# An approach for credit scoring based on sequential pattern mining

Yiting Guo[a], Dian Chen*[b]
[a]School of Economics and Management, Tsinghua University, Beijing, China; [b]Department of Psychology, Tsinghua University, Beijing, China

## ABSTRACT

Credit scoring is a significant research domain for both finance and computer science researchers. Customers' loan application behavior is highly related to their future default performance, which requires more studies. This study applies a modified approach of sequential-pattern-mining-based classification to mine important and discriminant behavioral patterns of customers. And then, we use the patterns to predict customers' likelihood to default on a future loan. Our approach specifically addresses the problem of data imbalance. Evaluation based on real business data shows that our approach outperforms a series of time series classification methods, including deep learning models. In addition, such pattern-based classification approaches have the merit of explainability in features. Specific patterns of customer loan applications are found to be related to their future default behavior. Therefore, our method enhances the business understanding, provides managerial insights, and thus is more likely to be accepted by the industry.

**Keywords:** Default prediction, sequential pattern mining, time series classification, loan application, imbalance data

## 1. INTRODUCTION

Credit scoring is an important issue for financial institutions because they need to decide whether to lend money to applicants with various credit risks. Credit scoring has become a typical setting of classification. It has attracted much attention from researchers in computer science this decade[1]. Machine learning methods, e.g., logistic regression, decision tree, Naive Bayes, support vector machine, have been applied to predict customers' default probabilities[2]. These methods usually take cross-sectional data as input. Nowadays, with the application of information systems and mobile platforms, more and more time series data are available, tracking customers' long-term behavior. However, prior studies on credit scoring seldom employ time series data and do not consider the dynamic features of customer behavior.

This study aims to predict customers' default likelihood by mining their loan application patterns. Statistics about customers' loan application behavior have been used in prior studies for default prediction (e.g., reference[3]). For example, ten percent of the FICO score comes from the number of the recent credit inquiry, which is a proxy for customers' loan application behavior.

Even though there are many time series analysis methods, most of them have inherent drawbacks. On the one hand, the traditional methods in finance, such as the autoregressive-moving average model, have limitations with respect to prediction and scalability. And these models usually have strict assumptions on data distribution. On the other hand, while deep learning models have advantages in prediction, they are hard to explain and require big data. Therefore, we propose a new approach based on sequential pattern mining (SPM).

Our approach firstly mines discriminant and frequent patterns from data. With these patterns, it is easy for us to understand customers' behavior. Then an XGBoost is used to predict customers' future default probability based on the sequential patterns. Using real-world data from a financial service company, we train and evaluate our method. The results demonstrate that our approach has higher performance than comparable methods in five types, including deep learning models.

Our study makes three-fold contributions to the literature. First, we extend the classification methods based on sequential pattern mining to fit continuous variables and imbalanced data sets. Specifically, we introduce a new metric to select sequential patterns. This method could be applied to other domains for time series classification. Second, we offer a way to meet the need of financial institutions for explainability. Compared to black-box methods (such as deep learning models), our approach is more understandable. Third, this study enhances our knowledge of how specific patterns of customer loan

*chen-d19@mails.tsinghua.edu.cn

International Conference on Computer Application and Information Security (ICCAIS 2021), edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 122600B · © The Authors.
 · doi: 10.1117/12.2637635

applications affect their future default intention.

## 2. METHOD

### 2.1 Data

We collaborate with one bank and one financial service institution in China. The bank provides the credit card business data, which consists of around 13000 customers' default information. In this dataset, approximately 20% of the customers default on a loan. The financial service institution provides the inquiry data of these customers. This institution works in the role of a credit bureau by aggregating different lenders' information. Inquiry is an action that a lender acquires a customer's information from the credit bureau, which presents that the customer applies for a loan from the lender. The objective of the current study is to predict the customers' default probability using historical loan application behavior (as shown in Figure 1).

