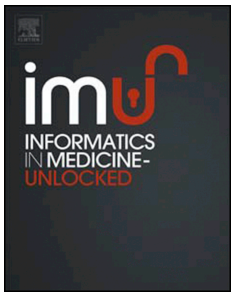

Type 2 diabetes mellitus prediction model based on data mining

Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang

PII: S2352-9148(17)30140-5

DOI: 10.1016/j.imu.2017.12.006

Reference: IMU 81

To appear in: *Informatics in Medicine Unlocked*

Received Date: 18 August 2017

Revised Date: 9 December 2017

Accepted Date: 10 December 2017

Please cite this article as: Wu H, Yang S, Huang Z, He J, Wang X, Type 2 diabetes mellitus prediction model based on data mining, *Informatics in Medicine Unlocked* (2018), doi: 10.1016/j.imu.2017.12.006.

Type 2 Diabetes Mellitus Prediction Model Based on Data Mining

Han Wu[1], Shengqi Yang[1,*], Zhangqin Huang[1], Jian He[1] and Xiaoyi Wang[1]

[1]Beijing Advanced Innovation Center for Future Internet Technology, Beijing Engineering Research Center for IoT Software and Systems, Beijing University of Technology, Beijing, 100124, China
syang@bjut.edu.cn

Abstract
Due to its continuously increasing occurrence, more and more families are influenced by diabetes mellitus. Most diabetics know little about their health quality or the risk factors they face prior to diagnosis. In this study, we have proposed a novel model based on data mining techniques for predicting type 2 diabetes mellitus (T2DM). The main problems that we are trying to solve are to improve the accuracy of the prediction model, and to make the model adaptive to more than one dataset. Based on a series of preprocessing procedures, the model is comprised of two parts, the improved K-means algorithm and the logistic regression algorithm. The Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis toolkit were utilized to compare our results with the results from other researchers. The conclusion shows that the model attained a 3.04% higher accuracy of prediction than those of other researchers. Moreover, our model ensures that the dataset quality is sufficient. To further evaluate the performance of our model, we applied it to two other diabetes datasets. Both experiments' results show good performance. As a result, the model is shown to be useful for the realistic health management of diabetes.

Keywords
Hybrid Prediction Model, Data Mining, Diabetes Mellitus

1. Introduction

Diabetes mellitus (DM) is a chronic disease that is characterized by high blood glucose. Nearly half of all diabetics have household heredity factors, which is one of the most important features of DM. Failure of the pancreas to produce enough insulin and the body's inefficient use insulin are both pathologic causes of DM. There are two types of DM. The pathogenesis of type 1 diabetes mellitus (T1DM) is that the pancreas secretes damaged β-cells, preventing it from lowering blood glucose level in time. Insulin resistance and insulin secretion deficiency are the pathogeneses of type 2 diabetes mellitus (T2DM), which is also called non-insulin dependent DM.

In the past 30 years of development in China, with rising number of diabetics, people have started to realize that this chronic disease has deeply impacted every family and everyone's daily life. There is an ascending trend in the proportion of diabetics in the general population, and the growth rate of male diabetics is higher than that of female diabetics, as shown in Fig. 1. According to some official statistics, the number of diabetics in China was nearly 110 million in 2017. This means that China has the largest diabetic population in the world.

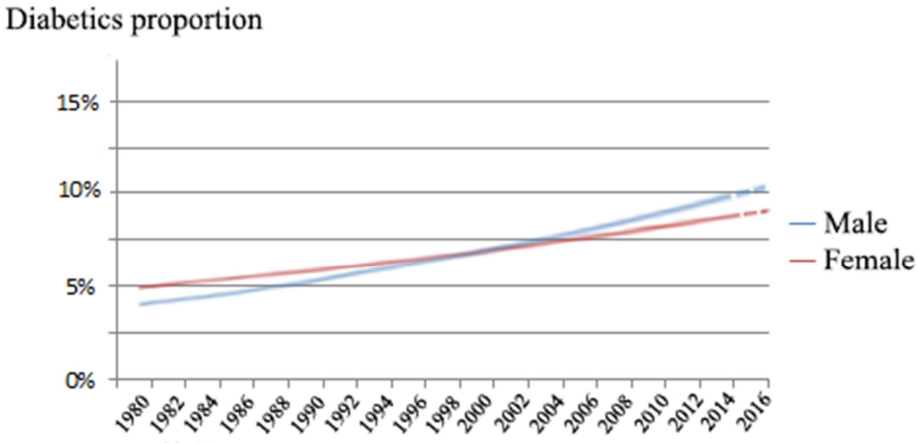

Figure 1. Trend of Diabetics proportion in China

The International Diabetes Federation (IDF) presents the latest data on DM in the Diabetes Atlas (Seventh Edition) [1]. It shows that in 2015, the number of diabetics worldwide was close to 415 million. In terms of the population growth trend of diabetics, it predicts that the number will approach to 642 million, or one in ten adults.

