International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 34, No. 4, August, 2020, pp. 501-506
Journal homepage: http://iieta.org/journals/ria

A Big Data-Based Anti-Fraud Model for Internet Finance

Fei Liu[1*], Yang You[2]

[1] Economics and Management School, Wuhan University, Wuhan 430072, China
[2] Institute of Finance, Chinese Academy of Social Sciences; China Banking and Insurance Regulatory Commission, Beijing 100033, China

Corresponding Author Email: lflios@whu.edu.cn

https://doi.org/10.18280/ria.340416

Received: 5 April 2020
Accepted: 12 June 2020

Keywords:
machine learning (ML), random forest (RF) algorithm, big data analysis, risk control model, internet finance, anti-fraud model

ABSTRACT

Fraud has become a serious risk to the burgeoning industry of Internet finance. To predict and prevent the fraud, this paper develops a risk control model that accurately predicts the fraud behaviors of online loan borrowers, based on their social network data and consumption information. Firstly, statistical analysis was carried out to clarify the relationship between user behavior and fraud, and that between social network and fraud. Next, suitable variables were selected through big data analysis, and used to build the risk control model, in combination with random forest (RF) algorithm and modern financial theory. The effectiveness of the proposed model was confirmed through contrastive experiments. Our model provides peer-to-peer (P2P) online loan platforms with an effective tool to prevent fraud.

1. INTRODUCTION

As a popular mode of Internet finance, peer-to-peer (P2P) online loan refers to the direct leading from individual investors to other individuals via third-party websites, i.e. online loan platforms, eliminating traditional financial intermediaries like banks. While bringing immense business opportunities, P2P online loan leads to various credit risks. With the help of the Internet, P2P online loan platforms can lend funds to ordinary borrowers through a rapid and intelligent audit process. In the meantime, the platforms are frequently exposed to fraud and other risks.

Statistics have shown that 18% of applications for P2P online loans are rejected because of fraud [1]. In fact, false information and credit default are common in many social networks. Therefore, it is very important for P2P online loan platforms to identify fraud in an accurate and timely manner.

Internet credit investigation, an emerging credit service, essentially evaluates the credit level of users based on the data on their historical activities online. The activities of Internet users involve a dazzling array of intercorrelated attributes. Compared with traditional credit evaluation system, Internet credit investigation can pinpoint the credit level and credit changes of each user on time, using the diverse user-generated data. This is a unique advantage of Internet finance.

The online data of user behaviors are complex, multidimensional, and huge in size. As a result, Internet credit investigation must make reasonable use, storage, and variable selection of the big data. It is urgent to predict the relationships generated in social networks, and apply them to the credit rating and credit behavior prediction of users. Only in this way can the possibility of fraud be foreseen and prevented in time.

This paper intends to screen the false information and fraud of Internet finance by integrating the fraud rules into mathematical models, and making full use of big data analysis and machine learning (ML) simulation. Specifically, the default risk of borrowers was evaluated based on the online data of their historical behaviors. The fraud probability of each loan applicant was calculated according to his/her attributes in the online data, and used to give a comprehensive credit score. Finally, countermeasures were formulated for borrowers with different levels of credit scores.

2. LITERATURE REVIEW

P2P online loan, especially its risk control, has attracted much attention in the academia. The risk control of P2P online loan is to evaluate the credit level and predict the behavior of users [2, 3]. The relevant studies mainly focus on two aspects: the selection of evaluation indices [4-6], and the establishment of the evaluation model [7-9].

On the selection of evaluation indices, most scholars evaluate the credit level of users based on their basic information, including standard data that can be accurately quantified, and nonstandard data that cannot be accurately quantified. The standard data include historical credit of the borrower, and verified financial data. The nonstandard data encompass physical features of the borrower (e.g. gender, age, and skin color), and the behavioral features of the borrower (e.g. online behavior features, occupation, and social relationship). Starting with physiological features, Galindo and Tamayo [10] explored whether a user can successfully borrow money, and which factors affect the interest rate to be paid by the borrower; the results show that race, gender and appearance are the borrower features that determine the successful access to P2P online loans. Based on the borrower's avatar, Gonzalez and Loureiro [11] how the borrower's personal features impact P2P loan decision-making, revealing that the success of the loan depends heavily on the borrower's gender, age, and appearance. On the establishment of the evaluation model, credit evaluation has been widely researched with multidimensional multivariate model and some artificial intelligence (AI) methods. In the era of big data, these methods are often combined with modern financial theory, as well as the latest ML algorithms for data mining [12-15], namely, support vector machine (SVM) [16-17], and neural network (NN) [18-19], providing effective solutions to the problem of Internet finance fraud.

