

Credit scoring with a feature selection approach based deep learning

Van-Sang Ha ¹, Ha-Nam Nguyen ²

¹ Department of Economic Information System, Academy of Finance, Hanoi, Viet Nam

² Department of Information Technology, University of Engineering and Technology, Hanoi, Viet Nam

Abstract. In financial risk, credit risk management is one of the most important issues in financial decision-making. Reliable credit scoring models are crucial for financial agencies to evaluate credit applications and have been widely studied in the field of machine learning and statistics. Deep learning is a powerful classification tool which is currently an active research area and successfully solves classification problems in many domains. Deep Learning provides training stability, generalization, and scalability with big data. Deep Learning is quickly becoming the algorithm of choice for the highest predictive accuracy. Feature selection is a process of selecting a subset of relevant features, which can decrease the dimensionality, reduce the running time, and improve the accuracy of classifiers. In this study, we constructed a credit scoring model based on deep learning and feature selection to evaluate the applicant's credit score from the applicant's input features. Two public datasets, Australia and German credit ones, have been used to test our method. The experimental results of the real world data showed that the proposed method results in a higher prediction rate than a baseline method for some certain datasets and also shows comparable and sometimes better performance than the feature selection methods widely used in credit scoring.

1 Introduction

The main purpose of credit risk analysis is to classify customers into two sets, good and bad ones [1]. Over the last decades, there have been lots of classification models and algorithms applied to analyse credit risk, for example decision tree [2], nearest neighbour K-NN, support vector machine (SVM) and neural network [3-7]. One important goal in credit risk prediction is to build the best classification model for a specific dataset.

Financial data in general and credit data in particular usually contain irrelevant and redundant features. The redundancy and the deficiency in data can reduce the classification accuracy and lead to incorrect decision [8-9]. In that case, a feature selection strategy is deeply needed in order to filter the redundant features. Indeed, feature selection is a process of selecting a subset of relevant features. The subset is sufficient to describe the problem with high precision. Feature selection thus allows decreasing the dimensionality of the problem and shortening the running time.

Credit scoring and internal customer rating is a process of accessing the ability to perform financial obligations of a customer against a bank such as paying interest or an original loan on due date, or other credit conditions for evaluating and identifying risks in the credit activities of the bank. The degree of credit risk changes over individual customers and is identified through the evaluation process. It is based on existing financial and non-financial data of customers at the time of credit scoring and customer rating.

Credit scoring is a technique using statistical analysis data and activities to evaluate the credit risk against customers. Credit scoring is shown in a figure determined by the bank based on the statistical analysis of credit experts, credit teams or credit bureaus. In Vietnam, some commercial banks start to perform credit scoring against customers but it is not widely applied during the testing phase and still needs to improve gradually. For completeness, all information presented in this paper comes from credit scoring experience in Australia, Germany and other countries.

Many methods have been investigated in the last decade to pursue even small improvement in credit scoring accuracy. Artificial Neural Networks (ANNs) [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 [20], stochastic optimization technique and support vector machine [21] have shown promising results in terms of prediction accuracy.

In this study, a new method for feature selection based on various criteria are proposed and integrated with a deep learning classifier in credit scoring tasks.

The rest of the paper is organized as follows: Section 2 presents the background of credit scoring, deep learning and feature selection. Section 3 is the most important section that describes the details of the proposed model. Experimental results are discussed in Section 4 while concluding remarks and future works are presented in Section 5.

2 Materials

2.1 Feature selection

Feature selection is the most basic step in data pre-processing as it reduces the dimensionality of the data. 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 modeling. However, the feature selection is usually a separate step in the whole process of data mining.

There are two different categories of feature selection methods, i.e. filter approach and wrapper approach. The filter approach considers the feature selection process as a precursor stage of learning algorithms. The filter model uses evaluation functions to evaluate the classification performances of subsets of features. There are many evaluation functions such as feature importance, Gini, information gain, the ratio of information gain, etc. A disadvantage of this approach is that there is no relationship between the feature selection process and the performance of learning algorithms.