In order to lower the morbidity and reduce the influence of DM, it is vital for us to focus on a high-risk group of people with DM. According to the latest World Health Organization (WHO) standard, the definitions of groups with a high risk of DM are as follows:

- Age $\geqslant$ 45 and seldom exercising
- BMI $\geqslant$ 24kg/m^2
- Impaired glucose tolerance (IGT) or impaired fasting glucose (IFG)
- Family history of DM
- Lower high-density lipoprotein cholesterol or hypertriglyceridemia (HTG)
- Hypertension or cardiovascular and cerebrovascular disease
- Gestation female whose age $\geqslant$30

In order to research the high-risk group of DM, we need to utilize advanced information technology. Therefore, data mining technology is an appropriate study field for us. Data mining, also known as Knowledge Discovery in Databases (KDD), is defined as the computational process of discovering patterns in large datasets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems [2]. The main purposes of these methods are pattern recognition, prediction, association, and clustering. Data mining contains a series of steps disposed automatically or semi-automatically in order to extract and discover interesting, unknown, hidden features from large quantities of data. The high quality of data and the properly applied method are two significant aspects of data mining.

Data mining has been successfully applied to various fields in human society, such as weather prognosis, market analysis, engineering diagnosis, and customer relationship management. However, the application in disease prediction and medical data analysis still has room for improvement. For example, every hospital possesses a plethora of patient's basic and medical information, and it is essential to revise, supplement, and extract meaningful knowledge from these data to support clinical analysis and diagnosis [3-4]. It is reasonable to believe that there are various valuable patterns and waiting for researchers to explore them.

As we all know, the number of diabetics is large, and it is continuously increasing. Additionally, most people know little about their health quality. Therefore, believe it is necessary to establish a model that can classify patients into either suspected patients or confirmed patients in 5 years from the first examination time for the high-risk DM group. In particular, we have focused on T2DM.

Section 2 presents the related work of data mining in the group of diabetics and potential patients. Section 3 details the experimental tools, dataset, and prediction model. Section 4 describes the results of the experiment. Section 5 discusses the results and the procedures of validation. Section 6 concludes the paper with some directions for future work.

2. Related Works

In recent years, using the data mining technique has been used with increasing frequency to predict the possibility of disease. Many algorithms and toolkits have been created and studied by researchers. These have highlighted the tremendous potential of this research field. In this section, a few important works that are closely related to the proposed issue are presented.

Based on several studies, we found that a commonly used dataset was the Pima Indians Diabetes Dataset from the University of California, Irvine (UCI) Machine Learning Database [5]. Patil [6] proposed a hybrid prediction model (HPM), which used a K-means clustering algorithm aimed at validating a chosen class label of given data and used the C4.5 algorithm aimed at building the final classifier model, with 92.38% classification accuracy. Ahmad [7] compared the prediction accuracy of multilayer perception (MLP) in neural networks against the ID3 and J48 algorithms. The results showed that a pruned J48 tree performed with higher accuracy, which was 89.3% compared to 81.9%. Marcano-Cedeño [8] proposed artificial metaplasticity on multilayer perceptron (AMMLP) as a prediction model for diabetes, for which the best result obtained was 89.93%. All the studies presented above used the same Pima Indians Diabetes Dataset as the experimental material. The Waikato Environment for Knowledge Analysis (WEKA) toolkit was the primary tool which most researchers chose.

In order to obtain more useful and meaningful data, we realized that the preprocessing methods and parameters should be chosen rationally. Vijayan V. [9] reviewed the benefits of different preprocessing techniques for predicting DM. The preprocessing methods were principal component analysis (PCA) and discretization. It concluded that the preprocessing methods improved the accuracy of the naive Bayes classifier and decision tree (DT), while the support vector machine (SVM) accuracy decreased. Wei [10] analyzed risk factors of T2DM based on the FP-growth and Apriori algorithms. Guo [11] proposed the receiver operating characteristic (ROC) area, the sensitivity, and the specificity predictive values to validate and verify the experimental results.

On the basis of an effective prediction algorithm, we need an appropriate way to make the model convenient for everyone [12]. We found that Sowjanya [13] had developed an android application-based solution to overcome the deficiency of awareness about DM in his paper. The application used the DT classifier to predict diabetes levels for users. The system also provided

information and suggestions about diabetes. It used a real world dataset collected from a hospital in the Chhattisgarh state of India. Shi et al. [14] considered that preventing T2DM should be directed toward individuals. Therefore, they focused on establishing a diabetes risk assessment model and developed a diabetes risk score system based on mobile devices.

Improving algorithm will be one main job of our paper. There are some papers focusing on improving the K-means algorithm. Juntao Wang [15] presented an improved K-means algorithm using noise data filter. Yanhui Sun [16] proposed a method to improve the selection of initial centers for k-means clustering based on extended Frobenius-norm (Efros) distance. And Shunye Wang [17] showed an improved k-means clustering algorithm with variance which selected the initial cluster centers using the Huffman tree structure. Most papers optimized the initialized procedure of cluster center.

For those people at risk of developing DM, it was necessary to develop a series of grading forecasting standards [18]. Chandrakar and Dr. Saini [19] proposed the Indian Weighted Diabetic Risk Score (IWDRS) as a diabetes screening tool to solve the problem of undetected pre-diabetes and late diagnosis. Han and Luo [20] proposed the pair-wise and size-constrained K-means (PSCKmeans) method to screen the high-risk population of DM. The method provided a tool for risk stratification of clinical disease.

