PERSONAL CREDIT SCORING MODEL RESEARCHBASED ON THE RF-GA-SVM MODEL

Zhang Qiuju
School of Management and Economics
Beijing Institute of Technology
Beijing
China
zhangqj1983@hotmail.com

Abstract. The importance measure of variables in the random forests algorithm is used to carry outa rank ordering to the importance of variables, so as to extract feature attributes on this basis. The feature attributes are regarded as inputs to conduct parameter optimization in order to support vector machine (SVM) model by using the genetic algorithm, building the classifier model by selecting the parameter with the highest accuracy of 5-fold cross-validation. The classifier model is utilized for empirical research, and the results show that the classifier is better than random forest classifier and support vector machine classifier in its higher classification accuracy.

Keywords: Genetic algorithm, random forest, support vector machine, data mining, credit scoring.

1. Introduction

The widespread use of credit cards has brought huge profits for credit card issuers, and also has brought huge risks. Through the personal measurement and control of credit risks, effectively avoiding risks and accurately dealing with the relationship between the benefits and risks is the key to success. Therefore, the measurement and control of personal credit risks has been always the important research topic in the development of personal banking business.

Along with the theory and technology of risk control of credit cards, especially with the rapid development of computer technology, more and more measures, such as statistics, operational research and other quantitative analysis tools, are applied to the field of credit scoring.Statistical methods mainly include linear regression, judgment analysis and Logistic regression. Operational research methods mainly include linear programming. Credit scoring model generally combines one or several of these methods for use.In recent years, some of the non-parametric statistical methods and artificial intelligence are gradually applied in the field of credit scoring [1-5], such as decision tree, neural network, genetic algorithm and support vector machine (SVM), etc. Currently, the common decision tree classification algorithms are divided into CLS algorithm,

C4.5 algorithm and CART algorithm as well as SPRINT algorithm, etc, and the common neural network algorithm is BP neural network algorithm.

However, although the decision tree algorithm possesses higher execution efficiency, the order of the attributes in the tree node is easy to influence performance. And even for small training sets, the decision tree also may be quite large, which affects the understanding of the decision tree. Neural network algorithm is easy to cause excessive fitting by excessive learning, so as to affect the prediction precision in practical application.

As a trainable machine learning method, SVM has been widely used in the field of data mining.In essence, SVM avoids the traditional processes from induction to deduction, efficiently implementing the transduction reasoning from training sample to forecast sample which, greatly simplifies the usual classification and regression problems.Final decision function of SVM is only determined by a few support vectors, which not only can help us to grasp the key sample, "eliminating" a large number of redundant samples, and the method is not only doomed to have simple algorithm, but also has good "robustness". But due to that the SVM generally can not simplify the space dimensions of input vectors, it needs a quadratic programming to solve the support vectors.Hence, when the sample size is large, the input variables will largely cost a lot of machine memory and computing time.In order to improve the operation efficiency, this paper firstly conducts the sample space dimensionality reduction by use of the principle of attribute importance ordering in the random forests, making full use of the flexibly nonlinear modeling capability of SVM on this basis, and the superior global optimization search ability of GA (Genetic Algorithm), establishing k-fold cross-validation optimal SVM model, so as to realize the recognition for "good customer" and "bad" customer.

2. Relevant theory of RF-GA-SVM model

Genetic Algorithm

Genetic Algorithm (GA) is a calculation model simulating natural selection of Darwin biological evolution theory and genetic mechanism of the biological evolution process, which is a kind of method in support of searching optimal solutions by simulating the natural evolution process. Since 1975, it has been widely used in the field of artificial intelligence. Basic operation process of genetic algorithm randomly generates individuals as initial group, and then calculates the fitness of each individual in the community, and chooses excellent individual to propagate directly to the next generation or the intersection of other individuals generates new individuals for propagating again to the next generation according to the fitness, producing the next generation of group in the genetic process through selection, crossover and mutation. If the presupposed termination conditions have been met at that moment, the calculation should be stopped to output the optimal solution; otherwise the calculation will continue to choose, cross, mutate until the termination conditions are achieved.

