Data-driven Risk Assessment for Peer-to-Peer Network Lending Agencies

Tianyuan Zhao, Lei Li, Yang Xie, Yue Lv

Beijing University of Posts and Telecommunications, Beijing 100876, China
zhaotianyuan13@163.com,leili@bupt.edu.cn,xieyangsp@163.com,lvyue@bupt.edu.cn

Abstract: With the rapid development of Peer-to-Peer(P2P) network lending in the financial field, more data of lending agencies have appeared. P2P agencies also have problems such as absconded with ill-gotten gains and out of business. Therefore, it is necessary to assess their risks based on P2P company data. This paper proposes a framework of Data-driven Risk Assessment for P2P(DRAP2P) network lending agencies based on unstructured natural language data. First, use the natural language processing technology, such as word segmentation, keyword, LDA topic model, word2vec and doc2vec, to process and extract features of company profile which reflect its business status. Then, seven machine learning classifiers and three deep learning models are used for analysis. Since keywords show good performance in machine learning models, we improve Convolutional Neural Network(CNN) with keywords and propose two CNN+Keyword models, namely CNN+Keyword(static+BP) and CNN+Keyword (Expand word embedding). Experiments have shown that CNN+Keyword(static+BP) can achieve the best performance. Finally, we use the method of meta-learning to integrate CNN+Keyword(static+BP) and logistic regression classifier to further strengthen the performance.

Keywords: Data-driven; P2P risk assessment; Machine learning; Deep learning; Meta-learning

1 Introduction

P2P network lending is a method of debt financing that enables individuals and companies to borrow and lend money without the use of official financial institution such as intermediary. It is a new type of financial business model which has developed rapidly worldwide in recent years due to its advantages of convenience, high interest rate, etc. However, P2P network lending agencies also have many problems, such as absconded with ill-gotten gains and difficult withdrawing. At present, the risk assessment of P2P lending agencies is still very scarce. In particular, P2P network lending has generated a lot of data, especially unstructured natural language text, which contain more plentiful information than structured one. It is of great significance to use these data to assess the lending risk effectively for strengthening market supervision, assisting in policy and decision-making, reducing the risk of lending and establishing a good environment of financial investment.

978-1-5386-6005-8/18/$31.00 ©2018 IEEE

In the current information age, data processing and analysis are crucial. Without machine learning(ML) dealing with a large number of chaotic data scientifically, much information will be useless. Deep learning(DL) is a new field in ML. It can automatically learn features from mass data, generalize rules that are generally difficult to identify, and apply the learned rules to similar data to deduce the expected results. As we can see that these characteristics just meet the application demand of P2P lending agencies' risk assessment.

In summary, this paper emphatically explores the risk assessment based on unstructured natural language text data. We combine ML, DL and natural language processing(NLP) technology and use company profile text data of the P2P corpus that can reflect the business status as the research object. The main contributions include three points:

(1) We consider the task of risk assessment as a kind of text classification and propose a data-driven risk assessment framework called DRAP2P for P2P network lending agencies.
(2) We introduce the keyword feature of company profile into CNN and propose two kinds of CNN+Keyword improved models, and their effectiveness is verified through experiments.
(3) Considering the advantages and disadvantages of different learning methods, we integrate DL and ML models with meta-learning to further strengthen the performance of P2P risk assessment.

2 Related Work

At present, ML and DL have been successfully applied in computer vision, NLP and other fields. For example, Support Vector Machine(SVM), Logistic Regression(LR) and Random Forest(RF) are used for text classification[1]. Also, CNN and its variants can be used in image recognition[2]and text classification[3,4].

At the same time, ML and DL have many applications in economics. For example, SVM is used to analyze financial time series[5]and DL is used to forecast stock price[6]. ML can also be used for financial news emotional analysis[7] and financial market recommendation system[8].

The existing P2P risk assessment is based primarily on economics and personal credit risk theory, using theoretical research and example analysis.

In summary, previous researches mainly focused on the analysis of text information as well as structured data, involving multiple technologies and algorithms in various fields of research, including ML, NLP, etc. However, at present, there are few researches specifically on risk assessment of P2P lending agencies and integration of different methods, especially combination of ML and DL. Although current methods show many advantages, many problems are also exposed, such as specific data dependencies, performance defects of supervised training mechanism, etc.

3 Proposed DRAP2P

Figure1 shows the framework of our proposed DRAP2P.

