ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(20), DOI: 10.17485/ijst/2016/v9i20/92299, May 2016* ISSN (Online) : 0974-5645

A Novel Credit Scoring Prediction Model based on Feature Selection Approach and Parallel Random Forest

Ha Van Sang[1*], Nguyen Ha Nam[2], Nguyen Duc Nhan[3]

[1]Department of Economic Information System, Academy of Finance, Hanoi, Viet Nam; sanghv@hvtc.edu.vn
[2]Department of Information Technology, VNU-University of Engineering and Technology, Hanoi, Viet Nam; namnh@vnu.edu.vn
[3]Department, Faculty of Telecommunications, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam; nhannd@ptit.edu.vn

Abstract

Background/Objectives: This article presents a method of feature selection to improve the accuracy and the computation speed of credit scoring models. **Methods/Analysis:** In this paper, we proposed a credit scoring model based on parallel Random Forest classifier and feature selection method to evaluate the credit risks of applicants. By integration of Random Forest into feature selection process, the importance of features can be accurately evaluated to remove irrelevant and redundant features. **Findings:** In this research, an algorithm to select best features was developed by using the best average and median scores and the lowest standard deviation as the rules of feature scoring. Consequently, the dimension of features can be reduced to the smallest possible number that allows of a remarkable runtime reduction. Thus the proposed model can perform feature selection and model parameters optimization at the same time to improve its efficiency. The performance of our proposed model was experimentally assessed using two public datasets which are Australian and German datasets. The obtained results showed that an improved accuracy of the proposed model compared to other commonly used feature selection methods. In particular, our method can attain the average accuracy of 76.2% with a significantly reduced running time of 72 minutes on German credit dataset and the highest average accuracy of 89.4% with the running time of only 50 minutes on Australian credit dataset. **Applications/Improvements:** This method can be usefully applied in credit scoring models to improve accuracy with a significantly reduced runtime.

Keywords: Credit Scoring, Feature Selection, Machine Learning, and Parallel Random Forest

I. Introduction

The credit risk analysis plays an important role in categorization of customers which allows the customers to be divided into two sets, those good and bad[1]. Many models and classification algorithms are applied to analyze credit risks over the last decades, for example the nearest neighbour K-NN, the decision tree, neural networks and support vector machine (SVM)[2–7]. An important goal of the credit risk prediction is constructing the best classification model for a particular data set. There are a lot of irrelevant and redundant features in financial data in general and credit data in particular. When the data is noisy and unreliable by the redundancy and the deficiency in data the accuracy of classification can be reduced that may lead to bad decisions[8,9]. In that case, a feature selection strategy is deeply needed in order to filter the redundant features. In order to select a subset of relevant features, feature selection is needed. The subset is sufficient to describe the problem with high precision.

* *Author for correspondence*

Feature selection thus reduces the dimension and the computational complexity of the problem and saves on the cost of measuring non selected features.

Today credit scoring and internal customer rating is widely used in banking activities to assess the ability to perform financial obligations of a customer against a bank. Beside normal activities the risk evaluation and identification functions are also very important in the credit activities of the bank. Credit risk level changes to individual clients and is identified through an assessment process. This process was based on financial data and existing non-financial customer's at the time of credit grading and evaluation.

Credit scoring is a statistical method used to evaluate the credit risk against customers through using customer data and activities. Credit scoring is performed by the bank based on judgmental view of credit experts, credit groups or credit bureaus. In Vietnam, some commercial banks began implementation of credit scoring for clients but it has not been widely applied in the test phase and still need to improve gradually. To complete, all the information adopted in this article to evaluate the predictive accuracy is obtained from the two real world datasets, the Australian and German credit datasets.

There are many methods that have been investigated in the last decade to improve the accuracy in credit scoring. Artificial Neural Networks (ANN)[10–13] and Support Vector Machine (SVM)[14–19] are two commonly soft computing methods used in credit scoring modelling. Recently, other methods like evolutionary algorithms, stochastic optimization technique have shown promising results in terms of prediction accuracy.

In this study, we proposed a new method for feature selection based on various criteria and integrated with a parallel Random Forest classifier in credit scoring tasks.

