A Hybrid Prediction Model for Type 2 Diabetes Using K-means and Decision Tree

Wenqian Chen, Shuyu Chen, Hancui Zhang
College of Software Engineering, Chongqing University, Chongqing, China
cwqseven@gmail.com, netmobilab@126.com, zhc_813@126.com

Tianshu Wu
College of Computer Science, Chongqing University, Chongqing, China
wutianshu@cqu.edu.cn

Abstract—**Type 2 diabetes has a quite high incidence all over the world. For the prevention and treatment of Type 2 diabetes, early detection is demanded. Nowadays, data mining techniques are gaining increasing importance in medical diagnosis field by their classification capability. In this paper, a hybrid prediction model is proposed to help the diagnosis of Type 2 diabetes. In the proposed model, K-means is used for data reduction with J48 decision tree as a classifier for classification. In order to get the experimental result, we used the Pima Indians Diabetes Dataset from UCI Machine Learning Repository. The result shows that the proposed model has reached better accuracy compared to other previous studies that mentioned in the literature. On the basis of the result, it can be proven that the proposed model would be helpful in Type 2 diabetes diagnosis.**

Keywords-diabetes diagnosis; data mining; classification; K-means; decision tree

I. INTRODUCTION

Diabetes is one of the most common diseases in recent years, and its global prevalence is growing rapidly. It is a general term for heterogeneous disturbances of metabolism for which the main finding is chronic hyperglycemia. The cause is either impaired insulin secretion or impaired insulin action or both [1]. The chronic hyperglycemia of diabetes is associated with long-term damage, dysfunction, and failure of various organs, especially the eyes, kidneys, nerves, heart, and blood vessels. The vast majority of diabetes can be divided into two categories, viz. Type 1 and Type 2. The cause of Type 1 diabetes is an absolute deficiency of insulin secretion. On the other hand, Type 2 diabetes is much more prevalent, and the cause is a combination of resistance to insulin action and an inadequate compensatory insulin secretory response [2]. The most common form of diabetes is Type 2 diabetes [3].

According to the six edition of IDF (International Diabetes Federation) Diabetes Atlas, an astounding 382 million people are estimated to have diabetes, with dramatic increases seen in countries all over the world and Type 2 diabetes constitutes the majority of all diabetes [4]. As a consequence, Type 2 diabetes is a severe health issue for the whole world. If we could diagnose and prevent diabetes as early as possible, millions of lives might be saved.

Along with the great progress of information technology, we could make use of the vast amount of data in health care industry to help doctors diagnosing diabetes. Data mining could predict the future by modeling. In recent years, a large number of computational methods and tools for data analysis are available. Data mining has been widely applied to the medical field and played an important role in medical research. Hence, this paper proposes a hybrid diagnosis model which could predict Type 2 diabetes in using multiple data mining methods. This model could assist doctors and medical professionals in making decisions and improve diagnostic accuracy.

II. BACKGROUND

In this section of the paper, we shall discuss data mining and some of the data mining tools and methods.

A. Data mining

We live in a world where vast amounts of data are collected daily. The traditional method of turning data into knowledge relies on manual data analysis. As data volumes grow rapidly, this form of data analysis is slow, expensive, and subjective. The traditional method is becoming completely impractical in many fields and could not meet the need of data analysis [5]. Data mining, also known as knowledge discovery in databases (KDD) could meet this need by providing tools to discover knowledge from data. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically [6].

During the past decades, data mining has been applied to a variety of areas, such as marketing, finance (especially investment), fraud detection, manufacturing, telecommunications and many scientific fields, including the analysis of medical data [5]. As medical data volumes grow dramatically, there is a growing pressure for efficient data analysis to extract useful, task-oriented information from the enormous amounts of data [7]. Such information may play an important role in future medical decision-making.

B. Data mining tools

For the implementation of the proposed model, it is necessary to make use of some data mining tools. An efficient data mining tool could assist us in transforming the huge data

The work of this paper is supported by National Natural Science Foundation of China (Grant No. 61272399 and No. 61572090) and Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20110191110038).

978-1-5386-0497-7/17/$31.00 ©2017 IEEE

into useful information. In the past few years, there are many open-source data mining tools and software available for use, such as Waikato Environment for Knowledge Analysis (WEKA), TANAGRA, Rapidminer, Orange, KNIME etc. Among all these data mining tools, WEKA is one of the most popular and fully-functional tools [8]. Thus, we decided to use WEKA as our data mining tool.