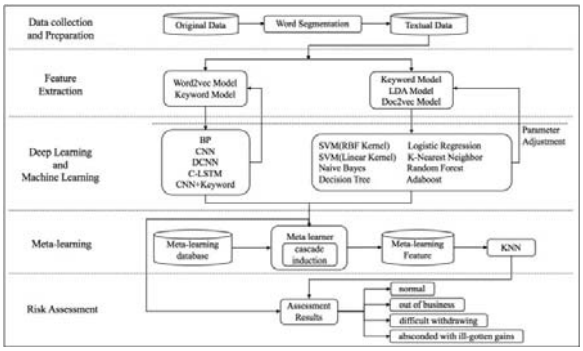

Figure1 Framework of DRAP2P

3.1 Data Collection and Preparation

We collected the raw data of P2P company profile, and preprocessed it with word segmentation, removal of punctuation and stop words. Then we partitioned the data set into two sets of 70% for training and 30% for test.

3.2 Feature Extraction

We mainly extracted the following text features.

Keyword: Keyword is well suit for text analysis since there are some important words with the same meaning in the texts of similar companies. We use information gain(IG) to select the keywords in company profile[9].

LDA: Latent Dirichlet Allocation(LDA) is a document topic generation model that is a popular way to identify the potential topic information[10] in a large document collection or corpus.

Word2vec and Doc2vec： Mikolov et al. proposed the efficient estimation of word representations in vector space in 2013[11], namely “Word2vec”. It can not only obtain the context of words, but also reduce the data size. Later, Mikolov et al. improved it and proposed a distributed representation of sentences and documents [12], namely “Doc2Vec”. Compared with Word2vec, Doc2vec can effectively express the semantic, grammatical and emotional information.

3.3 Machine Learning Model

We try the following classical ML models for risk classification and assessment.

(1) SVM is a kernel based ML algorithm. Its main idea is to transform the linear non-separable samples in low-dimensional input space into the high-dimensional feature space to make it linearly separable, and then construct the optimal hyperplane in the feature space, so that we can find the global optimal classifier[13]. We have chosen two commonly used kernel functions, namely RBF Kernel and Linear Kernel.

(2) The LR classifier usually uses the known independent variables to predict the value of a discrete dependent variable.

(3) The main idea of Naive Bayes(NB) classifier is to use Bayes' theorem to solve the posterior probability by combining probability model, and take the category with the largest posterior probability as the prediction category.

(4) The main idea of K-Nearest Neighbor(KNN) is to select k neighbors closest to input data point in the training set, and take the category of the most frequent occurrence of the k neighbors as result.

(5) The main idea of the decision tree(DT) based on information entropy measures is to construct a tree with the fastest decrease in entropy, and the entropy value at the leaf node is zero.

(6) The RF algorithm is mainly used to study the different new data sets by randomly sampling from the training data sets, and then average or vote on these predicted results.

(7) Adaboost can adjust the weight distribution of the samples adaptively, and raise the weight of wrong samples and lower the weight of right ones.

3.4 Deep Learning Model

In DL, the CNN model has a strong feature extraction capability, and we expect it can excavate deep features from the data of P2P company profiles that are not large enough. The commonly used input of CNN is the word embedding matrix of text[3]. According to whether adjusting the weight of word embedding, it is divided into rand, static and non-static ones, it can also be multi-channel. The input of DCNN (Dynamic CNN) is the word embedding matrix[14], the maximum pooling layer of dynamic k is proposed in this model, and the feature quantity extracted by the pooling layer is related to the length of input text. The C-LSTM (Long Short-Term Memory) model[15] combines CNN's ability to extract features with LSTM's memory characteristics.

During the experiment, we found that the keyword feature performed better in ML models, so we improved CNN with keywords, and proposed two kinds of CNN+Keyword models.

The first model, CNN+Keyword (static+BP), combines the feature extracted by CNN with keywords as Figure 2. Suppose the length of the sample sentence is n[3], let

$x_i \in R^k$ be the k-dimension word embedding of the i-th word in the sentence, and a sentence of length n is represented as

$$x_{1:n} = x_1 \oplus x_2 \oplus ... \oplus x_n, \quad (1)$$

where $\oplus$ is the concatenation operator. In general, let $x_{i:j} = x_i \oplus x_{i+1} \oplus ... \oplus x_j$. A convolution operation involves a filter $w \in R^{hk}$, where h is the size of window. A feature c_i is generated from a window of words $x_{i:i+h-1}$ by

$$c_i = f(w \cdot x_{i:i+h-1} + b), \quad (2)$$

where b is a bias and f is non-linear function such as relu or tanh. This filter is applied to each possible window of words in the sentence$\{x_{1:h}, x_{2:h+1},..., x_{n-h+1:n}\}$to produce a feature map $c = [c_1, c_2,..., c_{n-h+1}]$, with $c \in R^{n-h+1}$. Then, we apply a max-pooling operation for feature map, $\hat{c} = \max\{c\}$. If the number of filters with different window size are $m_1, m_2,...,m_M$, the final pooling of results are $c_{cnn} = [\hat{c}_1,\hat{c}_2,...,\hat{c}_{\sum_{i=1}^{M} m_i}]$.

