Dynamic Bayesian modeling for risk prediction in credit operations

Borchani, Hanen; Martinez, Ana Maria; Masegosa, Andres; Langseth, Helge; Nielsen, Thomas Dyhre; Salmerón, Antonio; Fernández, Antonio; Madsen, Anders Læsø; Sáez, Ramón

Published in:
The 13th Scandinavian Conference on Artificial Intelligence (SCAI'2015)

DOI (link to publication from Publisher):
10.3233/978-1-61499-589-0-17

Publication date:
2015

Document Version
Early version, also known as pre-print

Dynamic Bayesian modeling for risk prediction in credit operations

Hanen BORCHANI [a,1], Ana M. MARTÍNEZ [a,2,1], Andrés R. MASEGOSA [b,1], Helge LANGSETH [b], Thomas D. NIELSEN [a], Antonio SALMERÓN [c], Antonio FERNÁNDEZ [d], Anders L. MADSEN [a,e] and Ramón SÁEZ [d]

[a] *Department of Computer Science, Aalborg University, Denmark*
[b] *Department of Computer and Information Science, The Norwegian University of Science and Technology, Norway*
[c] *Department of Mathematics, University of Almería, Spain*
[d] *Banco de Crédito Cooperativo, Spain*
[e] *HUGIN EXPERT A/S, Aalborg, Denmark*

Abstract. In this paper we perform an exploratory analysis of a financial data set from a Spanish bank. Our goal is to do risk prediction in credit operations, and as data is collected continuously and reported on a monthly basis, this gives rise to a streaming data classification problem. Our analysis reveals some practical problems that have not previously been thoroughly analyzed in the context of streaming data analysis: the class labels are not immediately available and the relevant predictive features and entities under study (in this case the set of customers) may vary over time. In order to address these problems, we propose to use a dynamic classifier with a wrapper feature subset selection to find relevant features at different time steps. The proposed model is a special case of a more general framework that can also accommodate more expressive models containing latent variables as well as more sophisticated feature selection schemes.

Keywords. Streaming data, dynamic Bayesian modeling, variational Bayes, feature subset selection, credit operations

1. Introduction

An efficient and effective solution for risk prediction in banks can be crucial for reducing losses due to inefficient business procedures. Risk prediction solutions can be used as tools for monitoring the evolution of customers, in terms of credit operations risk, and thereby increase solvency of the banking institutions. From a machine learning perspective, this problem has traditionally been approached as a supervised classification problem [1,4,9].

In this paper we aim to explore the credit scoring problem based on a data set provided by Banco de Crédito Cooperativo (BCC), containing monthly informa-

[1] These 3 authors are considered as first authors and contributed equally to this work.
[2] Corresponding author: Ana M. MARTÍNEZ. E-mail: ana@cs.aau.dk

tion for a set of BCC clients for the period from April 2007 to March 2014. Since the customer information is received on a monthly basis, we can consider the credit scoring problem as a supervised classification problem within a streaming context. The problem does, however, also have some distinguishing characteristics that separate it from standard streaming problems: Firstly, instead of receiving a single sequence of data over time, we are faced with a stream of multiple sequences, each sequence representing a particular client. That is, at every time step t (which for the BCC data set corresponds to every month), we receive the data $\boldsymbol{D}_t$ containing information about all the clients. Secondly, in a conventional streaming data setting, the classification model would typically be trained with a subset of the observations collected up to time t, which would afterwards be used for predicting the class values of new instances received at time t. This is, however, not applicable in the BCC setting, since the class label for each sample/client corresponds to the client's defaulting behavior in the following twelve months and this information is therefore only available after a twelve month delay. Thus, the available data is a mixture of labeled and unlabeled samples. Thirdly, the domain exhibits a form of concept drift [3], where the set of feature variables relevant for classification may vary from one month to the next. Although these characteristics may at first seem as ad-hoc peculiarities of the BCC data set, they in fact apply to most credit scoring problems as well as many other domains. We will discuss this issue further in Section 5, which also serves to demonstrate the broader relevance of the above mentioned problems.

