustering, classification, regression etc . This study is making use of concepts from other disciplines like Operational research & Inventory Management. Using a Data Base of 9278 with several relevant attributes, an analysis is conducted to identify which patients have high probability of being diabetic.

IV. Other methods

A series of research work has been done in this area using many other techniques. The literature [13] analyzed the Pima Indian diabetes data sets using the tool Rapid Miner. Discovered the hidden relationships between Plasma Glucose and Class attributes which finds that the patients with higher Plasma- Glucose values are having more chance to develop diabetes and with low Plasma – Glucose values have less chance to develop diabetes nearby future.

Kavitha K at [14] presents an approach to designing a platform to enhance effectiveness and efficiency of health monitoring using data mining for early detection of any worsening in a patient's condition. The use of data mining has also been established in reducing rates of medication errors Diabetes severity assessment offering functions of data mining based on the classification and regression tree method. CART is a robust data mining and data analysis tool that searches for important patterns and relationships and quickly uncover hidden structure even in highly complex data. This study identified easily applicable diagnostic algorithms using early clinical test results obtained from the lab tests. The set of rules provided by this study is easily understandable by doctors.

Joseph L. Breault at [22] examined one diabetic data warehouse from a large integrated healthcare system in the New Orleans area with 30,383 diabetic patients. For this examination, the classification tree approach in CART with a binary target variable of HgbA1c >9.5 and 10 predictors: age, sex, emergency department visits, office visits, comorbidity index, dyslipidemia, hypertension, cardiovascular disease, retinopathy, end-stage renal disease is used. From this CART analysis, it is shown that those less than 65.6 years of age are almost three times as likely to have bad glycemic control as those who are older.

Bum Ju Lee, in paper [15], did a study among a total of 4870 subjects to predict the fasting plasma glucose status that is used in the diagnosis of type -2 diabetes by a combination of various anthropometric measures that are measured in a greater no of specific sites in the body can improve the predictive power of diagnosing type 2 diabetes. According to their findings it is indicated that a combination of anthropometric measures can clearly improve the predictive power for normal and high FPG status when compared with individual measures and prediction experiments using balanced data of normal and high FPG subject can improve the prediction performance and reduce the intrinsic bias of the model towards the majority class.

Sonukumari and Archana Singh, in paper [23], tried to propose an intelligent and effective methodology for the automated detection of Diabetes Mellitus based on neural network. A survey has been done among 100 data sets which include people from different age groups, gender and life style. Around 13 parameters like gender, age, weight, height, thirst increase, hunger increase, appetite increase, vomiting etc along with the possible valued are fed in the neural network system. The output is in the binary form. The value zero means the person is not affected from DM and if it is one, it reveals that the person is suffering from DM. As a result of this study the authors proved that their Neural network system is having an accuracy rate of 92.8 %. Asha A Aljarullah, in the study [24] ,used decision tree methods to predict patients with developing diabetes. The data set used was the Pima Indian diabetes Data sets. Pre processing was used to improve the quality of data. Next WEKA's J48 decision tree classifier was applied to the modified data set to construct the decision tree model, it can be seen in the figure 2.

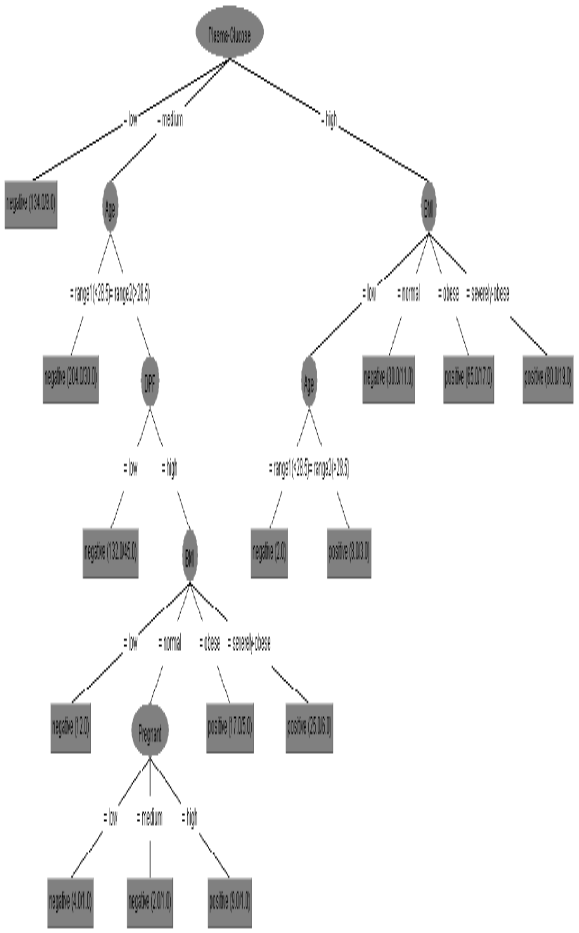

Figure 2: The J48 decision tree for diabetes diagnosis using the Pima Indian diabetes dataset

