A Review Paper on Track Diabetes using Classification Technique on an Android Application

Ms. K Sowjanya
Computer Science and Engineering
Rungta College of Engineering and Technology
Bhilai, Chhattisgarh, India

Ms. Chaitali Choudhary (Supervisor)
Computer Science and Engineering
Rungta College of Engineering and Technology
Bhilai, Chhattisgarh, India

***Abstract*— Medical Data Mining is the process of extracting hidden pattern from medical data. This paper develops an Android application using IF-THEN rules extracted from the Decision Tree (Classification Technique), which is constructed using classification algorithms (like C4.5). Once the tree was constructed, the production rules are obtained from that tree. In order to improve the coverage capacity and accuracy, rules from one or more trees are used to develop a predictive model on Android. This app provides user an easy tool which can predict diabetes and accordingly the user can maintain his/her diet and enhance some regular activities, this app provides some emotional support to users. The complete rule set which was extracted from the decision trees, was manually checked and conflicts are resolved which arises when the same consequents of the rule classify to different antecedent (class label).**

Keywords- Diabetes; Decision Tree; Classification Rules; C4.5 algorithm.

I. INTRODUCTION

Diabetes mellitus (DM) is the seventh leading cause of death in the United States and causes many serious complications including hypoglycemia, blindness, renal failure, cardiovascular disease [5].

Diabetes mellitus (DM) which is also often referred to as diabetes is a disorder that is caused by deficient production of insulin or by lacking ability to use insulin, for this reason glucose levels in the blood increases [2].

There are mainly four types of diabetes namely: Prediabetes, Type-1, Type-2 and Gestational Diabetes. The first type is Prediabetes which is described by the presence of blood glucose levels that are little bit higher than normal but not yet high enough to be diagnosed as diabetes [3].Type 1 DM results, when the body fails or unable to produce insulin. This diabetes previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes". Type 2 DM occurs because of insulin resistance, a condition in which cells fail or unable to use insulin properly, sometimes also occurs with an absolute insulin deficiency. This type of diabetes previously referred to as non-insulin-dependent diabetes mellitus (NIDDM). Another form of Diabetes is Gestational diabetes and occurs in a situation when pregnant women without a diabetes diagnosis previously develop a high blood glucose level [10].

Diabetes is classified into two more subgroups based on the specific mechanisms that cause the disease: diabetes in which genetic vulnerability is clarified at the DNA level and diabetes associated with other diseases or conditions [1].

II. CLASSIFICATION TECHNIQUES

Classification is perhaps the most familiar and most popular data mining technique. Various categories of classification as described below.

A. *Statistical-Based Algorithms*

- Regression- Regression problems deal with estimation of an output value based on input values. It is mainly used for Numeric Prediction. Regression can be performed using many different types of techniques, including Neural Networks. Regression actually takes a set of data and fits the data to a formula.

- Bayesian Classification - This classification technique based on Bayes Theorem, assuming that the contribution by all attributes are independent and that each contributes equally to the classification problem, a simple classification scheme called naïve Bayes classification has been proposed that is based on Bayes rule of conditional probability.

B. *Distance-Based Algorithms*- Each item that is mapped to the same class may be thought of as more similar to the other items in that class than it is to the items found in other classes. Therefore, similarity (or distance) measures may be used to identify the “alikeness” of different items in the database.

- Simple Approach- If we have a representative of each class, we can perform classification by assigning each tuple to the class to which it is most similar.

- *K*-Nearest Neighbor- One common classification scheme based on the use of distance measures is that of the K nearest neighbors (KNN). The KNN technique assumes that the entire training set includes not only the data in the set but also the desired classification for each item. In effect, the training data become the model. When a classification is to be made for a new item, its distance to each item in the training set must be determined, only the K closest entries in the training set are considered further. The new item is then placed in the class that contains the most items from the set of K closest items.

C. *Decision Tree Based Algorithm*- The decision tree approach is most useful in classification problems. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied each tuple in the data base and results in a classification for that tuple.

