A DBM-Based Ensemble Method for Improving Default Risk Prediction of Peer-to-Peer (P2P) Lending

Shan Gao*, Xuefeng Wang
Harbin Institute of Technology, Harbin and 150001, China.

* Corresponding author. Tel.: (86)18045660189; email: shangao@hit.edu.cn

Manuscript submitted May 28, 2020; accepted October 13, 2020.
doi: 10.17706/ijeeee.2021.11.3.85-92

Abstract: With the peer-to-peer lending (P2P) business growing up, the most important influencing factor for the healthy development of this industry is the default risk of borrowers. Because the behavior between lenders and borrowers is real time, naturally large amounts of transaction data are being generated all the time. However, it is difficult to extract useful representative features and choose an appropriate model to predict the default risk of the borrowing behavior. In this paper, we proposed a (Deep Boltzmann Machines) DBM-based ensemble method for the default risk prediction in p2p lending, which is based on the real data generated by Lending Club company. Experimental results on the real world data indicate that our model is more effective and powerful with a 0.9093 explanation power.

Key words: Peer to peer lending, default risk prediction, deep Boltzmann machines, deep learning.

1. Introduction

As the most important representative of the Fintech services, peer to peer lending (P2P) platforms works through the way of automatically matching borrowers and lenders which generates huge transaction data. How to deal with the big data to solve critical issues occurring in social, economic and technical spheres [1], [2] is becoming more and more important and significant in real world.

There are three major participants involving in the transaction procedure: borrowers, lenders and platforms. However, the biggest default risk is generated by borrowers, because the lenders suffer financial loss when the borrowers they invest do not pay or partially pay the money back in the duration. Research on the default risk prediction of peer to peer lending borrowers is popularly being studied. With the big transaction data emerging and several P2P lending platforms releasing their data to public, the labeled and unlabeled data were frequently used by applying different statistical methods or machine learning and deep learning methods in previous studies [3]-[5]. The Lending Club in the United States provides complete loan data containing the current loan status and latest payment information (https:// www.lendingclub.com). The company provides every item of borrowing information with 111 attributes. The structure of the data is huge and complex and difficult to select the features. So it is very significant to extract and classify the features to improve the performance of the default risk prediction.

In this study, we select 78 features and classify them into five categories, which is predictor, loan product information, borrower information, credit score and credit behavior. Then we propose a DBM-based ensemble method which improves the default risk prediction of lending club. The DBM-based ensemble method effectively determine whether or not different kind of data will work on the final default risk prediction. Contrast to the prior research on P2P lending with deep learning, we focus on the classification on different labels of variables also with a new learning program which is much more suitable for our reseach.

The remaining of the paper is organized as follows. Section 2 introduces the theoretical background of P2P lending and DBM-based SVM ensemble learning method. Subsequently, we introduce the experiment study in Section 3. And the result and conclusion will be shown in the following.

2. Theoretical Background

2.1. Prior Research on the Default Risk of Peer to Peer Lending

In the past five years, the default risk prediction is becoming popular reality topic with many methods. We will summarize the main research papers in Table 1 and the detail as following. At the beginning, the transaction data is still in its infant period. Some researchers apply the logistic regression to predict the risk performance. [6] proposed a method of predicting the number of obligations over a certain period using the borrower's social media account information. [7] tried to solve behavior scoring problem in social lending using the neural networks combined with logistic regression. With the database developing and machine learning method applying, credit behavior and risk performance are being solved in a new and much accurate way. Support vector machine and random forest prediction method for P2P loan combined with smooth information related to text description is used by [8]. [9] utilized the time series characteristics of social lending and proposed an ensemble method based on random forest to predict defaults. [10] used XGboost machine learning algorithm, Light GBM, using 'multi-observation' and 'multi-dimensional' data cleaning methods. [11] used two-stage model of wide and deep learning to predict the scoring approach. [12] heterogeneous ensemble learning to predict peer to peer lending default in China.