In this paper we present a first approach to address the BCC credit scoring problem[3] based on the use of a simple dynamic probabilistic graphical model [5]. A rough visual description of this model is given in Figure 1. Our preliminary approach is implemented based on the AMIDST Toolbox[4]. This toolbox provides an efficient implementation of approximate inference and learning methods for streaming data using the Bayesian networks modeling framework [5] as well as variational Bayes inference and learning procedures [6].

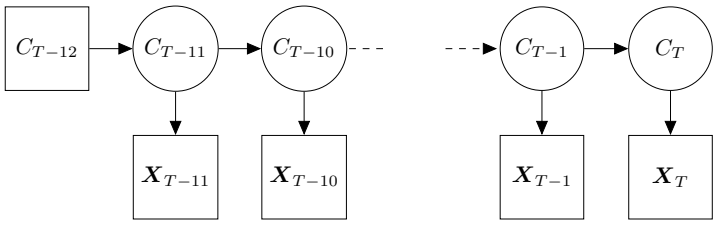

Figure 1. A dynamic probabilistic model for doing prediction in the BCC domain. At time T (assumed to be the current time) we wish to predict the defaulting status (C_T) of a particular customer based on previous socio-economical observations as well as the customer's known defaulting status $\lambda = 12$ months earlier. Note that due to the independence assumptions in the model, $\boldsymbol{X}_{T-12}$ and all observations prior to $T - 12$ become irrelevant, and are therefore not shown. Square/Round boxes indicate data which is available/non-available when predicting the defaulting status of the clients at month T.

[3]The presented models are not related to the current scoring models implemented in BCC.

[4]AMIDST is an open source toolbox available at `http://amidst.github.io/toolbox/` under the Apache Software License version 2.0.

The remainder of this paper is organized as follows. In Section 2 we describe the real-life financial data set from the Spanish bank in detail. Section 3 introduces the proposed framework for delayed-feedback stream classification using dynamic Bayesian networks. In Section 4 we examine the obtained results in terms of both predictive performance and feature selection. Finally, we present an overall discussion in Section 5 and we conclude in Section 6.

2. The financial data set

The data set, which was provided by Banco de Crédito Cooperativo (BCC), contains monthly aggregated information for a set of BCC clients for the period from April 2007 to March 2014. Only "active" clients are considered, meaning that we restrict our attention to individuals between 18 and 65 years of age, who have at least one automatic bill payment or direct debit in the bank. To make the data set as homogeneous as possible, we only retained clients residing in the Almería region (a largely agricultural area in the south-east of Spain), and excluded BCC employees, since they have special conditions. Finally, we reduced the resulting data set so that it only includes 50 000 clients each month.

We extracted 44 features, each of which encodes monthly aggregated information for each of the clients. Table 2 shows the variable ID for each of the so-called dynamic variables along with their descriptions. Two groups of variables are considered, 11 variables describing the financial status of a client (VARXX, where XX corresponds to a two digit number) and 33 socio-demographic variables identified by SOCXX. For space reasons, the semantics of the socio-demographic variables will only be presented when relevant.

Variable ID	Description	Variable ID	Description
VAR01	Total credit amount	VAR07	Unpaid amount in mortgages
VAR02	Income	VAR08	Unpaid amount in personal loans
VAR03	Expenses	VAR09	Unpaid amount in credit cards
VAR04	Account balance	VAR10	Unpaid amount in bank account deficit
VAR05	Risk balance in mortgages	VAR11	Unpaid amount in other products
VAR06	Risk balance in consumer loans	SOC01-33	Set of 33 socio-demographic variables

Table 1. Description of dynamic and socio-demographic variables.