3. BIG DATA ANALYSIS OF ONLINE LOAN BORROWERS

In the current Internet environment, it is very convenient to acquire all sorts of data. Then, how to make reasonable and effective use of the big data becomes an important question. In traditional finance, the application for most financial services (e.g. loan and credit card) is approved only with mortgage. The approval requires complicated procedures and proofs. Relying on advanced technology, the Internet finance greatly simplifies the approval process, eliminating the information asymmetry in traditional finance. With the aid of big data, the various traces left by users online can be used for risk rating and credit assessment, and the manual assessment can be completed automatically with machines.

3.1 Descriptive data analysis

Table 1 presents the three datasets used in this research. The consumption dataset stores the consumption data of users on an online loan platform. Four variables are included in this dataset, including user identity (ID), type of expense and income, amount of expense and income, and consumption time. The social dataset stores the social data of users, and involves two variables: user ID and friend ID. The status dataset stores the data on user type, and has two variables: fraud user, and normal user. The variables of consumption dataset and social dataset are listed in Tables 2 and 3, respectively.

Table 1. Dataset description

Name	Meaning	Number of variables	Number of observations
Consumption	Dataset of user consumption	4	256,258
Social	Dataset of user social information	2	346,920
Status	Dataset of user type	2	986

Table 2. Variables of consumption dataset

Dataset	Number of observations	Variables	Value
Consumption	256,258	user_id	1,568
		user_type	(0, 1, 2, 3)
		expense	a positive number represents an expense, and a negative number represents an income
		time	2015/01/01-2019/12/31

Table 3. Variables of social dataset

Dataset	Number of observations	Variables	Value
Social	346,920	user_id	1,568
		friend_id	224,741

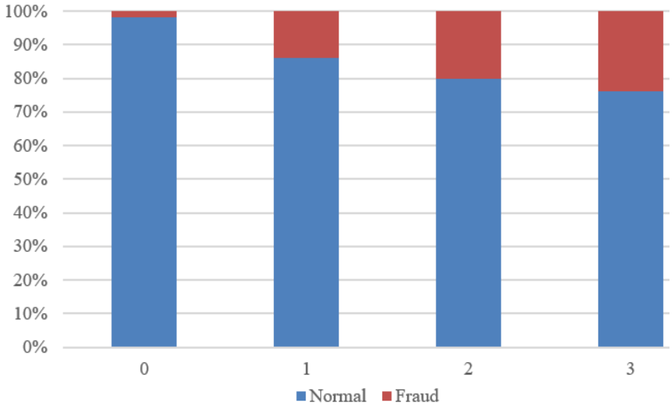

Figure 1. Distribution of fraud and normal users in different consumption types

Based on the big data, the risk control model for online loan platform needs to cover all kinds of information related to the borrowers, and differentiate fraud users from normal users based on their differences. Fraud users and normal users differ greatly in consumption behavior. If a user has committed a fraud, he/she usually faces a huge imbalance between income and expense, and tends to exhibit sudden and significant changes in consumption habits and amount. Therefore, the consumption of users was divided into four types (0, 1, 2, 3). Then, the distribution of fraud and normal users in the four consumption types was further investigated (Figure 1).

As shown in Figure 1, the fraud rates of types 0-3 were about 2%, 14%, 20%, and 24%, respectively. Judging by the proportions, the behaviors of fraudulent users are exactly the opposite to those of normal users.

The contingencies between consumption type and user type are presented in Table 4, and the relationship between them was tested by the Chi-Square test of independence. The test results show that the two data are independent of each other.