Table 1. Prior Research on the Default Risk of Peer to Peer Lending

Time	Author	Research content	Method	Data
2018	Wang et al.	Behavioral scoring	Ensemble mixture random forest	Time Series
2018	Ma et al.	Default prediction	LightGBM and XGboost	P2P transaction
2017	Jiang et al.	Default prediction	Support vector machine, random forest	P2P transaction
2017	Ge et al.	Default prediction	Logistic regression	Social media information
2017	Huo et al.	Default prediction	Logistic regression,Neural network	P2P transaction
2017	Zhang et al.	Default prediction	Long short-term memory	Time series, small amount
2016	SerranoCinca and GutiérrezNieto	Internal rate of return	Decision tree	P2P transaction
2016	Polena and Regner	Default prediction	Regression	P2P transaction
2015	Byanjankar et al.	Credit scoring	Neural network	P2P transaction

2.2. DBM-Based Ensemble Model

Compared with traditional methods, artificial intelligence is proved to be much more effective in risk prediction than before. We can see the point from the recent papers, like genetic algorithm (GA) [10], support vector machine [13], least square SVM [14], extreme learning machine [15] , to the latest deep learning [16]. As a typical deep learning algorithm, deep boltzmann machines (DBM) which has sufficient hidden layers has a strong capability of feature learning. There is only full connectivity between subsequent layers and no connections within layers or between non-neighbouring layers are allowed. A DBM-based ensemble method has three stages: stage 1 is the partition data then stage 2 is training base classifiers, stage 3 is final ensemble (the whole paradigm in Fig. 1).

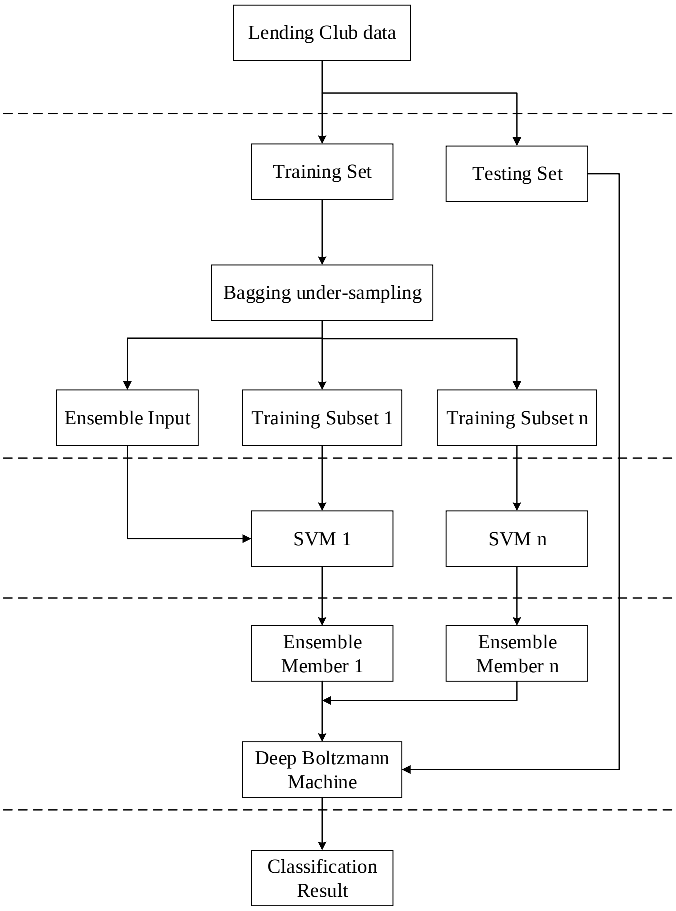


Fig. 1. General framework for DBM-based ensemble learning paradigm.