This paper is organized as follows: Section 2 describes the background of credit scoring, random forests and feature selection. The details of the proposed model are described in Section 3. Section 4 presents the experiments and the obtained results which show an accuracy improvement of the proposed model. Finally concluding remarks and future works are presented in Section 5.

2. Materials

A. Feature Selection

Feature selection is the important task in data preprocessing to choose a small subset of features that sufficient to predict the target labels well. Feature selection can be a part of the criticism that should focus on only related features, such as the PCA method or an algorithm modeling. However, in the whole process of data mining, feature selection is usually a separate step.

Feature selection methods can be categorized into two main types based on filter approach and wrapper approach. Filter methods consider the feature selection process as a precursor stage of learning algorithms. The irrelevant features are filtered out by using evaluation functions to evaluate the classification performances of subsets of features. Feature importance, Gini, information gain, the ratio of information gain, etc are common evaluation functions that can be used in the filter model. The main disadvantage of this approach is that they are not optimized for a specific classifier because there is no relationship between the process of feature selection and learning algorithm's performance.

Wrapper methods measure the goodness of a selected feature subset with the machine learning algorithm. Learning accuracy, recall and precision values are used to measure the performance of the learning algorithm. In the wrapper model the learning accuracy is used in evaluation to select the best features. The wrapper algorithm searches for the feature subset that generates the lowest error rate in the testing data set. On the other hand the feature subset that leads to the best correct classification rate is kept. The disadvantage of this approach is highly computational cost, hence the wrapper approach cannot be used for large data sets and time-consuming classification algorithm. Some methods that can accelerate the evaluation process were proposed to reduce costs. Common strategies are sequential wrapper Forward Selection (SFS) and reverse sequential Elimination (SBE). By searching on the feature space, the optimal features set is found. In this space, each state representing a subset of features and the size of the search space for the n features is $O(2^n)$, so it is not practical to search the whole not sterilization time, unless n is small.

B. H2O Random Forest

H2O is a platform for distribution in the analysis of memory and learning. H2O using pure Java that's easy to deploy with a single jar, automatic cloud detection. H2O does not analyze in memory on parallel clusters with famous machine learning algorithms are dispersed. Figure 1 shows H2O architecture:

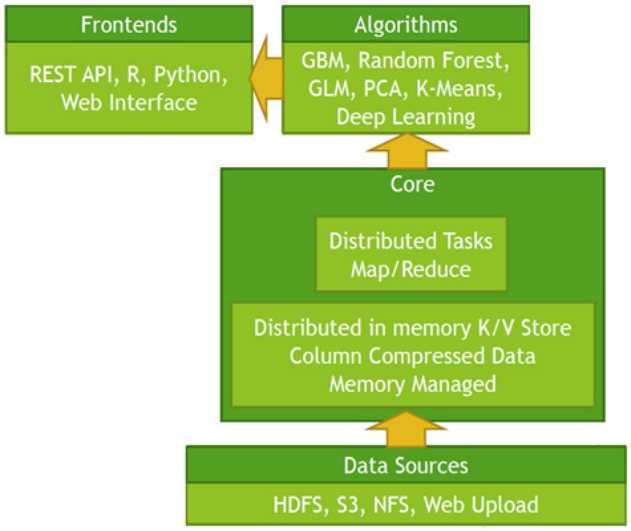

Figure 1. H2O architecture.

Random Forest (RF) is an ensemble classifier which uses bagging mechanism. RF consists of a set of CART classifiers. Each node of a tree only selects a small subset of features for a split, which enables the algorithm to create classifiers for highly dimensional data very quickly. In each section, the number of randomly selected features (*mtry*) must be determined. The default value is *sqrt(p)* for classification in which p is the number of features. The criterion of separation is Gini index as shown in Eq(1).

$$gini(N) = \frac{1}{2}\left(1 - \sum_j p(\omega_j)^2\right) \quad (1)$$

H2O's Random Forest algorithm is parallel processing which produces a dynamic confusion matrix. When each plant was built, the out of the bag error estimate (OOBE) is recalculated. The expected behavior is that the error rate increase before it decreases, so that is a natural result of the learning process of random forest. The error rate is expected to be relatively high if only a few trees is built on random subsets. When more trees were added, the resulting in more trees "voting" to correct classification of OOB data, the error rate will decrease.