For each sample, its keyword feature is $c_{keyword} = [k_1,k_2,...,k_K]$, where K is the number of keywords. $z = [c_{cnn}, c_{keyword}]$ contains the feature of CNN extraction and keyword. And the hidden layer is optional. If hidden layer is not added between the concatenation layer and output layer, the concatenation layer is connected to the softmax layer directly. If the hidden layer is added, the concatenation layer can be considered as the input layer of a three-layer BP neural network.

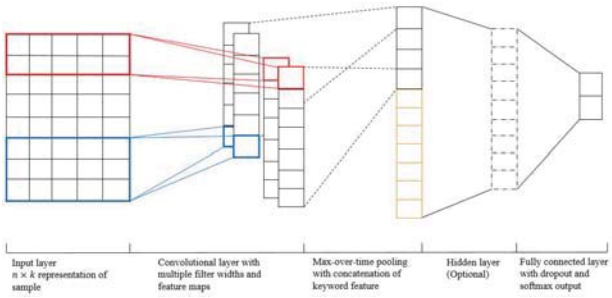

Figure2 CNN+Keyword (static+BP)

The second model, CNN+Keyword (Expand word embedding) expands the dimension of word embedding, as shown in Figure 3.

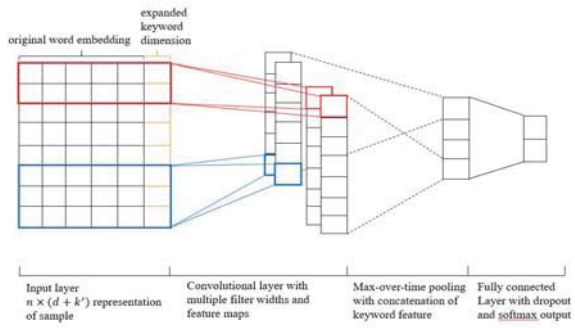

Figure 3 CNN+Keyword (Expand word embedding)

Suppose the original word embedding is represented by $[a_1, a_2, ..., a_d]$.Then we choose K' keywords according to IG, and the expanded word embedding is represented by $[a_1, a_2, ..., a_d, b_1, b_2, ..., b_{K'}]$, where K' is the number of expanded dimensions of keyword. If the word is a keyword, $b_1 = b_2 = ... = b_{K'} = 1$, otherwise $b_1 = b_2 = ... = b_{K'} = 0$.

3.5 Meta-learning

The meta-learning is based on D-S evidence theory and voting mechanism [16]. Through the method of cascade induction, the original best CNN and ML classifier are used as base learners and combined with meta-feature extracted from their results, then KNN is used for the final classification.

3.6 Risk Assessment

Risk assessment is carried out to judge the specific risk category based on the result of meta-learning module.

4 Experiments and Analysis

4.1 Data sources

We have collected data about 4,554 P2P companies from a third-party platform of network lending, the Home of Network Loan(https://www.wdzj.com). There are four categories graded from 0 to 3, and different categories represent different business status. The significance of all kinds of data is shown in Table 1.

Table 1 The status and numbers of the four categories

Grade	Business status	Number
0	normal	1849
1	out of business	1263
2	difficult withdrawing	595
3	absconded with ill-gotten gains	847

The company profile mainly includes business content, scope of business, operation philosophy, social responsibility, establishment background, company shareholders and management, also the legal information, honorary title, cooperation object, as well as registered capital, financing amount and detailed information of excellent products. That information can fully describe a company, and the description information of different types of companies varies greatly, which plays a significant role in the subsequent risk category assessment.

4.2 Data preprocess and feature extraction

First, we use the jieba module in Python for Chinese word segmentation, and then use regular expression and stop word list to delete punctuations and stop words.

In this paper we calculate the IG for each word and rank the words from large to small according to their IG values. After extracting top K keywords, we set up a K-dimensional vector for each company. In our experiment, we have tried the value of K according to experience from 5 to 200 while step length is 5. We also have tried different LDA topic numbers for each company's text according to experience from 10 to 100

while step length is 5. As to Word2vec, each sample is represented by a word embedding matrix, where the number of rows is the number of words. The number of words in all samples can be controlled to be similar by truncation and padding. The columns of the matrix represent the dimensions of the word embedding. In Doc2vec, according to experience, each sample is represented as a 200-dimensional paragraph vector.