Table 4. Contingencies between consumption type and user type

User type	Normal user	Fraud user	Fraud rate
0	155,086	3,165	1.999%
1	34,425	5,611	14.015%
2	19,905	4,977	20.002%
3	759	231	23.333%

3.2 Variable selection

The consumption dataset stores every consumption record of each user. It can be described by three variables: consumption type, consumption amount, and consumption time. Considering the sheer size of the dataset, it is necessary to select as few variables as possible to fully reflect the consumption features of users. For this purpose, feature extraction should be implemented, that is, choose a small amount of data to represent the knowledge of the entire definition domain.

Firstly, the feature variables of consumption type were derived from the user type variable. These variables need to reveal the situation of users in different consumption types, namely, the number of records for each consumption type and the total number of consumption records. As shown in Table 5, the feature variables of consumption type have an impact on the differentiation between normal and fraud users.

Table 5. Feature variables of consumption type

Original variable	New variable	Consumption feature
user_type	user_type0	Number of records for consumption type 0
	user_type1	Number of records for consumption type 1
	user_type2	Number of records for consumption type 2
	user_type3	Number of records for consumption type 3
	type_num	Total number of consumption records

Next, the feature variables of consumption amount were derived from the expense variable. These variables demonstrate the distribution and dispersion of consumption amount. As shown in Table 6, the feature variables of consumption amount either represent income or describe cost.

Table 6. Feature variables of consumption amount

Original variable	New variable	Consumption feature
expense	avg_cost	Mean cost
	std_cost	Standard deviation of cost
	max_cost	Maximum cost
	min_cost	Minimum cost
	avg_income	Mean income
	std_income	Standard deviation of income
	max_income	Maximum income
	min_income	Minimum income
	expense_deviation	Deviation of consumption amount
	expense_kurtosis	Kurtosis of consumption amount

The features of consumption amount vary with consumption types. Hence, 8 new variables were derived, combining the features of consumption amount and consumption type (Table 7).

Finally, the feature variables of consumption time were derived from the features of the time variable, which is contained in the original data. Specifically, the observations with obvious time anomalies were processed, and then divided by time distribution. Considering both income and cost as consumption records, the authors took the absolute value of income to study the features of consumption amount. Since 2015-2016 has fewer consumption records, these two years were treated as a whole. Since 2017, consumption records gradually increased until reaching the peak in 2019, when almost every user had multiple consumption records. As a result, the monthly consumption features of 2017-2019 were adopted as a variable, producing 37 variables.

Through the above analysis, a total of 60 variables were derived. These variables were combined with the 6 variables of the original data, forming a training set of 66 variables.

Table 7. Feature variables of consumption amount of different consumption types

Original variable		New variable	Consumption feature of different consumption types
expense >0	user_type =0	cost0	Mean cost with consumption type 0
	user_type =1	cost1	Mean cost with consumption type 1
	user_type =2	cost2	Mean cost with consumption type 2
	user_type =3	cost3	Mean cost with consumption type 3
expense <0	user_type =0	income0	Mean income with consumption type 0
	user_type =1	income1	Mean income with consumption type 1
	user_type =2	income2	Mean income with consumption type 2
	user_type =3	income3	Mean income with consumption type 3

4. MODEL CONSTRUCTION

The identification of Internet finance fraud is a classification problem. The essence of classification is to learn the historical data on user behaviors, and acquire the ability to classify and predict the behaviors of new users. Traditional classification algorithms include logistic regression algorithm, Bayesian algorithm, decision tree (DT) algorithm, etc. This paper chooses RF algorithm, a forest of randomly created DTs, to classify fraud users and normal users.

4.1 Random Forest (RF) algorithm

The RF is an ensemble learning algorithm [20] extended from DT algorithm and trained by the ensemble learning theory of bagging. In RF algorithm, once the tree classifier is established, each DT will receive a result about the class of a newly inputted user sample. The final result is decided by voting on the classification results of all trees in the RF.

DTs, as the bases of the RF, make full use of statistics and probability to realize classification. The key of each DT lies in the growth of its nodes, that is, how to choose the classification conditions that promote the speed and accuracy of classification decision-making. As each node of the DT grows, it is expected that some datasets will be labeled with their classes after a few steps, making the information purer than before. Generally, information purity is measured by information gain and Gini coefficient.

