Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 15, no. 6, June 2019, pp. 1591-1599
DOI: 10.23940/ijpe.19.06.p10.15911599

Learning P2P Lending Credit Evaluation Bayesian Network from Missing Data

Yali Lv[a,b,*], Jianai Wu[a], Junzhong Miao[a], Weixin Hu[a], and Tong Jing[a]

[a]*School of Information Management, Shanxi University of Finance and Economics, Taiyuan, 030006, China*
[b]*Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, 030006, China*

Abstract

Credit evaluation is an important issue for investors in the financial field. However, there is a large amount of missing data in the P2P lending platform. To evaluate borrowers' credit from missing data, a credit evaluation Bayesian network model learning algorithm is proposed based on domain knowledge. Specifically, we first give a credit evaluation Bayesian network (CEBN) model to represent the borrowers' attributions and the relationships between attributions, and then we design the CEBN learning algorithm based on domain knowledge. Furthermore, we analyze and discuss the time complexity of the algorithm. Finally, the experimental results demonstrate that the CEBN model has good interpretability, learning performance, and evaluation performance by comparing it with other methods.

Keywords: probabilistic inference; credit evaluation; Bayesian networks; domain knowledge; qualitative influences; P2P lending

(Submitted on March 20, 2019; Revised on April 3, 2019; Accepted on June 7, 2019)

1. Introduction

Peer to peer (P2P) lending is a kind of unsecured loan made by borrowers and lenders through Internet platforms without the involvement of traditional financial intermediaries. It has many advantages, such as convenience, low threshold, and high interest rate, that make the P2P lending platform very popular. It has been developed rapidly at home and abroad.

However, with the development of P2P lending, some disadvantages have been gradually revealed. One disadvantage is that the borrowing interest rate on the platform is much higher than that from traditional financial institutions. Additionally, the cost of default is low, which makes the borrower's credit risk increase. Another disadvantage is that the threshold of opening to borrowers is relatively low. Although there are many indicators that are designed to fill for borrowers, they are all not necessary to be filled; borrowers can borrow a certain amount of funds from the platform by simply providing their own identity information and income information. Thus, there are many missing values in P2P lending data, which severely affects the credit evaluation of borrowers. Therefore, to make P2P lending services run normally, it is necessary to build an effective borrowers' credit evaluation model from missing data.

Recently, many researchers have addressed methods for evaluating borrowers' credit in P2P lending, such as regression method, multiple instance learning, SVM, ensemble learning, and so on. Zhang et al. [1] analyzed the factors that influence the probability of the borrowers' obtaining funds online P2P lending based on binary logistic regression. Guo et al. [2] took the regression coefficient as the optimal weight of credit risk assessment to study a credit risk assessment model based on the mathematical framework of nuclear regression. For conventional data and dynamic transaction behavior data, Zhang et al. [3] addressed credit risk assessment by radial basis function multiple instance learning and showed that the method improves prediction performance. To address sample bias for new applicants, Li et al. [4] studied reject inference in credit scoring by semi-SVM and showed that this method improves the performance of scoring models. Xia et al. [5] addressed a heterogenous ensemble credit scoring model based on the bagging and stacking method. In these mentioned modeling methods, however, the case of evaluation data with missing values is not considered. Additionally, the uncertainty

* Corresponding author.
E-mail address: sxlvyali@126.com

relationship between attribution variables in the credit evaluation model cannot be described, which prevents these models from having good interpretability.

The Bayesian network (BN) [6] can be used to represent and deal with uncertainty relationships between variables and has good interpretability, so many efficient BN model learning algorithms [7-10] have been proposed. The BN model has been applied in many fields. Thus, in order to describe the uncertainty relationship between attributions and discuss that multidimensional information is helpful for evaluating correctly borrowers' credit status, Pei and Guo [11] studied a credit evaluation model based on BN and illustrated that the model has better accuracy and efficiency in evaluating the borrower's credit status. However, they did not consider the credit evaluation from missing data in this method. In our previous work [12], for missing data, we proposed the PQISEM algorithm to learn the structure of the BN model by the SEM algorithm and partial qualitative influences. In addition, we verified its performance by comparing it with other algorithms, but this algorithm has not been used to evaluate borrowers' credit in P2P lending. Therefore, for missing data in the P2P lending platform, we present a credit evaluation Bayesian network (CEBN) model to represent and evaluate the borrower's credit level, and then we propose the CEBN learning algorithm based on the PQISEM algorithm. This will reduce the investment risk of investors and guarantee the interests of investors in the P2P platform.