3. The Proposed Method

In the proposed method the cross validation accuracy and the importance of each feature as the performance parameters in the training data set are estimated by Random Forest algorithm first. Fast-trees are independent and can be built in parallel. Then we determine best features subset by choosing the best of Average score + Median Score and the lowest standard deviation (SD). In order to deal with over-fitting problem, n-fold cross validation technique is applied to minimize the generalization error. The evaluation procedures for feature selection are as follows:

Step 1: Train dataset by Parallel Random Forest classifier, calculate and sort median of variables important via 20 trails.

Step 2: Add each feature with best variables important and train dataset again by Parallel Random Forest with the cross validation.

Step 3: Calculate score for each feature F_i^{score} *where i=1..n (n is the number of features in current loop).*

Step 4: Select best feature subsets using selection rules which is presented below.

Step 5: Back to step 1 until reach the desired criteria.

In particular, we use Parallel Random Forest with n-fold cross validation to train the classifier in step 2. A set of (F_j, A_j^{learn}, $A_j^{validation}$) those are the feature importance, the learning accuracy and the validation accuracy respectively is obtained in the j^{th} cross validation.

By using above values the score criterion is computed in step 3. We use the results from step 1 and step 2 to build the score criterion in step 3 which will be used in step 4. The score of feature i^{th} is calculated by:

$$F_i^{score} = \sum_{j=1}^{n} F_{i,j} \times \left(A_j^{learn} + A_j^{validation}\right) \quad (2)$$

In the next step, the main step of our algorithm, the best of features using rules: the best of Average + Median Score and the lowest standard deviation (SD) will be selected by using following rules.

Rule 1: Select features with the best of median score

Rule 2: Select features with the best of average score

Rule 3: Select features with the lowest SD

Based on these rules we obtain the highest accuracy and the lowest Standard deviation. Thus the optimal set of features tends to reduce its dimension to the smallest number of output features. Then, the machine learning algorithms are used to calculate the RF relevance of the feature. From the calculated value of relevance, we find the subset of features having less number of features while achieving the objective of the problem.

4. Experiment and Results

The H2O Random Forest package in R language (http://www.r-project.org) has been used to demonstrate our proposed algorithm. This package is optimized to work "in memory" processing of distributed, parallel machine learning algorithms on clusters. A "cluster" is a software construct that can be fired up on your laptop, on a server or across the multiple nodes of a cluster of real machines, including computers that form a Hadoop cluster. Our experiment has been implemented to test the proposed algorithm with some datasets including two UCI public datasets, German credit and Australian credit.

In this paper, Random forest with the original dataset is used as the base-line method. Two methods, the proposed method and the base-line method, were performed on the same training and testing datasets to compare their efficiency. In order to test the consistency of obtained results, those implementations were repeatedly done 20 times.

C. German Credit Approval Dataset

The German credit dataset consists of 1000 loan applications, with 700 instances of creditworthy applicants and 300 instances of rejected applicants. For each applicant, 20 attributes describe the credit history, account balances, loan information and personal information. Figure 2 shows our final results that were averaged over these 20 independent trials. In our experiments, the default value for the *mtry* parameter was used and the *ntree* parameter was tried with value of 100.

As shown in Figure 2 the best subset contains 7 features and its accuracy is 76.2%.

Different classifiers over the German credit datasets were compared and their performances are shown in Table 1. Baseline is the classifier without feature selection. Classifiers used in our investigation include: Linear SVM, CART, k-NN, Naive Bayes, MLP. Various feature selection methods are used for comparison including filter approach and wrapper approach. The filter approach includes three methods: t-test, Linear Discriminant analysis (LDA), Logistic regression (LR). The wrapper approach includes two methods: Genetic algorithms (GA) and Particle swarm optimization (PSO).

