Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Credit risk modeling using Bayesian network with a latent variable

Khalil Masmoudi*, Lobna Abid, Afif Masmoudi

Laboratory of Probability and Statistics - PB 1171 Faculty of Sciences of Sfax Sfax University, Tunisia

ARTICLE INFO

Article history:
Received 1 November 2018
Revised 7 March 2019
Accepted 8 March 2019
Available online 8 March 2019

Keywords:
Credit risk
Bayesian network
Latent variable
EM algorithm
Mixture model
Payment default

ABSTRACT

Credit risk assessment is an important task for the implementation of the bank policies and commercial strategies. In this paper, we used a discrete Bayesian network with a latent variable to model the payment default of loans subscribers. The proposed Bayesian network includes a built-in clustering feature. A full procedure for learning its parameters, based on a customized Expectation-Maximization algorithm was provided. This model allows evaluating the payment default probability taking into account several factors and handling a multi-class situation. Relying on a real data set describing loans contracts, we calibrated the model and performed several analyses. The obtained results highlight a regime switching of the default probability distribution: Two classes were determined showing a change in credit risk profiles.

1. Introduction

The banking system is crucially affected by the credit risk which may lead to economic stagnation worldwide (Nkusu, 2011). Known as credit crisis, the 2007 sub-prime mortgage crisis had a significant effect on the economy as it predominantly triggered the global financial crisis of 2008 (Longstaff, 2010). To control the credit risk, banks have used both qualitative and quantitative methods in order to minimize households' payment defaults. To this end, many credit scoring procedures have been adopted to evaluate and analyze the credit risk. According to Thomas, Edelman, and Crook (2002), credit scoring is a set of decision models and techniques which allow lenders to appropriately select their customers. In this context, many methodologies have been developed (García, Marqués, & Sánchez, 2015) such as the statistical methods (Hand & Henley, 1997) and the artificial intelligence methods (Lessmann, Baesens, Seow, & Thomas, 2015; Louzada, Ara, & Fernandes, 2016). The statistical methods include the linear discriminant analysis (Altman, 1968) and the logistic regression (Abid, Masmoudi, & Zouari-Ghorbel, 2016) which are popular credit scoring techniques thanks to their accuracy and easy implementation (Lessmann et al., 2015). To illustrate the artificial intelligence techniques, we can cite the support vector machines (Harris, 2015; Tomczak & Zieba, 2015), artificial neural networks (Zhao et al., 2015), decision trees (Bijak & Thomas, 2012) and Bayesian networks (Pearl, 1988).

A Bayesian Network (BN) is a graphical representation of a probabilistic model that encodes a set of conditional independence relationships (Ghribi & Masmoudi, 2013; Pearl, 1988). It has become a popular tool for decision making systems in various fields such as biology (Hassen, Masmoudi, & Rebai, 2008), computer science (Bouchaala, Masmoudi, Gargouri, & Rebai, 2010) and finance (Abid, Zaghdene, Masmoudi, & Ghorbel, 2017). Indeed, the BNs are one of the most comprehensive and consistent formalisms for the acquisition and modeling of complex systems outperforming the logistic regression in terms of diagnostic prediction (Gevaert et al., 2006).

Based on credit worthiness, the authors of Abid et al. (2017) used a discrete BN model for personal loans prediction and classification. They set up the conditional relationships between the factors affecting the credit risk and used the calibrated conditional probability tables to analyze the payment default causes and effects.

In this paper, we introduced a new discrete BN model containing a latent variable that affects all the other observable variables. While the BN structure models the probabilistic relationships between factors leading to credit default payment, the latent variable allows representing different classes of probability distributions. A full procedure for learning this model was proposed relying on a customized Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). The proposed model was used to evaluate credit risk and cluster loans subscribers enabling a deeper analysis of customers' payment defaults.

* Corresponding author.
E-mail addresses: khalil.masmoudi@centraliens.net (K. Masmoudi), lobnabid@yahoo.fr (L. Abid), afif.masmoudi@fss.rnu.tn (A. Masmoudi).

https://doi.org/10.1016/j.eswa.2019.03.014
0957-4174/© 2019 Elsevier Ltd. All rights reserved.

The remaining of this paper was structured as follows: Section 2 detailed the previous studies related to the topic. Section 3 described the discrete BN with a latent variable and the proposed procedure for learning this class of BNs using a customized EM algorithm. The proposed method was applied in the context of loans classification and credit risk evaluation in Section 4. Finally, our main conclusions were drawn in the ultimate section.

2. Related work

The credit risk and bankruptcy prediction were extensively studied over the recent years. Various models and techniques were employed in these studies in the context of risk evaluation and debtors classification. For instance, Danenas and Garsva (2015) introduced a new approach based on linear SVM combined with external evaluation and sliding window testing. Their method addresses the imbalanced classes issue and is suitable for large data sets. They showed that their method provides equivalent results compared to the logistic regression and RBF network. In the same sphere of predicting financial distress, Geng, Bose, and Chen (2015) used different types of data mining techniques, such as neural network, Decision trees and SVM to build an early warning system that serves to forecast financial distress of Chinese companies. The obtained results show that the neural network is more accurate than the other considered classifiers. Alternatively, Bemš, Starỳ, Macaš, Žegklitz, and Pošík (2015) used an innovative approach based on a modified magic square in order to evaluate the corporations economic performance and detect potential companies that might be subject to default. One of the strengths of this method is the graphical interpretation of a company score. Contrary to other techniques, such as discriminant analysis, logistic regression, neural networks and survival analysis, Du Jardin and Séverin (2011) proposed a new approach to evaluate a company financial health based on a Kohonen map. This approach improves the financial failure prediction accuracy.

On the other hand, Leong (2016) compared the BN to other classification methods in the context of credit scoring. He pointed out the ability of the BNs to tackle both the imbalanced classes and data censoring issues. He also showed that the BN has an efficient performance when implemented onto a large data set compared to the logistic regression, SVM, decision trees and Multi-Layer Perceptron (MLP). Moreover, Sanford and Moosa (2015) used the BN in the context of operational risk evaluation. They described a BN-based tool developed within one of the major Australian banks to predict operational risk events, aggregate operational loss distributions, and Operational Value-at-Risk. Sousa, Gama, and Brandão (2016) used a new dynamic modeling framework for credit risk assessment consisting of sequential learning from the new incoming data. This technique applied on a Brazilian credit cards data set outperformed the static modeling methods. Another interesting technique for assessing credit risk, called the hybrid associative memory, was proposed by Cleofas-Sánchez, García, Marqués, and Sánchez (2016). This method was compared to several popular machine learning techniques using nine real-life financial databases. The obtained results suggest that this approach can be appropriate to predicting financial distress, especially in the case of data sets where the classes are strongly imbalanced and/or over-lapped. In the same context, Mselmi, Lahiani, and Hamza (2017) adopted five different techniques to predict the financial distress of small and medium French firms. Their study compared the Logit, ANN, SVM, Partial Least Squares models and a hybrid model integrating SVM with Partial Least Squares. They showed that while the SVM is the best classifier for one year prior to financial distress, the hybrid model outperforms all the others for a two year period prior to financial distress. Using credit default swaps data sets, Luo, Wu, and Wu (2017) used a deep learning technique in order to examine the performances of credit scoring models. When Compared to other traditional techniques such as logistic regression, MLP and SVM, the authors found out that deep belief networks give the best performance. Tavana, Abtahi, Di Caprio, and Poortarigh (2018) used two methods among the most recent machine learning techniques namely ANNs and BNs in order to evaluate financial problems related to key factors of liquidity risk. Relying on real-case numerical experiments, the authors came up to the conclusion that the two techniques are rather complementary in revealing the liquidity risk.

