

Improving Credit Risk Prediction in Online Peer-to-Peer (P2P) Lending Using Feature selection with Deep learning

Van-Sang Ha*, Dang-Nhac Lu**, Gyoo Seok Choi***, Ha-Nam Nguyen****, Byeongnam Yoon*****

*Department of Economic Information System, Academy of Finance, Hanoi, Vietnam

** Academy of Journalism and Communication, Vietnam

***Department of Computer Science, Chungwoon University, Incheon, South Korea

****Information Technology Institute, Vietnam National University, Hanoi, Vietnam

sanghv@hvtc.edu.vn, nhacld@gmail.com, lionel@chungwoon.ac.kr, namnh@vnu.edu.vn, tomayoon@vnu.edu.vn

Abstract— At the early of 21st century, In the UK Peer2Peer (P2P) lending began with Zopa at the start of 2005. After that, it has grown rapidly in United States, China and some other countries. The main challenge for individual investors in the P2P lending market is to allocate their money efficiently through different loans by accurately evaluating the credit of each loan. The traditional ranking models cannot conform to the needs of individual investors in P2P lending because they do not provide natural mechanisms for asset allocation. P2P loans do not have the presence of traditional financial institutions. In this study, we propose a novel method for analyzing data for this emerging market. We have designed a credit risk model based on advanced machine learning methods, capable of assessing the returns and risks of individual loans. We also applied a selection feature method to eliminate the irrelevant features for improving the efficient of machine learning models. We conducted experiments on real-world data sets from P2P lending markets. The experimental results show that the proposed model can improve the efficiency of investment compared to the existing methods of P2P lending.

Keywords— Credit scoring, Deep learning, Feature selection, P2P lending

I. INTRODUCTION

Traditional banking can have profound changes by p2p lending platform over the next decades. P2P lending is the largest online credit market in the world, facilitating personal loans, business loans, medical financing, and more. In 2005, the main objective of P2P lending platforms is democratizing consumer finance services. For more efficient distribution and guarantees of credit, technology was used by these platforms. To borrow money, individuals can make their pitches and credit can be provided by investor without the involvement of financial institutions. The earliest P2P lending platform is Zopa in UK which founded in 2005. Lending Club and Prosper in the US founded in 2006, and CreditEase founded in China in 2006.

Over the last few year, many researchers have developed and applied many models and algorithms to analyse the P2P

lending data. Lee-Eunkyoung et al [3] investigates herding behavior in the P2P lending market in Korea, they found strong evidence of herding and its diminishing marginal effect as bidding advances. In [2], a data mining (DM) approach to predict the performance of P2P loan before funded is proposed by Jin and Zhu. Byanjankar Ajay et al [1] proposed a credit scoring model using artificial neural networks in classifying peer-to-peer loan applications into default and non-default groups. They indicated that the neural network-based credit scoring model performs effectively in screening default applications. Malekipirbazari Milad and Aksakalli [5] proposed a random forest (RF) based classification method for predicting borrower status. Results on data from the popular social lending platform Lending Club (LC) indicate the RF-based method outperforms the FICO credit scores as well as LC grades in identification of good borrowers. H. Li [4] studied on predicting the potential lenders using an outlier detection method to find the abnormal lenders, and they found the detected outliers have bad credit scores with a high possibility. C. Serrano et al [6] focused on predicting the expected profitability of investing in P2P loans, measured by the internal rate of return. J.Yan et al [8] provided loan project recommendations by innovatively mining a huge amount of unstructured data, the text data of borrowers' and lenders' motivations. They solved the problem of mismatches between borrowers and lenders. Recently, Y.Xia et al [7] addressed the problem of reject inference in a specific P2P lending domain from the perspective of semi-supervised learning. They proposed a novel reject inference method (CPLE-LightGBM) by combining the contrastive pessimistic likelihood estimation framework and an advanced gradient boosting decision tree classifier (LightGBM).

P2P Lending data is collected by P2P lending platform, often contains irrelevant and redundant features. The classification accuracy is reduced by this redundancy, and deficiency data. A feature selection strategy should be applied to eliminate redundant features. In other words, feature selection will select an optimal subset of relevant features. With this optimal subset, we solve the problem with high performance. Feature

selection is one of the ways to reduce the dimensionality of the problem and shorten the runtime.