In addition, each client has an associated class variable, which indicates if that particular client will default during the following 12 months. We note that for technical reasons due to our data preparation, some clients may have missing values for some of the variables for a given month (e.g., because a client was not active during that particular month). However, the generative nature of the proposed model class (detailed in Section 3) ensures that these missing values are naturally handled within the model and do not need to be treated separately.

The delayed class-feedback inherent in the domain forces us to be careful when setting up the evaluation procedure. Since the financial data set covers the period from April 2007 to March 2014[5], we can only use the observations from

[5]Only class labels are retained from the data corresponding to April 2007 (the first month), since there is not class label information from the previous month (i.e., March 2007).

May 2007 until March 2013 for training, while we evaluate the learned models using observations from the period from May 2008 to March 2014.

3. Risk prediction using dynamic Bayesian networks

This section presents our preliminary methodology for addressing the problem of stream classification with delayed feedback based on probabilistic graphical models [5]. The risk prediction problem can be seen as a supervised classification problem in a dynamic domain, because both the class label (i.e., the defaulter/non-defaulter status of a client) as well as the set of predictive attributes (i.e., the socio-economical status) can evolve over time. In our case, the class label of client u at time step t is modelled as a random variable $C_t^{(u)}$ and the n predictive attributes are modelled as a set of random variables $\boldsymbol{X}_t^{(u)}$. For simplicity we will suppress the client index whenever making statements related to a single user u.

At a considered current time T, the provided data stream is denoted as $\boldsymbol{D}_{1:T} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_t, \ldots, \boldsymbol{D}_T\}$, where $\boldsymbol{D}_t = \{\boldsymbol{x}_t^{(u)}\}_{u=1,\ldots,m} \bigcup \{c_t^{(u)}\}_{u=1,\ldots,m}$ and m is the number of customers in the data set. Due to the delayed labelling, only the clients in $\boldsymbol{D}_{1:T-\lambda}$ have observed class labels, while the class labels in $\boldsymbol{D}_{T-\lambda+1:T}$ are not yet observed. In our setting, data is received on a monthly basis, and the delay is $\lambda = 12$ months.

Dynamic Bayesian classifiers

Our methodology is based on the use of dynamic Bayesian networks (DBNs) [8]. This class of models is an extension of the Bayesian networks framework [5], which is a widely used class of probabilistic graphical models for reasoning under uncertainty. This family of models encodes a set of conditional independence assumptions that are exploited to efficiently perform inference tasks such as prediction, marginal belief computation, belief updating, and most probable explanation. DBNs, or more specifically 2-Time-Slices DBNs in our case, model the joint probability $p(c_{1:T}, \boldsymbol{x}_{1:T})$ by exploiting the so-called *Markov assumption* (i.e., the future is independent from the past given the present) to factorize the joint probability as

$$p(c_{1:T}, \boldsymbol{x}_{1:T}) = \prod_{t=1}^{T} p(c_t, \boldsymbol{x}_t | c_{t-1}, \boldsymbol{x}_{t-1}).$$

Here $p(c_t, \boldsymbol{x}_t | c_{t-1}, \boldsymbol{x}_{t-1})$ defines the joint probability for all the variables at time t given the variables at the previous time step $t-1$ and, by convention, for $t = 1$, $p(c_t, \boldsymbol{x}_t | c_{t-1}, \boldsymbol{x}_{t-1})$ is defined as $p(c_1, \boldsymbol{x}_1)$. Furthermore, it is common to consider the *time invariance* assumption [8], which asserts that $p(c_t, \boldsymbol{x}_t | c_{t-1}, \boldsymbol{x}_{t-1}) = p(c_\tau, \boldsymbol{x}_\tau | c_{\tau-1}, \boldsymbol{x}_{\tau-1})$ for any τ.