Different techniques of artificial intelligence have been applied to diabetes problems. In the study [25], the artificial metaplasticity on multi layer perception (AMMLP) as a data mining technique is applied for the diabetes disease diagnosis. Artificial metaplasticity is a novel artificial neural network training algorithm inspired on the biological metaplasticity property of neurons and shannon's information theory[28]. The PIMA Indian diabetes data base was used to test the proposed model AMMLP. The results obtained by AMMLP were compared with decision tree, Bayesian classifier and other algorithms. The table 1 shows the matrix generated with the classification results obtained for each classifier in the best simulation .The performance measures obtained by AMMLP classifier are significantly superior to the obtained by decision tree and Bayesian classifier .

Table 1: Confusion Matrix generated with the classification results obtained for each classifier in the best simulation

Classifiers	Desired Result	Prediction	
		Diabetes	No Diabetes
AMMLP	Diabetes	82	26
	No Diabetes	5	195
DT	Diabetes	85	23
	No Diabetes	46	154
BC	Diabetes	70	38
	No Diabetes	28	172

Eleni Georga, in [26] aims to reveal the METABO diabetes management system and describe on the Data Mining techniques that are used for modeling patients metabolism and also for finding patients disease related awareness. METABO is a diabetes monitoring and management system which enables in registering and interpreting patients context as well as helping decision support to both patient as well as the doctor. The main system which are connected with METABO are patient mobile device, the control panel and the central system. In METABO, the data related with patient's life style which may either directly affect glucose levels or influence glucose controls is analyzed and metabolic modeling is achieved by considering the resent glycaemia profile of the patient for analyzing the insulin- glucose interaction. The techniques and methods employed within METABO helps in the prediction of clinically critical events in both Hypoglycemia and Hyper glycaemia. Appropriate suggestions are rendered to the patients to stabilize their diabetic profile.

V. Conclusion

Mechanical life style and obesity have contributed to the sudden rise of type 2 diabetes worldwide. The resent news regarding occurrence of type 2 diabetes shows that prevention measures are not up to the mark. In this context here I have done an analysis of various presentations and studies done by other researchers. From the analysis of different research papers it is evident that the occurrence of diabetes is having a strong relation with

- Deceases like Wheeze , Edema, Oral diseases
- Female pregnant
- Increase of age.

It is also evident that diabetes above certain age is also prone to have other complications. By using data mining techniques the chance of diabetes can be predicted which is helpful for early detection of the disease.

In future the data mining techniques have to be done more intrinsically to co relate diabetes with other diseases an enabling the doctors and patients for a much more accurate and early detection of diabetes

REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining : Concepts and Techniques", 2nd edition

[2] E. Papageorgiou, I. Kotsioni, A. Linos, "Data Mining: A New Technique In Medical Research", *HORMONES*, 4(4):pp 189-191,2005.

[3] Mohammed J. Zaki, Jason T. L. Wang,Hannu T.T. Toivonen, "BIOKDD01: Workshop on Data Mining in Bioinformatics",www.kdd.org/sites/default/files/issues/3-2-2002-01/zaki.pdf

[4] Khalid Raza,"Application of data mining in bioinformatics", *Indian Journal of Computer Science and Engineering*, Vol 1 No 2, pp 114-118

[5] Screening for type-2 diabetes: Report of a World Health Organization and International Diabetes Federation meeting.www.who.int/diabetes/publications/en/screening-mnc03.pdf.

[6] B. M. Patil, R. C. Joshi, Durga Toshniwal, "Association rule for classification of type -2 diabetic patients", *Proc. of the Second International Conference on Machine Learning and Computing*, pp 330-334, 2010

[7] S.M.Nuwangi, C. R. Oruthotaarachchi, J.M.P.P. Tilakaratna & H. A. Caldera, "Usage of Association rules and Classification Techniques in Knowledge Extraction of Diabetes", *Proc of the 6th International Conference on Advanced Information Management and Service(IMS)*, pp 372-377, 2010

[8] Milan Zormana, Gou Masudab, Peter Kokola, Ryuichi Yamamotob, Bruno Stiglica, "Mining Diabetes Database With Decision Trees and Association Rules", *CBMS,Proc. of the 15th IEEE symposium* , pp 134-139,2002.

