Clustering similar lenders in P2P lending

Haifeng Li[1, a]

[1]School of Information, Central University of Finance and Economics, Beijing, China

[a]mydlhf@139.com

Keywords: credit score, cluster, P2P.

Abstract. Online peer-to-peer (P2P) lending has been focused on since it is useful for small companies that are conducted on the website. In this study, we get the lenders and try to cluster them into groups, which can help the P2P company to trace and discovery the VIP customers.

1. Introduction

Online peer-to-peer (P2P) lending has been studied a lot since it is a useful finance method for small enterprises, which can get the emergency loan for turnover of capital, which is especially very important in China. The Online P2P can be implemented by the information technologies. And it allows users to lend and borrow funds directly through an online intermediary without the mediation of financial institutes.

When a lender wants to acquire capitals from the online P2P companies, they will be different for these companies. That is, some lenders maybe have the potential good crediting-rate and become the VIP users, others maybe cannot refund in time and become the unwelcome users. How to discovery them based on their information is an interesting problem. Many studies have focused on this problem and proposed some useful method.

[1] represented an extension of the expansive credit risk and credit migration literature, prominent in the corporate bond and securities risk pricing literature, to an analysis of the drift of consumer credit scores. A rich data set of residential mortgages was used to observe credit score migration post loan origination and in a test of the ability of credit score transition to serve as a precursor to potential default and prepayment. The results indicated credit scores provide signals and information to investors and servicing agents in a fashion similar to credit ratings on commercial paper as to default potential. Soner[2] presented a proposes a three stage hybrid Adaptive Neuro Fuzzy Inference System credit scoring model, which was based on statistical techniques and Neuro Fuzzy. The performance of the proposed model was compared with conventional and commonly utilized models. The credit scoring models were tested using a 10-fold cross-validation process with the credit card data of an international bank operating in Turkey. Results demonstrated that the proposed model consistently performed better than the Linear Discriminant Analysis, Logistic Regression Analysis, and Artificial Neural Network (ANN) approaches, in terms of average correct classification rate and estimated misclassification cost. [3] addressed the question of what determines a poor credit score. The authors compared estimated credit scores with measures of impulsivity, time preference, risk attitude, and trustworthiness, in an effort to determine the preferences that underlie credit behavior. Data was collected using an incentivized decision-making lab experiment, together with financial and psychological surveys. Credit scores were estimated using an online FICO creditscore estimator based on survey data supplied by the participants. Preferences were assessed using a survey measure of impulsivity, with experimental measures of time and risk prefer-ences, as well as trustworthiness. Controlling for income differences, the authors found that the credit score was correlated with measures of impulsivity, time preference, and trustworthiness. Based on trust theories, Chen et. al[4] the present study develops an integrated trust model specifically for the online P2P lending context, to better understand the critical factors that drive lenders' trust. The model is empirically tested using surveyed data from 785 online lenders of PaiPaiDai, the first and largest online P2P platform in China. The results show that both trust in borrowers and trust in intermediaries are significant factors influencing lenders' lending intention. Emerkter et. al[5] used data from the Lending Club, which is

one of the popular online P2P lending houses, to explore the P2P loan characteristics, evaluate the credit risk and measures loan performances. They found that credit grade, debt-to-income ratio, FICO score and revolving line utilization played an important role in loan defaults. Loans with lower credit grade and longer duration were associated with high mortality rate. The result was consistent with the Cox Proportional Hazard test. Also, they found that higher interest rates charged on the high risk borrowers were not enough to compensate for higher probability of the loan default; thus, the Lending Club must find ways to attract high FICO score and high-income borrowers in order to sustain their businesses. Harris[6] investigated the practice of credit scoring and introduced the use of the clustered support vector machine (CSVM) for credit scorecard development. This algorithm was well known that as historical credit scoring datasets get large while highly accurate becomed computationally expensive. Accordingly, he compared the CSVM with other nonlinear SVM based techniques and shows that the CSVM can achieve comparable levels of classification performance while remaining relatively cheap computationally.

In this paper, we also addressed this problem and proposed a clustering method by the online documents of the lenders. The rest paper is organized as follows: Section 2 presents the data related the lenders. Section 3 introduces our clustering method. Section 4 concludes this paper.

2. Data Preparation

In this paper, we used the data crawled from the website, which is the records from a BBS that provides the users to discuss the issues related to P2P lending. We preprocess the data and get the dataset with 18 properties. We describe it in Table 1. In this dataset, the title and the descriptions are string information; thus, it will be directly removed. In addition, we transform the continously changed property values, such as age, to the discrete values with an aequilate method. Also, we convert the credit rate and other string type properties to integer properties.

Table 1. The characteristics of the dataset

Properties	Record Count
Title, Amount, Annual interest rate, Repayment Time, Descriptions, Credit rate, Successful loan number, Failed loan number, Gender, Age, Borrowed credit score, Lending credit score, Overdue, Membership score, Prestige, Forum currency, Contribution, Group	20000

3. Clustering Method

In this section, we use a clustering method to build a group model to find the similar action lenders. We used the basic Euclidean distance to compute the similarity of two objects, and employed the K-Means algorithm to cluster the data, where k=3. We show the k-centers in Table 2. As can be seen, the second cluster has almost 2/3 objects.