WEKA is a Java based computer program for data mining and machine learning which was originally developed at the University of Waikato in New Zealand. WEKA offers four options for data mining, viz. Experimenter, command-line interface (CLI), Explorer and Knowledge flow. WEKA contains a large collection of the newest data mining and machine learning algorithms written in Java. It supports a diversity of standard tasks for data mining: data preprocessing, clustering, classification, regression, visualization and feature selection [9].

C. Data mining methods

Data mining is predicted to be one of the most revolutionary developments of the next decades. As a matter of fact, it was chosen as one of 10 emerging technologies that will change the world by the MIT Technology Review [10]. Researchers have been vigorously developing new data mining methodologies. Data mining methodologies should consider issues such as data uncertainty, noise, and incompleteness. Some data mining methods explore how user specified measures can be used to assess the interestingness of discovered patterns as well as guide the discovery process [11]. In this section, two of the common data mining methods that would be used in the proposed model are discussed.

1) K-means clustering algorithm: K-means has a rich and diverse history as it was independently discovered in different scientific fields by Steinhaus (1956), Lloyd (proposed in 1957, published in 1982), Ball & Hall (1965) and McQueen (1967). Although K-means was first proposed over 50 years ago, it is still one of the most widely used clustering algorithms. Ease of implementation, efficiency, simplicity and empirical success are the main reasons for its popularity [12]. The procedure of K-means follows a simple way to classify a given data set through a certain number of clusters (assume K clusters) fixed apriori. K-means algorithm randomly chooses K objects, representing the K initial cluster center. The following step is to take each point belonging to a given data set and associate it to the nearest center based on the closeness of the object with cluster center using Euclidean distance. When all the objects are distributed, it is time to recalculate new K cluster centers. The process would be repeated until there is no change in K cluster centers. K-means aims at minimizing an objective function known as squared error function that is given by the following [13].

$$J(C) = \sum_{k=1}^{k} \sum_{x_i \in C_k} \left\| x_i - \mu_k \right\|^2 \qquad (1)$$

2) Decision tree algorithm: In the last few years, a great number of algorithms have been developed for classification based data mining. Decision tree is an important classification algorithm in data mining. The main advantage of decision tree algorithms is that they are easy to construct and the resulting trees are readily interpretable. It is commonly used in various areas. Researchers have developed a variety of decision tree algorithms over a period of time with enhancement in performance and ability to handle different types of data. Popular decision tree algorithms including ID3, CART, C4.5, C5.0, J48 etc. C4.5 is developed by Ross Quinlan. It is an extension of Quinlan's earlier ID3 algorithm [14]. C5.0 and J48 are the improved versions of C4.5 algorithms. In the WEKA data mining tool, J48 algorithm is an open source Java implementation of the C4.5 algorithm. WEKA provides a number of options associated with tree pruning. J48 classifier creates a binary tree. By using this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for the tuple [15].

III. LITERATURE REVIEW

In recent years, predictive classification is one of the most essential and important tasks in data mining and machine learning. Its application to the medical diagnosis has received a strong boost due to earnest research activities in the medical big data field. Many researchers have highlighted the potential of predictive classification to provide decision support for doctors and medical professionals. Over the last few years, a great deal of research has been conducted on different datasets to predictive diabetes. Many of them showed good classification accuracy.

J. Pradeep Kandhasamy and S. Balamurali using data sample from UCI machine learning data repository to compare the performance of four common classifiers (J48 Decision Tree, K-Nearest Neighbors, Random Forest, and Support Vector Machines) to classify diabetes mellitus patients. The result shows that the J48 decision tree classifier achieves higher accuracy of 73.82 % than other three classifiers before data preprocessing and both KNN (k=1) and Random Forest show better performance than the other three classifiers after data preprocessing [16]. Rashedur M. Rahman and Farhana Afroz present a comparative study of different classification techniques by using three different data mining tools named WEKA, TANAGRA and MATLAB to analyze the performance of different classification algorithms for a large dataset. The study shows that the best algorithm in WEKA is J48graft with an accuracy of 81.33%, Naive Bayes classifier provides an accuracy of 100% in TANAGRA and ANFIS has 78.79% accuracy in MATLAB [17]. Xue-Hui Meng developed three predictive models (logistic regression, artificial neural networks and decision tree) then compare the performance by using 12 risk factors. The study suggests that the decision tree algorithm (C5.0) had the best classification accuracy of 76.13% [18]. The accuracy of data mining algorithms EM algorithm, KNN algorithm, K-means algorithm, amalgam

KNN algorithm and ANFIS algorithm is compared by Veena Vijayan V. and Aswathy Ravikumar. The experiment

shows that among these algorithms, amalgam KNN and ANFIS provides higher classification accuracy of 80% [19].

