2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Breast cancer diagnosis and prediction model based on improved PSO-SVM based on gray relational analysis

Chang Shuran
School of Artificil Intelligence and Computer Science
Jiangnan University
Wuxi,China
e-mail: 545824425@qq.com

Liu Yian
School of Artificil Intelligence and Computer Science
Jiangnan University
Wuxi,China
e-mail: lya_wx@jiangnan.edu.cn

***Abstract*—Early breast cancer diagnosis and prediction models use image data as input, which is likely to cause a large possibility of error in the conversion process of image data. Therefore, this paper proposes a PSO-SVM diagnostic prediction model called GP-SVM based on gray relational analysis (GRA) of a data set consisting of conventional sign data and blood analysis data. First of all, the original data set is optimized by gray relational analysis (GRA) to obtain a new data set. Secondly,the GP-SVM model composed of improved PSO and SVM,and uses the obtained data set as its input. The improvement point of its PSO algorithm is to dynamically adjust the inertial weights and learning factors to make the improved PSO The algorithm optimizes the parameters of SVM and balances the globality and speed of PSO algorithm convergence. On the breast cancer Coimbra data set in UCI, compared with other prediction models, the performance of the GP-SVM prediction model has better.**

Keywords-gray relational analysis; feature selection; improved PSO; GP-SVM model;

I. INTRODUCTION

Recent years, the incidence of breast cancer has continued to rise. It is a high-risk malignant tumor in the global female population. Its treatment strategy focuses on early detection and early treatment [1]. Therefore, the establishment of a diagnostic prediction model for the prevent-ion of breast cancer, It is very important to effectively reduce the incidence of breast cancer. Judging from the way of obtaining research data, the current research on the classification and prediction model of breast cancer is compartmentalized into the following two types.

(1) Diagnostic prediction model based on image data classification. At present, the main diagnostic method for breast cancer is still image diagnosis technology, so most of the diagnostic prediction model data sets are based on image data. Paper [2] constructed SFS-SVM prediction model. Paper [3] compared SVM, Naive Bayes, KNN and other classification algorithms based on the breast cancer data set, and finally concluded that SVM performance is better.Paper [4] constructed a hybrid method of K-Mean and SVM to classify and identify benign and malignant tumors. Paper [5] constructed a prediction model by XGBoost, and paper [6] used different feature selection methods in data mining and combined multiple classification algorithms for evaluation analysis, the results show that The time taken for feature selection and classification is significantly reduced, and the accuracy of classification using Bayesian network algorithm is higher.

(2) Diagnostic prediction model based on biomarker data classification. Paper [7] concluded that four biomarkers such as BMI and L/A ratio are reliable when used together to predict breast cancer. Paper [8] used the data collected in the clinical diagnosis process such as cancer antigens to predict early breast cancer through establishing an algorithm model. The experimental results indicate that the selected biomarkers are poor performance in diagnosis.The results of paper [9] shows that glucose, BMI, age and resistin are effective as predictors. It uses conventional physical data and blood analysis data to obtain conclusions through classification algorithms combined with Monte Carlo cross-validation technology.The study of the paper [10] is that age, glucose, and resistin are effective biomarkers to predict breast cancer. When using these features for classification, the classification accuracy of SVM can reach 83.684%, and the classification by the KNN classifier The accuracy can reach up to 92%.

In summary, when researching based on image data, the resulting image needs to be processed in multiple steps before it can be converted into numeric data. The overall processing process is relatively time-consuming. Because the image data itself will cause errors, leading to misdiagnosis or missed diagnosis of early patients, and the cost of imaging diagnosis is relatively high. Compared with the image data, the biomarker data processing error is small, and the detection cost is relatively low. It can be used not only for the prediction of early breast cancer, but also for in-depth correlation mining based on the biological characteristics themselves. Therefore, this paper uses a diagnostic prediction model based on biomarker data classification. In order to obtain an effective combination of powerful features, for the original data set, this paper uses gray relational analysis to optimize strong correlation features, and uses the improved PSO to optimize the parameters of SVM, and proposes GP-SVM breast cancer diagnosis prediction Model, which can avoid the delayed diagnosis or non-diagnosis caused by the inherent belief [11] of those who suffer from breast cancer risk, and also further protect personal privacy during the diagnosis process, and reduce the cost of diagnosis, to assist doctors in early breast cancer diagnosis provide a reference for making a quick and accurate diagnosis.