Table 1. Compare performances different classifiers over the German credit dataset

Classifier	Filter methods			Wrapper methods		Baseline
	t-test	LDA	LR	GA	PSO	
Linear SVM	76.74	75.72	75.10	76.54	73.76	77.18
CART	74.28	73.52	73.66	75.72	74.16	74.30
k-NN	71.82	71.86	72.62	72.24	71.60	70.86
Naïve Bayes	72.40	70.88	71.44	71.56	74.16	70.52
MLP	73.28	73.44	73.42	74.03	72.54	71.76
Random Forests						73.40
Our method			**76.20**			

As shown in Table 1 for comparing the performances of various methods, we saw that the accuracy of RF on

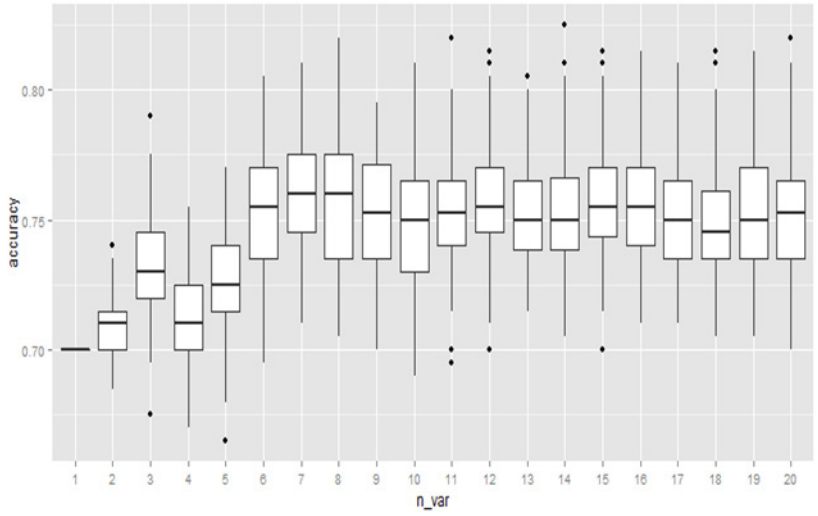

Figure 2. Accuracy in case of German dataset.

the subset of newly selected features has been obviously improved, and the number of features has been reduced by 35%. The average accuracy is 73.4% on the original data. After applying the feature selection, the average accuracy increases to 76.20%.

Furthermore, our method relying on a parallel processing strategy allows the time to run 20 trails with 5-fold cross validate taking only 4311 seconds (~72 minutes) while other methods must run several hours. This result emphasizes the efficiency of our method in terms of running time due to efficiently filtering the redundant features.

D. Australian Credit Approval Dataset

The credit data of Australia consists of 690 applicants, with 383 instances of credit worthy and 307 default examples. Each instance contains both numerical features, categorical features, and discriminant feature. We transferred sensitive information to the symbolic data for confidentiality reasons. Figure 3 shows the averages of classification results.

Table 2 shows the performances of different classifiers and selection methods over the Australian credit datasets for comparison. The obtained results indicate that the accuracy of RF on a subset of 9 selected features has been obviously improved. The average accuracy is 87.82% on the original data, while the average accuracy increases to 89.40% after applying the feature selection in our method. Based on parallel processing, time to run 20 trails with 5-fold cross validate taken by our method can be reduced to only 2974 seconds (~50 minutes).

Table 2. Performances of different classifiers over the Australian credit dataset

Classifier	Filter methods			Wrapper methods		Baseline
	t-test	LDA	LR	GA	PSO	
Linear SVM	85.52	85.52	85.52	85.52	85.52	85.52
CART	85.25	85.46	85.11	84.85	84.82	85.20
k-NN	86.06	85.31	84.81	84.69	84.64	84.58
Naïve Bayes	68.52	67.09	66.74	86.09	85.86	68.55
MLP	85.60	86.00	85.89	85.57	85.49	84.15
Random forests						87.82
Our method				**89.40**		

5. Conclusion

In this paper, we integrated feature selection and parallel Random Forest method in credit scoring model. Feature selection provides an effective method in determining the highest classifier accuracy of a subset or searching the acceptable accuracy of the smallest subset of features. We have introduced a new feature selection approach based on feature scoring. The accuracy of classifier using the selected features is improved compared with other methods. Fewer features allow a credit department to focus on collecting relevant and essential variables. As a result of the parallel processing procedure the runtime can be significantly reduced. Consequently, the workload