As shown in Fig. 1, it is a three-stage DBM-based ensemble learning model. From the top of the Fig, it is the first stage we input the original data into two parts called partitioning data phase. Coming to the second stage is training base classifiers. In this stage, the classical SVM model is selected as the base individual classifier due to its high accurate and less prone to over- fitting than other methods in binary classification problems [17]. In our study, the individual base classifiers are diverse via different re-sampling data in Stage 1, and resulting in the diversity of ensemble input members. The third stage is final ensemble, after getting the ensemble input set EI, the subsequent work is to choose a suitable ensemble strategy from a variety of different ensemble strategies for final classification. The DBM-based ensemble strategy is proposed in this study, for its unique advantage in seeking information that exists in the ensemble members.

Originated from artificial neural network, deep learning is a branch of machine learning which is featured by multiple non-linear processing layers and tries to learn hierarchical representations of data. In Section A, we have introduced a brief introduction to some major deep learning techniques that have been applied in P2P lending. In the following, three deep architectures including RBM, DBN, DBM and their corresponding variants are reviewed. Deep Boltzmann machine (DBM) can be regarded as a deep structured restricted Boltzmann machine (RBM) where hidden units are grouped into a hierarchy of layers instead of a single layer [2]. As a special type of Markov random field, restricted Boltzmann machine (RBM) is a two-layer neural network forming a bipartite graph that consists of two groups of units including input units **i** and hidden units **h** under the constraint that there exists a symmetric connection between input units and hidden units and there are no connections between nodes with a group.

Given the model parameters $\theta = [\mathbf{W}, \mathbf{b}, \mathbf{a}]$, the energy function can be given as:

$$E(\mathbf{i},\mathbf{h};\theta)=-\sum_{i=1}^{1}\sum_{j=1}^{J}w_{ij}i_i h_j-\sum_{i=1}^{I}b_i i_i-\sum_{j=1}^{J}a_j h_j \tag{1}$$

that $\boldsymbol{w_{ij}}$ is the connecting weight between input unit $\boldsymbol{i_i}$, whose total number is *I* and hidden unit $\boldsymbol{h_j}$ whose total number is $\boldsymbol{J}$, $\boldsymbol{b_i}$ and $\boldsymbol{a_i}$ denote the bias terms for input units and hidden units, respectively. The joint distribution over all the units is calculated based on the energy function $E(\mathbf{i},\mathbf{h};\theta)$ as:

$$p(\mathbf{i},\mathbf{h};\theta)=\frac{\exp(-E(\mathbf{i},\mathbf{h};\theta))}{Z} \tag{2}$$

where $Z=\sum_{\mathbf{h},\mathbf{i}}\exp(-E(\mathbf{i},\mathbf{h};\theta))$ is the partition function or normalization factor. Then, the conditional probabilities of hidden **h** and input units **i** can be calculated as:

$$p\left(h_j=1\,|\,i;\theta\right)=\delta\left(\sum_{i=1}^{I}w_{ij}i_i+a_j\right) \tag{3}$$

$$p\left(i_i=1\,|\,i;\theta\right)=\delta\left(\sum_{j=1}^{J}w_{ij}h_j+b_i\right) \tag{4}$$

where δ is defined as a logistic function, i.e. $\delta(X)=1/(1+\exp(x))$. RBM is trained to maximize the joint probability. The Fig. 2 shows how the DBM gradually formulated and difference between RBM, DBN and DBM. As shown in every note and the relationship between each other, the main difference between DBN (deep belief network) and DBM lies that DBM is fully undirected graphical model, while DBN is mixed directed/undirected one. Different from DBN that can be trained layer-wisely, DBM is trained as a joint model. Therefore, the training of DBM is more computationally expensive than that of DBN.