In this paper, we proposed a method based on feature selection integrated with Restricted Boltzmann Machines (RBMs) on credit scoring tasks. RBMs determine a non-linear transform the feature space in an unsupervised learning step. RBMs have been used successfully for learning and representing. Our proposed method select top features by RMSE ranking criterion. Our proposed method is applied to credit scoring problem. Integration with GPU tensorflow, the method showed better results and faster than original RBM.

Structure of this paper consists four main sections. The background of feature selection, deep learning and RBM is presented in Section 2. Section 3 presents and describes the proposed model. Section 4 discusses the obtained results while Section 5 highlights the concluding remarks and future research.

II. BACKGROUND

A. DEEP LEARNING

In recent years, as the computing power of computers has risen to new heights and vast amounts of data collected by large technology firms, Machine Learning has come a long way and one field. The new born was called Deep Learning. Deep Learning has enabled computers to perform seemingly impossible tasks ten years ago: categorize thousands of different objects in photos, create captions for images, mimic human speech and writing, communicating with people, or even writing or music. There are many factors that lead to this risen:

- Large data sets is labelled
- High-speed parallel computing capability of the GPU.
- The improvements of the architecture: GoogLeNet, VGG, ResNet, ... and transfer learning techniques, fine tuning.
- New regularization techniques: dropout, batch normalization, data augmentation.
- Many new libraries support deep network training with GPUs: theano, caffe, mxnet, Tensorflow, pytorch, keras, ...
- Many new optimization techniques: Adagrad, RMSProp, Adam, ...

Boltzmann Machines are a particular form of log-linear Markov Random Field. The energy function of Boltzmann Machine is linear in its free parameters. Limited parametric is set to a non-parametric to make function powerful enough to represent complicated distributions. Some of the attributes is hidden because they are never observed. The modelling capacity of the Boltzmann Machine can increase by having more hidden variables named hidden units. RBMs is a useful algorithm for dimensionality reduction, classification, regression, collaborative filtering and feature learning. RBMs are neural nets with two-layer. They constitute the building blocks of deep-belief networks. The first layer of the RBM is called the input layer (visible), and the second is the hidden layer.

B. FEATURE SELECTION

One of the most relevant topics in machine learning is defining what data attributes will be used in your forecasting models. Data scientists can develop stronger machine learning algorithms than ever before, but its core, garbage data in the garbage data. Today, more and more easily store large amounts of data. However, too many attributes can confuse the model and render the analysis useless. Therefore, before investing time in adjustment algorithms, we need to identify a set of predictive elements with high predictive accuracy without losing directionality. The benefits of reduced predictors include reduced computational complexity, lower storage costs, and easier explanation of the results.

Feature selection methods have three types: filter, wrapper, and embedded methods. In pre-processing steps, filter methods filter out irrelevant, redundant or noisy features based on some univariate or multivariate measure. In the wrapper methods, the predictor is a black box and the predictor performance as the objective function to evaluate the feature subset. Wrapper methods perform better than filter methods because the process of feature selection is optimized for the classifier. However, the wrapper method is expensive when used for large feature space because of high computational cost and each feature subset must be evaluated by a trained classifier. The filter method has lower computational cost and faster, but with less reliability in the classification efficiency than the wrapper methods and better suitable for high dimensional data sets. The embedded recently been developed in which these methods take advantage of both filter and wrapper methods. A hybrid approach using both an independent testing and performance evaluation function of the feature subset.

III. PROPOSED APPROACH

In this section, we presented the main workflow of our proposed RBM with feature selection. The framework diagram showing the sequence of tasks to perform the classification is shown in Figure 1.

1) Pre-processing Data

In this framework, credit data is first pre-processed by a scaling function, a normalization function and a transformation function in pre-processing module.

2) Feature selection with RBM

Our proposed method uses Tensorflow with GPU support to improve running time and accuracy. The proposed method is divided into two phases.