Similar to standard Bayesian networks, DBNs allow for further factorization of the transition probability $p(c_t, \boldsymbol{x}_t | c_{t-1}, \boldsymbol{x}_{t-1})$, and in this work we consider a dynamic extension of the well known *Naïve Bayes* classifier. This dynamic Naïve Bayes classifier assumes that only the class variables are connected across time and

that all the predictive variables at time step t are conditionally independent given the class variable at time t, i.e., $p(\boldsymbol{x}_t, c_t|\boldsymbol{x}_{t-1}, c_{t-1}) = p(c_t|c_{t-1}) \prod_{i=1}^{n} p(x_{i,t}|c_t)$, where $x_{i,t}$ denote the value of the i-th predictive attribute at time t. Thus, the joint probability further factorizes as follows:

$$p(c_{1:T}, \boldsymbol{x}_{1:T}) = \prod_{t=1}^{T} p(c_t|c_{t-1}) \prod_{i=1}^{n} p\left(x_{i,t}|c_t\right) \cdot \tag{1}$$

Learning the model

A common problem when analyzing data streams is the presence of concept drift [3]. To overcome this problem, we want to use as recent data as possible when evaluating the customer set at time $t = T$, and not let the model parameters be influenced by observations that happened a very long time ago. According to Equation (1), the probability parameters that must be learned are $p\left(x_{i,t}|c_t\right)$ and $p(c_t|c_{t-1})$. Since $t = T - \lambda$ corresponds to the last time step where C_t was observed, we choose to use the labeled data $\boldsymbol{D}_{T-\lambda}$ for learning the probabilities $p\left(x_{i,t}|c_t\right)$. We have found that this suffices since all client models share the same probabilities and the number of clients observed at each time-point is quite large.

Moreover, $p(c_t|c_{t-1})$ is learned looking at the class transitions from $\boldsymbol{D}_{T-\lambda-1}$ to $\boldsymbol{D}_{T-\lambda}$. Having learned these parameters, the model depicted in Figure 1 is rolled out using the time-invariance assumption. The actual learning was done using a Bayesian approach [2] with standard non-informative priors for multinomial and normally distributed data.

Predictions

Prediction in the model amounts to calculating the conditional probability for the class label for user u at time T given all the information collected so far, $\boldsymbol{D}_{1:T}$. Utilizing the conditional independence assumptions in the model, this simplifies to computing $p\left(c_T^{(u)}|\boldsymbol{x}_{T-\lambda+1:T}^{(u)}, c_{T-\lambda}^{(u)}\right)$, confer Figure 1. This posterior probability can be recursively computed using the dynamic Naïve Bayes' independence assumptions:

$$p\left(c_t^{(u)}|\boldsymbol{x}_{t-\lambda+1:t}^{(u)}, c_{t-\lambda}^{(u)}\right) \propto p\left(\boldsymbol{x}_t^{(u)}|c_t^{(u)}\right) \sum_{c_{t-1}^{(u)}} p\left(c_t^{(u)}|c_{t-1}^{(u)}\right) p\left(c_{t-1}^{(u)}|\boldsymbol{x}_{t-\lambda+1:t-1}^{(u)}, c_{t-\lambda}^{(u)}\right) \cdot$$

Feature subset selection

The data set consists of several different attributes describing the socio-demographic profile and the financial status of each client. The relevance of these variables may vary over time, therefore, we apply a feature subset selection technique to infer which variables are helpful in separating defaulters from non-defaulting customers. In order to capture the dynamics with which the relevant feature set changes, we propose to apply a feature subset selection method separately at each point in time. Specifically, we apply the simple wrapper feature subset selection method [7] with the Naïve Bayes model as the base classifier combined with greedy search. The area under the ROC curve (AUC) was used as

the objective function, because this metric usually performs well even if the data has class imbalance (as is the case for the financial data set).

4. Experimental results

4.1. Predictive performance analysis

Figure 2 displays the AUC obtained by the proposed dynamic Naïve Bayes (NB) classifier with and without feature selection. For each month on the x-axis, the result that is plotted corresponds to the AUC value obtained when the model is trained on that particular month, and tested on data for the following twelve months (only the class label for the twelfth month is considered when evaluating the predictive performance).