The wrapper approach uses a machine-learning algorithm to measure the good-ness of the set of selected features. The measurement relies on the performance of the learning algorithm such as its accuracy, recall and precision values. The wrapper model uses a learning accuracy for evaluation. In the methods using the wrapper model, all samples should be divided into two sets, i.e. training set and testing set. The algorithm runs on the training set, and then applies the learning result on the testing set to measure the prediction accuracy. The disadvantage of this approach is highly computational cost. Some researchers proposed methods that can speed up the evaluating process to decrease this cost. Common wrapper strategies are Sequential Forward Selection (SFS) and Sequential Backward Elimination (SBE). The optimal feature set is found by searching on the feature space. In this space, each state represents a feature subset, and the size of the searching space for n features is $O(2^n)$, so it is impractical to search the whole space exhaustively, unless n is small.

2.2 Deep Learning

Deep learning (deep machine learning, or deep structured learning) attempt to model high-level abstractions in data by using multiple processing layers with complex structures or otherwise, composed of multiple non-linear transformations. There are several theoretical frameworks for Deep Learning, but this research focuses primarily on the feed-forward architecture used by H2O. The basic unit in the model is the neuron, a biologically inspired model of the human neuron. In humans, the varying strengths of the neurons' output signals travel along the synaptic junctions and are then aggregated as input for a connected neuron's activation. In the model, the weighted combination $\alpha = \sum_{i=1}^n w_i x_i + b$ of input signals is aggregated, and then an output signal $f(\alpha)$ transmitted by the connected neuron. The function f represents the nonlinear activation

function used throughout the network and the bias b represents the neuron's activation thresh-old. Multi-layer, feed-forward neural networks consist of many layers of inter-connected neuron units, starting with an input layer to match the feature space, followed by multiple layers of nonlinearity, and ending with a linear regression or classification layer to match the output space. Multi-layer neural networks can be used to accomplish Deep Learning tasks. Deep Learning architectures are models of hierarchical feature extraction, typically involving multiple levels of nonlinearity. Deep Learning models are able to learn useful representations of raw data and have exhibited high performance on complex data such as images, speech, and text. The procedure to minimize the loss function $L(W,B | j)$ is a parallelized version of stochastic gradient descent (SGD). Standard SGD can be summarized as follows, with the gradient $\nabla L(W,B | j)$ computed via back propagation.

Parallel distributed and multi-threaded training with SGD in H2O Deep Learning

Initialize global model parameters W, B

Distribute training data T across nodes

Iterate until convergence criterion reached

For nodes n with training subset T_n , do in parallel:

Obtain copy of the global model parameters

W_n, B_n *Select active subset $T_{na} \subset T_n$ (user-given number of samples per iteration)*

Partition T_{na} into T_{nac} by cores n_c

For cores n_c on node n , do in parallel:

Get training example $i \in T_{nac}$

Update all weights $w_{jk} \in W_n$, biases $b_{jk} \in B_n$

$w_{jk} := w_{jk} - \alpha (\partial L(W, B | j)) / \partial w_{jk}$

$b_{jk} := b_{jk} - \alpha (\partial L(W, B | j)) / \partial b_{jk}$

Set $W, B := \text{Avg}_n W_n ; \text{Avg}_n B_n$

Optionally score the model on train/validation scoring sets

3 The Proposed Method

Our method uses Deep Learning to estimate the performance consisting of the cross validation accuracy and the importance of each feature in the training data set. In a multi-node system this parallelization scheme works on top of H2O's distributed setup, where the training data is distributed across the cluster. Each node operates in parallel on its local data. After that, we determine best feature set by choosing the best of Average score + Median Score and the lowest SD. To deal with over-fitting problem, we apply n-fold cross validation technique to minimize the generalization error.

Step 1: Train data by Random Forest via 20 trails, calculate and sort median of variables important

Step 2: Add each feature with best variables important and train data by Deep Learning 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 using selection rules

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

In step 2, we use deep learning with n-fold cross validation to train the classifier. In the j^{th} cross validation, we will obtain a set of $(F_j, A_j^{\text{learn}}, A_j^{\text{validation}})$ that are the feature importance, the learning accuracy and the validation accuracy respectively.

We will use those values to compute the score criterion in step 3.

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

$$F_i^{\text{Score}} = \sum F_{ij} \times (A_j^{\text{learn}} + A_j^{\text{validation}}) \quad (1)$$

The main of our algorithm is presented in step 4. In this step, we will select best of features using rules: the best of Average + Median Score and the lowest standard deviation (SD).