Random Forest

When using data mining technology to carry out personal credit scoring, there is a vast amount of information.But not all of the information has larger influence on personal credit scoring. When SVM model is set up, many independent variable inputs obviously will affect the operational efficiency of the model. Hence, it is necessary to conduct knowledge reduction firstly, and that can be realized by making use of the gain rate ordering of various attributes in the random forest algorithm. Random forest (RF) is a kind of integrated machine learning method, which employs randomly resampling technology of bootstrap and node randomly splitting technology to build decision trees, so as to obtain the final classification result by the way of voting.RF possesses the ability of analyzing the classified characteristics of the complex interaction, which has good robustness for the noise data and data that exist missing values, and have faster learning speed. The measure on the variable importance can be used as a feature selection tool for high-dimensional data, which has been widely applied in all kinds of classification and prediction, feature selection and anomaly detection problems in recent years [6-9].

Feature selection algorithm can be divided into two major categories of the Filter and Wrapper [10]based on the characteristics of evaluation strategy that has adopted.Filter method is independent of the follow-up machine learning algorithms, which can quickly eliminate some of the non-key noise characteristics, reducing the searching range of optimal feature subsets. However, it can not guarantee selecting out a smaller optimal feature subset. In the process of screening characteristics, wrapper method directly uses the selected feature subsets to train a classifier, evaluating the pros and cons of the feature subset according to the performance of the classifier in the test sets. Filter method is better than the method in computational efficiency, but the size of the selected optimal feature subsets is relatively larger.

Based on taking random forest algorithm as the basic tool, the article researches Wrapper feature selection method by using the classification accuracy of random forest classifiers as feature separability criterion, utilizing the variable importance measure to conduct characteristic importance sorting based on the random forest algorithm itself, adopting the sequence backward selection method generalized sequence backward selection method to select featured subsets.

Support Vector Machine (SVM)

Support vector machine (SVM) is a classification technique based on statistical learning theory proposed in 1995 by AT&T Bell laboratories research team led by Vapnik, which put forward the duality theory in the traditional optimization problem, mainly including the maximum and minimum duality and Lagrangian duality. The key to the SVM lies in the kernel function. Vector sets in low dimensional space are usually difficult to divide, and the solution is

to map them to the high dimensional space.Butthe difficulty the method has brought is the increase of the computational complexity, and kernel function just ingeniously solves the problem. That is to say, as long as choosing the appropriate kernel function, the classification function in the higher dimensional space can be calculated.General common kernel functions include linear kernel function, polynomial kernel function, the RBF kernel function and Sigmoid kernel function [11-12]. Linear kernel function is a special case of kernel functions, which is used to find the linear classifier with optimal generalization. As a global kernel function, polynomial kernel function needs a large amount of calculations.Gaussian radial basis kernel function is the most widely used kernel function with powerful locality and very good learning ability, regardless of the sample sizes and high or low dimensions. When sigmoid kernel function is applied to SVM, it will build up the multi-layer perceptron neural networks to achieve the global optimization through learning. However but when applied to the classification, the conditions are quite harsh.

3. Design of RF-GA-SVM Algorithm

Step 1 the data sets shall be randomly arranged, and are divided into five parts. In order to guarantee the stability of the experimental results, this paper uses the 5-fold cross-validation methods. In each iteration, the four pieces of data are regarded as the training sets in support of building the random forest classifier, and the remaining one piece of data is regarded as the validation set data for validation.

Step 2 set the maximum classification accuracy TGMaxAcc=0 in validation sets, and the corresponding characteristic collection FGSort is empty sets. Suppose the number of input attributes is $n; i1$.

Step 3: initialize the average classification accuracy $TLMeanAcc = 0$ of 5-fold cross validation, and the classification accuracy $TLAcc15 = 0$ in each iteration of cross validation. Run Random to create classifiers on data sets, and classify on the test sets. Compare the classification and the observed values and then calculate $TLAcc.$.

Step 4: $TLMeanAcc = TLMeanAcc + TLAcc[i]/5$.