3. Discrete Bayesian network with a latent variable

A BN consists of a directed acyclic graph (DAG) and a set of associated conditional probability distributions. The DAG reflects a set of conditional independence relationships between a set of variables (nodes). A finite Discrete BN $\mathcal{B}=(G,P)$ is a BN whose nodes are discrete random variables $(X_1,\ldots,X_d)$ taking a finite number of values. In this paper, the following notations were used:

- d denotes the number of variables in the BN.
- Each node X_i takes r_i possible values encoded as $1,2,\ldots,r_i$.
- The parents of each node X_i in G, denoted by $Pa(X_i)$ have $q_i=\prod_{m\in\Omega_i} r_m$ possible configurations encoded as $1,2,\ldots,q_i$, where $\Omega_i=\{m;\ X_m\in Pa(X_i)\}$.
- The conditional distribution of $X_i|Pa(X_i)$ is defined by the matrix of probabilities $\boldsymbol{\theta}_i=(\theta_{ijk})_{\substack{1\le j\le q_i\\1\le k\le r_i}}$, where:

  $\theta_{ijk}=P(X_i=k|Pa(X_i)=j)$.

  The set of all the parameters is denoted by $\boldsymbol{\Theta}=(\boldsymbol{\theta}_i)_{1\le i\le d}$ and the discrete BN model is depicted by $\mathcal{B}(G,\boldsymbol{\Theta})$. For each node X_i, $\boldsymbol{\theta}_{ij}=(\theta_{ij1},\ldots,\theta_{ijr_i})$ is the vector of conditional probabilities defining the distribution of $(X_i|Pa(X_i)=j)$ and satisfying $\sum_{k=1}^{r_i}\theta_{ijk}=1$.

  The joint probability distribution of $(X_1,\ldots,X_d)$ is given by

$$P(X_1,X_2,\ldots.,X_d)=\prod_{i=1}^{d}P(X_i|Pa(X_i)). \tag{1}$$

Including latent variables in a BN is, generally useful to model hidden effects or causes, which leads to simpler relationships between the nodes (Kwoh & Gillies, 1996). In other situations, the latent variable is a classifying or ranking one (Kim & Jun, 2013). In this case, we have a BN with a fixed DAG combined with a mixture distribution. The classifying variable, therefore, is used to manage the memberships to the different classes. In what follows, a detailed description of the proposed model is provided and an EM-based algorithm is proposed to learn the model parameters.

3.1. Model description

Consider a classifying latent variable C taking values in $\{1,\ldots,L\}$ and d discrete random variables $X_1,X_2,\ldots,X_d$ whose conditional dependencies are encoded by a DAG $G=(\boldsymbol{X},E)$, the BN with the latent variable C is defined as follows:

$$\text{For each } 1\le l\le L,\ \boldsymbol{X}|C=l\sim\mathcal{B}(G,\boldsymbol{\Theta}_l) \tag{2}$$

Let $\pi_1,\ldots,\pi_L$ be the prior probabilities of each class, i.e., $\pi_l=P(C=l)$ and $\mathcal{L}(G,\boldsymbol{\Theta}_l)$ is the joint probability distribution induced by the discrete BN $\mathcal{B}(G,\boldsymbol{\Theta}_l)$. Then, the model can be rewritten as a mixture model:

$$\boldsymbol{X}\sim\sum_{l=1}^{L}\pi_l\mathcal{L}(G,\boldsymbol{\Theta}_l) \tag{3}$$

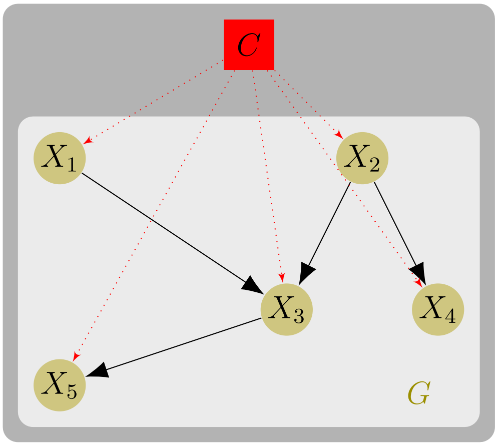

Fig. 1. An example of a BN with a classifying latent variable.

An example involving five observable variables is reported in Fig. 1. The number of classes L is assumed to be known. This information can be either provided by domain experts or determined through an advanced data analysis involving a model selection procedure (Njah, Jamoussi, Mahdi, & Masmoudi, 2015).

This kind of BNs offers several advantages: First, it provides a flexible tool to fit any probability distribution since it relies on a mixture of different distributions. Second, this model is naturally appropriate to deal with the data drawn from a population which consists of L groups. Finally, after calibrating the model from data, the main characteristics of each group can be easily extracted. These advantages are illustrated in Section 4 in the context of credit clients classification.

3.2. Model learning and inference

The main challenge with the BNs lies in the use of real data sets when calibrating the model by identifying the most appropriate *DAG* and the most likely parameters. Indeed, the model estimation consists of two parts: structure learning and parameters learning. The first task can be achieved using score-based (Korb & Nicholson, 2010) or constraint-based (Spirtes, Glymour, Scheines et al., 2001) algorithms. For parameters learning, we propose a customized version of the EM algorithm assuming a fixed structure.

Consider a random sample $\mathcal{D} = (\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(M)})$ drawn from (3). The data set contains only the observable variables $X_1, \ldots, X_d$. Here, the model is by construction formulated as an incomplete-data problem, since the classifying variable C is unobservable. Hence, it would be adequate to use the EM algorithm in order to estimate the model parameters. Thus, we introduce the component label vectors $(\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \ldots, \boldsymbol{Z}^{(M)})$ associated to the instances of $\boldsymbol{X}$ in $\mathcal{D}$, where $\boldsymbol{Z}^{(m)} = (Z_{m1}, \ldots, Z_{mL})$ is L-dimensional vector such that $Z_{ml} = 1$ if $\boldsymbol{X}^{(m)}$ belongs to l^{th} class an $Z_{ml} = 0$ otherwise. The sample's complete likelihood is given by:

$$l_c(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(M)}, \boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(M)} | \boldsymbol{\Theta} = (\boldsymbol{\Theta}_l)_{1 \le l \le L}, G)$$
$$= \prod_{m=1}^{M} \prod_{l=1}^{L} \left[\pi_l P(\boldsymbol{X} = \boldsymbol{X}^{(m)} | \boldsymbol{\Theta}_l)\right]^{Z_{ml}}$$

where $P(\boldsymbol{X} = \boldsymbol{X}^{(m)} | \boldsymbol{\Theta}_l) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{lijk}^{\delta_{ijk}^{(m)}}$ and $\delta_{ijk}^{(m)} = \begin{cases} 1 & if\ X_i^{(m)} = k\, and\, X_{pa(i)}^{(m)} = j \\ 0 & otherwise \end{cases}.$