Table 2. K-centers of the 3 clusters

PPROPERTY	1	2	3
amount	-0.02488	-0.09918	0.402455
annual interest rate	-0.0572	-0.4409	1.782759
payback time	-0.1694	0.070937	-0.26593
credit rating	-0.15999	0.46722	-1.86338
successful loan number	1.158259	-0.07857	0.180916
failed loan number	0.384325	-0.09233	0.326957

sex	-0.05339	0.006031	-0.01804
age	0.009662	0.03177	-0.12911
borrowed credit score	0.83178	-0.17946	0.62555
lending credit score	1.556155	-0.03667	-0.03443
membership score	4.955253	-0.13591	-0.03257
prestige	4.669313	-0.1125	-0.09339
forum currency	3.792114	-0.10883	-0.0055
contribution	4.74668	-0.11713	-0.0838
group	0.636604	-0.0386	0.080961
typecount	455	15658	3887

To analysis the feature of these users, we show the density of the second groups in Figure 1. As can be seen, the amount, the annual interest rate, the credit rating, successful loan number, sex, lending credit score, membership score, and group are the properties that performed significantly dense. This means these lenders can be regarded as the representations of all the lenders.

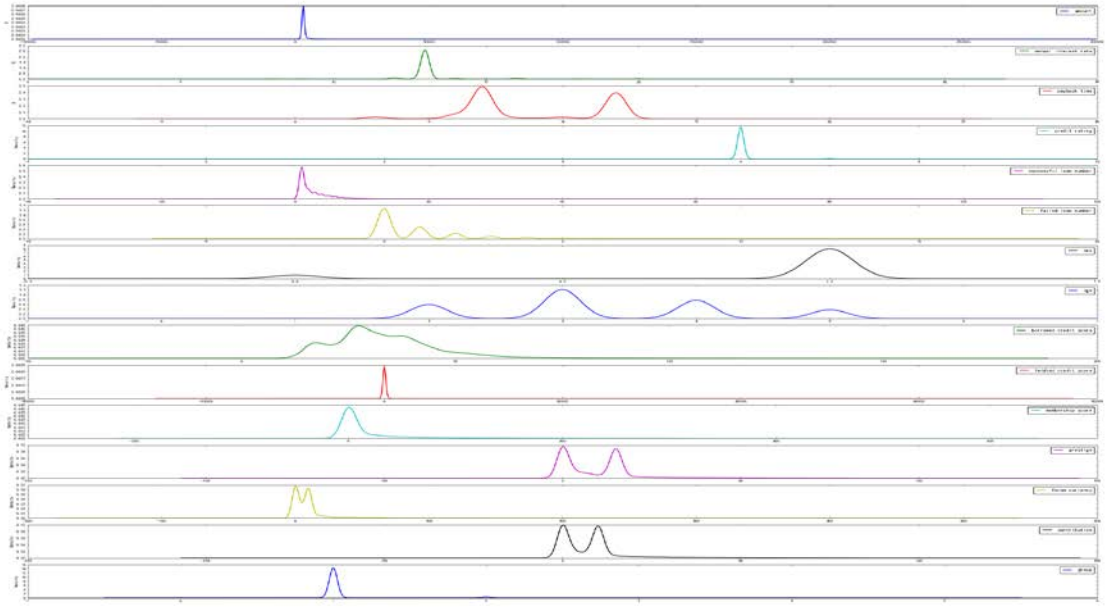

Fig.1 The density of cluster 2

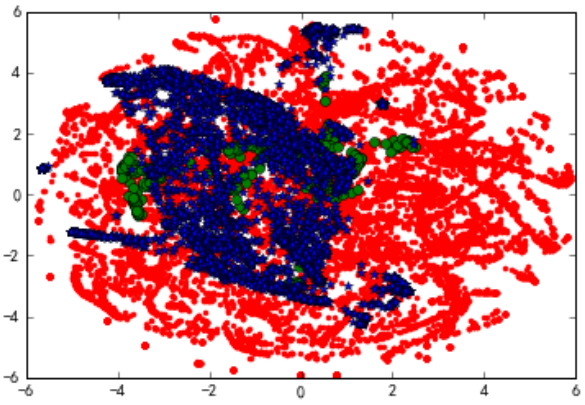

Fig. 2 The presentation of the clusters

We also use TSNE to reduce our analysis results to 2-demensions, which is shown in Figure 2. As shown, the red one is the second cluster, and it almost covers all the features.

4. Conclusions

In this paper, we used the social network data to find the similar lenders from peer-to-peer lending platforms. A clustering method K-Means was employed and we found a very large group from all the lenders. In addition, we used the TSNE method to visualize the clustering results, which showed our conclusions were reasonable.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (61100112, 61309030), Beijing Higher Education Young Elite Teacher Project (YETP0987). Key project of National Social Science Foundation of China(13AXW010), 121 of CUFE Talent project Young doctor Development Fund in 2014 (QBJ1427).

References

[1] B.C.Smith. Stability in consumer credit scores: Level and direction of FICO score drift as a precursor to mortgage default and prepayment. Journal of Housing Economics, 2011.

[2] A. Soner. An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. European Journal of Operational Research, 2012.

[3] S.Arya, C.Eckel, C.Wichman. Anatomy of the credit score. Journal of Economic Behavior & Ornanization, 2013.

[4] D.Chen, F.Lai, Z.Lin. A trust model for online peer-to-peer lending: a lender's perspective. Information Technology Management, 2014.

[5] R.Emerkter, Y.Tu, B.Jirasakuldech, M.Lu. Evaluating credit risk and loan performance in online Peer-to-Peer(P2P) lending. Applied Economics, 2014.

[6] T.Harris. Credit scoring using the clustered support vector machine. Expert Systems with Applications, 2015.