In summary, some studies of algorithm comparison and model establishing for DM prediction have been accomplished by these related works. However, the prediction accuracy and data validity were not high enough for realistic application. Besides, most models proposed by other researchers could only perform well in one specific dataset but not adapt to various datasets. We need to propose a novel prediction model for higher accuracy and adapt to more datasets. Therefore, we chose the same Pima Indians Diabetes Dataset and the same WEKA toolkit for further research. And two more datasets we collected were using to test the usability and adaptation of our model.

3. Model and Algorithm

This section is comprised of the dataset description, the preprocessing procedure, and the classification algorithm. All the experimental processes have been completed using the WEKA toolkit. The proposed model is shown in Fig. 2.

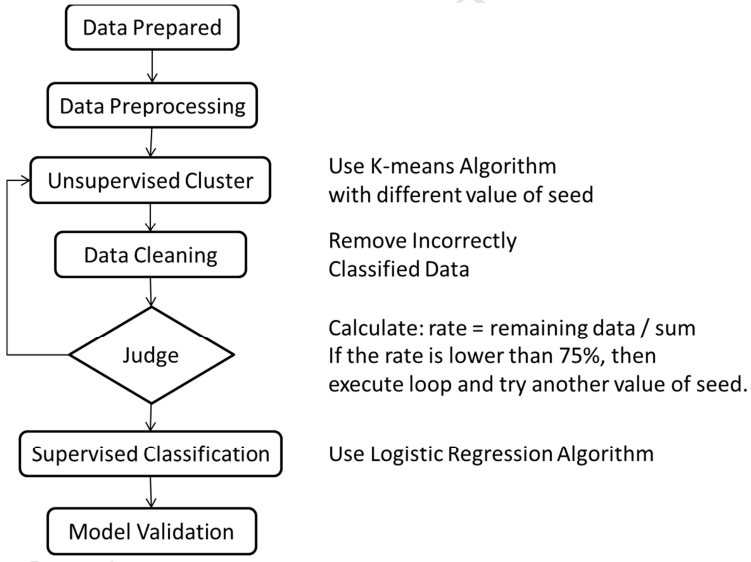

Figure 2. Algorithm Model

3.1 Data Mining Toolkit

WEKA is a free and non-commercial toolkit. It consists of standard machine learning and data mining algorithms, which are based on the JAVA environment. Using these preprocessing, classifying, clustering, associating algorithms, and the visual interface, we were able to obtain useful knowledge from databases easily and conveniently. Parts of those algorithms have been selected to establish the prediction model for T2DM.

In recent years, utilizing data mining algorithms in medical predictive analysis has increased due to earnest research in related areas. Over the last few years, several researchers have posited that it is possible to acquire clinically assistive supports and predictive models from basic patient data [21-23]. Most papers published in the field of disease predictive analysis for DM are aimed at improving accuracy. Some researchers have obtained considerable results by using this WEKA toolkit and the Pima Indian Diabetes dataset. However, the accuracy has room for improvement.

Extensive research has also been done on Pima Indian diabetes disease diagnosis, and the results obtained are presented in Table □ [24]. We used the preprocessing method which introduced in section 3.3 to deal with the original dataset and then

simulated the same experiments as other researchers'. After that, we updated the data in Table □. Most of the values of accuracy increased.

TABLE □ THE VALUES OF ACCURACY OF CLASSIFICATION MADE ON PIMA INDIAN DIABETES DATASET

Method	Accuracy（%）
Discrim	77.5
MLP	73.8
Logdisc	78.2
SMART	76.8
BayesNet	74.7
NaiveBay	74.9
RandomForest	76
J48	76.7
SGD	76.6
SMO	77
Backprop	75.2
RBF	75.7
LMT	76.6

3.2 Dataset Description

The Pima Indian Diabetes Dataset consists of information on 768 patients (268 tested_positive instances and 500 tested_negative instances) coming from a population near Phoenix, Arizona, USA. Tested_positive and tested_negative indicates whether the patient is diabetic or not, respectively. Each instance is comprised of 8 attributes, which are all numeric. These data contain personal health data as well as results from medical examinations. The detailed attributes in the dataset are listed as follows, and Table Ⅱ shows some samples extracted from the dataset.

- Number of times pregnant (preg)
- Plasma glucose concentration at 2 hours in an oral glucose tolerance test (plas)
- Diastolic blood pressure (pres)
- Triceps skin fold thickness (skin)
- 2-hour serum insulin (insu)
- Body mass index (bmi)
- Diabetes pedigree function (pedi)
- Age (age)
- Class variable (class)

3.3 Data Preprocessing

The quality of the data, to a large extent, affects the result of prediction. This means that data preprocessing plays an important role in the model [25]. The WEKA toolkit contains many kinds of filters for preprocessing purposes. In this study, we have selected some appropriate methods to optimize the original dataset.

First, we have analyzed each attribute's medical implication and its correlation to DM. We determined that the number of pregnancies has little connection with DM [6]. Therefore, we transformed this numeric attribute into a nominal attribute. The value 0 indicates non-pregnant and 1 indicates pregnant. The complexity of the dataset was reduced by this process.