- ID3- The ID3 (Iterative Dichotomizer) technique to building a decision tree is based on information theory and attempts to minimize the expected numbers of comparisons. The basic idea of the induction algorithm is to ask questions whose answer provides the most information.
- C4.5 - C4.5 is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees and rule derivation.
- CHAID- CHAID attempts to stop growing the tree before overfitting occurs, whereas the above algorithms generate a fully grown tree and carry out pruning as post-processing step. In that sense CHAID avoids the pruning phase.
- CART- CART (Classification And Regression Tree) is one of the popular methods of building decision trees in the machine learning community. CART builds a binary decision tree by splitting the records at each node, according to a function of a single attribute.

III. LITERATURE REVIEW

Data mining algorithms can be trained from past/known examples in clinical data and model the frequent times non-linear relationships between the independent and dependent variables. Classification is the often used technique in medical or in Health Care Systems data mining.

Comparing accuracy of algorithms: G.Visalatchi et al. compares the performances of various algorithms (SVM, k-NN, Naïve Bayes, C4.5 and Apriori) are evaluated based on the accuracy of predicting Diabetes using Indian Pima Diabetes (PIDD) data set. Algorithm used and Accuracy in percentage (%) is given as SVM: 74.8%, KNN: 78%, Naïve Bayes: 75%, C4.5: 86% and Apirori recorded 75% of accuracy [8].

Comparing Time taken to classify: VelidePhani Kumar and Lakshmi Velide examines the accuracy and time taken to give the results/output for various algorithms (C4.5, Jrip, Decision trees, Naïve Bayes, Neural- network, KNN, Fuzzy logic and Genetic Algorithms) are evaluated. Algorithms accuracy and corresponding time taken is given as: Jrip: 96.54%, 765min; C4.5: 100%, 658min; Decision Trees recorded 98.48% of accuracy and takes 875min of time for classification. Similarly Naïve Bayes: 95.85%, 845min; Neural Networks: 97.85%, 956min respectively. This evaluation is done with the help of TANGARA tool [9].

Bayesian Network: Mukesh kumara et al. developed Bayesian Network for the classification of Diabetes data set into three class labels: Pre-Diabetes, Non-Diabetic and Diabetic. The data is first pre-processed and make available to the WEKA tool, which classifies the data. This classifier gives output with 99.51% of accuracy [10].

Comparing error rate of algorithms: K. Rajesh et al. compares the error rate of various algorithms for classifying the Diabetic Data Set and concluded that the C4.5 algorithm gives minimum error rate of 0.0937 as compared to other algorithms (C-RT, CS-RT, ID3, K-NN, LDA, NAÏVE BAYES, PLS-DA and SVM) [11].

Logistic Regression: Jay Pedersen et al. developed a Logistic Regression model in which the Data Set classified either YES or NO class labels. In this case, a variable which is YES when a person has diabetes and NO when they do not is the variable of interest. The model can predict a YES or NO result for the variable of interest based on other variables [4].

P. Radha et al. evaluated the performance of five algorithms C4.5, SVM, k-NN, PNN, and BLR. Comparison of performance of data mining algorithms based on computing time, precision value, the data evaluated using 10 fold Cross Validation error rate, bootstrap validation and accuracy. TANGARA tool is used to achieve the best results.[12].

Rough Set: Joseph L. Breault uses Rough Set Theory and Rosetta Software for the analysis of Diabetic Data Set. The accuracy of predicting diabetic status on the PIDD (Pima Indian Diabetes Dataset) was 82.6% on the initial random sample [13].

SVM (Support Vector Machine): V. Anuja Kumari and R.Chitra proposed a classification model using Support Vector Machine with Radial Basis Function. The performance parameters such as the classification accuracy (78%), sensitivity (80%), and specificity (76.5%) of the SVM and RBF have found to be high thus making it a good option for the classification process [14]. Arwa Al-Rofiyee et al. proposed a method in which, the PIMA India diabetes dataset is converted to the ARFF format to be inserted in WEKA software. The WEKA (Classifier) is used to train and test the model that was build based on the training results, under the Multi-layer perceptron function [15].

DT and Regression model: Pardha Repalli in [16] built a Decision trees and Regression models to predict the binary target variable.

Murat Koklu and Yavuz Unal in [18] developed a decision support system for diagnosis of illness that make use of data mining and artificial intelligence classifier algorithms namely Multilayer Feed Forward Perceptron, Naïve Bayes Classifier and C4.5. Pima Indian dataset (PIDD) of UCI Machine Learning Repository was used.