The log-likelihood function is given by:

$$L_c = \log(l_c) = \sum_{m=1}^{M} \sum_{l=1}^{L} \left[Z_{ml} \log(\pi_l) + \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} Z_{ml} (\delta_{ijk}^{(m)} \log(\theta_{lijk})) \right] \tag{4}$$

The proposed algorithm for parameters estimation is as follows:

- **Initialization**: Choose the initial parameters $\Theta^{(0)}$.
- **E-Step**: The latent variable C is handled, in the tth iteration, by computing the conditional expectation of the complete-data log likelihood given the observed data D, using the current fit for $\Theta^{(t)}$

$$Q(\Theta || \Theta^{(t)}) = E_{\Theta}(L_c | \Theta^{(t)}, \mathcal{D}) \tag{5}$$
$$= \sum_{m=1}^{M} \sum_{l=1}^{L} \left[\tau_{ml}^{(t)} \log(\pi_l) + \tau_{ml}^{(t)} \sum_{i=1}^{n} \sum_{i=1}^{q_i} \left\{ \sum_{k=1}^{r_i} \delta_{ijk}^{(m)} \log(\theta_{lijk}) \right\} \right]$$

where $\tau_{ml}^{(t)} = E(Z_{ml} | D, \Theta^{(t)})$ is the current expectation of Z_{ml} given the observed data D. It can be viewed as the posterior probability that $\boldsymbol{X}^{(m)}$ belongs to the lth class:

$$\tau_{ml}^{(t)} = \frac{\pi_l^{(t)} P(\boldsymbol{X} = \boldsymbol{X}^{(m)} | \Theta_l^{(t)})}{\sum_{k=1}^{L} \pi_k^{(t)} P(\boldsymbol{X} = \boldsymbol{X}^{(m)} | \Theta_k^{(t)})} \tag{6}$$

- **M-Step** The M-step, at the $(t+1)$th iteration, requires the global maximization of $Q(\Theta || \Theta^{(t)})$ with respect to Θ over the parameters space to give the updated estimate.

$$\Theta^{(t+1)} = \arg\max_{\Theta} Q(\Theta || \Theta^{(t)}) \tag{7}$$

The parameters update that maximizes $Q(\Theta || \Theta^{(t)})$ is given by:

$$\pi_l^{(t+1)} = \frac{1}{M} \sum_{m=1}^{M} \tau_{ml}^{(t)} \tag{8}$$

and

$$\theta_{lijk}^{(t+1)} = \frac{\sum_{m=1}^{M} \tau_{ml}^{(t)} \delta_{ijk}^{(m)}}{\sum_{k=1}^{r_i} \sum_{m=1}^{M} \tau_{ml}^{(t)} \delta_{ijk}^{(m)}} \tag{9}$$

As within any EM-based algorithm, these steps are alternated until convergence. The data is then clustered using posterior probabilities τ_{ml}.

Proof. In order to maximize Q, we simply search the zero of its first derivative with respect to θ_{lijk}. Taking into account the constraint $\sum_{k=1}^{r_i} \theta_{lijk} = 1$, we rewrite $Q(\Theta || \Theta^{(t)})$ as follows:

$$\sum_{m=1}^{M} \sum_{l=1}^{L} \left[\tau_{ml}^{(t)} \log(\pi_l) + \tau_{ml}^{(t)} \sum_{i=1}^{d} \sum_{j=1}^{q_i} \left\{ \delta_{ij1}^{(m)} \log(1 - \sum_{k=2}^{r_i} \theta_{lijk}) \right.\right.$$
$$\left.\left. + \sum_{k=2}^{r_i} \delta_{ijk}^{(m)} \log(\theta_{lijk}) \right\} \right] \tag{10}$$

The first derivative of Q with respect to θ_{lijk} is given by:

$$\frac{\partial Q}{\partial \theta_{lijk}} = \sum_{m=1}^{M} \tau_{ml}^{(t)} \left(\frac{\delta_{ijk}^{(m)}}{\theta_{lijk}} - \frac{\delta_{ij1}^{(m)}}{1 - \sum_{k=2}^{r_i} \theta_{lijk}} \right) \tag{11}$$

Hence,

$$\frac{\partial Q}{\partial \theta_{lijk}} = 0 \Leftrightarrow \theta_{lijk} \sum_{m=1}^{M} \tau_{ml}^{(t)} \delta_{ij1}^{(m)} = \theta_{lij1} \sum_{m=1}^{M} \tau_{ml}^{(t)} \delta_{ijk}^{(m)} \tag{12}$$

Table 1
Descriptive statistics of the continuous variables: Amount, Duration, MBR and Age.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Amount	3001.00	8000.00	13232.00	17200.96	20000.00	149958.00
Duration	1.00	4.84	6.84	7.48	7.07	28.43
MRB	51.00	157.00	219.00	261.24	312.00	15550.00
Age	18.07	34.26	43.33	43.37	51.60	75.98

Finally, summing over k and using the fact that $\sum_{k=1}^{r_i} \theta_l ijk = 1$, we get:

$$\sum_{m=1}^{M} \tau_{ml}^{(t)} \delta_{ij1}^{(m)} = \theta_{lij1} \sum_{k=1}^{r_i} \sum_{m=1}^{M} \tau_{ml}^{(t)} \delta_{ijk}^{(m)}. \tag{13}$$

Thus, the maximum is achieved for $\theta_{lij1} = \frac{\sum_{m=1}^{M} \tau_{ml}^{(t)} \delta_{ij1}^{(m)}}{\sum_{k=1}^{r_i} \sum_{m=1}^{M} \tau_{ml}^{(t)} \delta_{ijk}^{(m)}}$. Similarly, this formula generalizes for any $k = 2, \ldots, r_i$. □

The proposed procedure for learning the discrete BN with latent variable can be summarized as follows:

(i) Use the incomplete data set to learn the structure of the discrete BN involving the observable variables $(X_1, \ldots, X_d)$.
(ii) Choose the appropriate number of classes L.
(iii) Learn the model's parameters and cluster the data using the proposed EM algorithm.
(iv) Use the conditional probability tables in order to determine the main characteristics of the obtained classes.

4. Application: loans subscriber classification

In this section, the proposed BN, described in Section 3, was used to model the credit risk and cluster the loans subscribers. This model evaluates the credit default probability taking into account several explanatory variables and a classifying latent variable C. The resulting model presents a default probability distribution with several regimes or classes. These regimes correspond for example to different market conditions (depending on economic and political environment). They can also indicate various clients subclasses. The built BN model is useful as a decision aid system to deal with new loan requests.

4.1. Data

A data set provided by the Tunisian Central bank was used to illustrate the proposed model. This data set contains 9 variables describing loan contracts granted by several Tunisian banks during the period 1990–2012. Among these variables, three provide personal information about the loan subscribers (Age, Job and Gender). The remaining variables specify the contract characteristics (Type, Amount, Duration, Monthly Reimbursement MBR, Default) and whether the credit provider is a public bank (isPublic). The Default variable indicates whether the credit was duly paid (Default=0) or not (Default=1): a customer is considered as a defaulter when a payment delay for more than 90 days is observed.

The data set includes 4 continuous variables whose statistical summary and histograms are reported in Table 1 and Fig. 2, respectively.