The remainder of this paper is arranged as follows. Section 2 reviews some preliminaries. In Section 3, we define a P2P borrowers' credit evaluation Bayesian network (CEBN) and propose the learning algorithm of CEBN model from missing data, and then we analyze the time complexity of the learning algorithm. In Section 4, experiments demonstrate the interpretability, learning performance, and evaluation performance of the CEBN model. Finally, Section 5 concludes this paper.

2. Preliminaries

Here, we mainly introduce the basic concepts of the Bayesian network (BN), qualitative influences, and PQISEM learning algorithm of Bayesian network structure from missing data.

2.1. BN Model

The BN model was proposed to represent and infer uncertainty knowledge by Professor Pearl [6], and it was developed by combining probability statistics with graph theory. Its main advantage is that it can reduce greatly the computational complexity of probabilistic inference.

The BN model can be formally expressed as $M = (G, \theta)$, where $G = (V, E)$ is a directed acyclic graph (DAG) structure and V is the node set that represents random variables. E expresses the set of directed edges, reflecting the interdependence between variables. θ is a set of each variable's conditional probability distribution (CPD) and reflects the dependence degree between parent and child variables.

To infer or evaluate borrowers' credit based on BN, we first need to learn the BN model (G, θ) from data set D, including its structure and parameters. Generally, the BIC score [8] can be used to measure the model; the higher the BIC score, the better the obtained BN model.

2.2. Qualitative Influence Knowledge

Qualitative influence (QI) [13] is a kind of representation formal of qualitative probabilistic knowledge. A QI between two nodes reflects how one node's value influences the probability of the other node's value.

Assume that each variable is a binary variable; they are TRUE and FALSE and $\mathrm{TRUE} > \mathrm{FALSE}$. The positive QI of variable A on B, written as $S^+(A, B)$, means that knowing a high value for A makes it more likely to have a higher value for B, regardless of other parent nodes C on B. In other words,

$$p(B = \mathrm{TRUE}|A = \mathrm{TRUE}, C = \mathrm{TRUE}) \geq p(B = \mathrm{TRUE}|A = \mathrm{FAlSE}, C = \mathrm{TRUE})$$

$$p(B = \mathrm{TRUE}|A = \mathrm{TRUE}, C = \mathrm{FAlSE}) \geq p(B = \mathrm{TRUE}|A = \mathrm{FAlSE}, C = \mathrm{FAlSE})$$

A negative QI, written as $S^-(A, B)$, can be obtained by replacing $\geq$ with $\leq$. It represents that knowing a high value for A makes it less likely to have a higher value for B, regardless of other parent nodes C on B.

2.3. PQISEM Algorithm for Learning DAG of Bayesian Network from Missing Data

The goal of DAG structure learning of the BN model is to obtain the best structure fitting with prior knowledge and data. For learning of the BN model structure from missing data, the most classical method is the SEM algorithm [9] proposed by Friedman. Further, we proposed a PQISEM algorithm based on partial QIs in [12].

In the PQISEM algorithm, there is a loop iteration for i times, and it includes three steps in the process of learning.

We firstly give an initial BN $(G^0, \theta^{0,0})$ and then optimize parameter $\theta^{0,0}$ for K times by the EM algorithm and get parameters $\theta^{0,k+1}$. Meanwhile, we make the missing data complete. Specifically, during the process of optimizing parameters, all parameters $\theta^{0,k+1}$ are checked; if the parameters do not meet the constraints of being given QIs, then we will modify the corresponding parameters by the weighted average method.

Secondly, the structure G^0 is fixed, and we get all candidate DAG models M_c of G^0 by three operators, such as adding edge, subtracting edge, and reversing edge operators. Then, the structure and parameter optimization are carried out, and we can obtain the best BN model $(G^{i+1}, \theta^{i+1,0})$. Specifically, during the process of optimizing parameters and the structure, we select the best DAG structure from M_c to optimize its parameters K times.

Thirdly, we judge if $\mathrm{BIC}(G^{i+1}, \theta^{i+1,0}|D) \leq \mathrm{BIC}(G^i, \theta^{i,K}|D)$, and then we get the current BN model $(G^i, \theta^{i,K})$. Otherwise, we take $(G^{i+1}, \theta^{i+1,0})$ as the current BN model. The process is carried out until the algorithm converges or the given iteration times are reached.