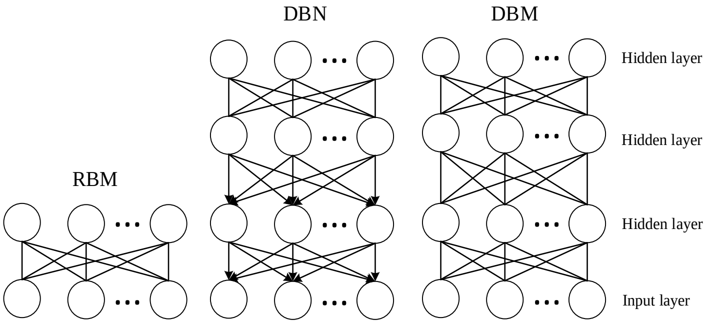


Fig. 2. Differences between RBM, DBN and DBM.

3. Experimental Study

3.1. Dataset

We used the dataset of lending club in the United States. According to the official explanation, Lending Club classify the loan status into 6 situations, which is “fully paid” “charged off” “in grace period” “Late 16-30 days” “Late 31-120 days”. In our study, we only choose the situation of fully paid and charge off to stand for the default risk prediction. We choose the database range from January 2018 to March 2019.

The dataset contains 110 attributes, according to the feature characteristic, we manually divided the attributes into 4 groups (except the predictor): loan product information, borrower information, credit history, credit behavior. After preprocessing, 78 attributes were selected to do the test. Table 2 lists the attributes used in the experiment. Because the features range from nominal to numeric, we need to scale all the variables in the [0.1] range as the inputs to the DBM model. According to [18], we used min-max normalize way to deal with continuous variables. The dataset is split in the ration 80:20 for the training and test as usual.

Table 2. List of Selected Features

Category	Features	Description	Type
predictor	Loan status	Fully paid ,charge off	binary
Borrow info	Employment length	Years(0-10)	nominal
	Home ownership	Rent, Own, Mortgage, Other	nominal
Credit behavior	Delinquent Ing last 6months And so on (59variables)	The past-due amount owed for the accounts on which the borrower is now delinquent.	numeric
predictor	Loan status	Fully paid ,charge off	binary

3.2. Results and Analysis

Different from other research [10], [19], [20], we classify the credit dataset of Lending club into two parts: credit score and credit behavior. Because the data of credit behavior is born with the process of loaning, it's significant to clear the working effect of behavior on final default risk. However, the credit score is independent from the peer to peer lending market which is past behavior of the borrowers in traditional financial market. That's the motivation of our initial attributes dealing base on the real world phenomenon.

Table 3. The Results of DBM Model

	Without behavior	Without score	All attributes
DBM+SVM	0.8402	0.8347	0.9093
SVM	0.8782	0.6659	0.6802

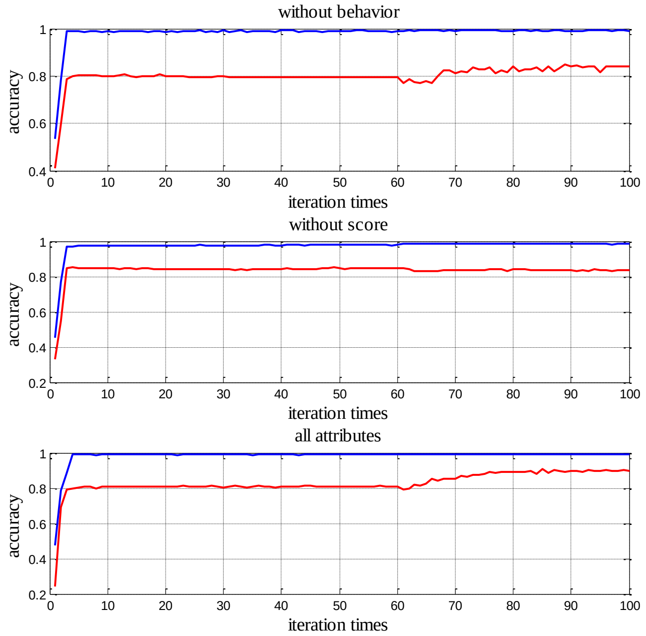


Fig. 3. Comparison between DBM+SVM and SVM.