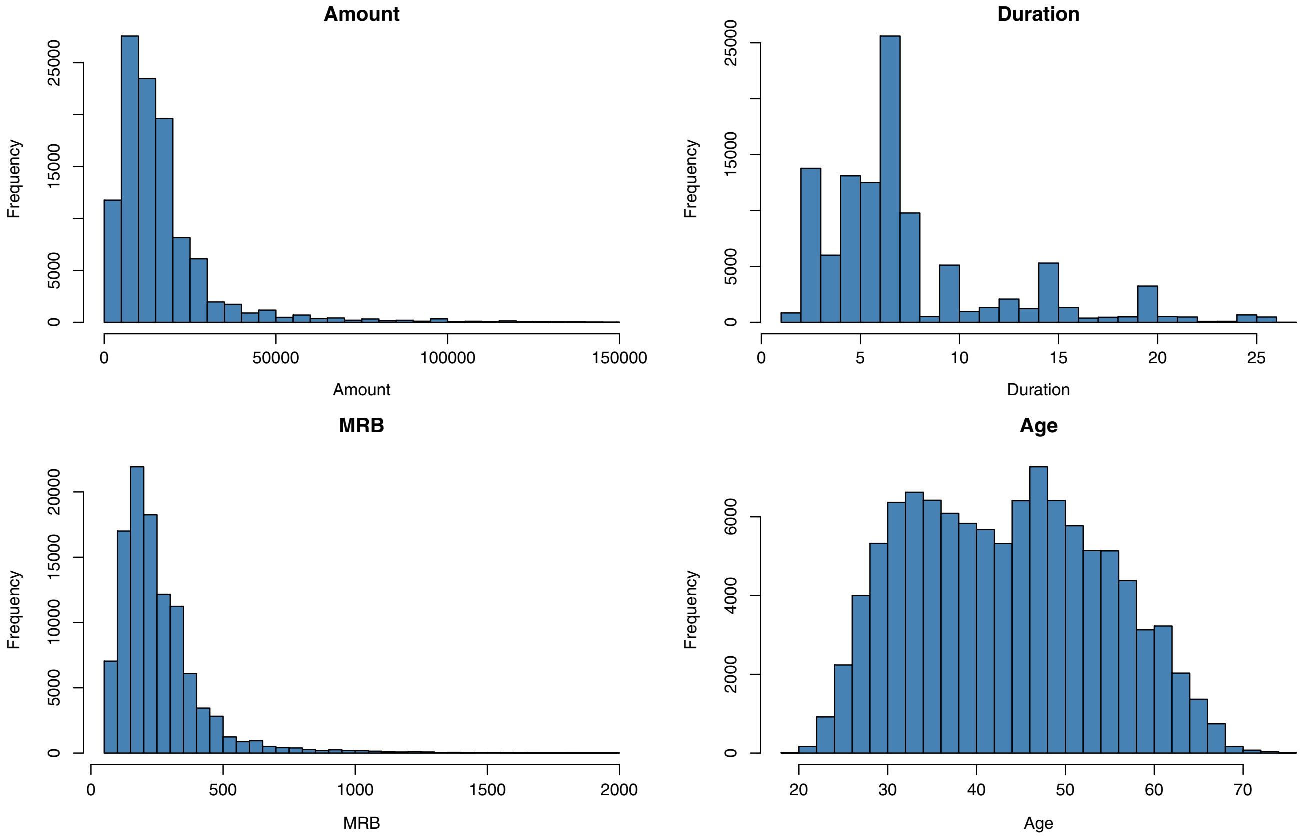

Fig. 2. Histograms of the continuous variables: Amount, Duration, MBR and Age.

Table 2
Discretized variables of the credit data set.

Variable	Brief description	Values (Frequency)
Default	Default payment	0: Non default (16%) 1: Default (84%)
isPublic	Is a state-owned bank?	0: No (87%) 1: Yes (13%)
Type	Credit type	1: Consumer Credit (23%) 2:Housing credit (77%)
Amount	Amount of credit	1:[13.6–30](38%) 2:[3–13.6](18%) 3:[30–500](44%)
MRB	Monthly repayment burden of credit	1:[0–261](26%) 2:[261,568](21%) 3:[568,2000](53%)
Duration	Credit Duration	1:[0–5.1](52%) 2:[5.1–30](24%) 3:[5.1–15.1](24%)
Gender	Men and Female	1: Women (39%) 1: Men (61%)
Age	Age in years	1: [0–35](25%) 2:[35–55](65%) 3:[55–99](10%)
Job	Job of households	1: Jobless (18%) 2:Middle Executive (26%) 3:Retired (14%) 4: Senior Executive (14%) 5:Student (14%) 6: Worker (14%)

In order to use the proposed model, these 4 variables are discretized taking the values indicated in Table 2. This table describes the 9 discrete variables included in the final data set: the discrete values and their empirical frequencies are reported in Table 2. It is worth noting that bad debtors proportion is 16%. The considered data set which contains $n = 106298$ instances was split into a training set composed of 79724 observations and a testing set of size 26574.

4.2. Model calibration

Using the above-described data set, a BN was built with a latent variable consisting of two classes. The learning procedure described in Section 3.2 was applied. First, a prior expert knowledge was combined with the Hill-climbing algorithm (Chow & Liu, 1968) to obtain the DAG reported in Fig. 3. Second, relying on the proposed EM algorithm, we estimated the model parameters. In particular, we learned the conditional probability tables within each class.

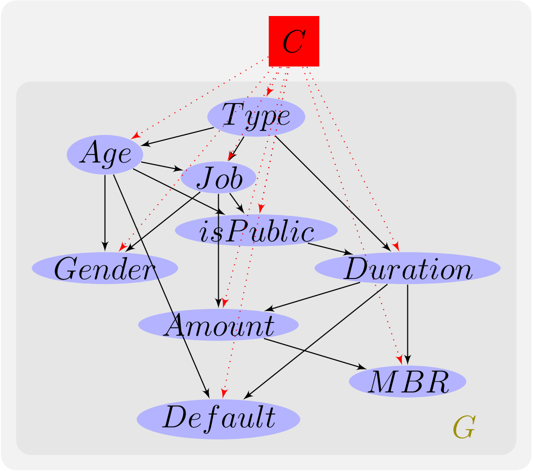

Fig. 3. The DAG structure from credit data.