II. GRAY RELATIONAL ANALYSIS

The theory of gray relational analysis is proposed by Chinese scholar Professor Deng Julong[12]. Gray relational analysis is not sensitive to the size and regularity of sample data. The algorithm process is as follows,

2473-3636/20/$31.00 ©2020 IEEE
DOI 10.1109/DCABES50732.2020.00067

Step1: Determine the reference sequence and comparison sequence:Let the data set have k attributes and m samples,the data set is denoted as $N = \{n_0, n_1, n_2, \cdots, n_i, \cdots, n_k\}$,where $n_i = \{n_i(1), n_i(2), \cdots, n_i(j), \cdots, n_i(m)\}$, n_0 is the reference sequence, n_i except n_0 is the comparison sequence, where $i \in [0, k]$, $j \in [1, m]$.

Step2: Dimensionless the number sequence: There are many ways for the data dimension. This article uses the averaging method, the formula is as follows,

$$n_i(j)' = n_i(j) / \text{mean}(n_i) \quad (1)$$

Where $n_i(j)$ is the j-th element of the i-th sequence , $\text{mean}(n_i)$ is the average value of the i-th sequence.

Step3: Calculate the gray relational coefficient of n_0 and n_i .

$$\xi_i(j) = \frac{\min_i \min_j |n_0(j) - n_i(j)| + \rho \max_i \max_j |n_0(j) - n_i(j)|}{|n_0(j) - n_i(j)| + \rho \max_i \max_j |n_0(j) - n_i(j)|} \quad (2)$$

ρ is the resolution coefficient, $\rho \in (0,1)$, it often takes the value 0.5.

Step4: Calculate the relevance.

$$r_i = \frac{1}{m} \sum_{j=1}^{m} \xi_i(j) \quad (3)$$

The $\xi i(j)$ which is too scattered information is concentrated into a value, called the degree of correlation r_i , r_i is convenient for overall comparison.

Step5: Relevance ranking: sort all the results obtained by formula (3), the larger the r value, the stronger the correlation between n_0 and n_i .

III. IMPROVED PSO ALGORITHM OPTIMIZES SVM

A. PSO and SVM

1) Improved PSO

The PSO algorithm is an evolutionary algorithm of swarm intelligence that finds the optimal solution through collaboration and information sharing between individuals in the swarm.

The relevant formula in the PSO algorithm is as follows,

$$v_{id}^{k+1} = w v_{id}^{k} + c_1 r_1 (p_{id}^{k} - x_{id}^{k}) + c_2 r_2 (p_{gd}^{k} - x_{id}^{k}) \quad (4)$$

$$x_{id}^{k+1} = x_{id}^{k} + a v_{id}^{k+1} \quad (5)$$

w is the inertial weight; v_{id}^{k+1} is the flight velocity vector of the k-th iteration of particle i ; c_1 and c_2 mean learning factors , usually set to 2; r_1 and r_2 mean random numbers, $r_1, r_2 \in [0,1]$; p_{gd} is the optimal position of the current particle in the d-dimensional search space, and p_{id} represents the global optimal position in the entire population.The relevant improvement points for the PSO algorithm are as follows,

a) Dynamically adjust the value of w

In the traditional PSO algorithm, the introduced inertial weight w is a fixed value, which describes the effect of the particle's previous generation speed on the operational speed. When w is larger, the particle flying speed is faster, the particle step length is longer, the global search capability is strong; when w is smaller, the particle step length is shorter, then tends to fine local search, local search ability is strong, and convergence is fast .