3. Credit Evaluation BN Model Learning

This section defines the credit evaluation BN (CEBN) of P2P lending and designs the learning algorithm of the CEBN model based on the PQISEM algorithm. Then, we analyze the complexity of the learning algorithm.

The goal of learning CEBN from missing data is to analyze the relationship between variables and evaluate the borrowers' credit. There are several problems that need to be considered. (1) For all variables, we need to give a CEBN model to reflect the dependent relationship and dependent degree between variables. (2) For missing data, we need know how to learn the CEBN model that can evaluate the borrowers' credit or judge who is the default borrower. What domain knowledge is helpful for learning the CEBN model?

3.1. Definition of Credit Evaluation BN Model

To describe the dependent relationship and dependent degree between variables when we evaluate borrowers' credit in P2P lending, we give the definition of the credit evaluation BN (CEBN) based on the formal representation of basic BN and the characteristics of problem in credit evaluation domain as follows.

Definition 1 (CEBN Model). A credit evaluation BN model of borrower on P2P platform, written as CEBN, consists of $(G_{\mathrm{CEBN}}, \theta)$.

- $G_{\mathrm{CEBN}} = (V \cup V_C, E \cup E_C)$ is also a DAG structure, where V_C is the node of borrower's "LoanStatus", also named the class node; V is the node set of general attributes of P2P borrowers' credit data; E is the set of directed edges of general attributes; and E_C is the set of all edges that V_C points to each general attribute node.
- $\theta = \{p(V_C) \cup p(X_i|\{\pi(X_i) \cup V_C\})\}$ is a set of CPDs corresponding to each node in the set $\{V \cup V_C\}$, reflecting the dependence degree between parent node and child node. $\pi(X_i)$ is the parent node set of variable X_i, and $X_i \in V$.

3.2. Leaning Algorithm based on PQISEM Algorithm

To learn the CEBN model of evaluating borrowers' credit from missing data, we address the learning algorithm based on the PQISEM algorithm that was proposed in the previous work [12]. The CEBN learning algorithm is described as follows.

Firstly, since the main purpose of CEBN model is to use other attribute information to study and predict a borrower's future "LoanStatus", it is assumed that the class label "LoanStatus" is the parent node of all other nodes, that is, we have some domain knowledge E_C. In addition, it is rather easy to give some qualitative influence (QI) knowledge according to domain knowledge. For example, if a person borrows money in the midnight, he is probably unemployed. The repayment

ability of borrowers is doubtful. We can obtain some qualitative influence knowledge as "*The default possibility of borrower whose borrowing time is in the midnight is higher* and *the default possibility of borrower who is unemployed is higher*". Thus, based on E_C and some QIs, we can accurately give the initial G^0.

Next, based on G^0, we randomly assign its parameters $\theta^{0,0}$ and then optimize these parameters for K times by the EM algorithm and get $\theta^{0,k+1} \leftarrow \arg\sup_\theta Q(G^0, \theta \,|G^0, \theta^{0,k})$, where Q is the expect BIC score of (G^0, θ) on complete data that is filled for missing data by the EM algorithm on $(G^0, \theta^{0,k})$. These parameters are modified using the given QIs as the PQISEM algorithm.

Thirdly, taking G^0 as the current DAG, we find its candidate DAGs using three operators that have been given in the PQISEM algorithm to continue optimize the current model. Moreover, we modify its parameters by QIs after model optimization every time. Note that we only optimize the parameters of the candidate optimal DAG structure, that is, only K times parameters are optimized, rather than every candidate DAG structure being optimized.

Finally, the above process is carried out until the algorithm converges or the given iteration times are reached.

The pseudo-code of the learning algorithm of the borrowers' CEBN model on the P2P platform based on the PQISEM algorithm is shown as Algorithm 1.