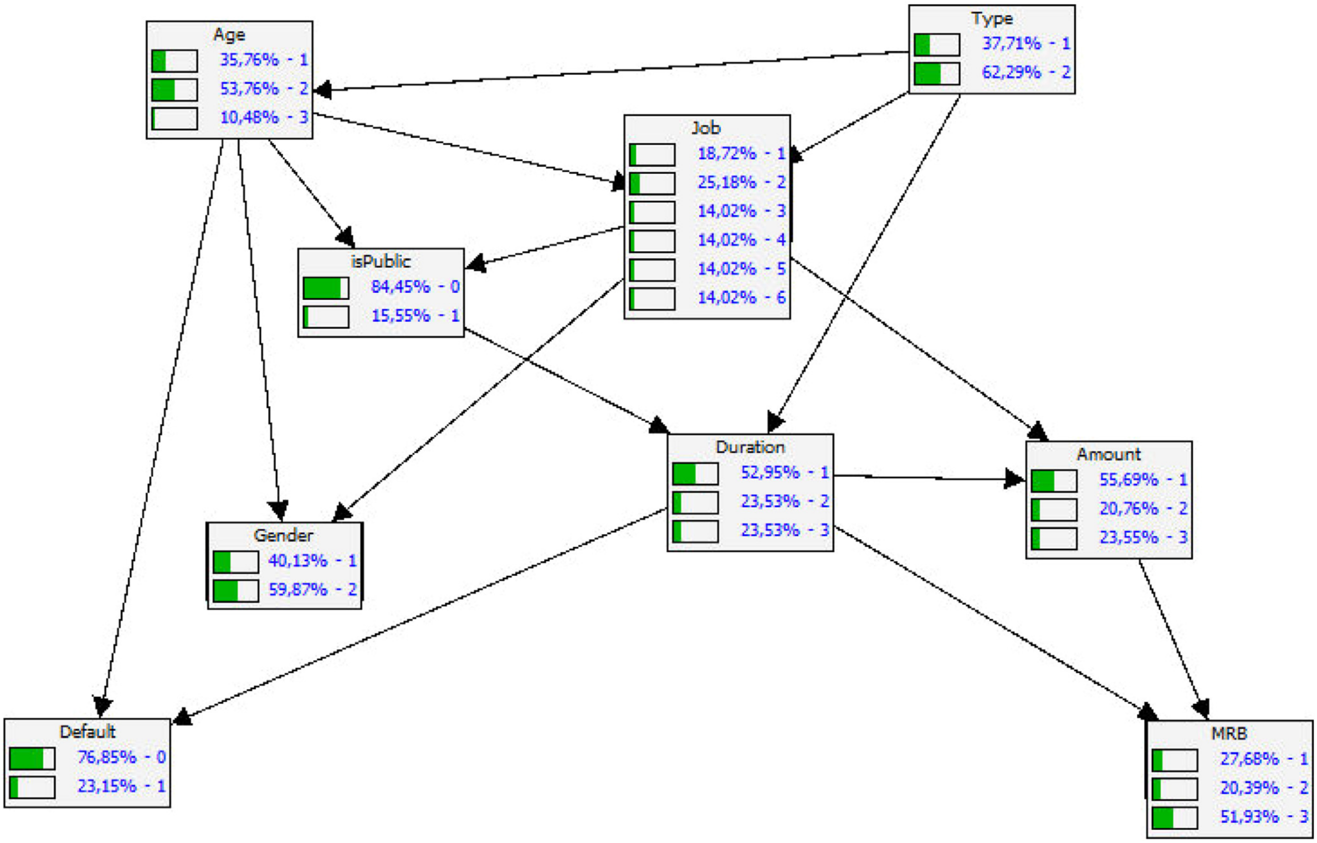

Fig. 4. Marginal distributions whithin class 1.

4.3. Inference and analysis

The DAG reported in Fig. 3 provides the relationships between the studied variables through the analysis of the active trails. Specifically, the Default variable has a direct connection with Age and Duration variables.The Type and Amount variables are both connected to the Default variable through different trails, e.g., Amount variable is connected to Default variable through Job variable, once Job is instantiated the connection via this trail is broken, and Type variable is connected to Default variable through Duration variable. Other possible trails exist.

Relying on the learned conditional probability tables, the joint probability distribution over the DAG nodes is a mixture of two BNs corresponding to two classes. In fact, the loan contracts are classified into 2 clusters having different properties and probability distributions. In order to highlight the characteristics of each class, we firstly compared the marginal distributions of each node; Then, we performed several inference tasks using various scenarios. Specifically, we used a step-by-step instantiation technique to track the default probability evolution within each class.

The marginal distributions for the two classes are displayed in Figs. 4 and 5, respectively. These figures particularly show that the probability of default in class 1 is 23.15% whereas it is 17.26% in class 2. For both classes, the marginal distributions show that the majority of borrowers are middle executives and got credit from private banks with a Duration not exceeding 5 years [0–5 years]. For class 2, the majority of borrowers received Housing credits (99.7%) whereas, in class 1, only 2/3 of borrowers obtained housing credits. Considering the Age variable, the majority of the borrowers (83.60%) are within the age interval 35–55 years old for class 2. As for class 1, only 53.76% of the borrowers belong to above-mentioned age group. It is worth noting that the highest proportion of debtors in class 2 got credits ranging between 30kTND and 500kTND, whereas the majority of households in class 1 obtained loans ranging between 13.6kTND and 30kTND.

We notice that the default probability is higher for housing credits compared to consumer credits. The most risky profile corresponds to jobless men below the age of 35 years having subscribed for a loan from a public bank.

A deeper analysis for both classes probability distributions was performed relying on step by step technique with two scenarios. This method consists in instantiating a variable at each single step and tracking the credit default probability. For scenario 1, the adopted steps are arranged by instantiating the five variables