In order to achieve the purpose of balancing the globality and convergence speed, the author dynamically adjusts the value of the inertia weight by adopting a method of linear reduction of the inertia weight proposed by Paper[13], whose formula is as follows,

$$w_i = w_{\min} - (w_{\min} - w_{\max}) * (i / \max gen) \quad (6)$$

i is the number of current iterations, $\max gen$ represents the upper limit of the iteration number.

b) Dynamically adjust the values of c_1 and c_2

c_1 and c_2 determine the degree of impact of the experience information of the particle itself and other particles for the particle trajectory of movement. So as to effectually control the flying speed of the particles, the algorithm achieves a valid balance between global exploration and local mining, a shrinkage factor φ is introduced, and it is brought into formula (4) to obtain the following speed update formula.

$$v_{id}^{k+1} = \varphi (w_i v_{id}^{k} + c_1 r_1 (p_{id}^{k} - x_{id}^{k}) + c_2 r_2 (p_{gd}^{k} - x_{id}^{k})) \quad (7)$$

$$\varphi = \frac{2}{\left| 2 - D - \sqrt{D^2 - 4D} \right|}, \quad D = c_1 + c_2, D > 4 \quad (8)$$

Usually, $D = 4.1$.

2) SVM

When using support vector machine SVM for classification, the setting of kernel function parameters σ and penalty coefficient C is very important. Therefore, it is necessary to adjust these important parameters in order to achieve the purpose of obtaining the best classification ability. The SVM parameter selection problem is actually equivalent to an optimization solution process. Each point in the search space may become a solution of the best model, and then the prediction value is evaluated through the ability promotion, and SVM has a strong generalization ability and Self-learning ability [14] is more suitable for high-dimensional, nonlinear data sets.

B. Improved PSO optimized SVM

Step 1: Initialize the scale of the PSO algorithm, set the algorithm weights, termination conditions and initial particle encoding;

Step 2: Initialize the position and speed of the PSO algorithm;

Step 3: After each iteration, calculate the initial fitness value of apiece particle in line with the fitness function of the particles;

Step 4: Improve the inertial weight and learning factor of each particle in line with formula (6), formula (7) and formula (8);

Step 5: Iteratively calculate according to the particle position and velocity update formula, update the particle position and velocity, and recalculate the fitness value;

Step 6: Compare the fitness value of apiece particle with the fitness value of its individual extremum,if it is better, then renew the individual extreme value, not so retain the former value;

Step 7: Compare the updated individual extremum of apiece particle with the global extremum, if it is better, update the global extremum, not so retain the former value;

Step 8: Determine whether the end term is satisfied. If the upper limit of the number of iterations is attained or the resulting solution has converged or the obtained solution has reached the expected effect, the iteration is terminated, otherwise return to Step 5;

Step 9: Obtain the parameter combination (σ , C) that makes the SVM classifier with the best performance, which is used to construct the sub-optimal model.

C. GP-SVM prediction model building steps

Step 1: Use gray relational analysis(GRA) to select features from the original data set D to obtain a new data set D';

Step 2: Normalize the data set D' according to the min-max standardization,and randomly extract 80% of the data is the training set,and the test set is the remaining 20% of the data;

Step 3: Use the improved PSO to optimize the parameters of SVM.The optimization process is as described in section B, and the optimal parameter combination (σ , C) is obtained;

Step 4: Construct the SVM model by combining the optimal parameter combination obtained in Step3 with the training set;

Step 5: The prediction result of the test data set is predicted according to the trained SVM model, and the classification accuracy is obtained.

IV. EXPERIMENT AND ANALYSIS

A. Data set

Breast cancer Coimbra data set in UCI is composed of routine sign data and blood analysis data.The feature order and feature code ID of the data set are shown in Table I.It has a total of 116 instance samples.