Algorithm 1. Leaning Algorithm of CEBN based on PQISEM Algorithm

Input: credit evaluation data set D, initial parameter θ, qualitative influences QI, the number of parameter optimization K
Output: $(G_{\text{CEBN}}, \theta)$
Begin
$E \leftarrow$ all directional edges of general attribution variables in QI;
$G^0 \leftarrow (V \cup V_C, E \cup E_C)$, $\theta^{0,0} \leftarrow \theta$;
for $i = 0$ **to** ∞ **do**
    for $k = 0$ **to** $K - 1$ **do**
        $\theta^{i,k+1} \leftarrow \arg\sup_\theta Q(G^i, \theta \,|G^i, \theta^{i,k})$ based on EM algorithm;
    end for
    while the parameters violate the constraints of the qualitative influence relations in QI **do**
        Modifying the order of corresponding parameters' probability distribution as PQISEM algorithm
    end while
    $M_C \leftarrow$ All candidate models obtained by any edge-adding, edge-subtracting or edge-reversing operator in G^t once as PQISEM algorithm;
    $G^{i+1}, \theta^{i+1,0} \leftarrow \arg\max_{G \in M_C} \sup_\theta Q(G, \theta | G^i, \theta^{i,K})$;
$D^* \leftarrow$ complete data that are filled the missing values in D;
    if $\text{BIC}(G^{i+1}, \theta^{i+1.0} | D^*) \leq \text{BIC}(G^t, \theta^{i,K} | D^*)$ **then**
        return $(G^i, \theta^{i,K})$
    end if
end for
end

The time complexity of Algorithm 1 is similar to that of the SEM algorithm. One difference is that we need to modify the optimized parameters according to the known constraints of QIs after that parameters have been optimized K times by the EM algorithm. Assume that the given number of QIs is N, the number of parent nodes corresponding to directed edges of QIs is up to u, and the value of these parent nodes is up to M, so we need to compare $N \times \left[(M-1) + (M-2) + \cdots + \left(M - (M+1)\right)\right] \times M^{u-1}$ times in the modifying process, that is, the time complexity of modifying is $O(N \times M^{u+1})$. Usually, $u \leq 5$.

4. Experiment Analysis

We will verify the performance of the CEBN model learning from missing data in the Prosper platform in this section.

4.1. Experiment Setting and Data Set

The experiments are conducted in R language by using package bnlearn, utils, rje, seewave, gRbase, gRain, and so on. The data set is the P2P lending data in the Prosper platform from July 13, 2013 to 2014, and it contains 81 attributes and 72844 samples. According to the node of "LoanStatus", the data set is divided into a training set and a test set. There are 47244 samples in the training set and 25600 samples in the test set. After data preprocessing, such as attribute selection and construction, normalization, discretization of continuous attributes based on the ChiMerge algorithm, and merging of

discrete values of discrete attributes, there are 43 attribute variables that are retained, one of which is named "LoanStatus" and is the class attribute. The overall missing rate of the statistical training set is 7.54%, and that of the test set is 6.79%. In addition, the given ten qualitative influences according to domain knowledge are shown in Table 1.

Table 1. The given ten qualitative influences

No.	Probability Relationship of Qualitative Influences
QI1	p(ListingCreationDate = FALSE| LoanStatus = TRUE) > p(ListingCreationDate = FALSE| LoanStatus = FALSE)
QI2	p(EmploymentStatus = not employed| LoanStatus = TRUE) > p(ListingCreationDate = not employed | LoanStatus = FALSE)
QI3	p(InGroupSituation = 1 and 2| LoanStatus = FALSE) > p(InGroupSituation = 1 and 2 | LoanStatus = TRUE)
QI4	p(OpenRevolveMonthlyPaymentRate = B and C | LoanStatus = TRUE) > P(OpenRevolveMonthlyPaymentRate = B and C | LoanStatus = FALSE)
QI5	p(TotalInquiries = B and C| LoanStatus = TRUE) > p(TotalInquiries = B and C | LoanStatus = FALSE)
QI6	p(Delinquent = TRUE | LoanStatus = TRUE) > p(Delinquent = TRUE | LoanStatus = FALSE)
QI7	p(DelinquenciesLast7Years = TRUE | LoanStatus = TRUE) > p(DelinquenciesLast7Years = TRUE | LoanStatus = FALSE)
QI8	p(PublicRecordsLast12Months = TRUE | LoanStatus = TRUE) > p(PublicRecordsLast12Months = TRUE | LoanStatus = FALSE)
QI9	p(StatedMonthlyIncome = B and C| LoanStatus = TRUE) > p(StatedMonthlyIncome = B and C | LoanStatus = FALSE)
QI10	p(Recommendations = TRUE | LoanStatus = TRUE) > p(Recommendations = TRUE | LoanStatus = FALSE)

4.2. The Learned Model Results and Analysis

In this subsection, we first analyze the learned results of CEBN model from missing data in four different missing rate cases, and then we show and analyze the selected model's structure and its parameters.