Phase 1: in order to select the top-ranked features, the dataset was trained by RBM. After training the model, we calculate ranking value for each feature using RMSE which capture the most significant information in the RBM, according to a measure of information loss. Based on the top of feature ranking value, we selected only the important features and eliminated the irrelevant feature(s). The output of this phase is top n features.

Phase 2: top n features in phase 1 is used as a filter of testing set.

3) Classification

In this research, we use some classification techniques which might be divided into the three groups based on the type of algorithm they use. The classifiers use a linear, non-linear or rule-based algorithm. Logistic regression (LR) and Linear discriminant analysis (LDA) are classification techniques based on linear algorithms. The two linearly based classifiers are presented first. The classifier using non-linear algorithm. These are Support Vector Machine (SVM), Artificial Neural Network (ANN), and k-Nearest Neighbor (k-NN). The last group of rule-based classifiers is Random forest (RF).

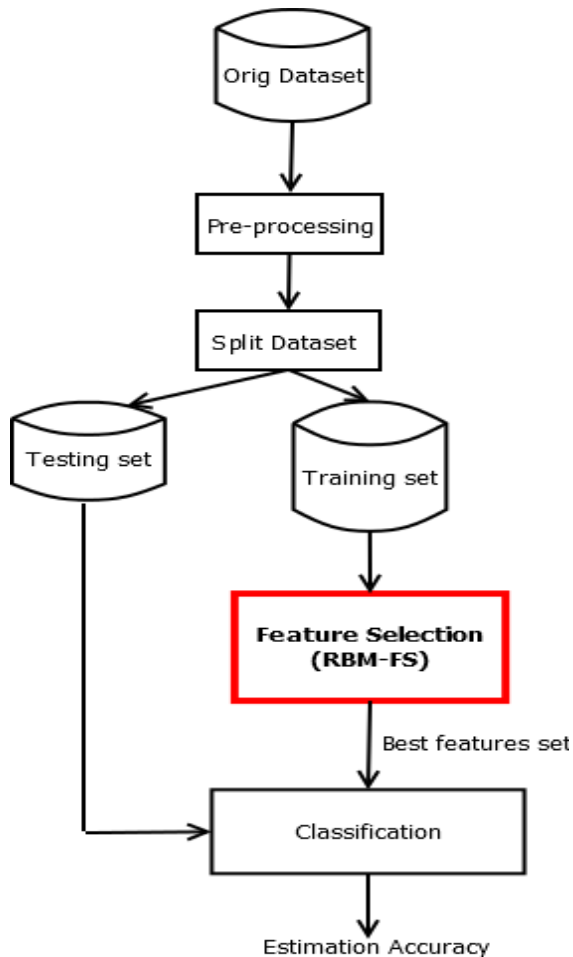


Figure 1. Proposed method

IV. EXPERIMENT AND RESULTS

Our proposed method was implemented using python with Tensorflow GPU. German credit, Australian and lending club data (USA) were used to test our method. LDA classifier is used with 5-fold cross validation in order to evaluate the selected features.

A. Data set

1) Australian credit

Australian credit dataset is composed of 690 applicants where 383 are good credit and 306 are bad credit. Each instance contains 16 features.

2) German credit dataset

German credit data consists of 1000 applicants, with 700 accepted and 300 rejected. Each applicant is described by 24 numeric input features which are finally obtained for classifier modeling.

3) Lending club dataset

The data used in the experiment is available at: <https://www.lendingclub.com/info/download-data.action>. After downloading, merging and pre-processing data, the final dataset is selected containing more than 60,000 records and 123 attributes between 2007 and 2015. It all comes from the personal information provided on their loan application.

B. Results

First, we will attempt to train machine learning algorithms on the full subset of predictors at our disposal. For all of the models, we use 5 fold cross validation to create train / test samples upon which we evaluate our accuracy.

TABLE 1. PERFORMANCE ON FULL MODEL

Models	Predictive Performance (%)		
	German	Australia	Lending Club
LDA	76.50	85.80	81.20
Logistic Regression	76.40	85.51	81.05
ANN	75.8	71.45	66.08
k-NN	67.10	65.94	72.55
Linear SVM	74.80	84.35	76.60
Random Forest	67.72	67.72	67.72

It's evident that the LDA model outperforms the other models on lending club data, Logistic Regression model with an accuracy score of 81.20%. The Linear SVM model comes in a close with 72.55%. The k-NN, ANN and RF Model follow slightly behind with scores of a lower 72.55% and 66%, respectively. So LDA is chosen for our model in classification phase.

