Risk prediction of early diabetes mellitus based on combination model

Haoxin Tang, *Yi* Zhang, *Baolin* Xiang, *Mingkun* Liu, and *Junming* Hu, and Cheng Liu*
Sichuan Agricultural University, Dujiangyan, China

Abstract. Aiming at the current low pre-diabetes detection rate, this paper proposes a PSO-SVM model to assist doctors in identifying the risk of patients with pre-diabetes. The paper uses the Support Vector Machine as the verification algorithm, takes the radial basis kernel as the kernel function, uses the adaptive Particle Swarm Optimization algorithm to optimize the penalty factor and kernel parameters of the Support Vector Machine, and establishes a PSO-SVM model, finally compares the model with Neural Network, Logistic Regression, and Naive Bayes model, and use Sensitivity, Specificity indicators and ROC curve to evaluate model performance. Empirical analysis proves that the combined model proposed in this paper can effectively identify the risk of patients with prediabetes.

1 Introduction

Pre-diabetes is a transitional state between normal blood sugar and diabetes. It is a transitional stage and early warning signal for developing diabetes [1]. With the continuous improvement of the economic level, the prevalence of pre-diabetes continues to increase. The latest data on diabetes in China shows: the pre-diabetes detection rate is only 35.7%, and there are 148 million pre-diabetes patients; the pre-diabetes prevalence rate in my country is 3 % To 4%, of which 7.7% to 8.95% develop diabetes every year [2]. Therefore, how to improve the pre-diabetes detection rate and provide medical guidance and prevention and control for pre-diabetes patients as soon as possible have become an urgent problem for pre-diabetes prevention and control.

With the application of big data and the development of machine learning, patient medical data has gradually received attention from the industry. Through the processing, analysis and clustering of massive data, they can provide prediction and decision support for disease prevention and control and medical management. Fang Hongxia, Wei Jincai, Li Hongjie, etc. established a (GDM) risk assessment model to make it more effective in establishing gestational diabetes with routine clinical indicators [3]; Wei Zhe, Zhang Yugang, Shi Dongdong, etc. were based on grid search and cross-validation, used Support Vector Machine to diagnose and predicted diabetes complications [4]; Naz Huma, Ahuja Sachin used data mining algorithms and in-depth learning to predict diabetes based on Indian data sets [5]. It can be seen that adopting data mining analysis methods for risk identification of pre-diabetes can assist doctors in risk diagnosis.

* Corresponding author: liucheng@sicau.edu.cn

To sum up, the main research directions of data mining analysis methods in pre-diabetes prediction include: first, how to mine patient medical data and disease prediction related information, and second, how to reduce the training cost of the constructed model and improve its accuracy. Based on these, this paper proposes a method to use Support Vector Machine algorithm to build prediction model and Particle Swarm algorithm to optimize model performance.

2 Related methods

Support Vector Machine (SVM) is a nonlinear black box model based on statistical theory [6].It is a strong binary classification model dealing with nonlinear, small sample and high latitude problems in machine learning. The basic principle is to map the selected feature as input vector to a higher latitude space through a nonlinear mapping function. The best hyperplane of classification samples is found in the high latitude space mapped to.

The basic principle of Particle Swarm Optimization (PSO) is inspired by the cooperative behavior of birds in the process of foraging [7]. PSO is to design a massless particle with only two attributes of velocity and position to simulate as an individual, where the velocity represents the speed and direction of movement. Each particle is constantly moving in space. After a certain number of iterations, we can obtain the optimal space position of particles and the optimal solution of particle swarm.

