EAI Endorsed Transactions
on Context-aware Systems and Applications

Research Article EAI.EU

A hybrid feature selection method for credit scoring

Sang Ha Van[1,*], Nam Nguyen Ha[2] and Hien Nguyen Thi Bao[3]

[1] Department of Economic Information System, Academy of Finance, Hanoi, Viet Nam
[2] Department of Information Technology, VNU-University of Engineering and Technology, Hanoi, Viet Nam
[3] Department of Corporate Finance, Academy of Finance, Hanoi, Viet Nam

Abstract

Reliable credit scoring models played a very important role of retail banks to evaluate credit applications and it has been widely studied. The main objective of this paper is to build a hybrid credit scoring model using feature selection approach. In this study, we constructed a credit scoring model based on parallel GBM (Gradient Boosted Model), filter and wrapper approaches to evaluate the applicant's credit score from the input features. Feature scoring expression are combined by feature important (Gini index) and Information Value. Backward sequential scheme is used for selecting optimal subset of relevant features while the subset is evaluated by GBM classifier. To reduce the running time, we applied parallel GBM classifier to evaluate the proposed subset of features. The experimental results showed that the proposed method obtained a higher predictive accuracy than a baseline method for some certain datasets. It also showed faster speed and better generalization than traditional feature selection methods widely used in credit scoring.

Keywords: Credit risk, Credit scoring, Hybrid Feature selection, GBM, RFE, Information Values, and Machine learning.

Received on 27 April 2016, accepted on 16 August 2016, published on 06 March 2017

doi: 10.4108/eai.6-3-2017.152335

1. Introduction

Credit scoring is one of the most essential issue for credit risk management which uses a technique using statistical analysis data and activities to evaluate the credit risk against customers. Retail banks build credit scoring model based on the statistical analysis of credit experts, credit teams or credit bureaus. In Vietnam, some retail banks start to perform credit scoring against customers but it is not widely applied during the testing phase and still needs to amend gradually. In this paper, we collect data and build models based on credit scoring experience in Germany, Australia, and other countries.

Over the last few year, Researchers have developed and applied many models and algorithms to analyse the credit risks, for example decision tree [1], nearest neighbour K-NN [2], support vector machine (SVM) and neural network [3]–[9]. The main purpose of credit risk prediction is to build a classification model with generalization performance, computational efficiency, especially fast processing speed.

Credit data, which is collected by credit department, often contains irrelevant and redundant features. The classification accuracy is reduced by this redundancy, and deficiency data. Obviously it will lead to incorrect decision [10][11]. In order to remove the redundant features, a feature selection strategy should be adopted. In other words, feature selection will select an optimal subset of relevant features. With this optimal subset, we solve the problem with high precision. Feature selection is one of the ways to reduce the dimensionality of the problem and shorten the runtime.

In the fields of machine learning and soft computing, Artificial Neural Networks (ANNs) and Support Vector Machine (SVM) are two commonly used in credit scoring modelling. In order to achieve higher classification performance, SVM recursive feature elimination (SVM-

[*] Corresponding author. Email: sanghv@hvtc.edu.vn

RFE)[12][13] filter relevant features and remove relatively insignificant variables. SVM-RFE uses numerical attribute but credit data sets has a lot of categorical attributes. How to deal with an SVM-RFE with categorical attributes? The conversion of categorical attributes into numerical attributes will lack information and reduce accuracy. GBM is a popular classification method which deal with this problem. Recently, many researchers have applied the optimization techniques like evolutionary algorithms [14], stochastic optimization with support vector machine[15] that have achieved good results in terms of prediction accuracy.

This study proposed a hybrid feature selection method based on RFE and integrated with a parallel GBM classifier in credit scoring tasks. Our proposed method removes features by complex ranking criterion. This criterion combines the importance of features, Information values and the correlation of training and testing accuracy which are obtained from GBM classifier. Our proposed method is applied to credit scoring problem. Integration with H2O parallel GBM, the method showed better results and faster than original GBM

Structure of this paper consists four main sections. Section 2 presents the background of feature selection, Information value and GBM. Section 3 is the most important section that describes the details of the proposed model. Section 4 discusses the obtained results while Section 5 highlights the concluding remarks and future research.

2. Background

2.1. Feature selection algorithms

In order to reduce the dimensionality of the data, researchers often use two common approaches: Feature selection and feature extraction. Feature selection can be a part of the criticism which needs to focus on only related features, such as the PCA method or an algorithm modelling. However, the feature selection is usually a separate step in the whole process of data mining.