Hybrid Technique (DT with K-Means clustering): Asha Gowda Karegowda et al. proposed a new methodology for improving the quality of the dataset: (PIDD). This methodology developed a model consists of two stages. In the first stage, the K-means clustering is used to identify and remove incorrectly classified instances. The continuous data is converted to categorical form by estimated width of the desired intervals, based on the opinion of diabetes expert. In the next stage a fine tuned classification is done using Decision tree C4.5 [19].

Divya Jain in [20] compares two Data Mining Tools namely: WEKA and TANAGRA for the comparison of classification accuracy of C4.5 algorithm for Diabetes Dataset.

Mohd. Mahmood Ali et al. [21] proposed methodology to improve the efficiency of C4.5 algorithm, and also the deficiencies exits in C4.5 classifier analyzed using WEKA tool by constructing decision trees with different types of data sets, especially with large data sets, resulted in few classification rules using C4.5 classifier.

Extracting Rules from DT: J. R. Quinlan in provides the information about importance of Rules extraction from Decision Tree [22].

Hybrid Model (DT with Agglomerative Hierarchical Clustering): Norul Hidayah Ibrahim et al. suggest one Hybrid Model to predict Diabetes. They developed the new hybrid model by exploring Agglomerative Hierarchical Clustering and Decision Tree Classifier on Pima Indians Diabetes dataset and concluded that the performance accuracy of the Decision Tree Classifier against the same classifier augmented with Hierarchical Clustering provides improved results [23].

Sam Drazin and Matt Montag [24] discusses applications of the Weka interface, the given data sets were tested using the J48 decision tree-inducing algorithm (Weka implementation of C4.5), which was published by Ross Quinlan in 1993.

Dr. Neeraj Bhargava et al. [25] focused on J48 algorithm which is used to create Univariate Decision Tree. They also discuss about the idea of multivariate decision tree with process of classifying instance by using more than one attribute at each internal node.

Mobile apps in medical field: C. Lee Ventola provides the benefits and uses of mobile apps in medical field. The use of mobile devices by health care professionals (HCPs) has transformed many aspects of clinical practice. Mobile devices have become common place in health care systems, leading to rapid growth in the development of medical software applications (apps) for these platforms. The ability to download medical apps on mobile devices has made a wealth of mobile clinical resources available to HCPs. Medical devices and apps are already invaluable tools for Health professional as well as for patients, and as their features and uses increase, they are expected to become more widely incorporated into nearly every aspect of clinical exercise [26].

Gunther Eysenbach et al. [27] proposes that the smartphone has a very bright future in the world of health care and medicine, while doctors, HCPs, engineers, and others alike continue to contribute more ingenuity to this dynamic field. Murray Aitken in [28] said that, creation of app products which aid users in condition management and provide emotional support and other patient success stories.

IV. CONCLUSION FROM LITERATURE REVIEW

Various algorithms are used for classification of diabetes data. Some authors compared the accuracy of the various algorithms, some considered the error rate and other focuses on the time taken to classify the data. Some authors uses *hybrid methods* to improve the accuracy and some pre-processed the data to increase the accuracy. The overview of the conclusion drawn from the literature review is given by the *Tables* 1-4 as shown below:

TABLE 1

Algorithms	**Paper References**				
	Murat Koklu et al. [18]		**V. Anuja Kumari et al. [14]**		
	Accuracy	*Time taken (sec)*	*Accuracy*	*Sensitivity*	*Speci ficity*
C4.5	73.823%	0.08			
SVM with Radial Basis function			78%	80%	76.5 %
Naïve Bayes	76.302%	0.01			
Neural Network	75.130%	1.13			

TABLE 2

Algorithms	**Paper References**				
	K. Rajesh et al. [11]	**P. Radha et al. [12]**			
	Error Rate	*Accuracy*	*Time Taken(ms)*	*Positive Recall*	*Error Rate*
C4.5	0.0938	86%	550	0.38	0.28
SVM	0.2253	74.8%	546	0.368	0.29
Naïve Bayes	0.2461				
KNN	0.1966	78%	640	0.473	0.34