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

These rules guarantee us to get the highest accuracy and the lowest Standard deviation. This proposed method tends to find the smallest optimal set of features in order to reduce the number of output features as much as possible. Then, machine-learning algorithms are used to calculate the relevance of the feature. Based on the calculated value of conformity level, we find the subset of features having less number of features while maintaining the objective of the problem.

4 Experiment and results

Our proposed algorithm was coded using R language (<http://www.r-project.org>), using H2O Deep Learning package. This package is optimized for doing “in memory” processing of distributed, parallel machine learning algorithms on clusters. A “cluster” is a software construct that can be can be fired up on your lap-top, on a server, or across the multiple nodes of a cluster of real machines, including computers that form a Hadoop cluster. We tested the proposed algorithm with several datasets including two public datasets, German and Australian credit approval, to validate our approach. In this paper, we used Random Forest with the original dataset as the base-line method. The proposed method and the base-line method were executed on the same training and testing datasets to compare their efficiency. Those implementations were repeatedly done 20 times to test the consistency of obtained results.

4.1. German credit approval dataset

The German credit approval 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 (Fig. 1). In our experiments, we use the default value for the hidden parameter and the number of epoch parameter was

tried with value of 10. The averages of classification results are depicted in Fig. 1.

The best subset contains 19 features and its accuracy is 74.68 %.

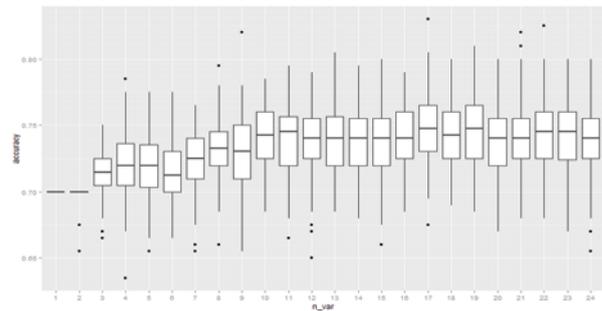


Figure 1. Accuracy in case of German dataset.

Table 1 shows the performances of different classifiers over the German credit datasets. Baseline is the classifier without feature selection. Classifiers used in [22] 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. 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						73.40
Our method	74.68					

Comparing the performances of various methods in Table 1, we saw that the ac-curacy of deep learning on the subset of newly selected features obviously in-creases, and the number of features has been reduced by 21%. The average accuracy is 73.4% on the original data. After applying the feature selection, the aver-age accuracy increases to 74.68%.

Moreover, relying on a parallel processing strategy, time to run 20 trails with 5-fold cross validate taken by our method is only 5286 seconds (~88 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.

4.2 Australian credit approval dataset

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 averages of classification results are depicted in Fig. 2.

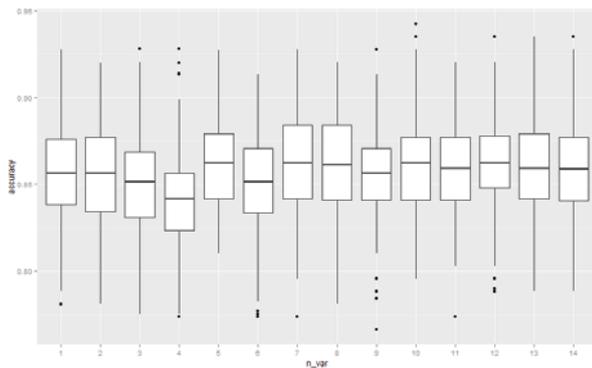


Figure 2. Accuracy in case of German dataset

Table 2 shows that the accuracy of Deep learning on a subset of 7 selected features obviously increases. The average accuracy is 85.82% on the original data. After applying the feature selection, the average accuracy increases to 86.24%. Relying on parallel processing, time to run 20 trails with 5-fold cross validate taken by our method is only 2769 seconds (~46 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						85.82
Our method	86.24					