Asma A. AlJarullah conducts a diabetes prediction model by using the decision tree algorithm. In this study, Weka's J48 decision tree classifier was applied to the dataset to construct the decision tree model. The accuracy of the resulting model was 78.1768% [20]. Wei Yu presents a potentially useful alternative approach based on support vector machine (SVM) techniques that can be used to classify persons with and without diabetes. The study used the data from the U.S. National Health and Nutrition Examination Survey to develop SVM models two classification schemes, one is diagnosed or undiagnosed diabetes vs. pre-diabetes or no diabetes, the other is undiagnosed diabetes or pre-diabetes vs. no diabetes. The results show the area under the receiver operating characteristic (ROC) curve were respectively 83.5% and 73.2%. The result indicates SVM modeling is a promising classification approach for detecting common diseases like diabetes [21]. Mira Kania Sabariah, Aini Hanifa and Siti Sa'adah combine Classification and Regression Tree method (CART) and Random Forest (RF) to build the classification model that can be used in the early detection of Type 2 diabetes. The study shows that the average accuracy of the proposed model is 83.8%, which is higher than the single classifier CART [22].

IV. DATA SOURCE

In order to conduct the research, we used the Pima Indian Diabetes Data (PIDD) set, which is publicly available from UCI repository [23]. The dataset contains females with at least 21 years old of Pima Indian heritage living around Phoenix, Arizona. There are 768 records in the dataset, out of which 268 cases in class "tested positive for diabetes" and 500 cases for "tested negative for diabetes" with 376 records contain missing values. The purpose of this research is to predict whether a person would test positive by using the eight physiological measurements and medical test results given in the dataset. It is a two-class problem with class value 1 being interpreted as "tested positive for diabetes" while class value 0 being interpreted as "tested negative for diabetes". The attribute information present in the dataset has been given in following Table I.

V. PROPOSED MODEL

A. Working Principle

For the purpose of prediction, a prediction model was defined. The working principle of the proposed model has been shown in Fig. 1. It comprises four steps:

1) Data preprocessing: Replace the missing values and impossible values with mean.

2) Data reduction: Remove the incorrectly classified data by using K-means algorithm to cluster the dataset.

3) Classification: Constructing decision tree by using the reduced data.

4) Performance evaluation: Evaluate the performance by using some of the classifier evaluation metrics.

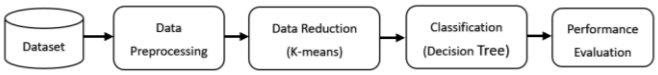

Figure 1. Block diagram of the proposed model

B. Data Preprocessing

The quality of the data is the key to the whole prediction model as it could influence the prediction result from the analysis. Hence, data preprocessing must be done before data analysis. The PIDD set contains a number of missing values and impossible values, such as 0 body mass index and 0 plasma glucose [17]. In this study, data preprocessing is done by replacing the missing values and impossible values with mean.

C. Data Reduction

Before the application of the classification algorithm, the clustering algorithm K-means implemented by WEKA is applied to remove the incorrectly classified samples. From the clustering result, we found 236 instances were incorrectly classified.

D. Classification

After the second step, it could be seen that 236 instances are incorrectly classified, these instances would be removed. Then, the classification algorithm could be applied. The J48 decision tree algorithm implemented by WEKA is used to build decision tree with 10-fold cross-validation method.

TABLE I. ATTRIBUTE INFORMATION

Number	Attribute	Mean	Standard Deviation	Type
1	Number of times pregnant	3.8	3.4	Numeric
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	120.9	32.0	Numeric
3	Diastolic blood pressure (mm Hg)	69.1	19.4	Numeric
4	Triceps skin fold thickness (mm)	20.5	16.0	Numeric
5	2-Hour serum insulin (mu U/ml)	79.8	115.2	Numeric
6	Body mass index (weight in kg/(height in m)^2)	32.0	7.9	Numeric
7	Diabetes pedigree function	0.5	0.3	Numeric
8	Age (years)	33.2	11.8	Numeric

It is important to test the validity and reliability of the model each time a model is constructed and trained. The result is usually optimistic when training and testing are performed on the same dataset since the training algorithm learns all the involved records. Hence, it is advisable to use an independent supplied dataset for testing. The decision tree was built up using 532 instances (obtained after data reduction of the 768 instances in which 236 were correctly classified by using the K-means algorithm) with 10-fold cross-validation.

E. Performance Evaluation

In this section, a number of measures for assessing how good or how accurate a classifier is at predicting the class label of tuples will be introduced [6].