4.2.1. The Learned Result Analysis in Different Missing Value Cases

In order to verify the effectiveness of the CEBN model, we randomly construct the missing values (except for class attributes) for the original training set. The rates of missing values are set as 10%, 12.5%, and 15%, respectively. Furthermore, we learn the CEBN model 100 times using Algorithm 1 in each missing value case. The qualitative influences that can work during the learning process are 7, 9, and 10. The learned results of the CEBN model are shown in Table 2.

Table 2. The learned results of CEBN model in different missing value cases

Missing rate	7.54%	10%	12.5%	15%
BIC score	-997364.0888	-993209.7385	-985374.9364	-989578.8804
Number of qualitative influences	7	7	9	10

From Table 2, we can see that with an increase in the proportion of missing values, the number of qualitative influences at work increases. Moreover, in the first three cases, the larger the proportion of missing values, the higher the BIC score. In other words, a better network structure can explain the information that is contained in the data more accurately. However, when the rate of missing values is 15%, the BIC score decreases and all ten qualitative influences are observed to work, which proves that ten qualitative influences are not enough to work better with so much missing data. Thus, to learn a better model from missing data with a higher missing rate, the number of qualitative influences needs to be increased appropriately for our proposed method.

4.2.2. The Selected Model Structure and Parameters

To further to evaluate borrowers' credit in original missing data, we select the learned CEBN model in the original missing rate case as the best model. It is shown in Figure 1. For DAG of the CEBN model, there are 43 nodes and 131 edges. The information statistics results in the selected model, including node names, directed edges, and each node's parent nodes and child nodes, are shown in detail in Table 3.

For the explanation of dependent degree between variables in the selected CEBN model, because there are too many parameter tables in the model, we only list the node parameter tables that are shown in Tables 4-6 for illustration.

For example, Table 4 is a parameter table between "ListingCreationDate" and "LoanStatus". These parameters indicate the probability distribution of "ListingCreationDate" during the day or at midnight in the case of "repayment" or "default". It can be observed that the "default " probability is higher when the borrowing time is at midnight than that during the day, which is consistent with the qualitative influence of QI1.

