Research on Diabetes Prediction Method Based on Machine Learning

Research on Diabetes Prediction Method Based on Machine Learning

Jingyu Xue[1st,a], Fanchao Min[2rd,b], Fengying Ma[3nd,c*]

Qilu University of Technology, Jinan, Shandong, China

[a]e-mail: daoxiang.mo@163.com, [b]e-mail: Min_FC@163.com

[c*] Corresponding author: mafengy@163.com

Abstract—Diabetes mellitus (DM) is a metabolic disease characterized by high blood sugar. The main clinical types are type 1 diabetes and type 2 diabetes. Now, the proportion of young people suffering from type 1 diabetes has increased significantly. Type 1 diabetes is chronic when it occurs in childhood and adolescence, and has a long incubation period. The early symptoms of the onset are not obvious, which may lead to failure to detect in time and delay treatment. Long-term high blood sugar can cause chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves. Therefore, the early prediction of diabetes is particularly important. In this paper, we use supervised machine-learning algorithms like Support Vector Machine (SVM), Naive Bayes classifier and LightGBM to train on the actual data of 520 diabetic patients and potential diabetic patients aged 16 to 90. Through comparative analysis of classification and recognition accuracy, the performance of support vector machine is the best.

1. INTRODUCTION

Diabetes mellitus is a metabolic disease with chronic hyperglycemia caused by multiple causes. The main reason is due to defects in insulin secretion and/or function. The typical symptoms are "three more and one less", that is, polyuria, polydipsia, polyphagia and weight loss, which may be accompanied by skin itching. Long-term carbohydrate, fat, and protein metabolism disorders can also cause a variety of chronic complications, such as chronic progressive disease, hypofunction, and failure of tissues and organs such as eyes, kidneys, nerves, heart, and blood vessels. Acute and severe metabolic disorders can occur in severe conditions or under stress, such as diabetic ketoacidosis (DKA), hypertonic hyperglycemia syndrome. At present, the classification criteria proposed by the WHO Diabetes Expert Committee are:

1.1 Type 1 diabetes mellitus (T1DM):

1.1.1 Immune-mediated (1A): Acute type and slow type.

1.1.2 Idiopathic (1B): No evidence of autoimmunity.

1.2 Type 2 diabetes mellitus (T2DM):

From insulin resistance with progressive insufficient secretion of insulin to the main plate insulin resistance.

1.3 Special types of diabetes, gestational diabetes, etc.

The etiology and pathogenesis of diabetes are extremely complex, and different types have different causes. Environmental factors play an important role in the pathogenesis. Environmental factors mainly include viral infections, chemical poisons and dietary factors. Other symptoms that may indicate diabetes include blurred vision, shortness of breath, chest tightness, slow wound healing, numbness, skin itching, sudden confusion, coma, periodontal disease, and sexual dysfunction. Common complications include kidney disease, nervous system disease, diabetic retinopathy, and macrovascular disease.This could be a clear evidence that, according to WHO, the number of the diabetic patient had been sharply increased from 108 million in 1980 to 422 million in 2014[1]. The World Health Organization predicts that by 2030, diabetes will become the seventh leading cause of death in the world. The global prevalence of diabetes among adults over 18 years of age increased from 4.7% in 1980 to 8.5% in 2014[1]

In the era of big data, and large amounts of data hide various useful information and knowledge. In the prediction of diabetes, a large amount of data filtered through relevant data sources integrates into a data set for data mining. After that, people can classify and analyze this data set by machine learning algorithms. This not only allows patients to prevent and treat diabetes at an early stage through prediction, but also greatly saves time and money costs. This paper uses several algorithms to train the integrated data set, and finally proposes an appropriate algorithm that can use the early symptoms of patients to predict diabetes.

2. METHODOLOGY

The algorithm process proposed in this paper shown in Figure 1. First, the data set as input to the prediction algorithm, and then, though the evaluation model which is the method of introducing a confusion matrix to verify the classification accuracy of the algorithm. Finally, we get the algorithm with the highest accuracy in predicting diabetes.

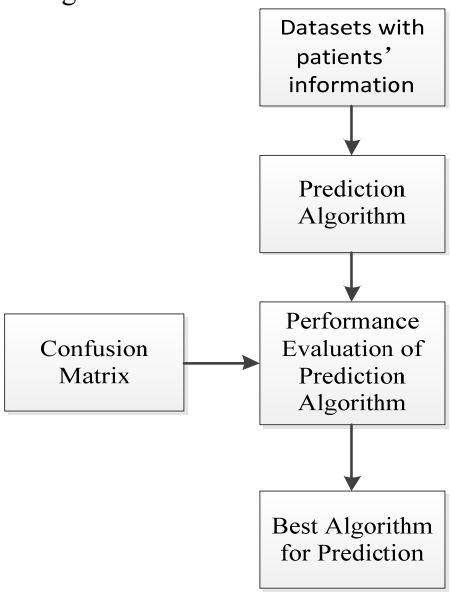


Fig.1 Process architecture

2.1 Dataset

The data set in this article comes from the open source standard test data set website UCI. The data set was obtained by direct questionnaires from 520 patients at the Sylhet Diabetes Hospital in Sylhet, Bangladesh, and was approved by doctors. The data set is divided into 17 attributes including age, gender, polyuria, depression, Sudden weight loss、Weakness、Polyphagia、Genital thrush、Visual blurring、Visual blurring、Itching、Irritability、Delayed healing、Partial paresis、Muscle stiffness、