Filter approach and wrapper approach are two main categories of feature selection methods, embedded/hybrid approach is another categories. The filter approach reduces features using properties of the data itself and independents to learning algorithm. The evaluation functions such as information gain (IG), GINI, feature important, information value (IV) are used to evaluate the classification performances of feature subset. Filter approach has some disadvantages such as the feature selection process and the performance of learning algorithms have no relationship.

The wrapper approach chooses the relevant features and remove noise features so we can improve the efficiency and accuracy of the classification.

Wrapper approach includes two phases: Phase 1 – Selecting feature subset, at this stage the best feature subset will be selected based on criteria class accuracy (of the training data); Stage 2 - learning and testing, a classifier will learn the knowledge from the training data through a set of best features are selected, and are checked using a testing data. When specific subsets are created in a systematic way (to seek), for each specific subset, a classifier is generated from data including the selected features. Accuracy of the classifier is recorded in each test and subset features the highest accuracy will be retained. When the selection process is over, the subset features the highest accuracy will be selected. Phase 2 is the process of learning and regular checks, at this stage we would have forecast accuracy on the test data. Sequential Forward Selection (SFS) and Sequential Backward Elimination (SBE) are common and well-known wrapper strategies. The searching process will find optimal feature set on the feature space.

2.2. Information Value

Information Value (IV) is used in credit scoring to determine the importance of each feature to the results of the classification process. The information value of feature X is calculated as follows:

$$InfoValue = \sum(X_{dist} - \bar{X}_{dist}) \ln\left(\frac{X_{dist}}{\bar{X}_{dist}}\right) \quad (1)$$

$$X_{dist} = f(Y = y_k | X = x_i) \quad (2)$$

$$\bar{X}_{dist} = f(Y \neq y_k | X = x_i) \quad (3)$$

As Eq. (2) shows, X_{dist} is the conditional probability density of f(Y = y_k) when the feature value is X = x_i. In the same way, equation (3) shows that $\bar{X}_{dist}$is the conditional probability density of f(Y ≠ y_k) when the feature value is X = x_i.

In order to identify the most importance features, other feature selection methods, such as Information Gain (IG) and principal component analysis (PCA), use a ranking algorithm. The IV determines the number of features that have power discrimination by using a pervasive threshold. It is difficult for other methods to determine the appropriate number of features. The information value calculation process is more quickly than other feature selection methods, so it is especially suitable for credit scoring applications.

The information value only provides the importance of each feature and does not consider redundancy between similar features.

2.3. H2O Gradient Boosted Model

H2O is a platform for distributed in memory predictive analytics and machine learning. H2O uses pure Java which easy deployment with a single jar, automatic cloud discovery. Figure 1 show H2O architecture:

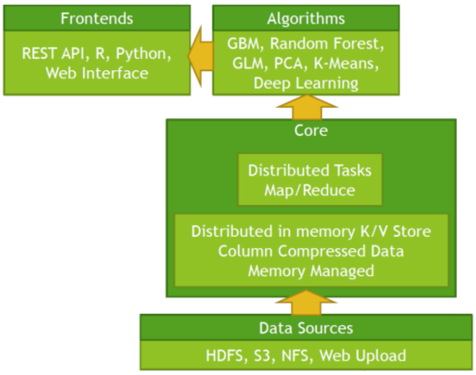
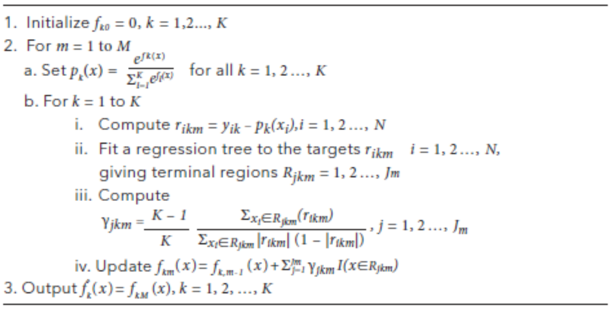


Figure 1. H2O Architecture

A gradient boosted model is an ensemble of tree models that can be either regression or classification. Boosting is a flexible nonlinear regression procedure that helps improve the accuracy of trees. Weak classification algorithms are sequentially applied to the incrementally changed data to create a series of decision trees, producing an ensemble of weak prediction models.