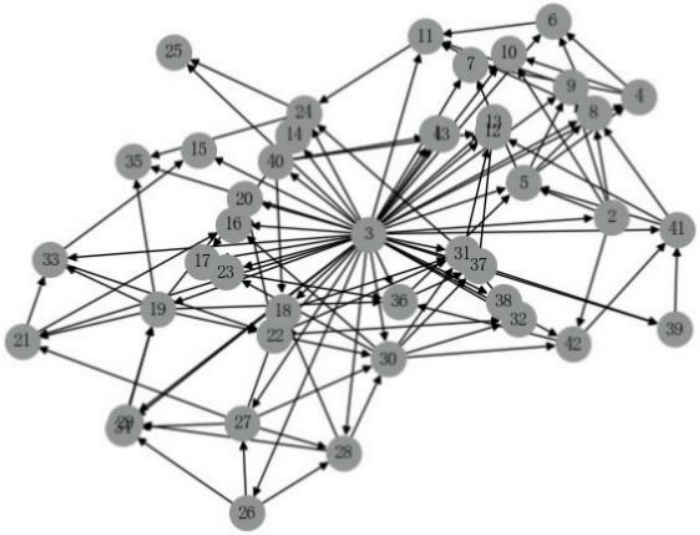

Figure 1. The learned CEBN model on Prosper flatform

Table 3. Information statistics results in the selected model

No.	Attributions	Parent node number	Child node number	Edge number	Sequence number of child nodes
1	ListingCreationDate	1	0	1	-
2	Term	1	5	6	5, 8, 9, 10, 42
3	LoanStatus	0	42	42	1, 2, 4-43
4	BorrowerAPR	2	3	5	6, 7, 10
5	BorrowerRate	4	4	8	4, 7, 8, 9
6	LenderYield	3	1	4	11
7	EstimatedEffectiveYield	3	0	3	-
8	EstimatedLoss	4	2	6	6, 11
9	EstimatedReturn	3	1	4	11
10	ProsperRating	3	0	3	-
11	ProsperScore	4	1	5	24
12	ListingCategory	4	0	4	-
13	BorrowerState	2	0	2	-
14	EmploymentStatus	1	0	1	-
15	EmploymentStatusDuration	2	0	2	-
16	IsBorrowerHomeowner	3	0	3	-
17	InGroupSituation	2	0	2	-
18	CreditScoreRangeMean	4	1	5	31
19	CurrentCreditLines	3	6	9	20, 21, 22, 33, 35, 36
20	OpenCreditLines	3	0	3	-
21	TotalCreditLinespast7years	3	2	5	16, 33
22	OpenRevolvingAccounts	3	4	7	20, 30, 31, 32
23	OpenRevolveMonthlyPaymentRate	2	1	3	36
24	InquiriesLast6Months	3	2	5	25, 35
25	TotalInquiries	2	0	2	-
26	Delinquent	1	3	4	27, 28, 34
27	DelinquenciesLast7Years	2	5	7	18, 21, 28, 30, 34
28	PublicRecordsLast10Years	3	3	6	18, 29, 30
29	PublicRecordsLast12Months	2	1	3	19
30	RevolvingCreditBalance	4	6	10	16, 23, 31, 32, 37, 42
31	BankcardUtilization	4	2	6	12, 32
32	AvailableBankcardCredit	4	0	4	-
33	TotalTrades	3	2	5	15, 22
34	TradesNeverDelinquent	3	1	4	19
35	TradesOpenedLast6Months	3	0	3	-
36	DebtToIncomeRatio	4	2	6	5, 37
37	IncomeRange	3	3	6	13, 24, 39
38	IncomeVerifiable	1	0	1	-
39	StatedMonthlyIncome	2	1	3	41
40	TotalProsperLoans	1	4	5	12, 17, 18, 43
41	LoanOriginalAmount	3	3	6	5, 8, 12
42	MonthlyLoanPaymentRate	3	2	5	36, 41
43	Recommendations	2	0	2	-

Table 4. The parameters of "ListingCreationDate" node

P(ListingCreationDate|LoanStatus)		LoanStatus	
		Repayment	Default
ListingCreationDate	In the day	0.24635	0.25581
	In the midnight	0.75365	0.74419

Table 5 shows the probability distribution between the "BorrowerState" node and its parent nodes "LoanStatus" and "IncomeRange". "LoanStatus" is classified as A, B, C, and D according to the GDP level. A is the most developed area, B is the developed area, C is the undeveloped area, and D is the most undeveloped area. From Table 5, we can see that the probabilities of A, B, C, and D decrease monotonously regardless of the status of the parent nodes. This means that the higher the economic level, the higher the usage rate for the online lending platform.

Table 5. The parameters of "BorrowerState" node

BorrowerState	IncomeRange	P(BorrowerState |LoanStatus, IncomeRange)			
		A	B	C	D
Repayment	$0 - $49999	0.61183	0.22485	0.12134	0.04198
Default		0.57201	0.25926	0.12169	0.04704
Repayment	$50000 -$99999	0.64730	0.21494	0.10170	0.03606
Default		0.60872	0.23746	0.11126	0.04256
Repayment	$100000+	0.72085	0.18151	0.07000	0.02764
Default		0.69082	0.22131	0.05738	0.03049

Table 6 reflects the relationship between the "Recommendations" node and its patent nodes "TotalProsperLoans" and "LoanStatus", which is also consistent with the qualitative influence QI10 selected in this paper. This means that the probability of repayment in the recommendations case is higher than that in the no recommendations case, regardless of the values of "TotalProsperLoans".

Table 6. The parameters of "Recommendations" node

LoanStatus	TotalProsperLoans	P(Recommendations | LoanStatus, TotalProsperLoans)	
		Yes	No
Repayment	Yes	0.00186	0.99814
Default		0.00068	0.99932
Repayment	No	0.03266	0.96734
Default		0.03158	0.96842

4.3. Model's Evaluation Results and Analysis

In addition to having good interpretability, another goal of learning the CEBN model is to evaluate borrowers' credit. Here, we will verify the evaluation performance in the CEBN model by comparing it with other structure learning methods of BN from missing data and comparing it with the other classification methods.