Information gain is the difference after the change of information entropy. Before selecting a user attribute as its classification condition, the information entropy of the data should be calculated. The information entropy of a node can be calculated by:

$$\text{Info}(D) = -\sum_{i=1}^{C} p_i log_2(p_i) \tag{1}$$

where, D is the sample dataset; C is the number of classes; p_i is the users of class i as a proportion of the total sample.

After a user attribute S has been selected as the node growth condition, the information entropy of attribute S to classify the dataset is $\text{Info}_S(D)$. Suppose attribute S divides the dataset into k parts. The $\text{Info}_S(D)$ value can be calculated by:

$$\text{Info}_S(D) = -\sum_{j=1}^{k} \frac{|D_j|}{D} \times \text{Info}(D_j) \tag{2}$$

Under the effect of attribute S, dataset D becomes purer, that is, the information entropy of dataset D is reduced. The information gain is a function to measure the difference of information entropy of dataset D after node selection:

$$\text{Gain}(S) = \text{Info}(D) - \text{Info}_S(D) \tag{3}$$

During the growth of a DT, the most suitable attribute on a node is the attribute with the largest $\text{Gain}(S)$.

The DT algorithm has a good classification effect. But a single DT often has limited accuracy and encounters overfitting. The ensemble algorithms have been designed to overcome these defects. One of the ensemble algorithms is the bagging algorithm [21].

The RF is a new ensemble algorithm which improves DT algorithm with bagging algorithm. The RF algorithm can be implemented in the following steps:

Step 1. A total of k new datasets of the same size are extracted from the original dataset in k times through bootstrapping. Then, k DTs are generated through the training based on the k new datasets. At the same time, the unselected data in the original dataset constitute another k datasets, denoted as out-of-bag (OOB) datasets.

Step 2. Assuming that there are W input features, when a single tree generates nodes, M variables are randomly selected from the W features. During variable selection, the features with the best classification ability are selected according to the impure nodes of the DT, and the entire DT is formed through recursive subprocess.

Step 3. To minimize the error and avoid overfitting, each tree branch and grow to the maximum extent without any pruning. RF is composed of these growing DTs. When a new sample needs to be classified and predicted, the classification results are determined by voting on the classification results of all trees.

4.2 Model construction

During RF-based modelling, the logistic regression was introduced to replace DT-based variable discrimination. Logistic regression is a generalized linear regression with simple and easy-to-implement mathematical model. Before applying logistic regression model to classification, the sigmoid function should be understood:

$$S(z) = 1/1 + e^{-z} \tag{4}$$

The logistic regression model is implemented as follows:

Step 1. Define hypothesis function $\text{h}(x)$: $h_\theta(x) = 1/1 + e^{-\theta^T x}$.

Based on the sigmoid function, the hypothesis function can be explained as:

$$h_\theta(x) = \text{P}(y = 1|\text{x}; \theta) \tag{5}$$

$$\text{P}(y = 1|\text{x}; \theta) + \text{P}(y = 0|\text{x}; \theta) = 1 \tag{6}$$

Step 2. Solve the model parameters.

The parameters of logistic regression model can be solved in two ways. One is the maximum likelihood estimation, which searches for the parameter value that maximizes the value of the likelihood function. When estimating the parameters of logistic regression model, the likelihood function can be expressed as:

$$\begin{aligned} L(\theta) &= P(D|\theta) = \prod P(y|\text{x}; \theta) \\ &= \prod S(\theta^T x)^y (1 - S(\theta^T x))^{1-y} \end{aligned} \tag{7}$$

Taking the logarithm, the likelihood function can be transformed into:

$$l(\theta) = \sum ylog\, S(\theta^T x) + (1 - y)\log\,(1 - S(\theta^T x)) \tag{8}$$

The other is a common approach in the field of machine learning (ML): solving the parameter value minimizing the loss function. In the logistic regression model, the loss function can be derived from maximum likelihood estimation:

$$C(h_\theta(x), \text{y}) = \begin{cases} -\log(h_\theta(x)) & if\ y = 0 \\ -\log(1 - h_\theta(x)) & if\ y = 1 \end{cases} \tag{9}$$

$$G(\theta) = -\frac{1}{m}[\textstyle\sum_{i=1}^{n} y_i log h_\theta(x_i) + (1 - y_i)\log\,(1 - h_\theta(x_i))] \tag{10}$$