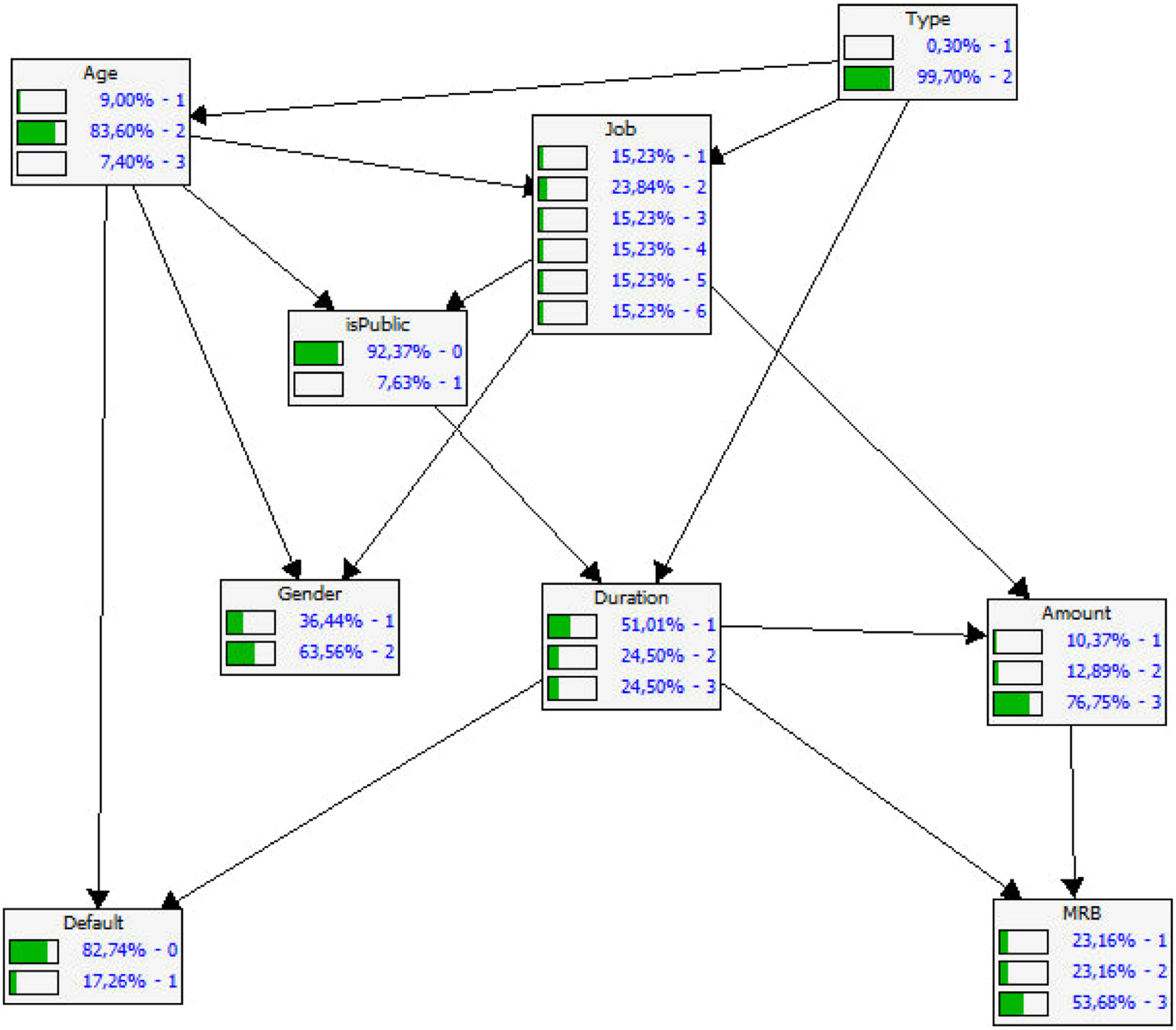

Fig. 5. Marginal distributions whithin class 2.

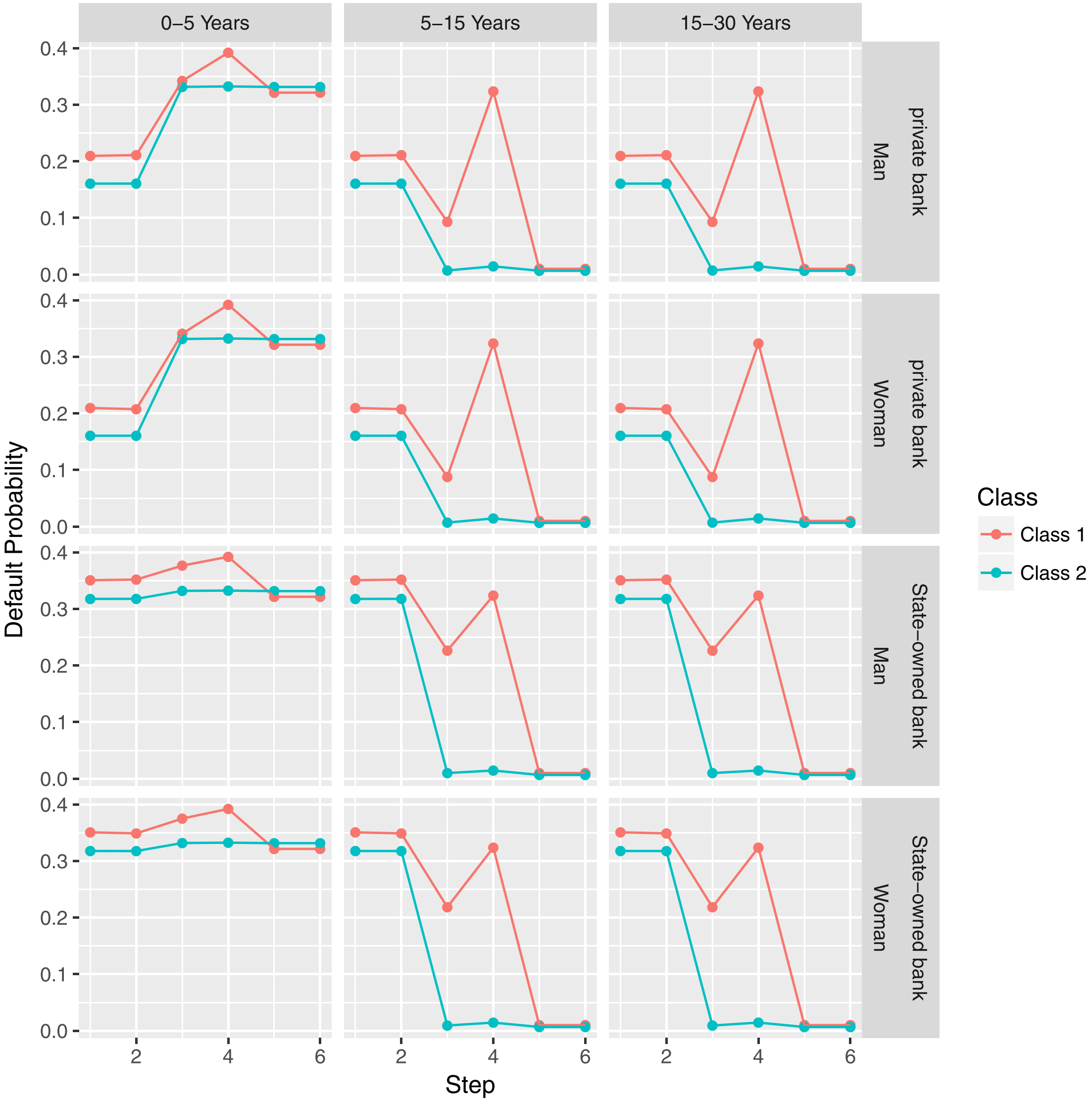

Fig. 6. Step by step inference to evaluate payment default probability in scenario 1.

as follows: Step 1, type of bank would be either Private or Public; Step 2 Gender variable is set either to Man or Woman; Step 3, the Duration variable is fixed either to [0–5years], [5–15 years] or [15–30 years]; And finally, in steps 4, 5 and 6, the age is set to 0–35, 35–55 and 55–99, respectively. For the second scenario, the first step is replaced by credit type instantiation and the remaining steps are the same.

For the two classes, the default probability evolution at each step is shown in Fig. 6. This figure proves that the default probability is higher for state-owned banks than private banks (35.07% for class 1 and 31.77% for class 2). This result seems logical since the state-owned banks are less restrictive when granting loans. The gender variable has little significant effect on payment default occurrence. However, Duration variable is closely related to the default occurrence. The probability of defaulting increases when Duration variable is instantiated in state [0–5 years] and decreases for the remaining states ([5–15 years] and [15–30 years]). Moreover, the riskiest clients are under 35 years old. For the remaining age categories, the probability of default shows a sharp downward trend.

Fig. 7 confirms that the highest probability of default is obtained from Housing credit. This probability of default is higher in class 1 (24.97%). This situation is stressed during the recent previous years as stated in the 2016 annual report of the Tunisian Central Bank. According to this report, the amount of unpaid housing loans has continued to evolve from 356 MTD in 2014 to 437 MTD in 2016, whereas the amount of unpaid consumer credits increased from 141 MTD to 168 MTD during the same period.

For both scenarios, we notice that the probability of default is higher in class 1 than class 2. Another difference between the two classes is the effect of the duration and age variables on the probability of default. In fact, this probability is significantly lower for class 2 when the duration variable is fixed to [5–15] years. It remains low when the age variable is instantiated for class 2 but clearly increases in class 1 for young loan subscribers (less than 35 years old).