TABLE 3

Algorithms	**Paper References**		
	Joseph L. Breault [13]	**Asha Gowda Karegowda [19]**	**Norul Hidayah Ibrahim et al. [23]**
	Accuracy	*Accuracy*	*Accuracy*
Rough Set Theory	82.6%		
C4.5 with K-Means		93.33%	
C4.5 with hierarchical Clustering			80.8%

TABLE 4

Algorithms	Paper References			
	G.Visalatchi et al. [8]	**VelidePhani et al. [9]**		**Mukesh kumara et al. [10]**
	Accuracy	*Accuracy*	*Time taken(min)*	*Accuracy*
C4.5	86%	100%	658	
SVM	74.8%			
Naïve Bayes	75%	95.85%	845	99.51%
KNN	78%			
Neural Network	-	97.85%	956	

V. PROPOSED METHODOLOGY

Figure 1. provides the overview of the proposed methodology used to implement Android Application. C4.5 algorithm is used to construct the Decision Tree. The whole data set is divided into two parts as *Training Dataset and Testing Dataset.*

At first data is collected from two districts, this data is cleaned and pre-processed in order to develop a Decision Tree using C4.5 algorithm. After the construction of Decision Tree, the production rules are extracted from this tree, this IF-THEN rule set is used to develop the Android Application which can predict the chances and risk level of Diabetes.

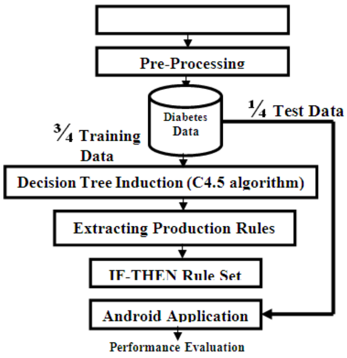

Figure 1 Proposed Methodology for developing Android Application

Note: The reader may wonder why we use the decision tree at all, instead of developing rules directly from the diabetes data set. Working from the tree has two major advantages: First, most interesting classification tasks involve attributes with continuous values which must be formed into tests by the development of appropriate thresholds (e.g. Plasma <=127 or above). Secondly, even a long path in a decision tree typically involves only a small proportion of the possible attributes. The space of potential rules is thus shrunk from $0(2^{23})$ to $0(2^{9})$ with a corresponding reduction in computational load [22].

VI. CONCLUSION

From this study we conclude that the C4.5 algorithm is best suited for our work, because its accuracy is good considering the overall performance and also the rules extracted from tree are most suitable and simple for developing the Android Application. Thus, the proposed methodology in this paper develops an Android application for predicting and managing Diabetes using IF-THEN rules extracted from Decision Tree constructed using C4.5 decision tree induction algorithm for Diabetes data set. This application helps the user by providing the result as the chances of Diabetes and helps in providing some guidance related to their regular activities and diet.