5 Conclusion

In this paper, we focused on studying feature selection and Deep Learning method. Features selection involves in determining the highest classifier accuracy of a subset or seeking 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 better than other methods. Fewer features allow a credit department to concentrate on collecting relevant and essential variables. The parallel processing procedure leads to a significant decrement in runtime. As a result, the workload of credit evaluation personnel can be reduced, as they do not have to take into account a large number of features during the evaluation procedure, which will be somewhat less computation-ally intensive. 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. E. I. Altman and A. Saunders, "Credit risk measurement: developments over the last 20 years," *Journal of Banking and Finance*, vol. 21, no. 11-12, pp. 1721–1742, (1997)
2. Z. Davoodabadi and A. Moeini, "Building C customers' Credit Scoring Models with Combination of Feature Selection and Decision Tree Algorithms," vol. 4, no. 2, pp. 97–103, (2015)
3. A. Khashman, "A neural network model for credit risk evaluation," *International Journal of Neural Systems*, vol. 19, no. 4, pp.285–294, (2009)
4. T. Bellotti and J. Crook, "Support vector machines for credit scoring and discovery of significant features," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3302–3308, (2009)
5. F. Wen and X. Yang, "Skewness of return distribution and coefficient of risk premium," *Journal of Systems Science and Complexity*, vol. 22, no. 3, pp. 360–371, (2009)
6. X. Zhou, W. Jiang, Y. Shi, and Y. Tian, "Credit risk evaluation with kernel-based affine subspace nearest points learning method," *Expert Systems with Applications*, vol. 38, no. 4, pp.4272–4279, (2011)
7. G. Kim, C. Wu, S. Lim, and J. Kim, "Modified matrix splitting method for the support vector machine and its application to the credit classification of companies in Korea," *Expert Systems with Applications*, vol. 39, no. 10, pp. 8824–8834, (2012)
8. Liu, and Motoda, "Feature selection for Knowledge Discovery and Data mining," *Kluwer Academic Publishers*, 1998.
9. Guyon, and Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, pp 1157-1182, 2003.
10. Oreski, S., Oreski, D., & Oreski, G., "Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment". *Expert Systems with Applications*, 39(16), 12605–12617, 2012
11. Saberi, M., Mirtalaie, M. S., Hussain, F. K., Azadeh, A., Hussain, O. K., & Ashjari, B. , "A granular computing-based approach to credit scoring modeling". *Neurocomputing*, 122, 100–115, (2013)
12. Lee, S., & Choi, W. S., "A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis". *Expert Systems with Applications*, 40(8), 2941–2946, (2013)
13. Ghatge, A. R., & Halkarnikar., "Ensemble Neural Network Strategy for Predicting Credit Default Evaluation", 2(7), 223–225, (2013)
14. Chaudhuri, A., & De, K., "Fuzzy Support Vector Machine for bankruptcy prediction. *Applied Soft Computing Journal*", 11(2), 2472–2486, (2011)
15. Ghodselahi, A., "A Hybrid Support Vector Machine Ensemble Model for Credit Scoring," *International Journal of Computer Applications*, 17(5), 1–5, (2011)
16. Huang L., Chen C., and Wang J., "Credit Scoring with a Data Mining Approach Based on Support Vector Machines," *Computer Journal of Expert*

- Systems with Applications*, vol. 33, no. 4, pp. 847-856, (2007)
17. G. Eason, B. Li T., Shiue W., and Huang H., "The Evaluation of Consumer Loans Using Support Vector Machines," *Computer Journal of Expert Systems with Applications*, vol. 30, no. 4, pp. 772-782, (2006)
 18. Martens D., Baesens B., Gestel T., and Vanthienen J., "Comprehensible Credit Scoring Models Using Rule Extraction from Support Vector Machines," *European Computer Journal of Operational Research*, vol. 183, no. 3, pp. 1466-1476, (2007)
 19. Wang Y., Wang S., and Lai K., "A New Fuzzy Support Vector Machine to Evaluate Credit Risk," *Computer Journal of IEEE Transactions on Fuzzy Systems*, vol. 13, no. 6, pp. 25-29, (2005)
 20. S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert System. Appl.*, vol. 41, no. 4, pp. 2052-2064, (2014)
 21. 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)
 22. Deron Liang, Chih-Fong Tsai, Hsin-Ting Wua, "The effect of feature selection on financial distress prediction," *Knowledge-Based Systems* 73 289-297 (2015)