While boosting trees increases their accuracy, it also decreases speed and user interpretability. The gradient boosting method generalizes tree boosting to minimize these drawbacks. Although GBM is known to be difficult to distribute and parallelize, H2O provides an easily distributable and parallelizable version of GBM in its framework, as well as an effortless environment for model tuning and selection.



```
1. Initialize $f_{k0} = 0, k = 1,2..., K$
2. For $m = 1$ to $M$
    a. Set $p_k(x) = \frac{e^{f_k(x)}}{\sum_{l=1}^{K} e^{f_l(x)}}$ for all $k = 1, 2 ..., K$
    b. For $k = 1$ to $K$
        i. Compute $r_{ikm} = y_{ik} - p_k(x_i), i = 1, 2 ..., N$
        ii. Fit a regression tree to the targets $r_{ikm}$ $i = 1, 2 ..., N$,
            giving terminal regions $R_{jkm} = 1, 2 ..., Jm$
        iii. Compute
            $\gamma_{jkm} = \frac{K-1}{K} \frac{\sum_{x_i \in R_{jkm}} (r_{ikm})}{\sum_{x_i \in R_{jkm}} |r_{ikm}| (1 - |r_{ikm}|)}, j = 1, 2 ..., J_m$
        iv. Update $f_{km}(x) = f_{k,m-1}(x) + \sum_{j=1}^{Jm} \gamma_{jkm} I(x \in R_{jkm})$
3. Output $\hat{f}_k(x) = f_{kM}(x), k = 1, 2, ..., K$
```

Figure 2. Gradient Boosting Algorithm

3. The proposed method

Our proposed method has two stages. In the first stage, we uses H2O parallel GBM to estimate performance and reduce running time. In order to select the best features, the training set was trained and tested by GBM classifier. The main objective of this stage is to estimate feature importance value for each feature. We use a recursive elimination approach to evaluated contribution of each feature to the classifier by removing each feature. After calculating means of feature ranking value, the procedure removed irrelevant features and only kept the important features. Result of this stage is a set of optimal features. N-fold cross validation technique was applied to deal with over-fitting problem.

In the second stage, optimal features subset in the first stage is used as a filter of test dataset.

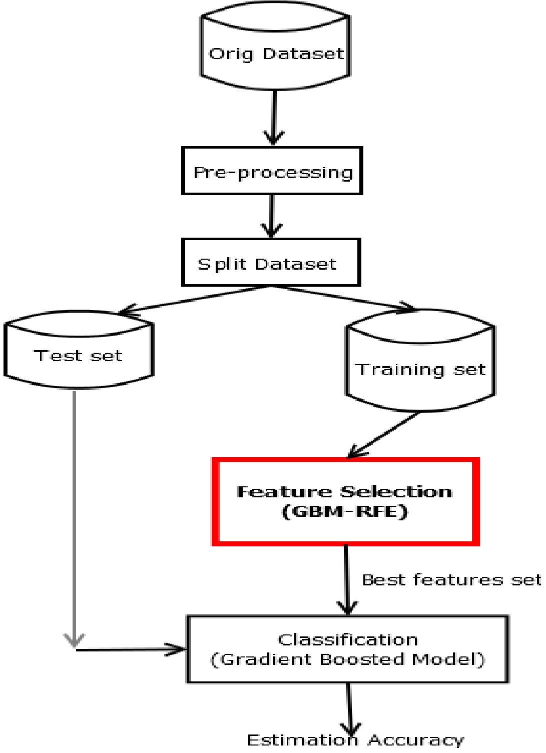


Figure 3. The proposed method

When computing the ranking criteria, the wrapper approaches focus on accuracies of the features, and don't care much about the correlation of the features. We construct a hybrid feature selection to solve this problem. First, we calculate the ranking criterion for all features F_i^{rank} where i=1..n (n is the number of features).

The ranking criterion is combined by the feature importance, the learning accuracy, the validation accuracy, the area under curve (AUC) and the Information Value. Then, we build the ranking criterion as follow:

$$F_i^{rank} = \sum_{j=1}^{n} (F_{i,j} + IV_{i,j}) \times \left(\frac{\left(A_j^{learn} + A_j^{validation}\right)}{\left|A_j^{learn} - A_j^{validation}\right| + \varepsilon} + AUC_j^{learn} \right) \quad (4)$$

where $j=1,..., n$ is the number of CV;

$F_{i,j}$:the feature importance which is calculated by GBM classifier .

A_j^{learn},,$A_j^{validation}$ the learning accuracy and the validation accuracy of feature.

ε is the real number with very small value.

AUC_j^{learn} : the area under curve (AUC)

$IV_{i,j}$: the information value

The first factor $F_{i,j}$ is the feature importance, which is obtained from training data by GBM. To solve over fitting problem, the n-folder cross validation technique was applied. The less difference between the learning accuracy and the validation accuracy, the more stability of accuracy.