In this paper, Particle Swarm Optimization algorithm is selected and combined with Support Vector Machine model. The parameters of Support Vector Machine model are optimized by using this algorithm, so that the performance of classification model is optimized. The PSO-SVM model was constructed to evaluate the individual characteristics of the data set and identify the risk index of patients with prediabetes. The specific implementation process is as follows:

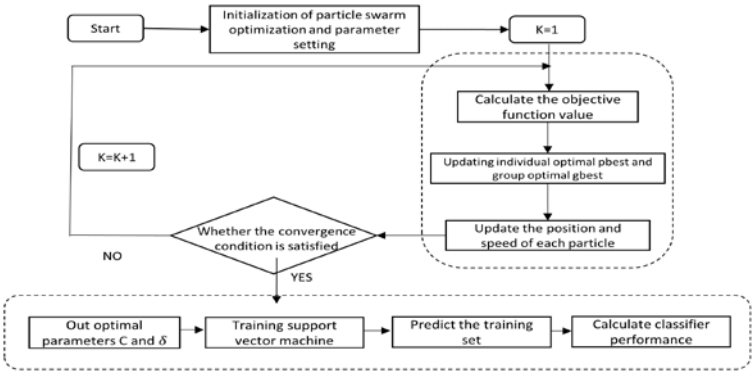

Fig. 1. PSO-SVM flow chart

1) The particle swarm of the PSO-SVM model is initialized and the parameters are set. According to the experimental needs, the number of particles, the maximum number of iterations and learning factor c_1 and c_2 are set, and the initial inertia weight is ω.

2) In the process of particle optimization, the spatial position of each particle can be regarded as a solution. It is necessary to update the particle velocity and position according to the optimal point and global optimal position, and evaluate the fitness of each particle.

3) After a round of iteration, the current fitness of each particle is compared with the $pbest_i$ position fitness, and the better position is regarded as the new $pbest_i$; for the whole group of particles, after the iteration, the position fitness of particles is better than the $gbest_i$ position fitness as the new $gbest_i$. And update the speed and position of each particle until

the maximum number of iterations is reached.

4) After optimization, the optimal parameters C and δ are brought into the SVM model to train the training set, and PSO-SVM model is obtained. The model is used to evaluate the data set, so as to predict the risk of diabetes.

3 Experimental process

3.1 Data sources

The original data uses the early diabetes risk data set in the machine learning database of the University of California, Irvine. The data set has a total of 520 samples. After preliminary analysis and generalization of the characteristics of the original data, A total of 14 features in each sample except age and gender are divided into two categories: main symptoms of diabetes and complications of diabetes.

Table 1. Feature classification.

	Main symptoms of diabetes	Complications of diabetes
Feature	Polyuria, Polydipsia, Polyphagia Weight loss, Fatigue, Loss of vision, Obesity, Hair loss	Itching, Irritability, Partial hemiplegia, Muscle tension, Genital fungal infection, Delayed healing

3.2 Risk assessment of pre-diabetes

256 cases of diabetic patients (80% of total diabetic patients) and 160 cases of non-diabetic patients (80% of non-diabetic patients) were randomly selected as the training set of the model. The radial basis function kernel was used as the kernel function of support vector machine, and particle swarm optimization algorithm was used to optimize. The accuracy of support vector machine model was set as the objective function, the number of particles was 45, the maximum number of iterations was 130, and the learning factors c_1=c_2=0.5, the initial inertia weight ω = 0.8, and the Linear Decreasing Weight (LDW) [8] strategy is introduced to improve the global optimization ability of the initial optimization and the local optimization ability of the final optimization. The optimized parameters of SVM are C= 4.663, δ= 2.859. The optimization process is shown in the figure:

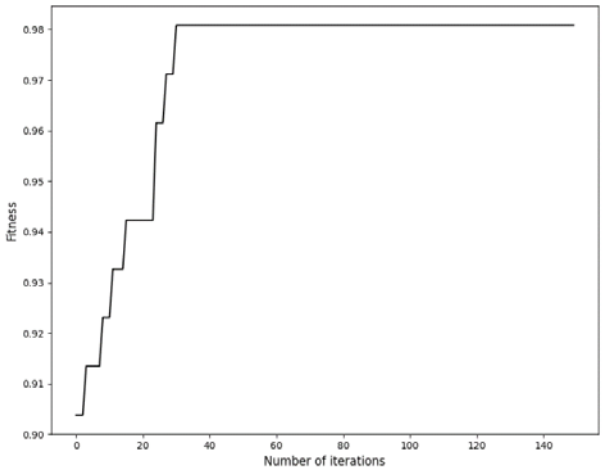

Fig. 2. Optimization process.

