Proceedings of 2012 4th International Conference on Machine Learning and Computing
IPCSIT vol. 25 (2012) © (2012) IACSIT Press, Singapore

Bayesian Network Based Causal Relationship Identification and Funding Success Prediction in P2P Lending

Xue Rui [1 +], Bingwu Liu [2] and Shaohua Tan [1]

[1] Department of Intelligence Science, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China

[2] School of Information, Beijing Wuzi University, Beijing 101149, China

Abstract. Peer-to-peer lending or P2P lending connects the people who want to borrow and the people who want to invest. To identify the determinant factors of funding success and to predict whether a listing will get funded or not are two key issues in P2P lending. In this study, Bayesian network model based on a new learning algorithm HEK2 (Hierarchy Exact K2) is proposed to solve these two key issues. With the DAG (directed acyclic graph) structure learned in our model, the causal relationships of the entire factor set can be revealed in a visible manner. Consequently, the determinants of funding success and several hidden patterns rarely discussed before are extracted directly. Comparison with earlier work shows that the prediction accuracy of our method is 7.5% higher than SVM and 13.5% higher than KNN, which are both popular classifiers. Empirical results show the effectiveness and flexibility of our model.

Keywords: P2P lending, causal relationship, funding success, Bayesian network, HEK2

1. Introduction

Peer-to-peer (P2P) lending, an emerging alternative to traditional institutional lending, is based on an online reverse auction. In P2P lending, people can either request loans by creating listings, taking the Borrowers role, or buy loans by making bids, taking the Lenders role [1][2].Compared with traditional financial services middlemen, P2P lending has several advantages [3]. For example, the returns are said to be higher (10.69%) and the borrow interest rate (rate starting at 6.59% for AA loans) to be lower [4].

In the study of P2P lending, to identify the determinant factors of funding success and to predict whether a listing will get funded or not are two key issues, which are valuable in providing decision support for borrowers. There are more and more studies concentrating on solving these two issues. For example, pair-wise correlation test is used to identify the determinants of funding success and then the regression model is used to predict the funding success [2][5]. However, there is a risk of multicollinearity in the regression model. As an example, factor *StartingRate* and factor *Amount* are both included in the regression model in [2], but the correlation between them is 0.55, which is statistically significant. To avoid multicollinearity, popular classification techniques, such as SVM, KNN and so on, are used in [1] to predict the funding success. However, it provides no explanation about the relationships among factors.

In this study, Bayesian network model is used to solve the two key issues mentioned above, which is believed to have several key novelties compared with earlier work. First, Bayesian network model can avoid multicollinearity as well as SVM and KNN. Second, with the DAG structure learned in our model, the causal relationships of the entire factor set can be revealed in a visible manner. However, correlation matrix in [5] only shows whether two factors are correlated or not and SVM and KNN in [1] provide no information about relationships among factors. Causal relationships discovered in our model directly reveal the hidden patterns

+ Corresponding author. Tel.: +86 10 62755745.
E-mail address: bjxuerui@gmail.com

buried in the data and identify the factors which actually drive the variation of funding success probabilities. Besides, we should not neglect that there is a skewing problem inside the meta-data. To solve this problem, a data filtering method is proposed in [1], but the samples of testing set are not randomly selected, which makes the method not useful in a practical environment. From a practical point of view, we use the weight adjustment technique as a solution.

The rest of the paper is organized as follows. Section 2 introduces how casual relationships among factors are modeled. Section 3 describes the processing of the meta-data. In Section 4, we illustrate and analyze the experimental results. Conclusions and discussions are in Section 5.