Second, there are some missing and incorrect values in the dataset due to errors or deregulation. Most of the inaccurate experimental results were caused by these meaningless values. For example, in the original dataset, the values of diastolic blood pressure and body mass index could not be 0, which indicates that the real value was missing. To reduce the influence of meaningless values, we used the means from the training data to replace all missing values.

TABLE Ⅱ SAMPLES OF DATASET

preg	plas	pres	skin	insu	bmi	pedi	age	class
1	89	66	23	94	28.1	0.167	21	tested_negative
0	137	40	35	168	43.1	2.288	33	tested_positive
3	78	50	32	88	31	0.248	26	tested_positive

2	197	70	45	543	30.5	0.158	53	tested_positive
1	189	60	23	846	30.1	0.398	59	tested_positive
5	166	72	19	175	25.8	0.587	51	tested_positive
0	118	84	47	230	45.8	0.551	31	tested_positive
1	103	30	38	83	43.3	0.183	33	tested_negative
1	115	70	30	96	34.6	0.529	32	tested_positive
3	126	88	41	235	39.3	0.704	27	tested_negative

After the above steps were applied, the unsupervised normalize filter for attribute was used to normalize all the data into the section [0, 1] by using (1), where x' is the mean or average value for the variable and s is the standard deviation for the variable. Value is the new normalized value. This avoids the complexity of calculation and accelerates the speed of the operation.

$$\text{Value} = \frac{value - x'}{s} \tag{1}$$

3.4 Data Classification

The model consists of double-level algorithms. In the first level, we used the improved K-means algorithm to remove incorrectly clustered data. The optimized dataset was used as input for next level. Then, we used the logistic regression algorithm to classify the remaining data.

3.4.1 Improved K-means Cluster Algorithm

Cluster analysis aims at partitioning the observations into disparate clusters so that observations within the same cluster are more closely related to each other than those assigned to different clusters [26]. The K-means is one of the most popular cluster algorithms. It is a typical distance-based cluster algorithm, and the distance is used as a measure of similarity, i.e., the smaller distance between objects shows the greater similarity. Fig. 3 shows a graphic procedure of the K-means algorithm, and the procedures of the K-means Cluster algorithm are as follows:

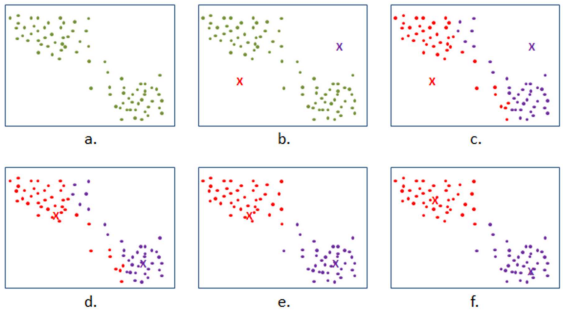

Figure 3. Procedures of the K-means algorithm

1) Show all objects (step a). Select K from provided N as the number of initial cluster center (step b). In Fig. 3b, the value of K is 2, and we use the '×' to present the categories.

2) Calculate distance between each object and cluster center. Cluster every object to the nearest cluster according to the distance using (2) extracted from [27] (step c).

$$S_i^{(t)} = \{x_p : \left\|x_p - m_i^{(t)}\right\|^2 \le \left\|x_p - m_j^{(t)}\right\|^2 \forall j, 1 \le j \le k\} \tag{2}$$

3) Recalculate every cluster center to verify whether they are changed using (3) extracted from [24] (step d).

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \epsilon S_i^{(t)}} x_j \tag{3}$$

4) Circulate step 2 and step 3 until the new cluster center is the same as the original one, i.e., convergence and end of the algorithm (step e and f).

In this study, we select 2 as the value of K because the 'Class' variable contains two results. We used the processed data after the preprocessing. Table III shows that cluster0 means negative class and cluster1 means positive class. One of the most important problems about K-means algorithm in the Weka toolkit is that the initial seed value is produced randomly and we need to set the value of seed according to our experience. The seed value directly affects the result of clusters. In order to avoid the deviation of experimental results caused by randomness of seed value, we take some steps. The first step is that we insert a program to record and sort the value called 'Within cluster sum of squared errors' by ascend order. In every experiment, a seed corresponds to the value called 'Within cluster sum of squared errors'. The smaller the value, the better the result. We record ten thousand values corresponding to the seed value from one to ten thousand. Those high quality seed value will be used first in the second step. So the initial value of the seed we chose in this experiment was 100. The second step is that we insert a loop at the

end of the algorithm. We removed those incorrectly clustered data and calculated the rate using the formula expressed as (4). If the rate was higher than 75%, then we moved to the next level. Otherwise, it should exit the loop and try another seed value. If an appropriate seed value could not be found to make the rate higher than 75% after 10 thousand loops or 60 seconds, we used the most proximate rate and corresponding seed for moving to the next level.

$$\text{rate} = \frac{\text{remaining data}}{\text{sum}} \tag{4}$$

After the removal procedure, we obtained 589 correctly classified patients, which all served as input to the logistic regression algorithm.