The calibrated model allows handling a default probability distribution with several regimes or classes taking into account various interdependent factors. In fact, using the posterior probability defined in (6), we clustered the loans contracts. The obtained classification is summarized in Fig. 8 which displays the frequency of each class by year. This figure shows in particular, a shift in the regime occurred in the year 2010: before 2010, the majority of contracts belonged to the second class whereas after this year the

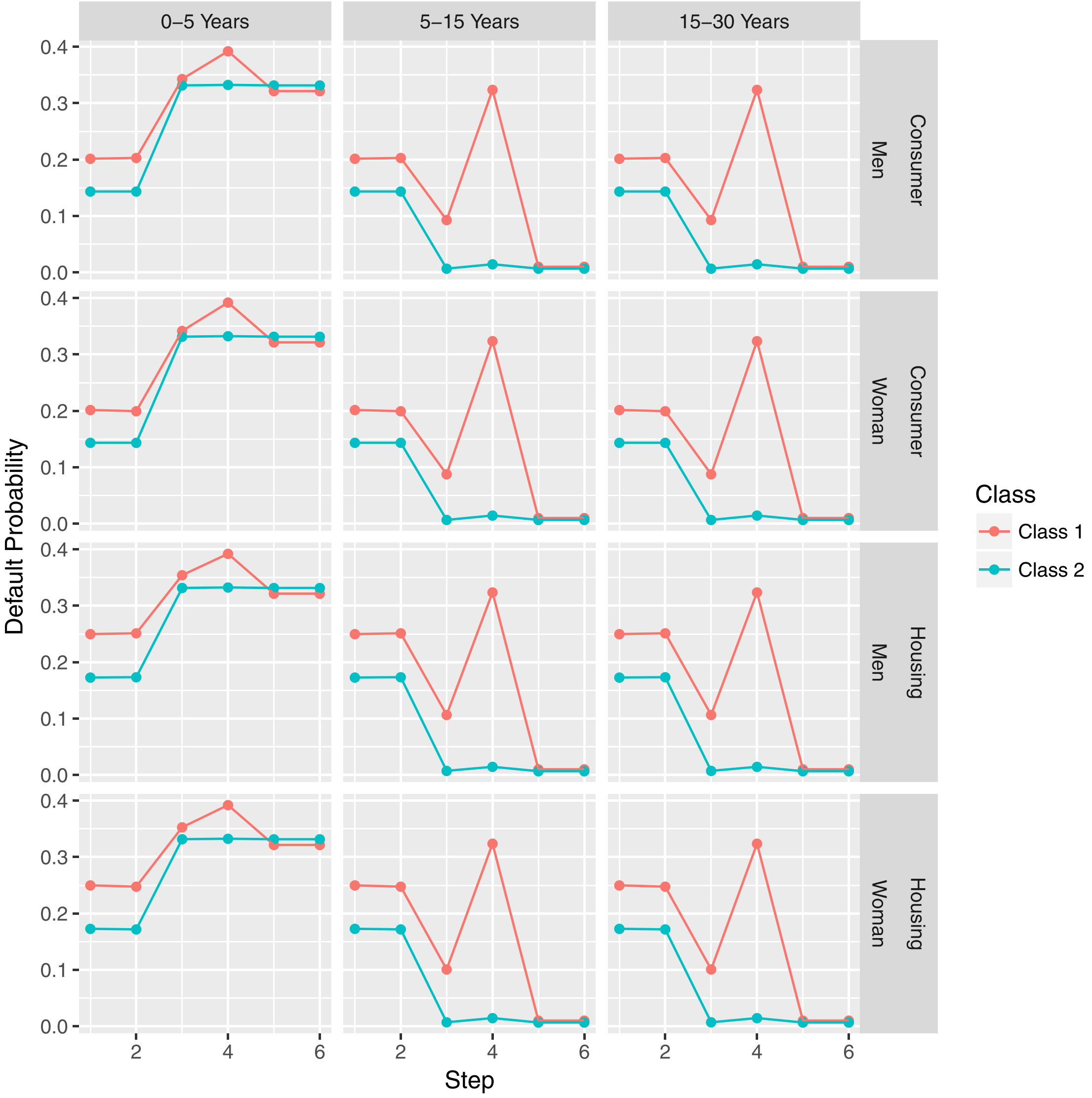

Fig. 7. Step by step inference to evaluate payment default probability in scenario 2.

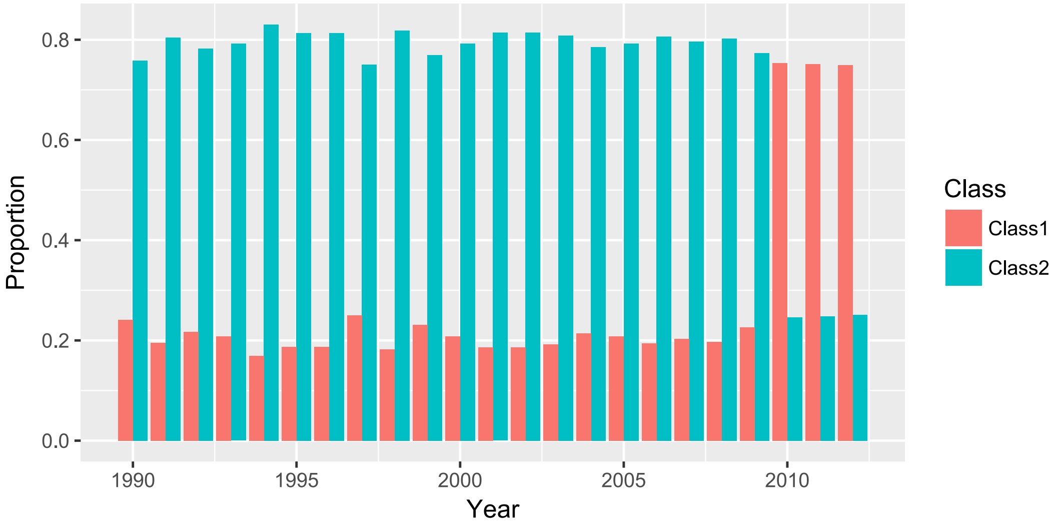

Fig. 8. Contract proportions per year for each class based on the EM algorithm clustering.

Table 3
Performance comparison of the four techniques.

Method	Accuracy	Precision	Recall	F_1
Discrete BN	93.53%	83.02%	74.87%	78.73%
Decision tree	87.39%	61.51%	65.12%	63.26%
Radial SVM	93.61%	83.25%	75.18%	79.01%
BN with latent variable	94,77%	85,31%	81,31%	83,26%

bigger part of the contracts belonged to class 1. This result is in concordance with the economic and political events happening during the period around 2010. First, there was the global crisis that had an important impact on the nonperforming loans. This period also witnessed a political instability and various social conflicts in Tunisia.

4.4. Classification and comparison

In this subsection, the calibrated BN with latent variable is used to cluster the customers through the probability of default. Its performance is compared, in terms of accuracy, precision and recall, to three other popular classifiers namely the decision tree, the discrete BN and the Radial SVM. The obtained results are reported in Table 3. The displayed evaluation metrics are defined as follows:

- Accuracy is the proportion of correctly classified items.
- Precision is the fraction of positive predictions that are correct.
- Recall is the fraction of all positive (default) instances the classifier correctly identifies as positive. It is also known as the True positive rate.
- F_1 score is calculated according to the following formula:

$$F_1 = 2.\frac{Precision.Recall}{Precision + Recall}$$

The reported results show that, for the considered data set, the discrete BN and SVM techniques have comparable performances and outperform the decision tree approach. Moreover, the addition of the latent variable in the BN improves the classification results.