The following observations can be made: Firstly, feature selection helps to improve the value of the AUC in general. Nevertheless, we can also see some time-points where the feature selection leads to a decrease the AUC. Later, we will discuss how these drops can be explained by the feature selection procedure failing to include key variables for those months. Secondly, the overall AUC value increases over time for both models. This indicates that the problem becomes easier to solve over time. If we analyze the evolution of some of the key variables (conditioned on the defaulting status of the clients) over time, we can detect a shift for both types of clients. This shift corresponds to an overall increase in the values for defaulting clients and an overall decrease in the values for non-defaulting clients for a particular variable (or the other way around depending on the variable), which makes the clients easier to categorize. For an example, see the evolution of VAR05 in the left part of Figure 5. This could be described as a virtual concept drift, since there is a shift in the probability distribution of the attributes given the class labels, but the decision boundaries of the class are not affected.

4.2. Analysis of relevant features

As described in the previous sections, feature selection has been performed independently for each month in order to determine the relevant features during the evaluation period. Figure 3 shows the selected attributes for each month of the testing period. Note that the attributes that are never selected are not displayed on the y-axis; this set of variables consists of VAR03, which encodes "Expenses", together with 15 out of the 33 socio-demographic variables. From the results we can make the following observations:

- Sociodemographic attributes play a minor role in terms of predictive performance. Many of these variables are consistently discarded (15 out of 33) and the remaining 18 sociodemographic variables seem to be the result of noisy selection as they are not consistently selected over time. Only variables SOC01 and SOC20 appear to add some value in the first half of the period. SOC01 encodes the clients' psychologically capacity for making banking decisions. SOC20 is used to label clients with a considerably amount of money in their accounts, as the bank offers these clients special financial counseling.

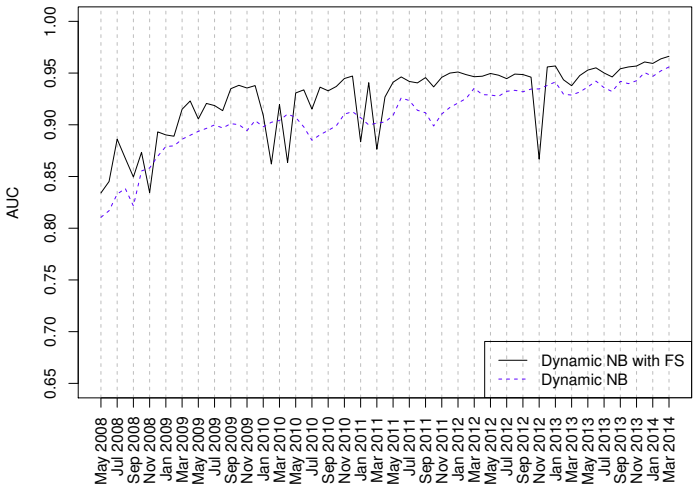

Figure 2. AUC results for the dynamic NB classifier with and without feature selection (FS). The x-axis corresponds to the training period.

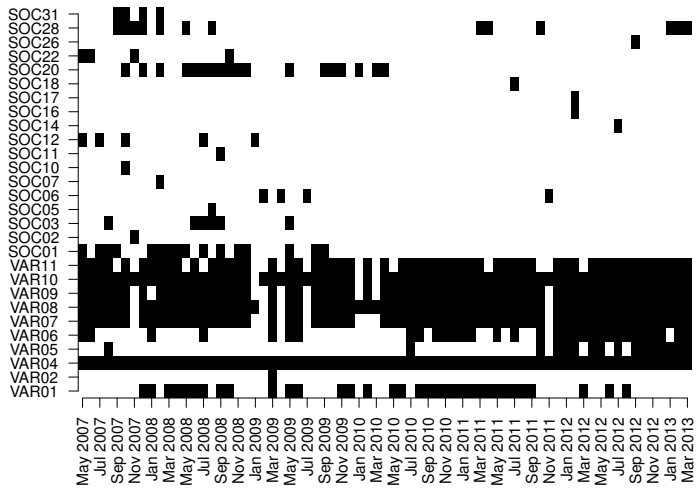

Figure 3. Selected attributes throughout the months. The x-axis corresponds to the testing period.