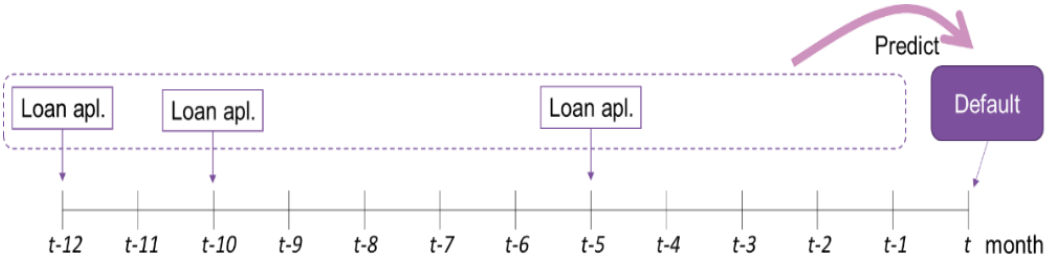


Figure 1. Presentation of the time series classification task of this paper.

The inquiry data of this study differ from former studies in that it is a multivariable time series. To be specific, in the time dimension, we have the loan application history of the customers in the recent 12 months (as shown in Table 1). In another dimension, we know the distribution of different types of institutions. The institutions include two types of banks (traditional banks and online banks) and six types of non-bank institutions (i.e., institutions for peer-to-peer (P2P) lending, microloan, cash loan installment, compensation, and others). And two types of data about loan applications are obtained: (1) the number of customers' loan applications (as presented in Table 1); (2) (unique) number of institutions applied by the customers. These two are related but distinguished. For example, one customer may apply for loans five times from three different banks in one month. These data presentations will be further processed in our application of PSM-based classification methods.

### 2.2 Classification based on sequential pattern mining

#### (1) Presentation of sequence data

Our original data is multivariable time series with continuous values (presented by Table 1). However, SPM methods usually process item-based sequences. A sequence *s* is traditionally defined as an ordered list of items $< e_1, e_2, e_3, \dots, e_l >$, where each element consists of at least one item, and *l* is the length of the sequence. A sequence $\mathrm{q} = < q_1, q_2, q_3, \dots, q_m >$ is defined as a subsequence of s if there exist integers $1 \leq t_1 < t_2 < \cdots < t_m \leq l$ such that $q_1 = e_{t_1}, q_2 = e_{t_2}, \dots, q_m = e_{t_m}$. In other words, a sequence *b* is a subsequence of *s* if each element in *b* is a subset of the respective mapped element of *s*. An example of s is $< a\ b\ c\ g\ h\ a\ d\ b\ d >$. And one subsequence of *s* is $< a\ c\ g\ d >$. The goal of SPM is to mine subsequences to appear frequently in the data.

Table 1. Sample of a customer' loan application number each month in different institutions.

| Application number | t-12 | t-11 | t-10 | t-9 | t-8 | t-7 | t-6 | t-5 | t-4 | t-3 | t-2 | t-1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bank | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 2 | 2 | 0 |
| Traditional bank | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 2 | 2 | 0 |
| Online Bank | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Non-bank | 2 | 3 | 5 | 4 | 2 | 7 | 4 | 1 | 0 | 1 | 2 | 2 |
| P2P | 2 | 3 | 5 | 3 | 0 | 7 | 2 | 1 | 0 | 1 | 2 | 2 |
| Microloan | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Cash loan installment | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Customer loan | 0 | 0 | 0 | 1 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| Compensation | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Others | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

In the current study, we firstly transfer the loan application sequences into the presentation as the sequences of institutions applied by the customer. For example, the sample in Table 1 is presented by $s1 = \{n1, n1, n1, (b1, n1, n4), (b1, n4), n1, (n1, n4), (b1, n1), (b1, n1), n1\}$, where b1, b2, n1, n2, n3, n4, n5, and n6 denote traditional banks, online banks, institutions of P2P lending, microloan, cash loan installment, customer loan, compensation, and others.

