Expert Systems with Applications 40 (2013) 665–676

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Personal bankruptcy prediction by mining credit card data

Tengke Xiong [a,*], Shengrui Wang [a], André Mayers [a], Ernest Monga [b]

[a] Department of Computer Science, University of Sherbrook, Sherbrooke, QC, Canada J1K 2R1
[b] Department of Mathematics, University of Sherbrook, Sherbrooke, QC, Canada J1K 2R1

ARTICLE INFO

Keywords:
Data mining
Personal bankruptcy prediction
Sequence mining

ABSTRACT

A personal bankruptcy prediction system running on credit card data is proposed. Personal bankruptcy, which usually results in significant losses to creditors, is a rapidly increasing yet little understood phenomenon. The most commonly used methods in personal bankruptcy prediction are credit scoring models. Some data mining models have also been investigated in this domain. Neither the scoring models nor the existing data mining methods adequately take sequence information in credit card data into account. In our system, sequence patterns, obtained by developing sequence mining techniques and applying them to credit card data from one major Canadian bank, are employed as main predictors. The mined sequence patterns, which we refer to as bankruptcy features, are represented in low-dimensional vector space. From the new feature space, which can be extended with some existing prediction-capable features (e.g., credit score), a support vector machine (SVM) classifier is built to combine these mined and already existing features. Our system is readily comprehensible and demonstrates promising prediction performance.

1. Introduction

Personal bankruptcy prediction has been of increasing concern both in the industry community and in academic investigation, as bankruptcy results in significant losses to creditors. In credit card portfolio management, bankruptcy prediction is a key measure to prevent the accelerating losses resulting from personal bankruptcy. There were 90,610 personal bankruptcy cases (excluding proposals) in Canada in 2008, more than four times the figure for 1988.[1] The total personal bankruptcy debt in 2008 was $7.414 billion, whereas it was less than $1 billion in 1987. It is also reported in *Industry Canada* that 87.4% of personal bankruptcy cases involved credit card debt, which is the most frequently reported type of debt. To address this problem, besides carefully evaluating the creditworthiness of credit card applicants at the very beginning, credit card issuers must make a greater effort to identify potential bad accounts whose owners will go bankrupt over the life of the credit, because many clients whose creditworthiness was good when they applied for credit ultimately went bankrupt. From the creditor's standpoint, the earlier bad accounts are identified, the lower the losses entailed, which can be seen in Fig. 1. The figure is computed from our project data (Master credit card data from one major Canadian bank), which shows the relationship between the debt of bankrupt accounts and the period before going bankrupt. We can see that the debt of bankrupt accounts increases linearly when approaching bankruptcy. However, early identification represents a greater challenge, which will be illustrated in the experimental section of our paper.

Personal bankruptcy is a phenomenon that is difficult to understand. There is little supporting theory related to this phenomenon. Compared to business clients, personal clients have much larger volumes of data, so it is difficult to resort to experienced and informed risk evaluation for predicting personal bankruptcy. The most commonly used methods in personal bankruptcy prediction are credit scoring models (He, Shi, & Xu, 2004; Jolson, 2007; Komorad, 2002; Mays, 2005; Thomas, Edelman, & Crook, 2002), especially the generic scoring models developed by credit bureaus (Equifax, TransUnion and Fair Isaac). These models are based on empirical knowledge, as they are developed by analyzing statistics and picking out characteristics that are believed to relate to creditworthiness.[2] Credit scoring systems, especially the generic ones, are run on multiple data sources from many creditors. In making a final decision, a customized system developed by the individual creditor is usually combined with generic scores purchased from a credit bureau.

Personal bankruptcy prediction involves discovering bankruptcy features that can distinguish bad accounts from good ones. This can also be treated as a binary classification problem with two class labels, 'bad account' and 'good account'. A big challenge in

* Corresponding author. Tel.: +1 819 5786798.
E-mail addresses: tengke.xiong@usherbrooke.ca (T. Xiong), shengrui.wang@usherbrooke.ca (S. Wang), andre.mayers@usherbrooke.ca (A. Mayers), ernest.monga@usherbrooke.ca (E. Monga).

[1] Industry Canada. Available: http://www.ic.gc.ca.

[2] Credit Scoring. Available: http://epic.org/privacy/creditscoring/.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.07.072

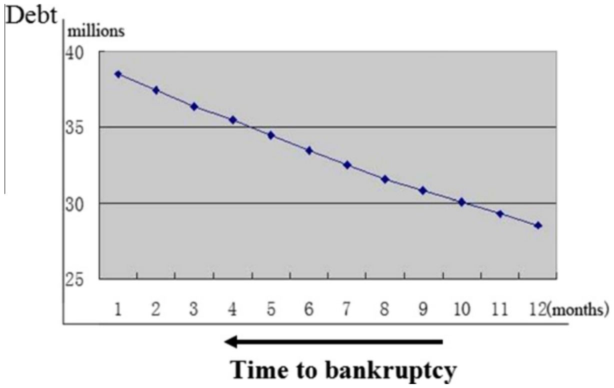

Fig. 1. The relationship between debt and period before bankruptcy.