TABLE □ RESULT OF THE 2-MEANS CLUSTER OF THE INITIAL DATASET

No.	Label	Count
1	cluster0	458
2	cluster1	310

3.4.2 Logistic Regression Algorithm

The classification algorithm aimed to establish a model that can map data items to a given category, based on the existing data. It was used to extract significant data items from the model or to predict the tendency of data. In most cases, the dependent variable of the logistic regression algorithm is binary-classification. It means that the logistic regression algorithm is always used to solve two-category problem. The main purpose of our experiment is to predict whether one person is diabetic or not, which is a typical binary-classification problem. Besides, the logistic regression algorithm is always used in data mining, disease automatic diagnosis and economic prediction, especially predicting and classifying of medical and health problem. In conclusion, we decided to use the logistic regression as one part of our proposed model. The logistic regression algorithm is based on the linear regression model expressed as (5).

$$P=\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_m x_m \tag{5}$$

The classification problem is very much like the linear regression problem. Linear regression problem can only predict a continuous value. It maintains consistent sensitivity throughout the real number field. The predictive value of the classification problem can only be 0 or 1, so we may set a critical point. The output is 1 if the value is greater than the threshold, otherwise the output is 0. The output variable range of logistic regression is always between 0 and 1. Logistic regression is a regression model that reduces the prediction range and limits the prediction value to [0, 1]. Based on linear regression, the logistic regression adds a layer of sigmoid function (non-linearity). The features are first summed linearly and then predicted using the sigmoid function. The main formulas of the logistic regression algorithm are shown in (6), (7), and (8).

$$\Pr(Y=+1|X) \sim \beta \cdot X \text{ and } \Pr(Y=-1|X)=1-\Pr(Y=+1|X) \tag{6}$$

$$\downarrow \sigma(x) := \frac{1}{1+e^{-x}} \in [0,1] \text{（the sigmoid function）} \tag{7}$$

$$\Pr(Y=+1|X) \sim \sigma(\beta \cdot X) \text{ and } \Pr(Y=-1|X)=1-\Pr(Y=+1|X) \tag{8}$$

In this study, we have two categories, i.e., the positive group and the negative group. The Y indicates that the patient is diabetic. X independent variables represent the 8 attributes in the original dataset. Every dependent variable X is assigned a coefficient value called β representing the weight. After being analyzed by the logistic regression algorithm, the dataset showed every variable's value of weight. Different weights represent diverse correlation between X and Y. Once the regression model has been settled, it is efficient to input new data and predict whether the outcome is positive or negative. We set the logistic regression algorithm as the final step. The output and result are discussed in the next chapter.

4. Experimental Result

Using the WEKA toolkit, it was convenient for us to study the result of the experiment through a visualized interface. We analyzed and evaluated our model based on the following aspects. The result is shown in Fig. 4.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         562               95.416  %
Incorrectly Classified Instances        27                4.584  %
Kappa statistic                          0.8975
Mean absolute error                      0.0947
Root mean squared error                  0.2093
Relative absolute error                 20.8655 %
Root relative squared error             43.9386 %
Total Number of Instances              589

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.982    0.098    0.950      0.982    0.965      0.899    0.979     0.990     tested_negative
               0.902    0.018    0.964      0.902    0.932      0.899    0.979     0.933     tested_positive
Weighted Avg.  0.954    0.070    0.954      0.954    0.954      0.899    0.979     0.970