2. Build Bayesian Network Model

A Bayesian network model is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a DAG (directed acyclic graph). Several algorithms, such as K2, HillClimbing, SimulatedAnnealing and so on, can be used to build the Bayesian network model. However, these algorithms only return approximate search results [6]. In this study, we propose a HEK2 (Hierarchy Exact K2) algorithm which returns exact search result finding the best matched structure. The HEK2 algorithm mainly consists of two steps: First, decide the level division of the factors collected from P2P lending marketplace. Second, use the score-search approach to find the best matched structure. Here we use Bayesian Dirichlet as our scoring criterion [6]:

$$P(B_s, D) = P(B_s)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N_{ij}^{'})}{\Gamma(N_{ij}^{'}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N_{ijk}^{'}+N_{ijk})}{\Gamma(N_{ijk}^{'})}$$

In our methodology, we take a hierarchical view based on Assumption 1: If the value of a factor v_i is determined before another factor v_j, then v_i can't be a descendant of v_j.

Under this assumption, we can divide the factors into three layers. Details about different layers can be seen in section 3.

HEK2 algorithm can be seen as an extension to the original K2 algorithm. In the original K2 algorithm, the order of factors is given as an input. However, the result relies heavily on the given order. It only returns approximate search result [8]. In HEK2 algorithm, every possible order of the factors in a same layer is considered. The parent set of a factor consists of the factors before it under a fixed order and the factors from the previous layer. Then our method searches through the space of all possible DAGs and the structure with highest score is returned. The pseudo code of HEK2 is in Algorithm 1. Dynamic programming can be used to accelerate. Assume that there are k_1 factors in layer 1, then in this study the time complexity is $[k_1 \times 2^{k_1-1} + (11-k_1)\times 2^{10} + 2^{11}]\times O(n)$.

As for the inference part, there have been well-developed algorithms for Bayesian network model [9].

Algorithm. 1

```
Input: FactorsSet, PreviousLayerFactorsSet
Output: BestStructure, BestScore
Algorithm:
List all the orders over the FactorsSet;
For each Order:
    For each Node in FactorsSet:
        List all the possible parent sets of the Node;
        For each ParentSet:
            Calculate the score of the Node and its ParentSet;
    Find the ParentSet with highest Score;
    Add the Node and its ParentSet to temp Structure;
    Add the Score to temp Score;
Find the BestStructure and associated BestScore;
```

3. Data Processing

Prosper.com is the world's largest peer-to-peer lending marketplace, with more than 1,170,000 members and $272,000,000 in funded loans. Cross-sectional annual data during 5 years from 2006 to 2010 are collected from Prosper.com in this study [4].

After removing irrelevant factors, there remain 12 factors including the class factor. Under Assumption 1, these factors are divided into three layers. Some factors need to be transformed. The status of *GroupKey* is entered as 'True' if the member has a group, otherwise as 'False'. The same transformation is done to *Description* and *Images*. As for the class factor *Status*, status 'completed' is entered as 'True', 'expired', 'withdrawn' and 'canceled' as 'False'. The other values are omitted. Instances with missing values are removed directly. Equal frequency discretization method is adopted to discrete the continuous variables. Details about factors can be seen in Table 1.

Table. 1: Factors.

Hierarchy	*Factor*	*Value Type*
First Layer	*DebtToIncomeRatio*	Nominal
	CreditGrade (ProsperRating)	Nominal
	GroupKey	Binary
	VerifiedBankAccount	Binary
	IsBorrowerHomeOwner	Binary
Second Layer	*AmountRequested*	Nominal
	BorrowerMaximumRate	Nominal
	Description	Binary
	Duration	Nominal
	FundingOption	Nominal
	Images	Binary
Third Layer	*Status*	Binary

4. Experimental Analysis

The HEK2 algorithm introduced in section 2 is applied to each of the 5 annual datasets. For clarity, we only show the learned structure of year 2006 as a representative (see Fig. 1). As can be seen from the graph, *CreditGrade* and *BorrowerMaxRate* are both determinants of the class factor *Status*. *GroupKey*, *AmountRequested* and *DebtToIncomeRatio* are ancestors of *Status*, which means that they have indirect influences. *DebtToIncomeRatio* has no significant influence as it's too far from the class factor *Status* in the graph. *Description* and *IsBorrowerHomeOwner* have no effect on *Status*. All these results are in line with earlier work [2][5].