The difference is equal to 0 when the learning accuracy is equal to the validation accuracy. To deal with this case, we use a small ε value to avoid the fraction to be ∞. We added AUC measure because the AUC is a commonly used evaluation metric for binary classification problems like predicting a Good (Buy) or Bad (Sell) decision (binary decision). The interpretation is that given a random positive observation and negative observation, the AUC gives the proportion of the time you guess which is correct. It is more affected by sample in-balance than accuracy. A perfect model will score an AUC of 1, while random guessing will score an AUC of around 0.5. AUC is in fact often predicted over accuracy for binary classification for a number of different reasons. The last factor $IV_{i,j}$ is the Information Value as presented in section 2.2. The information value can determine the importance of each feature to the results of classification. If IV is within the range of [0; 0,02], then this feature is not predictive; if IV is within the range of (0,02; 0,1], then this feature has weak predictive power; if IV is within the range of (0,1; 0,3], then this feature has medium predictive power; and if IV is within the range of (0,3; +∞), then this feature has strong predictive power. With using of the Information Values, we have a filter method for ranking features.

In our proposed method we execute the feature elimination strategy based on backward approach. Each feature will be eliminated by the ranking criterion while the validation accuracy is used to choose subset of features. The new subset is validated by H2O GBM module. The obtained validation accuracy plays a role of decision making. The subset of features from learning phase is used to reduce the dimension of the test dataset. Finally, we extend our previous work to a hybrid feature selection method that combined Information Value (Filter method) and RFE (wrapper method)

4 Experiment and results

Our proposed algorithm is executed on a cluster which can be can be fired up on a laptop, or across the multiple nodes of a cluster of real machines. We used H2O Gradient Boosted Model package. This package is optimized for doing “in memory” processing of distributed, parallel machine learning algorithms on clusters. The proposed algorithm was evaluated on two public datasets: German and Australian credit approval from UCI repository.

4.1 Australian credit

The Australian credit dataset is composed of 690 applicants, with 383 credit worthy and 307 default examples. Each instance contains eight numerical features, six categorical features, and one discriminant feature, with sensitive information being transferred to symbolic data for confidentiality reasons. The results are described in Fig 4.

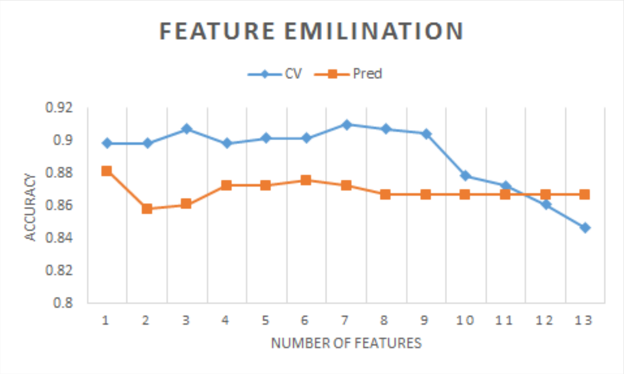


Figure. 4. Accuracy in case of Australian dataset

Table 1 shows the performances of different classifiers over the Australian credit datasets. Baseline is the classifier without feature selection. Classifiers used in [16] include: Linear SVM, CART, k-NN, Naïve Bayes, MLP. Filter methods include: t-test, Linear Discriminant analysis (LDA), Logistic regression (LR). The wrapper methods include: Genetic algorithms (GA) and Particle swarm optimization (PSO).

Table 1. Compare 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
RandomForests						86.67
Our method	87.04 (±0.81)					

The prediction the performances of different classifiers over the Australian credit dataset. The accuracy achieves 87.25% when performing on about 7 features retained. The table shows the classification accuracy of our method is much higher than these studies’ one. Relying on parallel processing, time to run program with 5-fold cross validate taken by our method is only 2974 seconds (~50 minutes).

4.2 German credit dataset

The German credit dataset consists of 1000 loan applications, with 700 accepted and 300 rejected. Each applicant is described by 20 attributes. Our final results were averaged over these 20 independent trials. In our experiments, the default value was used for all parameters. Fig 5 shows the averages of classification results.