=== Confusion Matrix ===

   a   b   <-- classified as
 377   7 |   a = tested_negative
  20 185 |   b = tested_positive
```

Figure 4. The result of the experiment

4.1 K-fold Cross Validation

K-fold cross-validation is a method we frequently use to verify the performance of a model. In this study, we used 10-fold cross validation. The initial sample was divided into 10 sub-samples. Each separate sub-sample was retained as the validation data, while the other 9 samples were used to train. The proposed model was trained and tested 10 times. The advantage of this method is that it reduces the bias associated with the random sampling method [5].

4.2 Detailed Accuracy

In general, the process of prediction contains four different results called true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The confusion matrix displays these four results of this study in Table □. Column A presents the tested positive results, and column B presents the tested negative results. The first row shows the predicted results for the positive class, and the second row shows the predicted results for the negative class.

TABLE □ CONFUSION MATRIX

A	B	Classified
377	7	Predicted Positive
20	185	Predicted Negative

From the outcome of detailed accuracy, we present some significant indicators as follows.

The precision is calculated by (9). In this experiment, the value was 0.954.

$$\text{Precision}=\frac{TP}{TP+FP} \tag{9}$$

The recall, also known as the specificity, is calculated by (10). In this experiment, the value was 0.954.

$$\text{Recall}=\frac{TP}{TP+FN} \tag{10}$$

The Mathews correlation coefficient (MCC) is used as a measure of the quality of binary classifications, calculated by (11). In this experiment, the value was 0.899.

$$\text{MCC}=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{11}$$

The ROC area is a graphical plot that illustrates the performance of a binary classifier system as shown in Fig. 5. In this experiment, the value was 0.979.

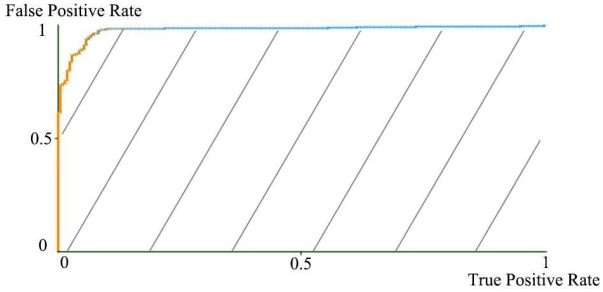

Figure 5. The ROC area

4.3 Kappa Statistic

The kappa statistic is a significant parameter to judge the consistency of the model. It compares the result of proposed model with a result generated by the randomly classified method. The value of the kappa statistic was between 0 and 1. The value close to 1 presents the expected effect of the model, while 0 means invalid. The equation of the kappa statistic is shown in (12), (13), and (14).

$$K = [P(A) - P(E)]/[1 - P(E)] \tag{12}$$

$$P(A) = (TP + TN)/N \tag{13}$$

$$P(E) = [(TP + FN) * (TP + FP) * (TN + FN)]/N^2 \tag{14}$$

The kappa value in this experiment was 0.8975, which means the proposed model attains great consistency.

5. Discussion

In this study, the model was evaluated based on the precision, recall, ROC area, and kappa statistic discussed above. Furthermore, the prediction accuracy was the most significant factor. In this section, we compare the results with those of some published works and apply our model in two relevant datasets.

5.1 Model Validation

In order to show that our model's accuracy of prediction has achieved a certain level of improvement, we compared our results with some researchers' experiments using the same dataset. After being processed by preprocessing and classified algorithms, the remaining 589 data only contain 27 instances which are classified into incorrect classes. The accuracy was up to 95.42%. The others' results are listed in Table □. Patil [6] showed an accuracy of 92.38%, which is the closest accuracy to ours, but the result came from a smaller number of samples. Only 433 instances left after they used K-means algorithm and deleted some incorrectly classified instances, while we obtained 589 correctly classified data after used improved K-means algorithm. Marcano-Cedeño [8] proposed a model named AMMLP, which achieved an accuracy of 89.93%. The confusion matrix shown in their paper only contained 308 instances, though it had 75.92% specificity and 97.5% sensitivity. Ahmad [7] used pruned and unpruned J48 algorithms to present accuracies of 89.3% and 86.6%. None of these experiments achieved a higher accuracy than ours.

TABLE □ COMPARISON WITH OTHERS' EXPERIMENTS

Method	Accuracy	Reference
Our Proposed Model	95.42%	This Paper
HPM	92.38%	B.M. Patil [6]
AMMLP	89.93%	Alexis Marcano-Cedeño [8]
J48 (pruned)	89.3%	Aliza Ahmad [7]
J48 (unpruned)	86.6%	Aliza Ahmad [7]
Hybrid model	84.5%	Humar Kahramanli [28]
MLP	81.9%	Aliza Ahmad [7]
Logistic	78.2%	Weka
J48	76.7%	Weka
SGD	76.6%	Weka
ELM	75.72%	Rojalina Priyadarshini [29]
NaiveBay	74.9%	Weka
BayesNet	74.7%	Weka
CART	72.8%	Ster & Dobnikar
KNN	67.6%	Statlog

5.2 Evaluated by New Datasets

5.2.1 Dataset Provided by Dr. Schorling

In order to provide more evidence to demonstrate the prediction accuracy and adaptability of our model, we applied the model in a new diabetes dataset, which was donated by Dr. Schorling from the Department of Medicine of the University of Virginia School of Medicine. It contains 1,046 instances divided into two opponent class. We choose 12 significant attributes from 19 original attributes according to comparison with the attributes of the Pima Indians Diabetes Dataset and some clinical experience. These attributes are shown as follows: total cholesterol, stabilized glucose, high-density lipoprotein (HDL), cholesterol HDL ratio, glycosylated hemoglobin, age, gender, height, weight, systolic blood pressure, diastolic blood pressure, and waist-hip ratio. The expanded coverage showed the great advantage of this dataset. The provided waist-hip ratio is a more

credible factor for diabetes research [30]. It is more reliable for containing both systolic blood pressure and diastolic blood pressure. The confusion matrix is displayed in Table □.

TABLE □ CONFUSION MATRIX

A	B	Classified
625	57	Predicted Positive