building such a classification model for credit card data is that the two classes are highly unbalanced, with typically no more than 5 in 1000 credit cardholders going bankrupt. Another challenge is that the credit card data are highly multi-dimensional in that they contain numerous monthly aggregation records, which consist of various data types, including numeric attributes, discrete attributes, date attributes, and even the attributes of sequence and time series data; as well as transaction records, which are time series data. For example, in our project on a major Canadian bank, there are more than 400 attributes in the monthly aggregation database and more than 50 attributes in the transaction database. Because of the high-dimensional character and complexity of the data, there is no existing data mining model that can handle all the data at one stroke. While it is well known that some data mining models, such as decision trees, neural networks and support vector machines, are applied to this domain, there is very little published literature concerning personal bankruptcy prediction in the data mining field. This is mainly due to two things. The first is commercial-in-confidence: corporations do not like to reveal their techniques to others. Second, large and interesting sources of data are not made available to the academic community. Even several of the papers that have been published (Donato, Schryver, & Hinkel, 1999; He et al., 2004; Peng, Kou, Shi, & Chen, 2005) do not present a practical prediction system, but just give some classification models where the prediction is simply treated as classification, without considering the prediction period, which is a critical issue for practical application. These existing models use a fixed-dimension vector consisting of "those attributes that were found to be most correlated to bankrupt behavior" (Donato et al., 1999) to represent a client, and the prediction models are trained in the vector space.

Comprehensibility is another issue that creates a gap between the academic and industry communities in the field of personal bankruptcy prediction. Both accuracy and comprehensibility are required for data mining techniques (Tsukimoto, 2005). The patterns discovered by the data mining models have little value in practical application if they have low comprehensibility. The creditor would not use a complex prediction model with low comprehensibility, even if its prediction results are advantageous to some extent. In summary, some data mining models concerning personal bankruptcy prediction have several drawbacks:

1. The attributes fed into the models need to be pre-selected, which is difficult even with knowledge and experience of the domain.
2. The format of data input into the models is vectorial, which means that original sequence data need to be aggregated into one value for each attribute, and this kind of aggregation leads to significant loss of useful information, such as sequence and sequential patterns.
3. The models can be difficult to interpret to creditors, especially if they are not experts in the data mining field.

Thus, employing these models directly as classifiers is difficult in industry applications. Furthermore, although these models have demonstrated some capacity to outperform basic classifiers such as decision trees, neural networks, etc, their predictive performance is not very satisfactory. For instance, in the hybrid model of Donato (Donato et al., 1999), 63.7% of bankrupt (bad) accounts are identified, with 17.3% of non-bankrupt (good) accounts mistaken for bad accounts; while in He's MCNP model (He et al., 2004), 82.8% of bad accounts are identified, but the percentage of good accounts that are misidentified is as high as 49.0%. These two systems lack appeal, due either to a low true positive ratio or a high false positive ratio.

In our investigation, we aim to design a prediction system running on a credit card data base, which is extensible, i.e., able to integrate existing prediction-capable features, either from data mining or domain expertise (e.g., credit scores); it is also readily comprehensible and can be used in industrial applications. The original purpose of our investigation was to complement existing prediction models, especially the credit scoring models, by identifying the bad accounts they tended to miss. To the best of our knowledge, neither the scoring models nor the existing data mining methods adequately take sequence information in credit card data into account. Sequence data are very common in data sources such as credit card data, which takes the form of ordered monthly aggregation or transaction records with a time stamp. Therefore, the temporal and sequence patterns are indicative information for identifying potential bankrupt accounts.

Instead of aggregating the original sequential and time series data into vectorical data, we explore the extent to which the use of sequence mining can help to solve the personal bankruptcy problem. Sequences corresponding to each attribute are first extracted, based on the attribute type. In general, for categorical attributes, the sequences can be constructed directly by extracting categorical values from the original database. For a numeric attribute, discretization is employed to transform the attribute to a categorical one before constructing sequences. The transformation obviously depends on the attribute. It is not possible to detail all the transformations. The process of these transformations can be considered as coding human knowledge about the information provided by each attribute. In our application, each original attribute is transformed either to a binary categorical attribute (meaning "good behavior" and "bad behavior") or a multi-valued ordinal attribute (meaning "good behavior" and "graded bad behaviors"). Consequently, we obtain two types of sequences, i.e., binary sequences and ordinal sequences. The binary sequences are the same as ordinary categorical sequences. However, the ordinal sequences differ from ordinary categorical sequences (such as protein sequences and text documents) because of the ordinal relationship between symbols in the sequences. In ordinal sequences, either two symbols are comparable. A description of how the sequences are built is given in Section 2.

We resort to a clustering technique to discover useful sequence patterns which can be used to distinguish bad accounts from good ones. To make our prediction system comprehensible, we do not build the classifier, which is commonly used to make predictions, directly on the sequences. Instead, we exploit sequence clustering to discover the sequence patterns, which are easy to understand, and use them as predictors in the final prediction system. Additionally, building the sequence classifier requires definition of the similarity between the sequences, which is difficult in our application, as our sequences are short and contain noise; furthermore, the sequence patterns vary for different variables. On the other hand, clustering is used to find useful variables on which