1) Accuracy, sensitivity and specificity: Firstly, there are four additional terms we need to know that are used in computing many evaluation measures.

a) True positives (TP): The positive tuples that were correctly labeled by the classifier.

b) True negatives (TN): The negative tuples that were correctly labeled by the classifier.

c) False positives (FP): The negative tuples that were incorrectly labeled as positive.

d) False negatives (FN): The positive tuples that were mislabeled as negative.

In this study, the following equations are used to measure the accuracy, sensitivity and specificity.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$\text{Sensitivity} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

2) Confusion matrix: The confusion matrix is a useful tool for analyzing how well a classifier can recognize tuples of different classes. TP and TN tell us when the classifier is getting things right, while FP and FN tell us when the classifier is getting things wrong. For a classifier that has good accuracy, ideally most of the tuples would be represented along the diagonal of the confusion matrix with the rest of the entries being zero or close to zero [6]. A confusion matrix is shown in Fig. 2.

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	P'	N'	$P+N$

Figure 2. Confusion matrix

3) k-fold cross-validation: Cross-validation is a useful and generally applicable technique which is often employed in machine learning, including decision tree. For a k-fold cross-validation, each single example occurs exactly k-1 times as a training example. Hence, the time needed to compute the statistics of all test examples is reduced by a factor k-1 compared to running the original algorithm k times. The time needed to sort examples into child nodes is reduced by k-1 if the same test is selected in all folds, otherwise a smaller reduction occurs. Besides this speed-up, there are no changes in the computational complexity of the algorithm [24]. In this study, we used 10-fold cross-validation in the proposed model. It can reduce the bias associated with random sampling method.

VI. EXPERIMENTAL RESULTS

In the present study, we firstly replaced the missing values and impossible values by mean and then removed the incorrectly classified samples by using K-means clustering algorithm. After this step, we had 532 samples left. Finally, we applied the J48 decision tree algorithm with 10-fold cross-validation to the dataset. Using this model we obtained the final results. The confusion matrix of the proposed model is shown in Table II

TABLE II. CONFUSION MATRIX OF THE RESULTS

Actual Class \ Predicted Class	Yes	No
Yes	144	21
No	32	335

According to the confusion matrix above, we could figure out the accuracy, sensitivity and specificity of the proposed model are 90.04%, 87.27% and 91.28% respectively.

Since we got the experimental results of the proposed model, we could compare it to some of the other former presented classification models with evaluation measures. From Table III, it can be proven that the proposed model has a better accuracy [16] [17] [18] [19] [20] [22] [25] [26] [27].

TABLE III. . COMPARISON OF THE PROPOSED WORK WITH THE EXISTING WORKS

Method	Accuracy	Reference
J48	73.82%	Kandhasamy J P, Balamurali S (2015)
J48	81.33%	Rahman R M, Afroz F (2013)
ANFIS (MATLAB)	78.79%	Rahman R M, Afroz F (2013)
C5.0	76.13%	Meng X H (2013)
Amalgam KNN algorithm	>80%	V Vijayanv , A Ravikumar (2014)
ANFIS algorithm with adaptive KNN	80%	V Vijayanv , A Ravikumar (2014)
J48	78.1768%	Jarullah A A A (2011)
CART and Random Forest	83.8%	Sabariah M T M K (2015)
Predictive model based on H-TSVM	87.46%	Tomar D, Agarwal S (2014)
Naive Bayes	83.37%	K.R. Ananthapadmanaban, G. Parthiban (2014)
Agglomerative Hierarchical Clustering and J48	80.8%	Norul Hidayah Ibrahim (2013)
Proposed Model	90.04%	This Study

VII. CONCLUSION

According to the results shown in Table III, we can figure out that the proposed model has better accuracy than other

classification models for Type 2 diabetes in the related studies we mentioned in this paper. Comparing with the above results, it is clear to see the proposed model obtains quite promising results in classifying the possible Type 2 diabetes patients. With the rapidly growing demand for medical data analysis, the proposed model can be fairly useful to the researchers and doctors for their decision-making on the patients as by using such an efficient model they can make more accurate decisions.

There are also few aspects of this study that could be improved further or extended in the future. For instance, the proposed model is proposed to apply to Type 2 diabetes diagnosis which is a two-class classification problem. It would be interesting to see its behavior on multi-class classification problems. The proposed model is applied to numeric data only, we could improve the model to see its behavior on different types of medical data, such as images and signals. Moreover, for practical implementation, future work is required to assess the effectiveness of the proposed method with a larger amount of data.