The method of minimizing the loss function is equivalent to the maximum likelihood estimation, except for the relatively high diversity of the function. The loss function was optimized iteratively through gradient descent:

$$\theta^* = arg \min_\theta(G(\theta)) \tag{11}$$

In the process of iteration, the objective function gradually approaches the optimal solution. In each iteration, the direction that maximizes the speed for the objective function to approach the optimal solution was taken as the basis for parameter adjustment. The final estimated values of parameters were obtained when the termination condition is reached or θ converges.

As stated in Section 3, the variables for our model cover two aspects of user information, namely, social information and consumption information. Finally, the authors selected a total of 986 samples, 66 basic variables and 1 objective attribute for the construction of the model.

For comparison, two models were created based on backpropagation neutral network (BPNN) and SVM, respectively, in addition to the RF-based model. The mean false rates of the three models on the training set and the test set were compared through cross validation. As shown in Table 8, the RF-based model achieved the lowest false rate, i.e. the highest classification accuracy.

Table 8. Mean false rate of different models

Model	Mean false rate on training set	Mean false rate on test set
BPNN	0.156	0.181
SVM	0.152	0.167
RF	0	0.161

5. EMPIRICAL ANALYSIS

This section mainly empirically tests the effectiveness of the proposed model. Because DT-based variable discrimination is discarded in model construction, rule importance index and fraud rate can be obtained from the final prediction results. These indices can be converted into credit scores, rather than the simple discrimination between fraud or non-fraud.

To evaluate the predictive ability of our model, the actual fraud rate and predicted fraud rate were calculated for each sample group in training set and test set:

Actual fraud rate = the number of fraud users / the number of samples contained in the group

Predicted fraud rate = the mean predicted fraud rate

The actual fraud rate and the predicted fraud rate of each sample group are compared in Tables 9-10, and Figures 2-3.

Table 9. Comparison of actual fraud rate and predicted fraud rate on the training set

Sample number	Number of samples	Number of fraud users	Actual fraud rate	Predicted fraud rate
1	99	38	0.3838	0.3841
2	99	16	0.1617	0.1502
3	99	10	0.1010	0.1063
4	99	9	0.0909	0.0889
5	99	7	0.0707	0.0716
6	99	6	0.0606	0.0629
7	99	6	0.0606	0.0622
8	99	4	0.0404	0.0429
9	99	2	0.0202	0.0231
10	99	1	0.0101	0.0175

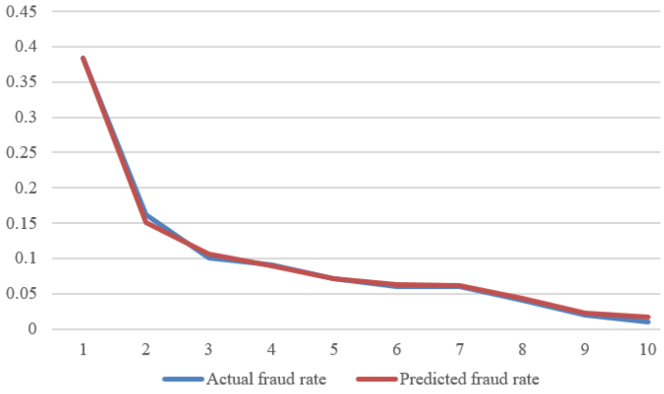

Figure 2. Prediction results on the groups in the training set

Table 10. Comparison of actual fraud rate and predicted fraud rate on the test set

Sample number	Number of samples	Number of fraud users	Actual fraud rate	Predicted fraud rate
1	99	40	0.4040	0.3819
2	99	15	0.1515	0.1531
3	99	14	0.1414	0.1223
4	99	9	0.0909	0.0972
5	99	8	0.0808	0.0752
6	99	5	0.0505	0.0611
7	99	4	0.0404	0.0462
8	99	3	0.0303	0.0389
9	99	2	0.0202	0.0261
10	99	1	0.0101	0.0189

Through comparison, it can be found that the RF-based model has good predictive ability on both balanced and unbalanced datasets. Note that, error rate (the proportion of incorrectly classified samples), a traditional measure of predictive ability, was not used to verify the effect of our model. The fraud rate could be converted into credit score. Then, the users could be classified according to the threshold of credit score.