Because we have continuous value (e.g., frequency) of customers' loan application behavior, item-based representation without frequency would lose important information. Therefore, this paper adopts an approach to utilize frequency information by discretizing the continuous values and constructing discretized item-frequency-based sequences. Each variable is discretized using chi-square[4]. A customer applying for loans $p$ times in one month is denoted as $\mathrm{b1\_p'}$, where $p'$ means that $p$ is in the $p'$ th bin. For example, the sample in Table 1 is presented by $s1 = \{n1\_2, n1\_2, n1\_3, (b1\_1, n1\_2, n4\_1), (b1\_2, n4\_2), n1\_3, (n1\_2, n4\_2), (b1_2, n1_1), (b1_2, n1_2), n1\_2\}$.

**(2) Supervised sequential pattern mining**

After obtaining the representations of customers' behavior sequences, SPM could be employed to mine customer behavior patterns. To obtain sequential patterns, we implement a typical SPM algorithm, PrefixSpam[5], which is especially efficient for large-scale datasets. The practice of PrefixSpam uses min_support, where support means the pattern's frequency divides the number of the whole sample as the threshold to select patterns.

Since the target of this paper is to do classification based on sequential patterns, we used supervised SPM. Most SPM applications use an unsupervised approach where the patterns found are important in that they frequently appear in the dataset. However, these frequent patterns may not necessarily be useful for the ultimate classification target. The overwhelming majority of patterns may appear equally frequent in all classes and thus useless for classification or interpretation[6]. Therefore, we need to apply supervised SPM to obtain discriminative sequential patterns. Such practice has been used in several prior studies (e.g., reference[7]). To be specific, we use the label of each sequence to calculate the chi-square value, which is one of the most common metrics for feature selection in machine learning tasks[8].

Default prediction is a typical context where the data is imbalanced. In other words, the borrowers that default is far fewer than those who pay in time. Therefore, in the SPM process, we need to consider the influence of imbalance. To solve this problem, we introduce a third metric, min_support of the small class (denoted as min_support_minor), to select patterns besides min_support and chi-square. When the target is highly imbalanced, a pattern meeting the min_support threshold may occur merely in the large class. Thus, it is meaningful to mine the patterns that frequently appear in the minor class.

**(3) Classification**

After obtaining the discriminant sequential patterns, any classifier could be applied to learn the features. In this paper, we employ XGBoost to do the task. XGBoost is one of the most popular machine learning algorithms, which outperforms

other algorithms in many data mining competitions[9]. Moreover, XGBoost would output feature importance, which is crucial for understanding the model.

## 3. RESULTS AND EVALUATION

### 3.1 Patterns

Based on the sequences and selection metrics described above, we conducted supervised SPM and obtained customers' behavioral patterns. We chose the min_support to be 0.1 and the min_support_minor to be 0.4. In this way, we could capture those features that appear frequently in the default class while did not meet the min_support. Moreover, chi-square was used to select the top 10% discriminant features. For example, the pattern {n4, (n3, n6)} (i.e., {*CustomerLoan*, (*CashLoan*, *Others*)}) had a chi-square value of 786. Therefore, if a customer applies for a customer loan in a month, then a cash loan and another type of loan in a later month, it is a signal that he/she is likely to default.

After fitting the XGBoost model, we found the key patterns according to feature importance. The top 6 features were as follows: {NonBankInstitution#_1, NonBankInstitution#_1, NonBankInstitution#_1}, {Others}, {P2P}, {CashLoan}, {NonBankInstitution#_3or4}, {NonBank_morethan4times}.

### 3.2 Prediction performance

To evaluate the performance of our method, we compared the classification result of our method with that of five types of time series classification methods. The benchmarks were as follows: (1) the SPM-based method proposed[7], which does not deal with imbalance; (2) the gold standard for time series classification: dynamic time warping-k-nearest neighbors (DTW-KNN)[10]; (3) proximity forest, which uses several distance metrics[11]; (4) Shapelets, which mines continuous subsequences[12]; (5) hidden Markov model (HMM), which is the typical generative model to fit time series; (6) deep learning models, i.e., long short-term memory (LSTM), fully convolutional network (FCN), and Multivariate LSTM-FCN[13]. We used AUC as the evaluation metric since it was widely used in default prediction tasks. Table 2 demonstrated that our method outperforms all the benchmarks.