Bring the obtained model optimal parameters C and into the support vector machine, and then obtain the support vector machine model PSO-SVM supported by the particle swarm algorithm, thereby assessing the risk index of each patient suffering from diabetes, and the risk index of each patient from small to large, it is divided into five categories: very low, low, medium, high, and very high. For different categories, different medical or preventive measures can be taken to respond. The risk index of diabetes in the data set is shown in the table:

Table 3. Diabetes Risk Index.

Patient number	Diabetes risk index
001	0.998
002	0.344
……	……
518	0.459
519	0.265
520	0.756

Table 4. Pre-diabetes Risk classification.

Risk level	Frequency
Very low	104
Low	87
Medium	184
High	91
Very high	54
Total	520

Based on the training set, the PSO-SVM model is constructed to quantify the diabetes risk index (probability of 0 to 1) for each patient. The closer to 1, the higher the risk index. It can be seen that the number of patients at moderate risk is the largest. The order is very low, low, high, low and very high. For patients with low and very low risk, education and publicity of diabetes-related prevention knowledge can be carried out. For patients with medium risk, certain medical measures should be taken to control them. Risks should be taken seriously for patients with high and very high risk levels, and various indicators of the patient's body should be controlled in time.

4 Result analysis

Use the remaining part of the patient data (20%) to test and evaluate the performance of the PSO-SVM model. The ROC curve and AUC index were used to evaluate the model [9], and combine the PSO-SVM model with Logistic Regression, Neural Network and Naive Bayes [10]-[12]. After comparing, get the ROC curve as shown in the figure:

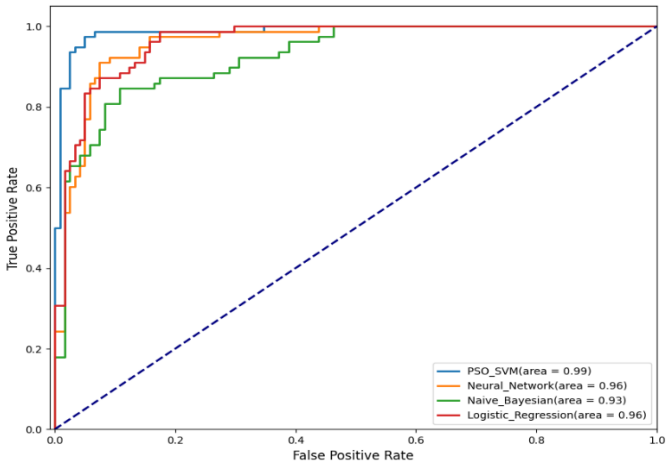

Fig. 3 ROC curve.

The test set shows that the AUC area of the PSO-SVM model is 0.989, and the AUC areas of the neural network, naive Bayes and logistic regression are 0.958, 0.926 and 0.955, respectively. It can be seen from the figure that the performance of the PSO-SVM model is significantly higher than that of the comparison algorithm, and it has better predictive performance in diabetes risk prediction.

In addition, the cost of incorrectly classifying diabetic patients into non-diabetic patients and non-diabetic patients into diabetic patients is different. The cost of the former is higher than that of the latter, so the specificity and sensitivity of the model must be calculated. In this paper, Confusion Matrix is used to evaluate the performance of PSO-SVM model in detail [13], define 0 as a diabetic patient (positive) and 1 as a non-diabetic patient (negative). The calculated specificity is 0.96 and the sensitivity is 1.00, which is within an acceptable range.

Table 5. PSO-SVM Confusion Matrix.