Table 3 shows that compared with SVM, DBM-based ensemble method achieved a higher performance of the attributes. That means it's much more predictive in the default risk with a higher value and applicant. The results of DBM accuracy is shown directly in Fig. 3. The blue line stands for the training set, the red is testing dataset. The figure shows the training phase of all three experiments steady and the test results are local convergence in the end.

In the phase of DBM model, there are two parts, the supervised part and the unsupervised part. The precise result of DBM model depends on reasonable parameter adjustment. The parameters of this proposed model are shown in Table 4. On the basis of certain generalization ability and less cumulative error, two hidden layers are selected after a lot of experiments. Because more hidden layers can produce more cumulative errors and generate a bad result. In the table 4, the fourth column has five different parameters, that '78' ,'73' and '21' are the number of attributes of 'all attribute' 'without score' 'without behavior', then '2' is the number of the new class. The iteration time of RBM and the fine tune time of DBM should be small to prevent over-fitting. Iteration number of classification is required to be larger to make the training error as small as possible.

Table 4. The Parameters of DBM Model

Lending club data	Number of hidden layers	Number of input attributes	Node number of every layer	Iteration time of RBM	Fine tune time of DBM	Iteration number of classification
All attributes	3	78	78-50-50-200-2	50	20	100
Without score	3	73	73-50-50-200-2	50	20	100
Without behavior	3	21	21-50-50-200-2	50	20	100

4. Conclusion

The industry of peer to peer lending is increasing all over the world. Except Lending club, we also can see other companies like prosper, kiva leading the way up. No matter what the fund will go for, the basic bottom line to keep the industry blooming is to accurately predict and calculate the default risk of every loaning. In this paper, we proposed a DBM-based ensemble structure for predicting the default risk of Lending club using the database of 2018-2019. Deep Boltzmann machines have the potential of learning internal representations that become increasingly complex ， which is considered to be a promising way of solving objective problems. And the high-level representations can be built from a large supply of labeled and unlabeled inputs. Unlike deep belief networks, the inference procedure can incorporate top-down feedback, allowing DBM to better propagate uncertainty inputs. Hence the result is very obvious that the model has a strong explanation no matter the variables chosen comparing to the tradition SVM. In the experiment study, we can see DBM-based ensemble model performs very smoothly, with more credit score and credit behavior variables, it is gaining a much powerful explanation and default prediction.

Because peer to peer lending is traded online every minute, it's possible to collect and observe the data from all aspects to prevent the default risk happening. Some p2p lending companies also provide the access to customers' social network accounts, it's possible to obtain more inputs in our proposed model. Therefore, it's very practical and significant to design a good model of default risk prediction which can be extended applied in the future. That's the point and the first step in our paper. We expect with the data expanding the model will achieve much better performance.

As for the limitation and constraint about our research, the public dataset is just part of the borrowers' information, especially some important variable like salary is self-reported. That absolutely will affect the final result. Beyond that, the generalization of our proposed model will be another obstacle for the research because the public dataset offered by Lending Club is different from other leading P2P lending companies like Prosper. So we will apply our proposed DBM method in more practical dataset.

5. Recommendation

Improve the credit background investigation of borrowers in P2P lending market: As our research result shows that credit score and credit behavior play the same significant role in the default prediction of P2P lending market. Therefore, more detailed information about the borrowers' credit behavior and score will benefit the whole industry healthy development.

Improve and apply more new deep learning methods in Fintech industry: with the generation of big data in emerging financial market, it seems new deep learning method can do more than the traditional way. Our research shows that compared to SVM method, DBM proves a stronger prediction power of default.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Shan Gao and Xuefeng Wang analyzed the data and constructed the model; Shan Gao collected data and made programming experiment; Xuefeng Wang wrote and proofread the paper; all authors had approved the final version.

Acknowledgment

This work was supported by a grant from the National Natural Science Fund China (No. 71771066).