Images also has a direct influence on *Status*. *VerifiedBankAccount* doesn't have relationship strong enough with any other factor. These interesting findings are barely shown before.

A high correlation between *IsBorrowerHomeOwner* and *Status* is expected in both [2] and [5], but in fact the correlation between them is relatively low, which is hard to explain. However, it can be seen clearly under our learned structure that they are both resulting factors of *CreditGrade*. There is no direct relationship between them.

If an edge with the same direction appears at least three times out of the 5 cross-sectional datasets, we confirm it as a credible relationship (see Fig. 2). To summarize the 5 cross-sectional datasets, *VerifiedBankAccount* and *Description* has no relationship strong enough with any other factor. *CreditGrade(ProsperRating)*, *AmountRequested* and *BorrowerMaxRate* are determinant factors of the class factor *Status*. *GroupKey* is an important factor influencing other listing options. *CreditGrade (ProsperRating)* has the most widely effect on other factors.

Soft margin SVM with different kernels and KNN are applied to the annual dataset of year 2007 to predict the funding success in [1]. The result shows that SVM with Radial Basis Kernel has the highest accuracy 85%. The prediction accuracy of KNN is 79%. The prediction accuracy of our model is 7.5% higher than SVM, and 13.5% higher than KNN. The prediction performance of our model can be seen in Table 2.

However, the prediction sensitivity, which indicates the proportion we truly recognized of the successful listings, is too low to accept. This is because the data skews towards the failure listings heavily. For example, only 9% of all the listings in 2006 got funded. The weight adjustment technique is used to solve this problem. We enhance the relative weight of successful listings to promote the sensitivity. Since there is a tradeoff between the sensitivity and accuracy (see Table 3), the relative weight can be decided according to the relative importance of different classes. In the case of 2006, 4.6 may be a proper value for the weight. The sensitivity rises up to 67.60% while the accuracy and specificity stay on 86.72% and 88.49%.

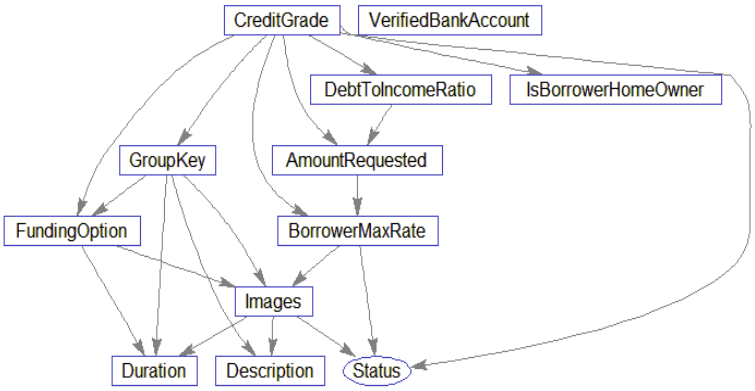

Fig. 1: Bayesian network structure for year 2006. A directed edge in the graph represents the causal relationship between two factors. *CreditGrade*, *BorrowerMaxRate* and *Images* are believed to have direct influences on *Status*.