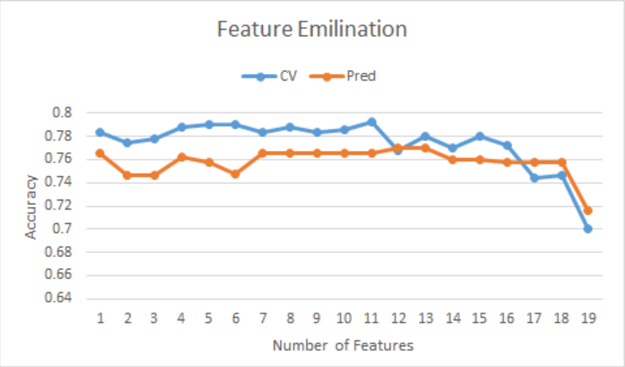


Figure. 5. Accuracy in case of German credit dataset

Table 2 shows the performances of different classifiers over the German credit datasets. The accuracy achieves 77% when performing on about 13 features retained after using the RFE procedure. The table shows the classification accuracy of our method is much higher than other methods. Moreover, the standard deviation is significantly lower.

Table 2. Performances of 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
RandomForests						74.20
Our method	75.52 (±1.25)					

Moreover, relying on a parallel processing strategy, time to run program with 5-fold cross validate taken by our method is only 4311 seconds (~72 minutes) while other methods must run several hours. This result highlights the efficiency in terms of running time of our method when filtering the redundant features.

5. Conclusion

In the present paper, we focused on studying feature selection and GBM method. Our features selection method determined the optimal feature subset with highest accuracy. We have introduced a hybrid feature selection approach based on filter and wrapper methods. The accuracy of classifier using the selected features is better than other methods. With fewer features, a retail bank concentrates on collecting relevant and essential input. The parallel processing procedure leads to a significant decrement in runtime. Therefore, the workload of the staff credit assessment may be reduced, because they do not have to take into database a large number of features during the evaluation procedure. The experimental results show that our method is effective in credit risk analysis. It makes the evaluation more quickly and increases the accuracy of the classification.

References

[1] [1] Y. Jiang, “Credit Scoring Model Based on the Decision Tree and the Simulated Annealing Algorithm,” in *World Congress on Computer Science and Information Engineering*, 2009, no. 2007, pp. 18–22.

[2] [2] A. Laha, “Building contextual classifiers by integrating fuzzy rule based classification technique and k-nn method for credit scoring,” *Adv. Eng. Informatics*, vol. 21, no. 3, pp. 281–291, 2007.

[3] [3] S. Oreski, D. Oreski, and G. Oreski, “Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment,” *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12605–12617, 2012.

[4] [4] M. Saberi, M. S. Mirtalaie, F. K. Hussain, A. Azadeh, O. K. Hussain, and B. Ashjari, “A granular computing-based approach to credit scoring modeling,” *Neurocomputing*, vol. 122, pp. 100–115, 2013.

[5] [5] S. Lee and W. S. Choi, “A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis,” *Expert Syst. Appl.*, vol. 40, no. 8, pp. 2941–2946, 2013.

[6] [6] A. R. Ghatge and P. P. Halkarnikar, “Ensemble Neural Network Strategy for Predicting Credit Default Evaluation,” vol. 2, no. 7, pp. 223–225, 2013.

[7] [7] A. Chaudhuri and K. De, “Fuzzy Support Vector Machine for bankruptcy prediction,” *Appl. Soft Comput. J.*, vol. 11, no. 2, pp. 2472–2486, 2011.

[8] [8] A. Ghodselahi, “A Hybrid Support Vector Machine Ensemble Model for Credit Scoring,” *Int. J. Comput. Appl.*, vol. 17, no. 5, pp. 1–5, 2011.

[9] [9] C.-L. Huang, M.-C. Chen, and C.-J. Wang, “Credit scoring with a data mining approach based on support vector machines,” *Expert Syst. Appl.*, vol. 33, no. 4, pp. 847–856, 2007.

[10] [10]H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. 1998.

[11] [11]I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[12] [12]J. Wang, X. Li, and H. Fan, “Classification of lip color based on multiple SVM-RFE,” pp. 769–772, 2011.

[13] [13] A. R. Hidalgo-Muñoz, M. M. López, I. M. Santos, A. T. Pereira, M. Vázquez-Marrufo, A. Galvao-Carmona, and A. M. Tomé, "Application of SVM-RFE on EEG signals for detecting the most relevant scalp regions linked to affective valence processing," *Expert Syst. Appl.*, vol. 40, no. 6, pp. 2102–2108, 2013.

[14] [14] S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 2052–2064, 2014.

[15] [15] Y. Ling, Q. Y. Cao, and H. Zhang, "Application of the PSO-SVM model for credit scoring," *Proc. - 2011 7th Int. Conf. Comput. Intell. Secur. CIS 2011*, pp. 47–51, 2011.

[16] [16] D. Liang, C.-F. Tsai, and H.-T. Wu, "The effect of feature selection on financial distress prediction," *Knowledge-Based Syst.*, vol. 73, pp. 289–297, 2015.