23	161	Predicted Negative

The results are shown in Table □. We use some algorithms integrated in the Weka toolkit to test our proposed model with this new dataset. All the data prove that the proposed model is suitable for predicting DM based on this new dataset.

TABLE □ RESULT OF THE NEW DATASET

Item	Value	Method	Accuracy
Prediction Accuracy	0.907	Our model	0.907
Precision	0.916	RandomForest	0.79
Recall	0.964	MLP	0.78
MCC	0.752	BayesNet	0.77
ROC Area	0.957	J48	0.72
Kappa Statistic	0.752	Logistic	0.72

5.2.2 Dataset Collected from Online Questionnaire

Our proposed model has been shown to have a high accuracy for predicting diabetes. Considering the large number of DM patients in China, which is mentioned in chapter □, we collected more basic healthy parameters. The questionnaire we designed consists of 14 contributes: age, gender, pregnant, family factor, BMI, sleep time, sleep quality, snoring, diuresis, hunger, smoking and drinking, blood pressure, blood glucose, and OGTT. The dataset contains 384 instances which are divided into two groups, 68 positive and 316 negative. It was meaningful for us to understand the practicability of our proposed model by applying it in a realistic dataset of Chinese populations. The confusion matrix of the experiment result is displayed in Table Ⅷ.

TABLE Ⅷ CONFUSION MATRIX

A	B	Classified
49	4	Predicted Positive
10	291	Predicted Negative

After being processed by the steps of preprocessing and classification, the dataset revealed some significant results, which are demonstrated in Table Ⅸ. We use some algorithms integrated in the Weka toolkit to test our proposed model with this new dataset. The result shows as below. The predictive accuracy was approximately 94%, which proves the proposed model is reliable and effective.

TABLE Ⅸ RESULT OF THE NEW DATASET

Item	Value	Method	Accuracy
Prediction Accuracy	0.939	Our model	0.935
Precision	0.925	RandomForest	0.896
Recall	0.929	BayesNet	0.88
MCC	0.786	Logistic	0.859
ROC Area	0.962	J48	0.859
Kappa Statistic	0.786	MLP	0.854

6. Conclusion and Future Work

This paper aimed to establish an appropriate prediction model for the high-risk T2DM group. Based on a number of researchers' experiences, we proposed a novel model, which consists of double-level algorithms, i.e., the improved K-means and logistic regression algorithms. In order to make a valid comparison with others' results, it was necessary to conduct this model using the WEKA toolkit and use the same Pima Indian Diabetes Dataset. Proper filters were utilized to improve the validity and rationality of the dataset. The proposed model that consisted of both cluster and class method ensured the enhancement of prediction accuracy. In Section 4, another realistic dataset provided by Dr. Schorling was used to test and verify the model. Our proposed model has proven to be appropriate for predicting T2DM. One of our proposed model's benefits is that it avoids deleting overmuch original data. It ensures the high quality of experimental data. The other benefit is that our model can apply in the Pima Indian Diabetes Dataset as well as other various datasets. While the limitation is that it consumes more time during the part of preprocessing.

We described that some papers focus on improving K-means by optimizing the initialized procedure of cluster center in Section 2. But our improved model is based on the purpose of predicting DM2 and matches up with the logistic regression algorithm. It assures less time consuming and maximum retention of original data. Although the improved model is not so complicated, it attained well effect according to plenty of experiments.

The main problems we solved are improving accuracy of prediction model and making the model to adapt to different datasets. In this paper, we conclude that our proposed model showing higher prediction accuracy than other researchers' experimental results. And the improved K-means algorithm we proposed contributed a lot to the prediction model. Moreover, there are two more dataset applied in our proposed model and all of them obtained well effect.

For future work, it is necessary to bring in hospital's real and latest patients' data for continuous training and optimization of our proposed model. The quantity of the dataset should be large enough for training and predicting [31-32]. Some advanced algorithms and models should be applied in the study of DM. Grading forecasting standards are also necessary for potential diabetes patients. Developing a series of rules and standards is a valid method to prevent people from developing DM. Based on that, a more effective model for predicting DM and grading potential patients is presented. This will help to lower the growth rate of diabetes and eventually decrease the risk of developing DM.

It is more convenient and efficient for people to obtain an application about health management of DM on their mobile devices [33-37]. We are currently developing an application that will provide reasonable and rational health suggestions to the high-risk group. Diabetes patients can conveniently use this application to test their blood glucose level, blood pressure, and heart rate. Furthermore, this medical data will be saved in a database for further procedures about data visualizing and model optimization. This will not only help people understand their health conditions, but will also help them create a healthy lifestyle.

References

[1] International Diabetes Federation (IDF) DIABETES ATLAS (Seventh Edition), 2015.

[2] http://en.wikipedia.org/wiki/Data_mining#cite_note-acm-1.

[3] Riccardo, B., & Blaz, Z. (2008). Predictive data mining in clinical medicine: Current issues and guidelines, International Journal of Medical Informatics, 77, 81–97.