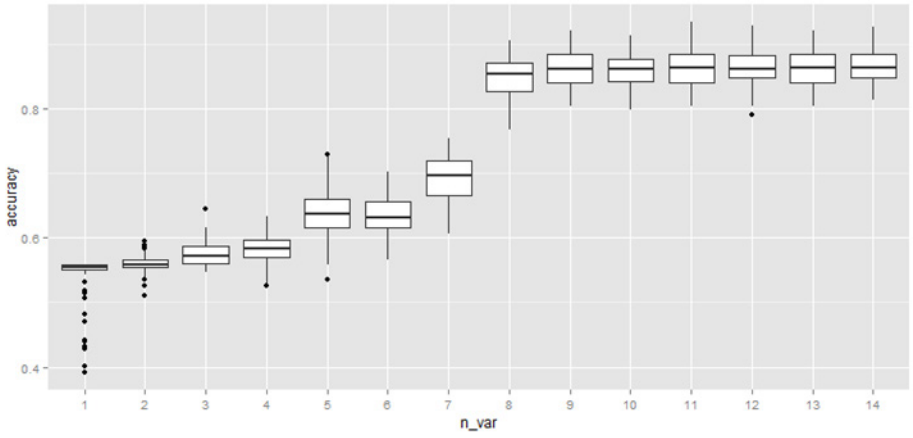

Figure 3. Accuracy in case of Australian credit dataset.

of credit evaluation personnel can be reduced because our model does not have to take into account a large number of features in the assessment process, which requires much less effort in computation. This paper has investigated and compared different methods over two real world credit datasets. Experimental results show that our method is effective in credit risk investigation. The method offers a quick assessment with improved accuracy of the classification.

6. References

1. Altman EI and Saunders A. Credit risk measurement: Developments over the last 20 years. J. Bank. Financ. 1997; 21, 1721–42.
2. Wu, X. et al. Top 10 algorithms in data mining. 2008. Doi: 10.1007/s10115-007-0114-2.
3. Angelini E, di Tollo G, & Roli, A. A neural network approach for credit risk evaluation. Q. Rev. Econ. Financ. 48, 733–755 (2008).
4. Bellotti, T. and Crook J. Support vector machines for credit scoring and discovery of significant features. Expert Syst. Appl. 2009; 36, 3302–08.
5. Wen F and Yang X. Skewness of return distribution and coefficient of risk premium. J. Syst. Sci. Complex. 2009; 22:360–71.
6. Zhou X, Jiang W, Shi Y and Tian Y. Credit risk evaluation with kernel-based affine subspace nearest points learning method. Expert Syst. Appl. 2011; 38:4272–79.
7. Kim G, Wu CH, Lim S and Kim J. Modified matrix splitting method for the support vector machine and its application to the credit classification of companies in Korea. Expert Syst. Appl. 2012; 39:8824–34.
8. Liu H and Motoda H. Feature Selection for Knowledge Discovery and Data Mining. 1998.
9. Guyon I and Elisseeff A. An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 2003; 3:1157–82.
10. Oreski S, Oreski D and Oreski G. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert Syst. Appl. 2012; 39:12605–617.
11. Saberi M. et al. A granular computing-based approach to credit scoring modeling. Neurocomputing. 2013; 122:100–15.
12. Lee S and Choi WS. A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. Expert Syst. Appl. 2013; 40:2941–46.
13. Ghatge AR and Halkarnikar PP. Ensemble Neural Network Strategy for Predicting Credit Default Evaluation. 2013; 2:223–25.
14. Chaudhuri A and De K. Fuzzy Support Vector Machine for bankruptcy prediction. Appl. Soft Comput. J. 2011; 11:2472–86.
15. Ghodselahi A. A Hybrid Support Vector Machine Ensemble Model for Credit Scoring. Int. J. Comput. Appl. 2011; 17:1–5.
16. Huang C-L, Chen M-C and Wang C-J. Credit scoring with a data mining approach based on support vector machines. Expert Syst. Appl. 2007; 33:847–56.
17. Li ST, Shiue W and Huang MH. The evaluation of consumer loans using support vector machines. Expert Syst. Appl. 2006; 30:772–82.
18. Martens D, Baesens B, Van Gestel T and Vanthienen J. Comprehensible credit scoring models using rule extraction from support vector machines. Eur. J. Oper. Res. 2007; 183:1466–76.
19. Wang Y, Wang S and Lai KK. A new fuzzy support vector machine to evaluate credit risk. IEEE Trans. Fuzzy Syst. 2005; 13:820–31.