4.3.1. Experimental Result Analysis by Comparing with other Learning Methods for Missing Data

There are many Bayesian network structure learning algorithms, such as the SEM algorithm, Tabu algorithm, and MMHC algorithm [14]. In these experiments, we verify the learning and evaluation results through comparison with these three algorithms. The SEM algorithm is a classical algorithm for missing data. However, the Tabu algorithm and MMHC algorithm are structure learning algorithms for complete data, so we use the EM+Tabu and EM+MMHC algorithms to learn the CEBN model and then evaluate borrowers' credit. We take "LoanStatus" as a query variable and other variables that are not null as evidential variables and calculate the posterior probability of "LoanStatus" based on our CEBN model inference method on the training set and test set. Measures include BIC score, edge number, accuracy in the training set, and accuracy in the test set. The learned results and evaluation results are shown in Table 7.

From Table 7, in terms of BIC scoring, we can see that the Tabu algorithm and MMHC algorithm have similar scores because they both use the EM algorithm to fill data values first and then learn the model's structure. Meanwhile, the SEM algorithm and our proposed algorithm fill data according to current parameters in each iteration process. Therefore, the latter two algorithms' scores are higher than those of the former two algorithms. Moreover, the BIC score of our algorithm is about 3.47% higher than that of the SEM algorithm. This is because we introduce some qualitative influence knowledge, which strengthens the model structure and data. The evaluation accuracy of the four algorithms is the highest in our proposed algorithm in both the training set and test set, which shows that the CEBN model outperforms the other three algorithms.

Table 7. The learned results and evaluation results of four algorithms

Algorithms	EM+Tabu	EM+MMHC	SEM	CEBN
BIC score	-1393551.2159	-1387133.9377	-1033254.3143	-997364.0888
Edge number	152	145	126	131
Accuracy in training set	83.58%	84.27%	88.59%	93.86%
Accuracy in test set	80.44%	81.63%	84.30%	89.20%

4.3.2. Evaluation Result Analysis by Comparing with other Classification Methods

In order to further to verify borrowers' credit evaluation results of our proposed algorithm, we compare the CEBN model with other classification methods, such as Na ïve Bayes, C4.5, SVM, and Adaboost.

The evaluation measures include *Accuracy* (*Ac*), *Precision* (*Pr*), and *F-measure* (*Fm*). *Ac*, listed as Equation (1), is the proportion of the number of predicted results that are in accordance with the actual situation. *Pr*, shown as Equation (2), means the proportion of having been judged as non-default samples and having been the actual number of non-default samples in the model. *Fm*, shown as Equation (3), is a comprehensive evaluation index and a weighted harmonic average of accuracy and recall. The higher the three evaluation indexes, the better the results. The equations for each index are shown as follows.

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Pr = \frac{TP}{TP + FP} \tag{2}$$

$$Fm = \frac{(\alpha^2 + 1) \times Pr \times Re}{\alpha^2 Pr + Re} \tag{3}$$

Where

$$\text{Recall}(Re) = \frac{TP}{TP + FN}$$

is the recall rate and the probability of correctly predicting positive classes. *TP* is the number of positive classes that are consistent with actual and predicted cases. *TN* means the number of negative classes that are consistent with actual and prediction cases. *FP* is the number of samples that belong to a negative class in the actual case but belong to a positive class in the prediction case. *FN* is the number of samples that belong to a positive class in both the actual case and the prediction case. α is the weight value between *Pr* and *Re*.

For missing data, we fill the missing values using the mode in statistics for other classification algorithms. Furthermore, we take ten-fold cross-validation to obtain classifiers and evaluate the borrowers' credit in the test set. Meanwhile, in the test set, we use the CEBN model to infer the borrowers' credit. The obtained evaluation results are shown in Table 8.

Table 8. Evaluation results of five algorithms

Algorithms	*Accuracy* (*Ac*)	*Precision* (*Pr*)	*F-measure* (*Fm*)
Na ïve Bayes	0.7882	0.9262	0.8726
C4.5	0.7974	0.9190	0.8799
SVM	0.8274	0.9269	0.8989
Adaboost	0.8606	0.9382	0.9191
CEBN	0.9386	0.9736	0.9648

From Table 8, we can see that the performance of the P2P CEBN model constructed by our proposed algorithm outperforms the other four classification methods in terms of *Ac*, *Pr*, and *Fm*. The classification accuracy is nearly 8% higher than that of the Adaboost algorithm, which has the highest accuracy among the other four algorithms.