[4] Mechelle Gittens, Reco King, Curtis Gittens and Adrian Als, Post-diagnosis Management of Diabetes through a Mobile Health Consultation Application, 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom).

[5] http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.

[6] B.M. Patil, Hybrid prediction model for Type-2 diabetic patients. Expert Systems with Applications 37 (2010) 8102–8108.

[7] Aliza Ahmad and Aida MustaphaH, Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus. ICDIPC 2011, Part I, CCIS 188, pp. 537–545, 2011.

[8] Alexis Marcano-Cedeño, Joaquín Torres, and Diego Andina, A Prediction Model to Diabetes Using Artificial Metaplasticity. IWINAC 2011, Part II, LNCS 6687, pp. 418–425, 2011.

[9] Veena Vijayan V. and Anjali C., Decision Support Systems for Predicting Diabetes Mellitus –A Review. Proceedings of 2015 Global Conference on Communication Technologies (GCCT 2015).

[10] Zhe Wei, Guangjian Ye and Nengcai Wang. Analysis for risk factors of type 2 diabetes mellitus based on FP-growth algorithm. China Medical Equipment, 2016. 13(5):45-48.

[11] Yirui Guo. Application of artificial neural network to predict individual risk of type 2 diabetes mellitus. Journal of Zhengzhou University, 2014.49(3):180-183.

[12] Shuaishuai Li, Enke Zhang, Min Li and Wei Pan, Research on the Effectiveness of Application of Diabetes Management APP, China Medical Devices, 2015. Vol 30. No.08.

[13] Ms. K Sowjanya, MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices. 2015 IEEE International Advance Computing Conference (IACC).

[14] Gang Shi, Shanshan Liu and Ding Ye, Design and Implementation of Diabetes Risk Assessment Model Based On Mobile Things, 2015 7th International Conference on Information Technology in Medicine and Education.

[15] Juntao Wang and Xiaolong Su, An improved K-Means clustering algorithm, 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN).

[16] Yanhui Sun, Liying Fang and Pu Wang, Improved k-means clustering based on Efros distance for longitudinal data, 2016 Chinese Control and Decision Conference (CCDC).

[17] Shunye Wang, Improved K-means clustering algorithm based on the optimized initial centroids, 2013 3rd International Conference on Computer Science and Network Technology (ICCSNT).

[18] Phattharat Songthung and Kunwadee Sripanidkulchai, Improving Type 2 Diabetes Mellitus Risk Prediction Using Classification, 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE).

[19] Omprakash Chandrakar, Dr. Jatinderkumar R. Saini. Development of Indian Weighted Diabetic Risk Score (IWDRS) using Machine Learning Techniques for Type-2 Diabetes, ACM COMPUTE '16, October 21-23, 2016, Gandhinagar, India.

[20] Longfei Han, Senlin Luo. An Intelligible Risk Stratification Model based on Pairwise and Size Constrained Kmeans, 2016 IEEE Journal of Biomedical and Health Informatics.

[21] Aruna Pavate and Nazneen Ansari, Risk Prediction of Disease Complications in Type 2 Diabetes Patients Using Soft Computing Techniques, 2015 Fifth International Conference on Advances in Computing and Communications.

[22] Naganna Chetty, An Improved Method for Disease Prediction using Fuzzy Approach, 2015 Second International

Conference on Advances in Computing and Communication Engineering.
[23] Saad Masood Butt and Karla.FelixNavarro, Using Mobile Technology to improve Nutritional Information of Diabetic Patient's, New Advances in Information Systems and Technologies(2016).
[24] Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine learning, neural and statistical classification. Ellis Horwood.
[25] Karim M. Orabi1, Yasser M. Kamal, and Thanaa M. Rabah. Early Predictive System for Diabetes Mellitus Disease. ICDM 2016, LNAI 9728, pp. 420–427, 2016.
[26] Guojun, G., Chaoqu, M. and Jianhong, W.. Data clustering theory algorithm and application (1st Ed.). ASA-SIAM.M (2007).
[27] https://en.wikipedia.org/wiki/K-means.
[28] Humar, K. and Novruz, A., Design of a hybrid system for the diabetes and heart diseases. Expert Systems with Applications, 35, 82–89, 2008.
[29] Rojalina Priyadarshini, Nilamadhab Dash and Rachita Mishra, A Novel approach to Predict Diabetes Mellitus using Modified Extreme Learning Machine.
[30] Murari Devakannan Kamalesh, Predicting the Risk of Diabetes Mellitus to Subpopulations Using Association Rule Mining. The International Conference on Soft Computing Systems 2016.
[31] Huan Li, Qi Zhang and Kejie Lu, Integrating Mobile Sensing and Social Network For Personalized Health-Care Application, Health care information systems(2016).
[32] Yan Luo, Charles Ling, Jody Schuurman and Robert Petrella, GlucoGuide: An Intelligent Type-2 Diabetes Solution Using Data Mining and Mobile Computing, 2014 IEEE International Conference on Data Mining Workshop.
[33] Rebecca Schnall and Marlene Rojas, A user-centered model for designing consumer mobile health (mHealth) applications (apps). Journal of Biomedical Informatics 60 (2016) 243–251.
[34] Md Abul Basar, Hassan Nomani Alvi, Gazi, A Review on Diabetes Patient Lifestyle Management Using Mobile Application, 18th International Conference on Computer and Information Technology (ICCIT), 21-23 December, 2015.
[35] Qasim Majeed, Hayder Hbail and Abdolah Chalechale, A Comprehensive Mobile E-Healthcare System, IKT2015 7th International Conference on Information and Knowledge Technology.
[36] Muhammad H. Aboelfotoh, Patrick Martin and Hossam S. Hassanein, A mobile-based architecture for integrating personal health record data, 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom).
[37] Ki-Hyun Kim, Ehsanul Kabir and Shamin Ara Jahan, The use of cell phone and insight into its potential human health impacts, Environ Monit Assess (2016) 188: 221.

Author Contributions

Han Wu and Shengqi Yang wrote the main manuscript. Zhangqin Huang, Jian He and Xiaoyi Wang optimized and reviewed the manuscript.

Additional Information

Competing Interests: The authors declare that they have no competing interests.