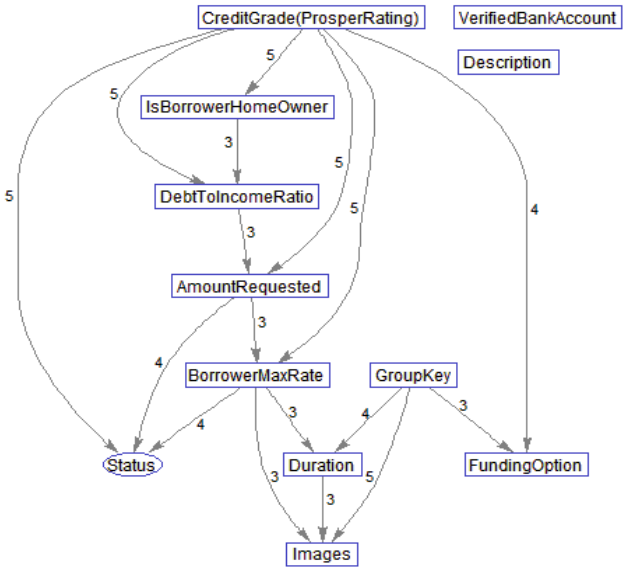

Fig. 2: General model for 5 cross-sectional annual datasets. A directed edge represents the causal relationship between two factors. The number besides the edge represents the times this relationship appears in 5 annual datasets. A relationship with 3 appearances or above is confirmed to be credible. *CreditGrade(ProsperRating), AmountRequested* and *BorrowerMaxRate* are three stable factors influencing *Status* across 5 years.

Table. 2: Prediction accuracy for cross-sectional annual dataset

Year	*#Training Instances*	*#Testing Instances*	*Accuracy*
2006	43,322	21,837	91.69%
2007	95,210	47,611	92.50%
2008	66,272	33,103	89.88%

2009	8,304	3,996	84.38%
2010	14,714	7,600	78.33%

Table. 3: Prediction performance with different weight

Weight / *Prediction*	*1.0*	*1.9*	*2.8*	*3.7*	*4.6*	*5.4*
Accuracy(%)	91.69	90.88	89.33	88.66	86.72	86.72
Sensitivity(%)	10.85	42.06	55.62	60.64	67.60	67.60
Specificity(%)	99.18	95.40	92.45	91.26	88.49	88.49

5. Conclusion and Discussion

In this study, we propose a HEK2 algorithm to build the Bayesian network model on empirical data collected from P2P lending marketplace. The method is effective in discovering the complicated causal relationships among various factors. With the DAG structure learned in our model, important factors which actually drive the variation of funding success probabilities are clearly illustrated. Empirically, our basic results are in line with earlier work. The difference is that our model reveals more hidden patterns. The prediction accuracy of our model is 7.5% higher than SVM and 13.5% higher than KNN, compared with earlier work. Our model has the practical significance with the help of the weight adjustment technique.

However, our algorithm has an exponential time complexity. To find a more efficient exact search method is one of the future research directions.

6. Acknowledgements

Supported by the Key Project of Beijing Natural Science Foundation (category B, No. KJ201210037037).

7. References

[1] Herrero-Lopez, A Sheng-Ying Pao, R Bhattacharyya. The Effect of Social Interactions on P2P Lending. *media.mit.edu.*

[2] L Puroa, JE. Teichb, H Walleniusa, J Wallenius. Borrower Decision Aid for people-to-people lending. *Decision Support Systems*. Volume 49, Issue 1, April 2010, Pages 52-60.

[3] M Klafft. Online peer-to-peer lending: A lender's perspective. *Proceedings of the International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government*, EEE 2008.

[4] http://www.Prosper.com

[5] J Ryan, K Reuk, C Wang. To Fund Or Not To Fund: Determinants Of Loan Fundability in the Prosper.com Marketplace. *Stanford Graduate School of Business.*

[6] R Daly, Q Shen, S Aitken. Learning Bayesian networks: approaches and issues. *The Knowledge Engineering Review (2011)*, 26: pp 99-157.

[7] F. M. Malvestuto. Approximating discrete probability distributions with decomposable models. *STATISTICS AND COMPUTING*, Volume 6, Number 2, 169-176.

[8] GF. Cooper and E Herskovits. A Bayesian method for the induction of probabilistic networks from data. *MACHINE LEARNING*, Volume 9, Number 4, 309-347.

[9] A Darwichek. Recursive conditioning. *Artificial Intelligence*, Volume 126, Issues 1-2, February 2001, Pages 5-41.