Model with feature selection

With such a large number of variables, we can intuitively sense that not all of them will be useful in predicting if a loan will default. A solution to this problem would be to reduce the number of attributes we use in the model. A reduced model would ideally be easily interpretable while retaining predictive accuracy.

First, we select 2 to n attributes we believe to be most informative regarding loan default.

After running feature selection, our best subset includes 12 selected features and the method achieves 86.09% accuracy in case of Australian credit. The best feature subset in case of German credit includes 22 selected feature and accuracy is 76.7%.

In case of Lending club data, the best subset includes 80 features and accuracy achieves 81.53%. Our model outperform the more modern, and powerful, machine-learning algorithms such as Linear SVM, ANN, K-NN and Random Forests.

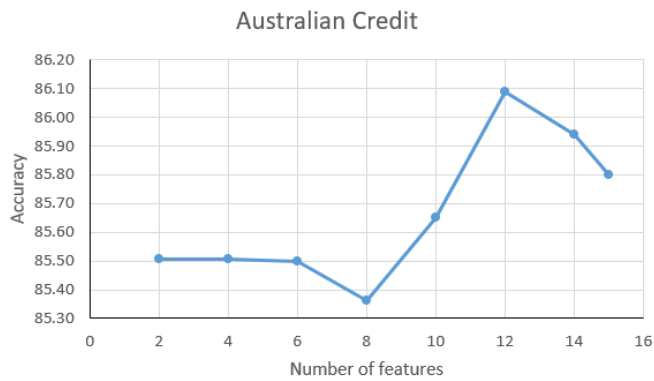


Figure 2. Accuracy on Australian credit

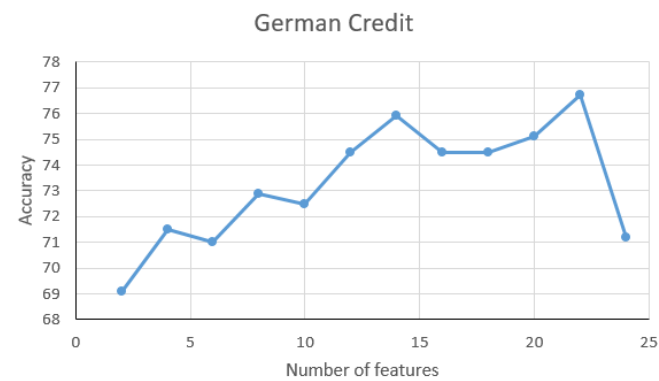


Figure 3. Accuracy on German credit

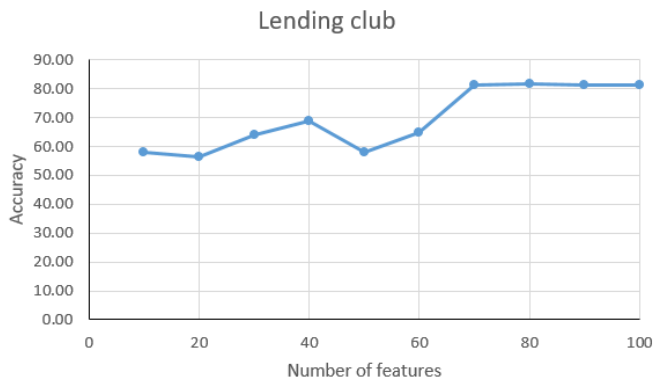


Figure 4. Accuracy on Lending club

V. DISCUSSION AND CONCLUSIONS

In this paper, we have introduced an effective method based on feature selection and deep learning. Our method determined top n features with highest accuracy. Accuracy of LDA classifier using the selected features is better than other methods. The performance of P2P lending platforms can be improved with fewer features. The processing procedure using GPU leads to reduced 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.