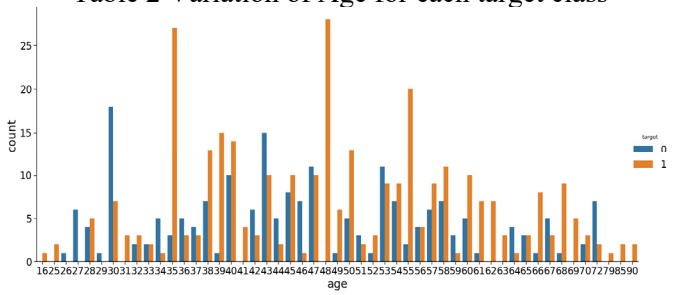
Alopecia、and Obesity.

Table 1 Description of attribute

	Attributes	Values
1	Age	16-90
2	Sex	1.Male, 0.Female
3	Polyuria	1.Yes, 0.No.
4	Polydipsia	1.Yes, 0.No.
5	Sudden weight loss	1.Yes, 0.No.
6	Weakness	1.Yes, 0.No.
7	Polyphagia	1.Yes, 0.No.
8	Genital thrush	1.Yes, 0.No.
9	Visual blurring	1.Yes, 0.No.
10	Itching	1.Yes, 0.No.
11	Irritability	1.Yes, 0.No.
12	Delayed healing	1.Yes, 0.No.
13	Partial paresis	1.Yes, 0.No.
14	Muscle stiffness	1.Yes, 0.No.
15	Alopecia	1.Yes, 0.No.
16	Obesity	1.Yes, 0.No.
17	Class	1.Positive, 0.Negative.

In Table 1, "1" is to indicate "diseased" and "positive", and "0" is to indicate "not diseased" and "negative". The above attributes are distributed by age among all surveyed patients as shown in Table 2.

Table 2 Variation of Age for each target class



2.2 Support Vector Machine (SVM)

SVM is a generalized linear classifier that performs binary classification of data according to supervised learning. Its decision boundary is the maximum-margin hyperplane for solving learning samples [2-4]. SVM uses the hinge loss function to calculate empirical risk and adds a regularization term to the solution system to optimize structural risk. It is a classifier with sparsity and robustness [3]. SVM can perform non-linear classification through the kernel method, which is one of the common kernel learning methods [5].

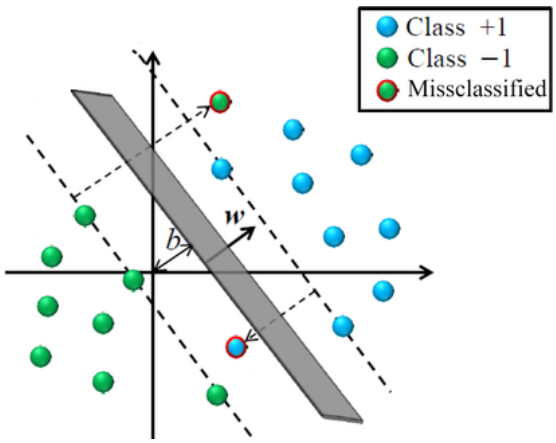


Figure 2. Support Vector Machine

SVM is an algorithm suitable for binary classification. Zayrit Soumaya [6] and others apply genetic algorithms and SVM to extract features from speech signals to detect some neurological diseases such as Alzheimer's disease, depression and Parkinson's disease. The best accuracy they got was 91.18%. Agrawal, Dewangan [7] and others used the data of 738 patients for experimental analysis. Combining the SVM with the current discriminant analysis algorithm, the best accuracy rate of is 88.10%. The classification capabilities of support vector machines are excellent, especially when a large number of features are involved.

2.3 Naïve Bayes Classifier

Naive Bayes classifier is a series of simple probability classifiers based on the use of Bayes' theorem under the assumption of strong (naive) independence between features. The classifier model assigns class labels represented by feature values to problem instances, and class labels are taken from a limited set. For the given item to be classified, the probability of each category appearing under the condition of the occurrence of the item is solved, whichever is the largest, and the category to be classified is considered to be. This prediction of the most likely class by probability is suitable for diabetic prediction. The specific classification formulas are shown in (1) to (4). Where x_p represents people who are at risk of diabetes, x_n represents people who are not at risk of diabetes, and X is the data set.

$$P(X|x_p) = \prod_{a=1}^{D} P(x_a|x_p) = P(x_1|x_p)P(x_2|x_p)\dots P(x_D|x_p) \tag{1}$$

$$P(X|x_n) = \prod_{a=1}^{D} P(x_a|x_n) = P(x_1|x_n)P(x_2|x_n)\dots P(x_D|x_n) \tag{2}$$

$$P(x_a|x_p) = \frac{Total\,(x_a|x_p)}{Total\ \ x_p} \tag{3}$$

$$P(x_a|x_n) = \frac{Total(x_a|x_n)}{Total\ \ x_n} \tag{4}$$

Here D is the attribute with D dimension.