TABLE I. DATA SET FEATURES AND NUMBER OF SAMPLES

ID	*Feature*	*Number of samples*
1	age	116
2	BMI	
3	glucose	
4	insulin	
5	HOMA	
6	leptin	
7	adiponectin	
8	resistin	
9	MCP-1	

B. Data Preprocessing

In this thesis, when preprocessing the data, max-min standardization is used, and its relevant function formula is as follows,

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \quad x' \in [0,1] \tag{9}$$

Normalizing the original data can reduce influence of the inconsistent dimensionality of the eigenvalues, so that the indicators are in the same order of magnitude, which is suitable for comprehensive comparative evaluation.

C. Feature Selection

Find the degree of feature correlation according to the gray relational procedure in section II. Set the threshold of its strong correlation degree to 0.88. When the correlation degree indicates that the two have strong correlation, the eligible features are arranged in descending order according to the correlation degree. The sequence 3, 2, 1, 8, 7, 9, the relevance ranking is as shown in Table II.

TABLE II. RELEVANCE RANKING OF STRONGLY CORRELATED FEATURES

ID	*Feature*	*Correlation Degree(r)*
3	glucose	0.9347580147711
2	BMI	0.9225385269235
1	age	0.9188338960023
8	resistin	0.8936782344402
7	adiponectin	0.8898993457811
9	MCP.1	0.8883079679364

According to the correlation results, the data sets of the first $n(3 \le n \le 6)$ feature combinations are put into the prediction model respectively. The result of the comparison is that when the data sets obtained by the feature combinations numbered 3, 2, 1, 8 are used as the model input, the model prediction accuracy of the model is better and takes less time.

D. Experimental results and comparative analysis

The original data set is input into the GP-SVM prediction model, and the confusion matrix is used as the evaluation index of the GP-SVM algorithm. The evaluation result is shown in Figure 1.

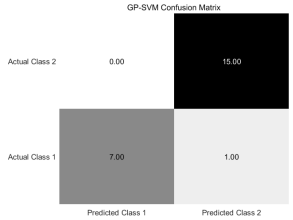

Figure 1. GP-SVM confusion matrix.

The prediction classification result of GP-SVM algorithm is shown in Figure 2.In order to better show the excellent predictive classification effect of the GP-SVM algorithm, this article compares the GP-SVM algorithm with SVM and grid search optimization SVM (GS-SVM) algorithm.Figures 3 and 4 are the prediction classification results of SVM algorithm and GS-SVM algorithm respectively.

From the comparison of Figure 1,Figure 2,and Figure 3, we can see that the prediction effect of GP-SVM is the best. Table III describes the prediction accuracy of each algorithm,which shows that GP-SVM has the highest classification accuracy.

TABLE III. THE FORECAST ACCURACY OF THE THREE ALGORITHMS

Algorithm	*Accuracy (%)*
GP-SVM	95.65
SVM	82.61
GS-SVM	86.96

V. CONCLUSION

This paper presents a PSO-SVM (GP-SVM) breast cancer diagnosis and prediction model based on gray relational analysis (GRA) of a data set composed of conventional sign data and blood analysis data. The use of biomarker data can more accurately reflect the organism's body change information. Gray relational analysis can perform correlation analysis on each feature and target factor to obtain more effective input, reduce classification time, and use inertia weights and learning factors as After the dynamic adjustment improves the PSO algorithm, it optimizes the parameters of the SVM and improves the prediction accuracy. Experimental results show that the GP-SVM prediction model has better performance than the SVM and grid search optimized SVM prediction models. From the perspective of computer-aided diagnosis technology, it provides a low-cost and effective diagnostic prediction model for early medical diagnosis of breast cancer.

REFERENCES

[1] Moura Daniel C,Guevara López Miguel A. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis.[J]. International journal of computer assisted radiology and surgery,2013,8(4).

[2] Lai Shengsheng, Liu Qiancheng, Yu Liling, et al. Construction of breast cancer prediction model based on SFS-SVM[J]. Chinese Journal of Medical Physics, 2019, 36(7): 826-829. DOI: 10.3969/j .issn.1005-202X.2019.07.