4.3 The optimization method and parameters of the deep learning models

With BP neural network as the baseline of deep learning, we implemented six models, including two kinds of improved models of CNN+Keyword we proposed. Table 2 shows the parameters for all models.

Table 2 The parameters of the DL models and BP neural network

Model	seq_length	embedding_dim	filter_sizes	num_filters	extra
BP	600	150			hidden_units:300
CNN	600	150	(3, 4, 5)	150	
DCNN	600	150	(7, 5)	(6，14)	
C-LSTM	600	150	3	150	
CNN+Keyword (static+BP)	600	150	(3, 4, 5)	150	keyword_dim：50
CNN+Keyword (Expand word embedding)	600	150	(3, 4, 5)	150	keyword_dim：50

4.4 Experiments of ML model

Table 3 The precisions of the ML models

Model	Keyword(%)	LDA(%)	Doc2vec(%)
SVM(RBF Kernel)	75.58	32.58	16.5
SVM(Linear Kernel)	76.63	32.92	20.08
NB	66.54	37.95	20.07
DT	67.12	36.76	33.08
LR	76.81	33.91	21.18
KNN	68.38	37.1	25.76
RF	73.97	38.6	38.57
Adaboost	73.25	40.02	36.07

The precision results of ML models are shown in Table 3, where the value is the maximum precision within the corresponding parameter range, and the precision is the average of four categories. We can see that the precision of Keyword feature is significantly better than that of LDA and Doc2vec, and the precision of LR model can reach 76.81%.

4.5 Experiments of DL model

The results of BP, CNN, DCNN and C-LSTM are shown in Table 4. Compared with Table 3, the results of CNN are generally higher than that of ML models, but the results of DCNN and C-LSTM are lower than ML models using keywords, which is higher than LDA and Doc2vec. For CNN, although the effect of rand model is better, it is very easy to cause overfitting in experiments. After reaching a higher training precision, the precision of the test set drops rapidly. The precision of static model is lower than that of rand model, but it is more stable and not easy for overfitting. For DCNN, the two-layer deep network needs more data to train, and for deep learning, the data of P2P company profiles are still less, causing the precision of DCNN generally lower than that of CNN. The training process of C-LSTM and DCNN models is slow, but the effect of the DL models is better than that of BP, which verifies that the DL models are superior for feature extraction.

Table 4 The results of BP、CNN、DCNN、C-LSTM

Model	Precision(%)	Recall(%)	F1-score(%)
BP	68.52	69.79	68.37
CNN-rand	79.11	79.62	79.14
CNN-static	77.59	78.15	77.61
CNN-non-static	77.32	77.71	77.04
CNN-multi channel	77.56	78.01	77.07
DCNN	70.02	70.23	68.98
C-LSTM	74.72	75.07	74.8

Some results of CNN+Keyword(static+BP) are shown in Table 5. The range of Hidden_units is from 100 to 150, and the step length is 10. It can be seen that CNN with keywords is better than CNN-static model, and some F1-score have increased to more than 0.78.

Table 5 The results of the CNN+Keyword(static+BP)

Hidden_units	Precision(%)	Recall(%)	F1-score(%)
100	77.51	77.57	77.19
110	78.48	78.89	78.43
120	77.02	77.57	76.99
130	77.73	78.37	77.82
140	77.56	77.93	77.16
150	77.79	78.01	77.21

Some results of CNN+Keyword(Expand word embedding) are shown in Table 6. The range of Expand_dim is from 1 to 5 and the step length is 1. It can be found that the effect of most models is worse than that of CNN-static. The reason may be that the expanded word embedding affects the distribution of the original word embedding in space, causing the loss of semantic information.

Table 6 The results of the CNN+Keyword(Expand word embedding)

Expand_dim	Precision(%)	Recall(%)	F1-score(%)
1	76.44	77.05	76.48
2	76.03	76.69	76.08
3	78.29	78.74	77.94
4	77.19	77.49	76.63
5	76.94	77.05	76.6

The best results of two kinds of CNN+Keyword models are shown in Table 7.