5. Conclusions

To evaluate and analyze P2P borrowers' credit, in this paper, a credit evaluation Bayesian network (CEBN) has been addressed based on domain knowledge. We mainly defined the ACBN model, studied its learning algorithm, and verified its

good performance in P2P lending data on the Prosper platform. The proposed method also raises other issues. We only discuss the borrowers' credit that is expressed as "LoanStatus" or "default or not"; in the future, we can divide the credit into different grades to address the probability of borrower default. Additionally, because of the limitations of conditions, the model can be learned from data from foreign platforms, and it could be interesting to address a CEBN model that is suitable for our country's situation.

Acknowledgments

This work has been funded by the National Natural Science Foundation of China (Nos. 61432011, U1435212, 61322211, and 61672332), the Natural Science Foundation of Shanxi Province, China (Nos. 201801D121115, 2013011016-4), and the Postdoctoral Science Foundation of China (No. 2016M591409).

References

1. Y. Zhang, H. Li, M. Hai, J. Li, and A. Li, "Determinants of Loan Funded Successful in Online P2P Lending," *Procedia Computer Science*, Vol. 122, pp. 896-901, 2017
2. Y. Guo, W. Zhou, C. Luo, C. Liu, and H. Xiong, "Instance-based Credit Risk Assessment for Investment Decisions in P2P Lending," *European Journal of Operational Research*, Vol. 249, No. 2, pp. 417-426, 2016
3. T. Zhang, W. Zhang, X. Wei, and H. Hao, "Multiple Instance Learning for Credit Risk Assessment with Transaction Data," *Knowledge-based Systems*, Vol. 161, pp. 65-77, 2018
4. Z. Li, Y. Tian, K. Li, F. Zhou, and W. Yang, "Reject Inference in Credit Scoring using Semi-Supervised Support Vector Machines," *Expert Systems with Applications*, Vol. 74, pp. 105-114, 2017
5. Y. Xia, C. Liu, B. Da, and F. Xie, "A Novel Heterogeneous Ensemble Credit Scoring Model based on bstacking Approach," *Expert Systems with Applications*, Vol. 93, pp. 182-199, 2018
6. J. Pearl, "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference," Morgan Kaufmann Publishers, San Mateo, California, 1988
7. M. Scanagatta, G. Corani, C. P. de-Campos, and M. Zaffalon, "Approximate Structure Learning for Large Bayesian Networks," *Machine Learning*, Vol. 107, pp. 1209-1227, 2018
8. C. P. De-Camposab, M. Scanagatta, G. Corani, and M. Zaffalon, "Entropy-based Pruning for Learning Bayesian Networks using BIC," *Artificial Intelligence*, Vol. 260, pp. 42-50, 2018
9. N. Friedman, "The Bayesian Structural EM Algorithm," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 129-138, Morgan Kaufmann Publishers Inc., 1998
10. Y. W. Park and D. Klabjan, "Bayesian Network Learning via Topological Order," *Journal of Machine Learning Research*, Vol. 18, pp. 1-32, 2017
11. P. Pei and Y. Guo, "P2P Borrower's Credit Evaluation Model based on Bayesian Network," *China Economic Studies*, Vol. 2, pp. 29-41, 2017
12. Y. Lv, J. Wu, and T. Jing, "PQISEM: BN's Structure Learning based on Partial Qualitative Influences and SEM Algorithm from Missing Data," *International Journal of Wireless and Mobile Computing*, Vol. 14, No. 4, pp. 348-357, 2018
13. M. P. Wellman, "Fundamental Concepts of Qualitative Probabilistic Networks," *Artificial Intelligence*, Vol. 44, pp. 257-303, 1990
14. I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm," *Machine Learning*, Vol. 65, No. 1, pp. 31-78, 2006

Yali Lv is an associate professor at Shanxi University of Finance & Economics of China. She received her Ph.D. from Tianjin University. Her research interests include probabilistic reasoning, concept learning, data mining, and machine learning.

Jianai Wu is a master's candidate at Shanxi University of Finance & Economics of China. Her research interests include data mining and financial data analysis.

Junzhong Miao is a master's candidate at Shanxi University of Finance & Economics of China. His research interests include Bayesian machine learning and concept learning.

Weixin Hu is a master's candidate at Shanxi University of Finance & Economics of China. Her research interests include data mining and machine learning.

Tong Jing is a master's candidate at Shanxi University of Finance & Economics of China. His research interests include Bayesian machine learning and concept learning.