2.4 LightGBM

LightGBM is a gradient Boosting framework that uses a learning algorithm based on decision trees. It can be said to be distributed and efficient, and has the following advantages: faster training efficiency,

low memory usage, higher accuracy, support for parallel learning, and can handle large-scale data. Compared with common machine learning algorithms, its speed is very fast. LightGBM uses histogram algorithm. The basic idea of the histogram algorithm is to discretize the continuous floating-point eigenvalues into k integers, and at the same time construct a histogram with a width of k. When traversing the data, use the discretized value as the index to accumulate statistics in the histogram. After traversing the data once, the histogram accumulates the necessary statistics, and then traverse to find the optimal value according to the discrete value of the histogram.

3. Result & Discussion

In order to compare the pros and cons of the classification models, it is necessary to provide metrics to evaluate the performance of the models. Here we divide the sample into four classes like true examples (True Positive, TP), false positive (FP), true negative examples (True Negative, TN), and false negative examples (False Negative, FN)[3]. Let TP, FP, TN, and FN respectively denote the corresponding number of samples, TP+FP+TN+FN=n, n is the sample size, and the confusion matrix of the classification result is shown in the following table 3.

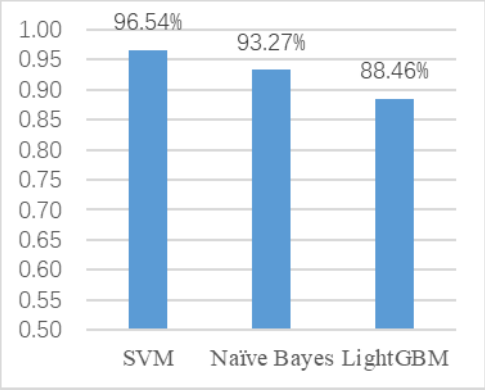
Table 3 Confusion Matrix

Real Classes	Forecasts	
	True Examples	False Examples
True Examples	TP	FN
False Examples	FP	TN

This article divides the characteristic results into two categories, using "1" for positive results and "0" for negative results. First, we split the data into two parts. In this experiment, the ratio of training set to prediction set is 80:20. Using the training set data for model to train, and then use the trained model and prediction set as input in the prediction component.

We summarize the results of the above three classification algorithms as shown in Table 2. Although the naive Bayes classifier is the most popular classification algorithm, the final accuracy rate on our data set is only 93.27%. SVM has the highest accuracy rate, with an accuracy rate of 96.54%. The accuracy of LightGBM is only 88.46%. This shows that the most suitable classification algorithm for diabetes prediction is SVM.

Table 4 Accuracy of classification algorithm



4. Conclusion

Although there is no clear research showing that there is an exact relationship between diabetes and age, there is a clear trend of younger diabetes now. Early detection of diabetes plays a vital role in treatment, and the emergence of machine learning has revolutionized the study of diabetes risk prediction. With the continuous advancement of data mining methods, we have studied various methods of diagnosing diabetes. We found that SVM has the highest accuracy through the confusion matrix evaluation test. However, this kind of research needs to be updated regularly with more instance data sets. Finally, we

can see that data mining algorithms through research, machine learning techniques and various other technologies have made outstanding contributions in the medical field and disease diagnosis. It is hoped that it can help clinicians make better judgments on disease status.

ACKNOWLEDGMENTS

This work was supported by the Shandong University Undergraduate Teaching Reform Research Project (Approval Number: M2018X078), and the Shandong Province Graduate Education Quality Improvement Program 2018 (Approval Number: SDYAL18088).The work was partially supported by the Major Science and Technology Innovation Projects of Shandong Province (Grant No.2019JZZY010731)

REFERENCE

[1] Diabetes, World Health Organization (WHO): 30 Oct 2018.
[2] Vapnik, V.. Statistical learning theory. 1998 (Vol. 3). . New York, NY：Wiley，1998：Chapter 10-11, pp.401-492
[3] Zhou Zhihua. Machine learning. Beijing: Tsinghua University Press，2016：pp.121-139, 298-300
[4] Li Hang. Statistical learning methods. Beijing: Tsinghua University Press, 2012: Chapter 7，pp.95-135
[5] Qin, J. and He, Z.S., 2005, August. A SVM face recognition method based on Gabor-featured key points. In Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on (Vol. 8, pp. 5144-5149). IEEE.
[6] Zayrit Soumayaa, Belhoussine Drissi Taoufiqa, Nsiri Benayadb, Korkmaz Yunusc, Ammoumou Abdelkrim, The detection of Parkinson disease using the genetic algorithm and SVM classifier, Elsevier Ltd Applied Acoustics:2021.doi:10.1016/j.apacoust.2020.107528
[7] Agrawal, P., Dewangan, A.: A brief survey on the techniques used for the diagnosis of diabetes-mellitus. Int. Res. J. Eng. Technol. (IRJET).02(03) (2015). e-ISSN: 2395-0056; p-ISSN: 2395-0072