- The variables VAR02 and VAR03 ("Income" and "Expenses", respectively) are rarely selected or not selected at all. We believe that this is due to the presence of variable VAR04 ("Account balance"), which summarizes information from both variables. Hence, adding VAR02 and VAR03 may duplicate information that can entail a decrease in the dynamic classifier performance.

• The most frequently selected variables are VAR04, VAR10, VAR08, VAR09, VAR07, and VAR11. Figure 4 shows the evolution of VAR04 ("Account balance") and VAR08 ("Unpaid amount in personal loans") for non-defaulting and defaulting clients. Although the ranges of the attributes have been jointly normalized for all defaulting and non-defaulting clients due to confidentiality reasons, we can see how the variables consistently separate the two types of clients.

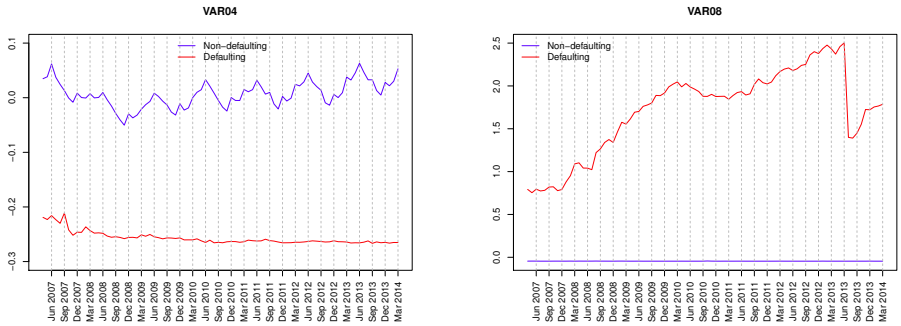

Figure 4. Time-dependent averages of variables VAR04 ("Account balance") and VAR08 ("Unpaid amount in personal loans") for non-defaulting and defaulting clients.

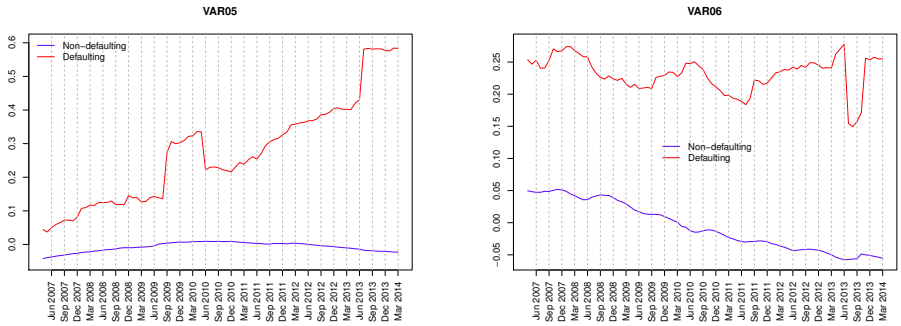

Figure 5. Time-dependent averages of variables VAR05 ("Risk balance in mortgages") and VAR06 ("Risk balance in consumer loans") for non-defaulting and defaulting clients.

• There exist other attributes that are not consistently popular, but still play an important role during certain periods. This is the case for both VAR05 and VAR06, which correspond to "Risk balance in mortgages" and "Risk balance in consumer loans", respectively. These two attributes, shown in Figure 5, become relevant in the second half of the period. It is also interesting to observe that when VAR05 is not selected, then VAR01 ("Total credit amount") is chosen as a relevant feature.