Step 5: if $TLMaxAcc <= TLAcc[i]$, then $TLMaxAcc = TLAcc[i]$.

Step 6: measure the importance of each attribute according to the computation formula of gain rates of calculation information in the decision tree C4.5. The characteristics are arranged according to the variable importance and store this attribute set as $FSort$. If $(TGMaxAcc <= TLMeanAcc)$, then $TGMaxAcc = TLMeanAcc, FGSort = FSort$.

Step 7: $i = i + 1.If i < n?1$ then execute Step 8otherwise execute Step 9.

Step 8 take out a characteristic with the lowest score in its importance from FSort, and get the new data set, then perform Step 3.

Step 9 output the highest global classification accuracy $TGMaxAcc$, and the corresponding. characteristics of collection $FGSort$.

Step 10 serve *FGSort* as the input data set, choose kernel function to build up SVM classification model.

Step 11 make use of the genetic algorithm c, g to determine the optimal parameters for searching the parameter space, perform 5-fold cross-validation on the training sets , build the classifier model through the selection of the parameter with the highest accuracy of cross-validation.

4. Positive analysis

Download from UCI data sets to Australia and Germanys related data of personal credit scoring, including Australias personal credit scoring data set that conatins14 input attributes and a output attribute with a total of 690 data. There are 307 and 383 of "good credit" and "bad credit" samples respectively. German personal credit scoring data set contains 24 input attributes and a output attribute with a total of 1000 data, which contains 700 samples of "good credit", 300 samples of "bad credit".

Take x_i as the representative of the i input attributes in data set, y is used to represent the output properties of the data set. The methods in step 1 $\sim$ *step*9 are used to conduct attribute reduction on two data sets respectively, and Australia's data reduction result of credit cards is $\{x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}\}$; while Germany's data reduction result of credit cards is $\{x_1, x_2, x_6, x_7, x_8, x_9, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, x_{23}, x_{24}\}$.

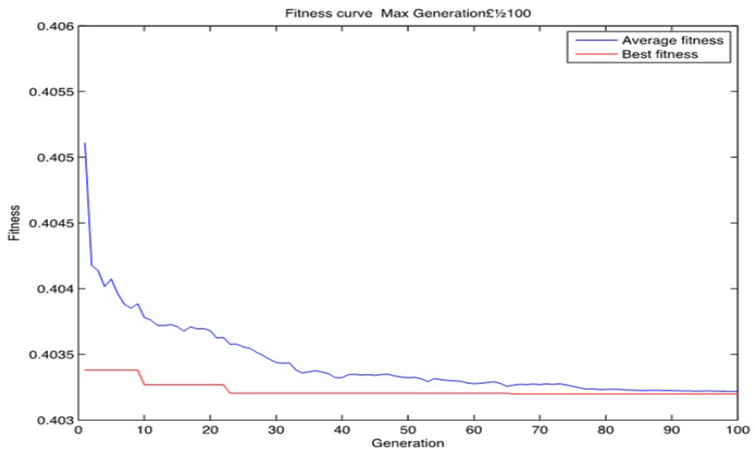

Figure 1: Fitness curve of Germany personal credit scoring data set based on RF - GA - SVM classifier

Take x_i as the representative of the i input attributes in data set, y is used to represent the output properties of the data set. The methods in step 1 $\sim$ step 9 are used to conduct attribute reduction on two data sets respectively, and Australia's data reduction result of credit cards is $\{x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10},$

Figure 2: Fitness curve of Austrilia personal credit scoring data set based on RF-GA-SVM classifier

Model	Classification accuracy of good customer	Classification accuracy of bad customer	Overall classification accuracy
RF	76.71%	85.88%	81.29%
SVM	44.00%	90.93%	67.41%
GA-SVM	86.88%	84.52%	84.71%
RF-GA-SVM	86.05%	86.90%	86.47%

Table 1: Classification accuracy of Australia's credit card data in all kinds of classification models

$x_{11}, x_{12}, x_{13}\}$; while Germany's data reduction result of credit cards is $\{x_1, x_2, x_6, x_7, x_8, x_9, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, x_{23}, x_{24}\}$.