REFERENCES

[1] Abdullah A. Aljumah, Mohammed Gulam Ahamad and Mohammad Khubeb Siddiqui, "Application of data mining: Diabetes health care in young and old patients,"1319-1578 @ 2013 Production and hosting by Elsevier B.V. on behalf of King Saud University.
[2] Emirhan Gulcin Yildirim, Adem Karahoca and Tamer Ucar, "Dosage planning for diabetes patients using data mining methods," @2010 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of the Guest Editor, from www.elsevier.com/locate/procedia.
[3] http://www.diabetes.co.uk/pre-diabetes.html.
[4] Jay Pedersen, Fangyao Liu, Fahad Alfarraj and Harry Ngondo, "Examining Disease Risk Factors by Mining Publicly Available Information," © 2013 The Authors. Published by Elsevier B.V., from www.sciencedirect.com.
[5] Myung-kyung Suh1, Jonathan Woodbridge1, Tannaz Moin et.al, "Dynamic Task Optimization in Remote Diabetes Monitoring Systems," 2012 IEEE Second Conference on Healthcare Informatics, Imaging and Systems Biology.
[6] Jiawei Han, Micheline Kamber and Jian Pei, "DATA MINING Concepts and Techniques," Morgan Kaufmann Publishers, An imprint of Elsevier.© 2012, by Elsevier Inc.
[7] Sunita Soni, "Lecture notes," @2011 BIT, Durg .
[8] G.Visalatchi, S.J Gnanasoundhari and Dr.M.Balamurugan, "A Survey on Data Mining Methods and Techniques for Diabetes Mellitus," International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February- 2014, pg. 100-105.
[9] VelidePhani Kumar and Lakshmi Velide, "A DATA MINING APPROACH FOR PREDICTION AND TREATMENT OF DIABETES DISEASE," @ IJSIT, 2014, 3(1),073-079.
[10] Mukesh kumari, Dr. Rajan Vohra and Anshul arora, "Prediction of Diabetes Using Bayesian Network," @(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5174-5178.
[11] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis," @International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
[12] P. Radha and Dr. B. Srinivasan, "Predicting Diabetes by cosequencing the various Data Mining Classification Techniques," @IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.
[13] Joseph L. Breault, MD, MPH, MS, "Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?," @Department of Health Systems Management, Tulane University Department of Family Practice, Alton Ochsner Medical Foundation joebreault@tulanealumni.net.
[14] V. Anuja Kumari and R.Chitra, "Classification Of Diabetes Disease Using Support Vector Machine," @International Journal of Engineering Research and Applications, Vol. 3, Issue 2, March -April 2013, pp.1797-1801.
[15] Arwa Al-Rofiyee, Maram Al-Nowiser, Nasebih Al-Mufadi and Dr. Mohammed Abdullah AL-Hagery, "USING PREDICTION METHODS IN DATA MINING FOR DIABETES DIAGNOSIS,"@Qassim University, College of Computer, Department of IT, Kingdom of Saudi Arabia.
[16] Pardha Repalli, "Prediction on Diabetes Using Data mining Approach" @2013:Oklahoma State University.
[17] WEKA, by university of Waikato, http://www.cs.waikato.ac.nz/ml/weka/
[18] Murat Koklu and Yavuz Unal, "Analysis of a Population of Diabetic Patients Databases with Classifiers," @World Academy of Science, Engineering and Technology International Journal of Medical, Health, Pharmaceutical and Biomedical Engineering Vol:7 No:8, 2013.
[19] Asha Gowda Karegowda, Punya V, M.A.Jayaram and 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.
[20] Divya Jain, "A Comparison of Data Mining Tools using the implementation of C4.5 Algorithm," @International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Impact Factor (2012): 3.358.

[21] Mohd. Mahmood Ali, Mohd. S. Qaseem, Lakshmi Rajamani and A. Govardhan, "EXTRACTING USEFUL RULES THROUGH IMPROVED DECISION TREE INDUCTION USING INFORMATION ENTROPY," @International Journal of Information Sciences and Techniques (IJIST) Vol.3, No.1, January 2013.

[22] J. R. Quinlan, "Generating Production Rules From Decision Trees," @Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 545 Technology Square, Cambridge MA 02139 USA.

[23] Norul Hidayah Ibrahim, Aida Mustapha, Rozilah Rosli and Nurdhiya Hazwani Helmee, "A Hybrid Model of Hierarchical Clustering and Decision Tree for Rule-based Classification of Diabetic Patients," @International Journal of Engineering and Technology (IJET).

[24] Sam Drazin and Matt Montag, "Decision Tree Analysis using Weka," @2012 Machine Learning – Project II, University of Miami.

[25] Dr. Neeraj Bhargava, Girja Sharma, Dr. Ritu Bhargava and Manish Mathuria, "Decision Tree Analysis on J48 Algorithm for Data Mining," @International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013. Available online at: www.ijarcsse.com.

[26] C. Lee Ventola, "Mobile Devices and Apps for Health Care Professionals: Uses and Benefits," @P T. May 2014; 39(5): 356–364. © 2014, MediMedia USA, Inc. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4029126/.

[27] Gunther Eysenbach, Rob Wu and Felasfa Wodajo, "The Smartphone in Medicine: A Review of Current and Potential Use Among Physicians and Students," @J Med Internet Res. 2012 Sep-Oct; 14(5): e128. Published online Sep 27, 2012.

[28] Murray Aitken, "Patient Apps for Improved Healthcare From Novelty to Mainstream, "@ oct-2013 IMS Institute for Healthcare Informatics, 11 Waterview Boulevard, Parsippany, NJ 07054 USA info @theimsinstitute.org and www.theimsinstitute.org.

[29] Arun K Pujari, “DATA MINING TECHNIQUES, ”© University Press (India) Private Limites 2001.