References

[1] Kim, T. Y., & Cho, S. B. (2018). Web traffic anomaly detection using C-LSTM neural networks. *Expert Systems with Applications, 106*, 66-76.

[2] Ronao, C. A., & Cho, S. B. (2016). Anomalous query access detection in RBAC-administered databases with random forest and PCA. *Information Sciences*, *369*, 238-250.

[3] Li, W., Ding, S., Chen, Y., & Yang, S. (2018). Heterogeneous ensemble for default prediction of peer-to-peer lending in China. *IEEE Access*, *6*, 54396-54406.

[4] Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. (2018). A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electronic Commerce Research and Applications*, *27*, 74-82.

[5] Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2017). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, 1-19.

[6] Ge, R., Feng, J., Gu, B., & Zhang, P. (2017). Predicting and deterring default with social media information in peer to peer lending. *Journal of Management Information Systems*, *34*, 401–424.

[7] Huo, Y., Chen, H., & Chen, J. (2017). Research on personal credit assessment based on neural network-logistic regression combination model. *Open Journal of Business and Management*, *5*, 244.

[8] Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2017). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, 1-19.

[9] Xu, J., Chen, D., & Chau, M. (2016). Identifying features for detecting fraudulent loan requests on P2P platforms. *Proceedings 2016 IEEE Conference on Intelligence and Security Informatics (ISI)* (pp. 79-84).

Tucson: IEEE.

[10] Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, *31*, 24-39.

[11] Bastani, K., Asgari, E., & Namavari, H. (2019) Wide and deep learning for peer-to-peer lending. *Expert Systems with Applications*, *134(15)*, 209-224.

[12] Li, W., Ding, S., Wang, H., Chen, Y., & Yang, S. (2019). Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in China. *World Wide Web*, 1-23.

[13] Chen, M. C., & Huang, S. H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications*, *24(4)*, 433-441.

[14] Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, *37(4)*, 543-558.

[15] Lai, K. K., Yu, L., Zhou, L., & Wang, S. (2006). Credit risk evaluation with least square support vector machine. *Proceedings International Conference on Rough Sets and Knowledge Technology* (pp. 490-495). Berlin, Heidelberg: Springer.

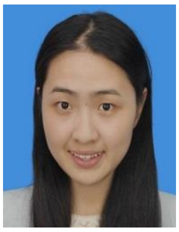
[16] Zhou, H., Lan, Y., Soh, Y. C., Huang, G. B., & Zhang, R. (2012). Credit risk evaluation with extreme learning machine. *Proceedings 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 1064-1069). Seoul: IEEE.

[17] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20(3)*, 273-297.

[18] Kantardzic, M. (2011). *Data Mining: Concepts, Models, and Algorithms*. New Jersey: John Wiley& Sons.

[19] Fu, Y., (2017). Combination of random forests and neural networks in social lending. *Journal of Financial Risk Management*, *4(6)*, 418–426.

[20] Kim, J. Y, & Cho, S. B. (2019). Predicting repayment of borrows in peer-to-peer social lending with deep dense convolutional network. *Expert Systems*, e12403.



Shan Gao is a doctoral candidate in school of management, at Harbin Institute of Technology, Harbin, Heilongjiang province, China. She received her BS degree major in economics from Shandong University in Jinan, Shandong province, China. She received her MS degree major in economics from Harbin Institute of Technology, Harbin, Heilongjiang province, China. Her research focuses on identifying and understanding key pathways involved in the development of deep learning in peer to peer lending with big data development and how the macro-economic environment works on the micro-loan behavior.

Xuefeng Wang is a professor of economics in school of management with 17 years of teaching and research experience, at Harbin Institute of Technology, Harbin, Heilongjiang province, China. He has a wide range of expertise in financial economics and financial math analytical technology, along with strong background of math. His research interests focus on the option and stock pricing model and investment strategy, dynamic searching algorithm for multi-layer neural network.