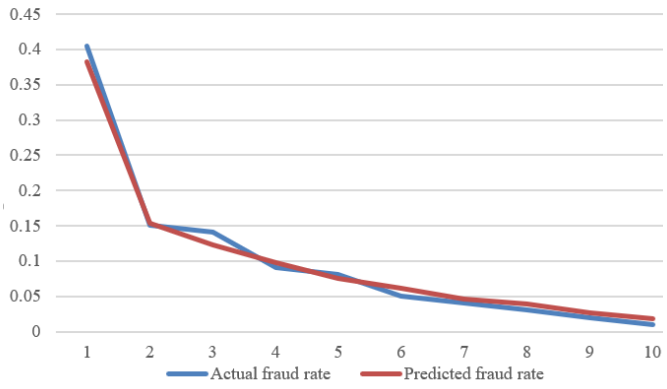

Figure 3. Prediction results on the groups in the test set

Next, the fraud rate predicted by our model was sorted and divided equally into five parts to get the mean of the expected fraud rate of each part. Moreover, the mean response variable of each part was calculated. For each part, the predicted fraud rate was compared with the actual fraud rate. It is not difficult to find that the predicted fraud rate is positively correlated with the actual fraud rate. The predicted and actual fraud rates after sorting are shown in Table 11. The trend of actual and predicted fraud rates after grouping is shown in Figure 4.

Table 11. Predicted and actual fraud rates after sorting

Sample number	Predicted fraud rate	Actual fraud rate
1	0.03952	0.02769
2	0.05018	0.04386
3	0.07116	0.06032
4	0.11296	0.12085
5	0.22983	0.25903

6. CONCLUSIONS

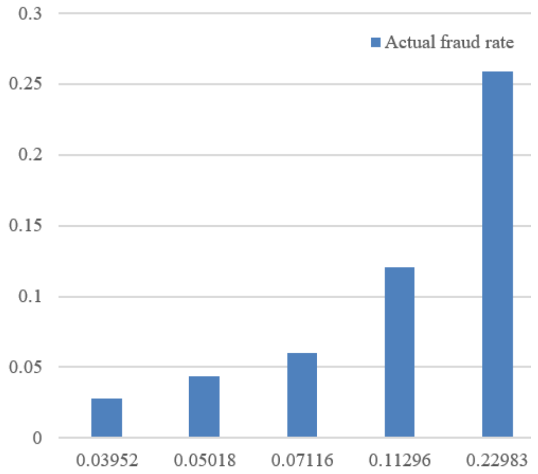

Figure 4. Trend of actual and predicted fraud rates after grouping

Aiming at the fraud in the Internet financial industry, this paper establishes a prediction model for user fraud rate based on big data analysis. Based on the collected data on user consumption, social network, and fraud, multiple variables were screened, and combined with advanced ML technologies like RF to establish a preliminary model. The effectiveness of the model was demonstrated through empirical analysis on actual datasets. However, the proposed model might not be able to deal with multiple independent variables, and ultra-high-dimensional samples. To overcome the limitation, the future research needs to further optimize the proposed model to select suitable indices and compute sparse matrix, aiming to make accurate predictions in the face of multiple independent variables, and ultra-high-dimensional samples.

REFERENCES

[1] Creti, A., Verdier, M. (2014). Fraud, investments and liability regimes in payment platforms. International Journal of Industrial Organization, 35: 84-93. https://doi.org/10.1016/j.ijindorg.2014.06.003

[2] Levi, M. (2008). Organized fraud and organizing frauds: Unpacking research on networks and organization. Criminology & Criminal Justice, 8(4): 389-419. https://doi.org/10.1177/1748895808096470

[3] Chiu, C., Ku, Y., Lie, T., Chen, Y. (2011). Internet auction fraud detection using social network analysis and classification tree approaches. International Journal of Electronic Commerce, 15(3): 123-147. https://doi.org/10.2753/JEC1086-4415150306