ACKNOWLEDGMENT

The work of this paper is supported by National Natural Science Foundation of China (Grant No.61572090), Research Fund of the Doctoral Program of Higher Education of China (Grant No.20110191110038) and the Science and Technology Research Project of Chongqing Municipal Education Committee (grant no. KJ1704081).

REFERENCES

[1] Kerner W, Brückel J. Definition, classification and diagnosis of diabetes mellitus.[J]. Experimental and clinical endocrinology & diabetes : official journal, German Society of Endocrinology [and] German Diabetes Association, 2014, 122(7):384.

[2] Malchoff C D. Diagnosis and classification of diabetes mellitus.[J]. Diabetes Care, 2011, 34(Suppl 1):S62-S69.

[3] Rajendra A U, Tan P H, Subramaniam T, et al. Automated Identification of Diabetic Type 2 Subjects with and without Neuropathy Using Wavelet Transform on Pedobarograph[J]. Journal of Medical Systems, 2008, 32(1):21-29.

[4] Aguiree F, Brown A, Cho N H, et al. IDF Diabetes Atlas : sixth edition[J]. International Diabetes Federation, 2013.

[5] Fayyad, Usama M, PiatetskyShapiro, et al. From data mining to knowledge discovery: an overview[J]. Ai Magazine, 1996, 17(3):37-54.

[6] Han J, Kamber M. Data Mining: Concepts and Techniques[J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2012, 5(4):1 - 18.

[7] Sumathi S, Sivanandam S N. Introduction to Data Mining and its Applications[J]. Studies in Computational Intelligence, 2006, 26(25):236-238.

[8] Hasim N, Haris N A. A study of open-source data mining tools for forecasting[C]// International Conference on Ubiquitous Information Management and Communication. ACM, 2015:79.

[9] Jovic A, Brkic K, Bogunovic N. An overview of free software tools for general data mining[C]// International Convention on Information and Communication Technology, Electronics and Microelectronics. IEEE, 2014:1112-1117.

[10] Larose D. Data mining methods and models[M]. Wiley-Interscience, 2006.

[11] Liang M. Data Mining: Concepts, Models, Methods, and Algorithms[J]. IIE Transactions, 2004, 36(5):495-496.

[12] Anil K. Jain. Data clustering: 50 years beyond K-means ☆[J]. Pattern Recognition Letters, 2010, 31(8):651-666.

[13] Velmurugan T. Efficiency of k-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points[J]. International Journal of Computer Technology & Applications, 2012, 03(05):1758-1764.

[14] Patel, B. R., & Rana, K. K. (2014). A Survey on Decision Tree Algorithm For Classification. International Journal of Engineering Development and Research, 2(1), 1–5.

[15] Patil, T. R., & Sherekar, S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, 6(2).

[16] Kandhasamy J P, Balamurali S. Performance Analysis of Classifier Models to Predict Diabetes Mellitus ☆[J]. Procedia Computer Science, 2015, 47:45-51.

[17] Rahman R M, Afroz F. Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis[J]. Journal of Software Engineering & Applications, 2013, 06(3):85-97.

[18] Meng X H, Huang Y X, Rao D P, et al. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors[J]. Kaohsiung Journal of Medical Sciences, 2013, 29(2):93.

[19] Vijayanv V, Ravikumar A. Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus[J]. International Journal of Computer Applications, 2014, 95(17):12-16.

[20] Jarullah A A A. Decision tree discovery for the diagnosis of type II diabetes[C]// International Conference on Innovations in Information Technology. IEEE, 2011:303-307.

[21] Yu W, Liu T, Valdez R, et al. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes[J]. BMC Medical Informatics and Decision Making, 2010, 10(1):16.

[22] Sabariah M T M K, Hanifa S T A, Sa'Adah M T S. Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART)[C]// Advanced Informatics: Concept, Theory and Application. IEEE, 2015:238-242.

[23] UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

[24] Blockeel H, Struyf J. Efficient algorithms for decision tree cross-validation[M]. JMLR.org, 2003.

[25] Tomar D, Agarwal S. Predictive model for diabetic patients using hybrid twin support vector machine[C]//Proceedings of the 5th International Conferences on Advances in Communication Network and Computing (CNC'14). 2014: 1-9.

[26] Ananthapadmanabhan K R, Parthiban G. Prediction of Chances - Diabetic Retinopathy Using Data Mining Classification Techniques[J]. Indian Journal of Science & Technology, 2014, 7(10).

[27] Ibrahim N H, Mustapha A, Rosli R, et al. A hybrid model of hierarchical clustering and decision tree for rule-based classification of diabetic patients[J]. International Journal of Engineering & Technology, 2013, 5(5).