Suppose the maximum iterative algebra in genetic algorithm is 100, the initial group population is 30, the crossover probability is 0.7 and the mutation probability is 0.1. The RBF kernel function is chosen to establish the SVM model, the Matlab programming is utilized to realize the algorithm designed in this article. The data after reduction are used to conduct 5-fold cross validation, so as to obtain the optimal parameters of the SVM classifier of the German credit cards $c = 54.8582, g = 99.6882$, and the optimal parameters of the SVM classifier of the Australian credit cards $c = 7.3611, g = 3.7405$.

Based on RF-GA-SVM classifier, Australia and Germany's credit card data are classified, and the optimal parameters that have achieved are used to test the data of validation set, which achieves satisfactory results. The results are shown in table 1 and table 2. As can be seen from the classification results, the classification accuracy rates of two data sets in the RF-GA-SVM model are very high, which are higher than single RF model, the SVM model and the GA - SVM model.

Model	Classification accuracy of good customer	Classification accuracy of bad customer	Overall classification accuracy
RF	86.66%	86.33%	86.50%
SVM	85.09%	84.85%	84.87%
GA-SVM	99.36%	85.98%	92.50%
RF-GA-SVM	100%	85.98%	92.81%

Table 2: Classification accuracy of Germany's credit card data in all kinds of classification models

5. Conclusions

This article constructs the credit scoring model based on random forest, genetic algorithm and support vector machine (SVM), and adopts the Australia and Germanys credit data in UCI data set for empirical testing, and the results show that:

(1) The use of random forests for data reduction can eliminate some noise properties, which not only reduce the SVM model input data and improve the efficiency of learning and prediction, as well as enhance the model prediction accuracy rates.

(2) The GA - the SVM model is built for the data after reduction, and this model is utilized to make a classification of bank credit card users in Australia and Germany, and the results show that the model established in this paper is efficient and precise, which possesses practical application prospect.

References

[1] T.S. Lee, C.C. Chiu, Y.C. Chou, et al., *Mining the customer credit using classification and regression tree and multivariate adaptive regression splines,* Computational Statistics & Data Analysis, 50 (2006), 1113-1130.

[2] T.S. Lee, I.F. Chen, *A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines*, Expert Systems with Applications, 28 (4) (2005), 743-752.

[3] N.C. Hsieh, *Hybrid mining approach in the design of credit scoring models*, Expert Systems with Applications, 28 (4) (2005), 655-665.

[4] M.K. Lim, S.Y. Sohn, *Cluster-based dynamic scoring mode*, Expert Systems with Applications, 32 (2) (2007), 427-431.

[5] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, et al., *Conditional variable importance for rand om forests,* BMC Bioinformatics, 9 (1) (2008), 1-11.

[6] David M. Reif, Alison A. Motsinger, Brett A. McKinney, et al., *Feature selection using a rand om forests classifier for the integrated analysis of*

multiple data types, IEEESymposium on Computational Intelligence and Bioinformatics and Computational Biology, (2006), 171-178.

[7] Khalilia Mohammed, Chakraborty Sounak, Mihail Popescu, *Predicting disease risks from highly imbalanced data using random forest*, BMC Medical Informatics and Decision Making, 11 (7) (2011), 51-58.

[8] A. Verikas, A. Gelzinis, M. Bacauskiene, *Mining data with random forests: a survey and results of new tests*, Pattern Recognition, 44 (2) (2011), 330-349.

[9] I. Inza, P. Larranaga, R. Blanco, *Filter versus wrapper gene selection approaches in DNA microarray domains*, Artificial Intelligence in Medicine, 31 (2) (2004), 91-103.

[10] T. Bellotti, J. Crook, *Support vector machines for credit scoring and discovery of significant features*, Expert Systems with Application, 36 (2) (2009), 3302-3308.

[11] C.L. Huang, M.C. Chen, C.J. Wang, *Credit scoring with a data mining approach based on support vector machines,* Expert Systems with Applications, 33 (4) (2007), 847-856.

Accepted: 17.02.2017