[4] Lam, S.S., Liu, H. (2006). Failure recovery for structured p2p networks: Protocol design and performance under churn. Computer Networks, 50(16): 3083-3104. https://doi.org/10.1016/j.comnet.2005.12.009

[5] Fan, H., Sun, X. (2010). A multi-state reliability evaluation model for P2P networks. Reliability engineering & System Safety, 95(4): 402-411. https://doi.org/10.1016/j.ress.2009.11.011

[6] Zhou, G., Zhang, Y., Luo, S. (2018). P2P network lending, loss given default and credit risks. Sustainability, 10(4): 1010. https://doi.org/10.3390/su10041010

[7] Tan, K.C., Wang, M.L., Peng, W. (2005). A P2P genetic algorithm environment for the internet. Communications of the ACM, 48(4): 113-116. https://doi.org/10.1145/1053291.1053297

[8] Huang, R.H. (2018). Online P2P lending and regulatory responses in China: Opportunities and challenges. European Business Organization Law Review, 19(1): 63-92. https://doi.org/10.1007/s40804-018-0100-z

[9] Ma, L., Wang, Y., Ren, C., Li, H., Li, Y. (2020). Early warning for internet finance industry risk: An empirical investigation of the p2p companies in the coastal regions of china. Journal of Coastal Research, 106(SI): 295-299. https://doi.org/10.2112/SI106-069.1

[10] Galindo, J., Tamayo, P. (2000). Credit risk assessment using statistical and machine learning methods as an ingredient for risk modeling of financial intermediaries. Computing in Economics & Finance, 15(1-2): 107-143.

[11] Gonzalez, L., Loureiro, Y.K. (2014). When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans. Journal of Behavioral and Experimental Finance, 2: 44-58. https://doi.org/10.1016/j.jbef.2014.04.002

[12] Lin, W.Y., Hu, Y.H., Tsai, C.F. (2011). Machine learning in financial crisis prediction: a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4): 421-436. 10.1109/TSMCC.2011.2170420

[13] Chirra, V.R.R., Uyyala, S.R., Kolli, V.K.K. (2019). Deep CNN: A machine learning approach for driver drowsiness detection based on eye state. Revue d'Intelligence Artificielle, 33(6): 461-466. https://doi.org/10.18280/ria.330609

[14] Pandey, T.N., Jagadev, A.K., Choudhury, D., Dehuri, S. (2013). Machine learning–based classifiers ensemble for credit risk assessment. International Journal of Electronic Finance, 7(3-4): 227-249. https://doi.org/10.1504/IJEF.2013.058604

[15] Pierdzioch, C., Risse, M. (2018). A machine-learning analysis of the rationality of aggregate stock market forecasts. International Journal of Finance & Economics, 23(4): 642-654. https://doi.org/10.1002/ijfe.1641

[16] Dai, H.L. (2015). Class imbalance learning via a fuzzy total margin based support vector machine. Applied Soft Computing, 31: 172-184. https://doi.org/10.1016/j.asoc.2015.02.025

[17] Sun, J., Fujita, H., Chen, P., Li, H. (2017). Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble. Knowledge-Based Systems, 120: 4-14. https://doi.org/10.1016/j.knosys.2016.12.019

[18] Zhao P. (2018). RETRACTED: Quantitative analysis of portfolio based on optimized BP neural network. Cognitive Systems Research, 52:709-714. https://doi.org/10.1016/j.cogsys.2018.08.024

[19] Xu, C., Li, P. (2019). On finite-time stability for fractional-order neural networks with proportional delays. Neural Processing Letters, 50(2): 1241-1256. https://doi.org/10.1007/s11063-018-9917-2

[20] Mihelčić, M., Džeroski, S., Lavrač, N., Šmuc, T. (2018). Redescription mining augmented with random forest of multi-target predictive clustering trees. Journal of Intelligent Information Systems, 50(1): 63-96. https://doi.org/10.1007/s10844-017-0448-5

[21] Lee, S.J., Xu, Z., Li, T., Yang, Y. (2018). A novel bagging C4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making. Journal of Biomedical Informatics, 78: 144-155. https://doi.org/10.1016/j.jbi.2017.11.005