ACKNOWLEDGMENT

REFERENCES

- [1] Byanjankar, A., Heikkilä, M., & Mezei, J. (2015). Predicting credit risk in peer-to-peer lending: A neural network approach. *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*, 719–725.
- [2] Jin, Y., & Zhu, Y. (2015). A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending. *Proceedings - 2015 5th International Conference on Communication Systems and Network Technologies, CSNT 2015*, 609–613.
- [3] Lee, E., & Lee, B. (2012). Herding behavior in online P2P lending: An empirical investigation. *Electronic Commerce Research and Applications*, 11(5), 495–503.
- [4] Li, H., Zhang, Y., Zhang, N., & Jia, H. (2016). Detecting the Abnormal Lenders from P2P Lending Data. *Procedia Computer Science*, 91(Ictm), 357–361.
- [5] Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10), 4621–4631.
- [6] Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems*, 89, 113–122.
- [7] Xia, Y., Yang, X., & Zhang, Y. (2018). A rejection inference technique based on contrastive pessimistic likelihood estimation for P2P lending. *Electronic Commerce Research and Applications*, 30, 111–124.
- [8] Yan, J., Wang, K., Liu, Y., Xu, K., Kang, L., Chen, X., & Zhu, H. (2017). Mining social lending motivations for loan project recommendations. *Expert Systems with Applications*.



Gyoo Seok Choi received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Yonsei University, Seoul Korea, in 1982, 1987, and 1997, respectively. He worked at the laboratory of DACOM Company as a researcher from 1987 to 1990. He also worked at the laboratory of SK Telecom Company as a senior researcher from 1991 to 1996. He is currently a professor at the Dept. of Computer Science in Chungwoon University. He is a vice-president of the Korea Institute of Internet, Broadcasting & Communication (IIBC). His current research interests include Artificial Intelligence, Telematics, Mobile Computing, etc.



Ha-Nam Nguyen received his BSc in Information Technology from Hanoi University of Science and Technology (HUST) in 1998, MSc in Computer Science from Chungwoon University (CWU), Korea in 2003, and PhD in Software Applications from Korea Aerospace University (KAU), Korea in 2007, respectively. At present, he is a vice-president of Information Technology Institute, Vietnam National University in Hanoi. He has more than 10 years of teaching and research experience in the areas of financial risk analysis, bioinformatics, behavior analysis, maritime logistics/transport and developing information systems using techniques from data analysis, modelling, and software engineering.



Byeongnam Yoon (M'97) He became a Member of IEEE in 1997. He was born in Seoul Korea 1949. He got the PhD in computer science, Chungnam National University, Korea, 1997. He worked for the Sperryrand UNIVAC as a Computer Specialist 1974 -1978, Samsung as a Manager of Telecommunications Section 1978 -1982, Electronics & Telecommunications Research Institute (ETRI) as a Principal Researcher 1982 – 1999, National Information society Agency (NIA) as a Senior Executive Director General 1999 – 2010, Kyonggi University (KGU) as an Associate Professor, Faculty of Computer Science 2010 - 2016. Global IT Research Institute GIRI as a President 1999 – current. His research area include a Telecommunications, Internet, Software, Web programming & security, e-Government, Enterprise Architecture, Work Flow, Information Control Nets. System Work Method. SPiCE, CMMI, BPM, etc.



Van-Sang Ha received his BSc in Information from Hanoi University of Education (HUE) in 2004, MSc in Information System from VNU - University of Engineering and Technology ,Vietnam in 2009, and PhD in Information System from VNU - University of Engineering and Technology, VietNam in 2018, respectively. At present, he is a lecturer of Department of Economic Information System, Academy of Finance, Hanoi, Vietnam.

He has more than 10 years of teaching and research experience in the areas of financial risk analysis, behavior analysis, and developing information systems using techniques from data analysis, modelling.



Dang-Nhac Lu received his BSc in Information from Hanoi University of Education (HUE) in 2004, MSc in Information Technology from VNU - University of Engineering and Technology, Vietnam in 2009. At present, he is a lecturer at Academy of Journalism and Communication, Vietnam. He has more than 10 years of teaching and research experience in the areas of behavior analysis, and developing information systems using techniques from data analysis, modelling.

ing information systems using techniques from data analysis, modelling.