[3] Lu Wei, Zheng Yandi, Zhang Xuenong. Comparison of classification methods based on breast cancer data [J]. Medical Information, 2016, 29(3): 278-279. DOI: 10.3969/j.issn.1006-1959.2016.03.225.

[4] Bichen Zheng,Sang Won Yoon,Sarah S. Lam. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms[J]. Expert Systems With Applications,2014,41(4).

[5] Shen Qianqian, Shao Fengjing, Sun Rencheng. Breast cancer prediction model based on XGBoost[J].Journal of Qingdao University (Natural Science Edition),2019,32(1):95-100. DOI:10.3969/j.issn.1006-1037.2019.02.18

[6] Wu Chenwen, Qi Chenhong, Gao Shengpeng.Evaluation and analysis of breast cancer data based on feature selection and data classification[J].Journal of Ningxia University (Natural Science Edition),2018,39(2):155-159. DOI:10.3969/j. issn.0253-2328.2018.02.013.

[7] Santillán-Benítez Jonnathan G,Mendieta-Zerón Hugo,Gómez-Oliván Leobardo M,et al. The tetrad BMI, leptin, leptin/adiponectin (L/A) ratio and CA 15-3 are reliable biomarkers of breast cancer.[J]. Journal of clinical laboratory analysis,2013,27(1).Electronic Publication: Digital Object Identifiers (DOIs):

[8] Opstal-van Winden Annemieke W J,Rodenburg Wendy,Pennings Jeroen L A,et al. A bead-based multiplexed immunoassay to evaluate breast cancer biomarkers for early detection in pre-diagnostic serum.[J]. International journal of molecular sciences,2012,13(10).

[9] Patrício Miguel,Pereira José,Crisóstomo Joana,et al. Using Resistin, glucose, age and BMI to predict the presence of breast cancer.[J]. BMC cancer,2018,18(1).

[10] Bikesh Kumar Singh. Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm[J]. Biocybernetics and Biomedical Engineering,2019,39(2).

[11] Xiang Wang, Zhao Fanghui, Shi Jufang,et al.Feasibility of joint screening for cervical cancer, breast cancer and reproductive tract infections in rural areas[J].Acta Academiae Medicinae Sinicae,2009,31(05)

[12] Tan Xuerui, Deng Julong. Gray relational analysis: a new method of multi-factor statistical analysis[J]. Statistical Research, 1995(03): 46-48.

[13] S.Naka, T. Genji, T. Yura and Y. Fukuyama, "Practical distribution state estimation using hybrid particle swarm optimization," 2001 IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No.01CH37194), Columbus, OH, USA, 2001, pp. 815-820 vol.2, doi: 10.1109/PESW.2001.916969.

[14] Qin Bo, Sun Guodong, Zhang Liqiang,et al.. Research on fault diagnosis of rolling bearing based on singular value of Hilbert envelope spectrum and IPSO-SVM [J]. Journal of Mechanical Transmission, 2017, 41(3): 166～171.

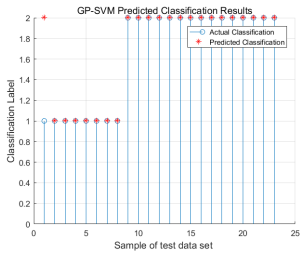

Figure 2. GP-SVM Predicted Results

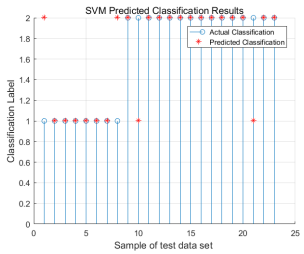

Figure 3. SVM Predicted Results

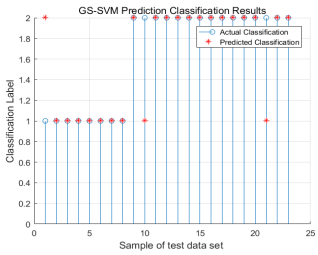

Figure 4. GS-SVM Predicted Results