Test set		Predictive classification		Total	Correct rate /%	Error rate /%
		1(Positive)	0(Negative)			
Actual classification	1(Positive)	56	0	56	1.00	0.00
	0(Negative)	2	46	48	0.96	0.04

5 In conclusion

This paper constructs a PSO-SVM model on the early diabetes risk data set, predicts the patient's diabetes risk, assists doctors in identifying the patient's diabetes risk, carries out medical prevention and control in advance. After particle swarm optimization, the optimal parameters of the support vector machine are obtained: C= 4.663, δ= 2.859, the AUC area of the PSO-SVM model is 0.989, the specificity and sensitivity are 0.96 and 1.00, which are significantly better than other comparison algorithms, and can be used in practical applications. However, in the process of model construction, this article sets the optimization goal of particle swarm optimization as the optimization of model accuracy. In order to make the model have better practical application value, we can consider using the weighted specificity and sensitivity combination function as the optimization goal. This further reduces the cost of misclassification.

References

1. W.Y. Yang, Diagnosis of diabetes and prediabetes[J]. *Chinese Journal of Endocrinology and Metabolism*,**36**,401-404 ,(2005).
2. *Guidelines for the prevention and treatment of type 2 diabetes in China* (2013 edition) [J]. *Chinese Journal of Medical Frontiers* (Electronic Edition), **7**, 26-89, (2015).
3. H.X. Fang, J.C. Wei, H.J. Li, H.L Zhao, H. Chang. Establishment and evaluation of the prediction model of gestational diabetes[J]. *Chinese Journal of Maternal and Child Health*, **11**, 13-18, (2020).
4. Z. Wei, Y.G. Zhang, D.D. Shi, N.C. Wang, G. Zhao. Support vector machine diagnosis and prediction of diabetes complications based on grid search and cross-validation[J]. *Chinese Medical Equipment,* **17**, 8-11, (2020).
5. Naz Huma, Ahuja Sachin. Deep learning approach for diabetes prediction using PIMA Indian dataset.. **19**, 391-403, (2020).

6. S.J. Wu. Research and implementation of topic crawler based on support vector machine classification algorithm [D]. Central China Normal University, (2009)
7. B.W. Yang, W.Y. Qian. Summary of inertia weight improvement strategies in particle swarm optimization algorithm[J]. Journal of Bohai University (Natural Science Edition), **40**, 274-288 (2019).
8. L.L. Guo, Y. Liu, W.X. Wang. Improvement of the inertia weight decreasing strategy of particle swarm optimization algorithm[J]. *Engineering Journal of Heilongjiang University*, **10** , 67-71 (2019).
9. M. Wang, C.Y. Xu, X.Z. Shi. Examples of application errors of ROC curve in medical papers[J]. *Acta Editor*,**31**, 159-161 (2019).
10. A. E. Adegboyegun, E. Ben-Caleb, A. O. Ademola, et al. Fair Value Accounting and Corporate Reporting in Nigeria: A Logistics Regression Approach. **11** (2020)
11. X.K. Yu, T.H. Xu, J.T. Wang. Sound Velocity Profile Prediction Method Based on RBF Neural Network[A]. *Academic Exchange Center of China Satellite Navigation System Management Office*. *Proceedings of the 11th China Satellite Navigation Conference——S10 PNT system and multi-source integrated navigation*[C]. Academic Exchange Center of China Satellite Navigation System Management Office: Zhongke Beidouhui (Beijing) Technology Co., Ltd.,1 (2020).
12. F. Zeng, L. Yao, B.L. Wu, et al. Dynamic human contact prediction based on naive Bayes algorithm in mobile social networks. **50**, 2031-2045 (2020).
13. Wasim A. Bagwan, Ravindra S. Gavali. Delineating changes in soil erosion risk zones using RUSLE model based on confusion matrix for the Urmodi river watershed, Maharashtra, India. Modeling Earth Systems and Environment, 1-14 (2020).