5. Discussion and conclusion

In this paper, we described the BN with a latent variable and proposed a procedure for its calibration. This model was used to evaluate the payment default probability of loans subscribers. The calibrated model takes into account several risk factors including clients attributes (Age, job,...) and contract characteristics (amount, duration,...). It also describes the relationships between these factors relying on their probabilistic conditional dependencies. Finally, the loan contracts can be classified using the Bayes criterion.

The obtained BN involved two classes which have similarities and differences in terms of probability distributions. For both classes, the gender variable seems to have little significant effect on the probability of payment default.The second similarity lies in the fact that the payment default is more likely to occur when the household is under 35. Finally, housing loans have a higher probability of default compared to other loan types. However, the payment default rate is higher in class 1 and the impact of Age and Duration variables is different for the two classes. Focusing on the relationship established between the loans classification and their subscription dates showed a change in the probability distribution occurring after the year 2010. This shift is tightly connected to the economic and political environment: growth stagnation, inflation and rising interest rates.

Apart from the reached conclusions, the BN with a latent variable proved to be useful as a decision making tool in loans attribution and clients' selection. The use of computational models may be informative on a more adequate management of nonperforming loans.

The proposed network can be further developed to involve other variables than those discussed in this research study. It would also be interesting to use the proposed model in other contexts relying on other data sets.

Credit authorship contribution statement

Khalil Masmoudi: Software, Conceptualization, Methodology, Data curation, Writing - review & editing. **Lobna Abid:** Conceptualization, Visualization, Investigation, Resources, Formal analysis. **Afif Masmoudi:** Conceptualization, Methodology, Supervision, Validation.

References

Abid, L., Masmoudi, A., & Zouari-Ghorbel, S. (2016). The consumer loans payment default predictive model: An application of the logistic regression and the discriminant analysis in a tunisian commercial bank. *Journal of the Knowledge Economy*, 1–15.

Abid, L., Zaghdene, S., Masmoudi, A., & Ghorbel, S. Z. (2017). Bayesian network modeling: A case study of credit scoring analysis of consumer loans default payment. *Asian Economic and Financial Review, 7*(9), 846.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609.

Bemš, J., Starỳ, O., Macaš, M., Žegklitz, J., & Pošík, P. (2015). Innovative default prediction approach. *Expert Systems with Applications, 42*(17–18), 6277–6285.

Bijak, K., & Thomas, L. C. (2012). Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications, 39*(3), 2433–2442.

Bouchaala, L., Masmoudi, A., Gargouri, F., & Rebai, A. (2010). Improving algorithms for structure learning in Bayesian networks using a new implicit score. *Expert Systems with Applications, 37*(7), 5470–5475.

Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory, 14*(3), 462–467.

Cleofas-Sánchez, L., García, V., Marqués, A., & Sánchez, J. S. (2016). Financial distress prediction using the hybrid associative memory with translation. *Applied Soft Computing, 44*, 144–152.

Danenas, P., & Garsva, G. (2015). Selection of support vector machines based classifiers for credit risk domain. *Expert Systems with Applications, 42*(6), 3194–3204.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B* , 1–38.

Du Jardin, P., & Séverin, E. (2011). Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of a financial failure model. *Decision Support Systems, 51*(3), 701–711.

García, V., Marqués, A. I., & Sánchez, J. S. (2015). An insight into the experimental design for credit risk and corporate bankruptcy prediction systems. *Journal of Intelligent Information Systems, 44*(1), 159–189.

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed chinese companies using data mining. *European Journal of Operational Research, 241*(1), 236–247.

Gevaert, O., De Smet, F., Kirk, E., Van Calster, B., Bourne, T., Van Huffel, S., et al. (2006). Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. *Human Reproduction, 21*(7), 1824–1831.

Ghribi, A., & Masmoudi, A. (2013). A compound poisson model for learning discrete Bayesian networks. *Acta Mathematica Scientia, 33*(6), 1767–1784.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A, 160*(3), 523–541.

Harris, T. (2015). Credit scoring using the clustered support vector machine. *Expert Systems with Applications, 42*(2), 741–750.

Hassen, H. B., Masmoudi, A., & Rebai, A. (2008). Causal inference in biomolecular pathways using a Bayesian network approach and an implicit method. *Journal of Theoretical Biology, 253*(4), 717–724.

Kim, J.-S., & Jun, C.-H. (2013). Ranking evaluation of institutions based on a Bayesian network having a latent variable. *Knowledge-Based Systems, 50*, 87–99.

Korb, K. B., & Nicholson, A. E. (2010). *Bayesian artificial intelligence*. CRC press.

Kwoh, C.-K., & Gillies, D. F. (1996). Using hidden nodes in Bayesian networks. *Artificial Intelligence, 88*(1–2), 1–38.

Leong, C. K. (2016). Credit risk scoring with Bayesian network models. *Computational Economics, 47*(3), 423–446.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136.

Longstaff, F. A. (2010). The subprime credit crisis and contagion in financial markets. *Journal of Financial Economics, 97*(3), 436–450.

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: systematic review and overall comparison. *Surveys in Operations Research and Management Science*.

Luo, C., Wu, D., & Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence, 65*, 465–470.

Mselmi, N., Lahiani, A., & Hamza, T. (2017). Financial distress prediction: The case of french small and medium-sized firms. *International Review of Financial Analysis, 50*, 67–80.

Njah, H., Jamoussi, S., Mahdi, W., & Masmoudi, A. (2015). A new equilibrium criterion for learning the cardinality of latent variables. In *Tools with artificial intelligence (ICTAI), 2015 IEEE 27th international conference on* (pp. 958–965). IEEE.

Nkusu, M. M. (2011). *Nonperforming loans and macrofinancial vulnerabilities in advanced economies*. International Monetary Fund.

Pearl, J. (1988). Morgan Kaufmann series in representation and reasoning. probabilistic reasoning in intelligent systems: Networks of plausible inference.

Sanford, A., & Moosa, I. (2015). Operational risk modelling and organizational learning in structured finance operations: A bayesian network approach. *Journal of the Operational Research Society, 66*(1), 86–115.

Sousa, M. R., Gama, J., & Brandão, E. (2016). A new dynamic modeling framework for credit risk assessment. *Expert Systems with Applications, 45*, 341–351.

Spirtes, P., Glymour, C., Scheines, R., et al. (2001). Causation, prediction, and search. *MIT Press Books, 1*.

Tavana, M., Abtahi, A.-R., Di Caprio, D., & Poortarigh, M. (2018). An artificial neural network and Bayesian network model for liquidity risk assessment in banking. *Neurocomputing, 275*, 2525–2554.

Thomas, L. C., Edelman, D., & Crook, J. (2002). *Credit scoring and its applications: Siam monographs on mathematical modeling and computation*. Philadelphia: University City Science Center, SIAM.

Tomczak, J. M., & Zieba, M. (2015). Classification restricted Boltzmann machine for comprehensible credit scoring model. *Expert Systems with Applications, 42*(4), 1789–1796.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications, 42*(7), 3508–3516.