[9] Weidong Mao, Jinghe Mao, "The Application of Apriori-Gen Algorithm in the Association Study in Type 2 Diabetes", *Proc. of the 3rd International Conference Bioinformatics and Biomedical Engineering*(*ICBBE 2009),* pp 1-4, 2009

[10] Palivela Hemant, Thotadara Pushpavathi, "A novel approach to predict by cascading clustering and classification", *Proc. of the 3rd International Conference on Computing Communication & Networking Technologies*, pp 1-7. 2012

[11] S.M. Nuwangi, C. R. Oruthotaarachchi, J.M.P.P. Tilakaratna, H. A. Caldera, "Utilization of Data Mining Techniques in Knowledge Extraction for Diminution of Diabetes", *Proc. of the Second Vaagdevi International Conference on Information Technology for Real World Problems(VCON),pp. 3-8* ,2010.

[12] Asha Gowda Karegowda, Punya V, M A Jayaram,A S Manjunath,"Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5", *International Journal of Computer Applications* (0975 – 8887) Volume 45– No.12, May 2012

[13] Jianchao Han, Juan C. Rodriguze, Mohsen Beheshti, "Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner", *Proc. of the Second International Conference on Future Generation Communication and Networking*, Vol. 3,pp 96-99, 2008.

[14] Kavitha K , Sarojamma R M, " Monitoring of Diabetes with Data Mining via CART Method", *International Journal of Emerging Technology and Advanced Engineering*, Website: www.ijetae.com ISSN 2250-2459, Volume 2, Issue 11, November 2012.

[15] Bum Ju Lee, Boncho Ku, Jiho Nam, Duong Duc Pham, Jong Yeol Kim, "Prediction of Fasting Plasma Glucose Status using Anthropometric Measures for Diagnosing Type 2 Diabetes", *IEEE Journal of Biomedical and Health Informatics*, Vol. pp, Issue. 9,page 1,TITB-00020-2013

[16] World guide to IDF BRIDGES 2012.

[17] www.biomedcentral.com/1741-7015/9/103

[18] P Kasemthaweesab, W Kurutach, "Association Analysis of Diabetes Mellitus (DM) with Complication states Based on Association Rules", *Proc. of the 7th*

IEEE Conference on Industrial Electronics and Applications, pp. 1453-1457, 2012.

[19] Chien-Lung Chan, Chien – Weichen, Ban – Jhiune Liu, "Discovery of Association Rules in Metabolic Syndrome Related Diseases", *Proc. of the International Joint Conference on Neural Network*, pp 856-862, 2008.

[20] C M Velu, K R Kashwan, "Visual Data Mining Techniques for Classification of Diabetes Patients". *Proc. of the IEEE 3rd International Advance Computing Conference,* pp. 1070-1075, 2013.

[21] Xiao Fang, "Are You Becoming a Diabetic? A Data Mining Approach". *Proc. of the 6th International Conference on Fuzzy Systems and Knowledge Discovery*, Vol.5, pp 18-22, 2009 .

[22] Joseph L. Breault, Colin R. Goodall, Peter J. Fos," Data mining a diabetic data warehouse", *Journal of Artificial Intelligence in Medicine 26 (2002)*, pp. 37–54, 2002

[23] Sonukumari, Archana Singh, " A data mining approach for the Diagnosis of Diabetes Mellitus" Proc. of the *7th International Conference on Intelligent Systems and Control*,pp. 373-375,2013.

[24] Asha A Aljarullah, "Decision Tree Discovery of the Diagnosis of Type II Diabetes". *Proc. of the International Conference on Innovations in Information Technology*, pp. 303 -307, 2011 .

[25] Alexis Marcano – Cedefio, Diego Andina, " Data Mining for the diagnosis of type II diabetes", WAC 2012 1569535615.

[26] Eleni Georga, Vasilios Protopappas, Alejandra Guillen, Guiseppe Fico, " Data Mining for Blood Glucose Prediction and Knowledge Discovery in Diabetic Patients: The METABO diabetes Modeling and Management System". *Proc. of the 31st Annual International Conference of the IEEE EMBS,* pp. 5633- 5636, 2009.

[27] www.wekipedia.org/wiki/Metabolic-syndrome

[28] Andina D, "Testing Artificial Metaplasticity in MLP applications", *Proc. of the IEEE International Conference on Systems, Man and Cybernetics(SMC 2009) ,* Pp. 4256- 4261, 2009.