Table 7 The best results of the CNN+Keyword

Model	Precision(%)	Recall(%)	F1-score(%)
CNN-static	77.59	78.15	77.61
CNN+Keyword (static+BP)	78.48	78.89	78.43
CNN+Keyword (Expand word embedding)	78.29	78.74	77.94

As we can see that both of the two improved CNN+Keyword models with keywords can improve the effect of classification based on the original CNN-static model, and the method of CNN+Keyword(static+BP) performs better and more stable.

4.6 Meta-learning

The results of meta-learning are shown in Table 8. We use the prediction results of CNN+Keyword(static+BP) and LR classifier on the test set as input to another KNN classifier. The results show that compared with the original CNN+Keyword(static+BP) and LR classifier, the meta-learning has increase in Precision, Recall and F1-score. Thus it can be seen that the meta-learning method of integrating ML and DL is effective.

Table 8 The result of the Meta-learning

Model	Precision(%)	Recall(%)	F1-score(%)
LR	76.81	77.27	76.50
CNN+Keyword (static+BP)	77.43	77.93	77.37
Meta-learning	78.09	78.45	77.91

To sum up, all the results have shown that both ML and DL play an important role in the data-driven risk assessment of P2P network lending agencies. At the same time, the method of meta-learning can effectively integrate advantages of ML and DL models.

5 Conclusions

In this paper, we propose a data-driven risk assessment framework for P2P network lending agencies, which can effectively evaluate the business status of P2P companies, thus reducing the risk of P2P lending. Results show that both ML and DL can assess company business risks, CNN can be improved with keyword, and the meta-learning integrating ML and DL can also strengthen the effect. We will further study the risk assessment of P2P lending agencies in the future, especially to analyze more unstructured data, extract better semantic features and optimize the classification model.

Acknowledgements

This work was supported by National Social Science Foundation of China [grant number 16ZDA055]; National Natural Science Foundation of China [grant numbers 91546121, 71231002]; EU FP7 IRSES MobileCloud Project [grant number 612212]; the 111 Project of China [grant number B08004]; Engineering Research Center of Information Networks, Ministry of Education; Beijing BUPT Information Networks Industry Institute Company Limited; the project of Beijing Institute of Science and Technology Information; the project of CapInfo Company Limited.

References

[1] Chen P H, Zafar H, Galperinaizenberg M, et al. Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports.[J]. Journal of Digital Imaging, 2017, 31(1):1-7.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,pages 1–9,2015.

[3] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.

[4] Wang R, Donghong J I, School C, et al. Twitter Sentiment Classification Method Based on Convolutional Neutral Network and Multi-feature Fusion[J]. Computer Engineering, 2018.

[5] Das S P, Padhy S. A novel hybrid model using teaching–learning-based optimization and a support vector machine for commodity futures index forecasting[J]. International Journal of Machine Learning & Cybernetics, 2018, 9(1):97-111.

[6] Minami S. Predicting Equity Price with Corporate Action Events Using LSTM-RNN[J]. Journal of Mathematical Finance, 2018, 08(1):58-63.

[7] Shuhidan S M, Hamidi S R, Kazemian S, et al. Sentiment Analysis for Financial News Headlines using Machine Learning Algorithm[M]// Proceedings of the 7th International Conference on Kansei Engineering and Emotion Research 2018.

[8] Hendricks D, Roberts S J. Optimal Client Recommendation for Market Makers in Illiquid Financial Products[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2017:166-178.

[9] W Dong, X Liu, N I. Hong, "Adaptive feature selection method based on information gain[J]", Computer Engineering & Design, 2014.

[10] H E Jin-Qun, P J Liu, "The documents classification algorithm based on LDA[J]", Journal of Tianjin University of Technology, 2014.

[11] T Mikolov, K Chen, G Corrado et al., "Efficient Estimation of Word Representations in Vector Space[J]", Computer Science, 2013.

[12] Q V Le, T. Mikolov, "Distributed Representations of Sentences and Documents[J]", Eprint Arxiv, vol. 4, pp. 1188-1196, 2014.

[13] B E Boser, "A training algorithm for optimal margin classifiers[C]", PROCEEDINGS OF THE 5TH ANNUAL ACM WORKSHOP ON COMPUTATIONAL LEARNING THEORY, pp. 144-152, 1992.

[14] Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences[J]. Eprint Arxiv, 2014, 1.

[15] Zhou C, Sun C, Liu Z, et al. A C-LSTM Neural Network for Text Classification[J]. Computer Science, 2015, 1(4):39-44.

[16] Mandler E, Schuermann J. Combining the Classification Results of Independent Classifiers Based on the Dempster/Shafer Theory of Evidence[C]// Pattern Recognition and Artificial Intelligence. Towards an Integration. 1988:381-393.