Table 2. Classification results.

| Type | Model | AUC |
|---|---|---|
| SPM-based | Our method | **77.22%** |
| | Method[7] | 75.48% |
| Distance-based | DTW-KNN | 53.23% |
| | Promixity Forest | 56.34% |
| Feature-based | Shapelets | 74.31% |
| Model-based | HMM | 65.50% |
| Deep-learning | LSTM | 28.00% |
| | FCN | 75.50% |
| | Multivariate LSTM-FCN | 70.00% |

## 4. CONCLUSION

Leveraging an approach based on sequential patterns analysis (SPM), this paper found that customers' default likelihood was related to their loan application behavior in particular financial institutions in a certain way (related to number and order). We modified the SPM-based classification methods to address data imbalance. Using real-world data of customers' default and loan application behavior, we demonstrated that our method outperforms a series of benchmarks, such as Shapelets, HMM, and LSTM.

Our approach has several strengths: (1) it is especially useful for discrete sequences or event analysis (with timestamps/order); (2) compared to traditional time series analysis methods in finance, our method is scalable and effective in prediction; (3) compared to most machine learning and deep learning models for time series classification, our method is

easy to understand and explain; (4) compared to methods using raw time series, our approach could avoid overfitting; (5) unlike most time series classification method, our approach can deal with high-dimensional and varying length series.

In future research, we will evaluate this approach in more cases and other domains. And large-scale datasets could be used to verify the ability of this method further. Moreover, since our approach is flexible, the sequential patterns obtained by SPM could be integrated with other formats of data for prediction. Overall, this paper inspires future research on time series classification and credit scoring.

## REFERENCES


[1] Louzada, F., Ara, A. and Fernandes, G. B., "Classification methods applied to credit scoring: Systematic review and overall comparison," Surv. Oper. Res. Manag. Sci., 21, 117-34 (2016).

[2] Lessmann, S., Baesens, B., Seow, H. V. and Thomas, L. C., "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," Eur. J. Oper. Res., 247, 124-36 (2015).

[3] Malekipirbazari, M. and Aksakalli, V., "Risk assessment in social lending via random forests," Expert Syst. Appl., 42, 4621-31 (2015).

[4] Lavangnananda, K. and Chattanachot, S., "Study of discretization methods in classification," 2017 9th Inter. Conf. on Knowledge and Smart Technology (KST), 50-5 (2017).

[5] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U. and Hsu, M. C., "Mining sequential patterns by pattern-growth: The prefixspan approach," IEEE Trans. Knowl. Data Eng., 16, 1424-40 (2004).

[6] Fradkin, D. and Mörchen, F., "Mining sequential patterns for classification," Knowl. Inf. Syst., 45, 731-49 (2015).

[7] Nowozin, S., Bakir, G. and Tsuda, K., "Discriminative subsequence mining for action classification," Proc. IEEE Int. Conf. Comput. Vis., (2007).

[8] Liu, H. and Setiono, R., "Chi2: Feature selection and discretization of numeric attributes," Proc. of 7th IEEE Inter. Conf. on Tools with Artificial Intelligence, 388-91 (1995).

[9] Chen, T. and Guestrin, C., "Xgboost: A scalable tree boosting system," Proc. of the 22nd ACM SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining, 785-94 (2016).

[10] Lines, J. and Bagnall, A., "Time series classification with ensembles of elastic distance measures," Data Min. Knowl. Discov., 29, 565-92 (2015).

[11] Lucas, B., Shifaz, A., Pelletier, C., O'Neill, L., Zaidi, N., Goethals, B., Petitjean, F. and Webb, G. I., "Proximity forest: an effective and scalable distance-based classifier for time series," Data Min. Knowl. Discov., 33, 607-35 (2019).

[12] Lines, J., Davis, L. M., Hills, J. and Bagnall, A., "A shapelet transform for time series classification," Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining," 289-97 (2012).

[13] Karim, F., Majumdar, S., Darabi, H. and Harford, S., "Multivariate LSTM-FCNs for time series classification," Neural Networks, 116, 237-45 (2019).