• As above-mentioned, we observe sudden drops in AUC for some specific months (i.e., Nov 2008, Feb 2010, April 2010, Jan 2011, Mar 2011, and Nov 2012). If we analyze the features selected when training these models, we note that they correspond to months where only two features are selected, namely VAR04

together with either VAR08 or VAR10. Even though these three variables are relevant, they do not suffice without the support from other variables.

5. Discussion

Compared to a traditional streaming context, our financial data set presents two peculiarities: Firstly, we have severely delayed feedback regarding the true class labels of the objects. This forces us to make long-term predictions using a sound probabilistic framework, and prevents us from using "immediate" feedback to monitor the concept drift. Secondly, the stream contains a high number of independent sequences that are observed at each time step, with the added difficulty that the set of observable sequences varies with time. These characteristics can be found in many domains, like fusing sensor data from a variety of sources and churn prediction for different customers.

This paper presents a first exploratory analysis of the financial data using a dynamic Bayesian network approach with feature selection. We have developed and used the AMIDST Toolbox to produce the results. The toolbox performs inference and learning under a Bayesian framework, and provides functionality to make improvements to the presented model: Firstly, a fully Bayesian approach can be employed for learning in a semi-supervised learning environment. Secondly, more expressive network structures, e.g., in the spirit of [10], can be used to take complex variable relationships into consideration.

The feature selection method considered is simplistic, in that the set of features selected at one time-step is chosen without taking the set of selected features from the previous time-steps into account. We have seen that this approach, although useful to get an idea of the most relevant features over time, can provide sub-optimal selections. There are features that are (spuriously) selected only once or twice, and there are months for which several feature that are selected in months in the vicinity are discarded. We are currently working towards a feature selection methodology that also put weight on the set of features selected at the previous time-steps.

6. Conclusions

In this paper we have taken a first step towards analyzing risk prediction in credit operations for the bank Banco de Crédito Cooperativo. Given the peculiarities of the domain, we have resorted to dynamic Bayesian networks as a sound and well-established class of models for reasoning over time. As a proof of concept of the applicability of this framework, we have developed a simple dynamic Naïve Bayes classifier, which learns from the observed data and then predicts each customer's defaulting status twelve months into the future using computationally efficient calculations.

We compare the results of this classifier with those obtained when performing monthly feature selection. The feature selection helps to improve the results, and also gives insight into which attributes are most relevant as a function of time.

We observe that while some attributes have consistently high or consistently low effect on the predictive capability throughout the observation period, others are relevant only for a part of that period. We see this as a clear indication of the importance of properly handling concept drift, and attribute the effect to the volatile economic climate resulting from the economic crisis in Spain during the observation period.

Risk prediction is a simple instantiation of a more general set of problems where one wants to deal with dynamic classification of streaming data. We have developed the AMIDST Toolbox (`http://amidst.github.io/toolbox/`) to support expressive and powerful dynamic Bayesian networks equipped with Bayesian learning, with the aim of providing a general purpose system for handling this type of problems.

Acknowledgments

This work was performed as part of the AMIDST project. AMIDST has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 619209. The data set has been provided by Banco de Crédito Cooperativo.

References

[1] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, 2003.

[2] José M. Bernardo and Adrian F.M. Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[3] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014.

[4] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33(4):847 – 856, 2007.

[5] Finn V. Jensen and Thomas D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer-Verlag, Berlin, Germany, 2007.

[6] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Laurence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[7] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[8] Kevin P. Murphy. *Dynamic Bayesian networks: Representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.

[9] Lyn C. Thomas. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2):149 – 172, 2000.

[10] Shengtong Zhong, Helge Langseth, and Thomas D. Nielsen. A classification-based approach to monitoring the safety of dynamic systems. *Reliability Engineering and System Safety*, 121:61–71, 2014.