



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## Default feature selection in credit risk modeling

**Citation for published version:**

Chai, N, Shi, B, Meng, B & Dong, Y 2023, 'Default feature selection in credit risk modeling: Evidence from Chinese small enterprises', *SAGE Open*, vol. 13, no. 2. <https://doi.org/10.1177/21582440231165224>

**Digital Object Identifier (DOI):**

[10.1177/21582440231165224](https://doi.org/10.1177/21582440231165224)

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Publisher's PDF, also known as Version of record

**Published In:**

SAGE Open

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# Default Feature Selection in Credit Risk Modeling: Evidence From Chinese Small Enterprises

SAGE Open  
 April-June 2023: 1–15  
 © The Author(s) 2023  
 DOI: 10.1177/21582440231165224  
[journals.sagepub.com/home/sgo](https://journals.sagepub.com/home/sgo)  


Nana Chai<sup>1</sup>, Baofeng Shi<sup>1</sup> , Bin Meng<sup>2</sup>, and Yizhe Dong<sup>3</sup>

## Abstract

This paper aims to design a novel AFCM-SMOTENC-APRIORI model to mine the default feature attributes of small enterprises. It can overcome the problem that the data characteristics of “small defaulting small enterprises and large non-defaulting small enterprises” make it difficult to mine the defaulting feature attributes of existing small enterprises. We used 1,231 small enterprise credit data from a city commercial bank in China to make an empirical analysis. We found that 23 feature attributes are strongly associated with default and 87% of the association rules are the same between the extended data and the original data mining. It shows that the data mining results with SMOTE-NC are highly consistent with the results of the original data mining, and the model is robust and reliable. It can be used as a reference for the credit risk identification of small enterprises in commercial banks.

## Keywords

feature attributes, association rule, small enterprise, AFCM-SMOTENC-APRIORI algorithm, imbalanced data

## Introduction

Small enterprises are the cornerstone of the global economy and employment. There are many of them, and they provide a large number of jobs (Chai et al., 2019). With the spread of COVID-19 across China, forcing the economic model to shift from a “contact economy” to a “distance economy” poses a great challenge for small enterprises (Belitski et al., 2022). Small enterprises are engaged in labor-intensive and “touch” services. The transformation of the economic model hinders the further development of small enterprises. With their reliance on long-tail customers, small enterprises have always faced the problems of difficult and expensive financing (Nizaeva & Coskun, 2019; Y. Sun et al., 2022). From 2019 to the end of 2021, the inclusive loans provided by Chinese financial institutions to small and micro enterprises increased from 31,973.4 billion yuan to 72,111.8 billion yuan (China Banking and Insurance Regulatory Commission [CBIRC], 2019). With the increase in the number of inclusive loans provided by financial institutions to small and micro enterprises, financial institutions also need to control their default risk and grasp the fundamental reasons for their default. However, due to the unbalanced characteristics of more non-default samples and fewer default samples in the credit evaluation data of

small enterprises, it is easy to focus on high-frequency events (non-default rules) and fail to effectively mine important low-frequency events (default rules) in association rule mining (Mahdi et al., 2022). In practice, mining the feature attributes strongly associated with defaulting small enterprises is the key to the identification of credit risk features, that is, the exploration of default rules is more important (Calabrese et al., 2019).

Feature attributes can reflect the characteristics of small enterprises with different loans under the same indicator. For example, the indicator of “Residence status of business owners” in the credit evaluation of small enterprises includes five types: “Self-owned or mortgage,” “Family building,” “Rent,” “Collective dormitory or common housing,” “Other, unknown or missing data.” Each type is a feature attribute of small loan enterprises. It is a finer dimension relative to indicators, which can

<sup>1</sup>Northwest A&F University, Yangling, Shaanxi, China

<sup>2</sup>Dalian Maritime University, Liaoning, China

<sup>3</sup>University of Edinburgh, UK

## Corresponding Author:

Baofeng Shi, College of Economics and Management, Northwest A&F University, No. 3, Taicheng Road, Yangling, Shaanxi 712100, China.  
 Email [shibaofeng@nwsuaf.edu.cn](mailto:shibaofeng@nwsuaf.edu.cn)



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of

the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

not only reflect the characteristics of small loan enterprises, but also reveal the fundamental causes of enterprise default. Existing research on credit risk focuses on indicator screening, score solving and credit rating grading (Ashofteh & Bravo, 2021; Shi et al., 2020; Xia et al., 2018; L. Yu et al., 2021). Some scholars have also made a detailed analysis of the indicators, including the comparison of default probability between blacks and whites, and the comparison of default probability between customers with high loan amounts and low loan amounts (Karlan & Zinman, 2010; Pope & Sydnor, 2012). However, few can reveal the relationship between the feature attributes of small enterprises and default status, especially the relationship between different intervals and the default status under quantitative indicators, reflecting the fundamental causes of small enterprises' default. Inadequate understanding of the default feature attributes and non-default feature attributes of small enterprises can easily lead financial institutions to make incorrect credit decisions, thus triggering small enterprise credit risk and increasing the pressure on financial institutions' credit risk.

To solve the above problems, this paper designs a novel AFCM-SMOTENC-APRIORI model to mine the default feature attributes of small enterprises. Then, we use the credit data of 1,231 small enterprises from a city commercial bank in China to make an empirical analysis. We found that among the 23 indicators that can significantly identify the default status of small enterprises, there are 23 feature attributes strongly associated with non-default. This means that the default risk of small enterprises with these feature attributes is relatively low, which can be used as a reference for credit granting. At the same time, we also found that among these 23 indicators, there are 23 feature attributes strongly related to default, showing that the default risk of small enterprises with those feature attributes is relatively high. This was especially true when: the return on total assets of a small enterprise is in the range [0.0000, 0.0344); the turnover speed of accounts receivable of a small enterprise is in the range [0.0484, 3.9500); the enterprise has less than five patents in the same industry; the product sales scope of small enterprises is unclear, or it is not in the two ways of domestic sales and export sales; the personal credit card of the legal representative (or the person in charge of the enterprise) of a small enterprise has a default record or missing data; and the legal person has held the position for less than 2 years. The financial information of small enterprises is not perfect, and financial institutions often do not know much about. The feature attributes of large enterprises can't help financial institutions make credit decisions for small enterprises. The features found in this paper, especially the default features, can help credit decision makers understand small businesses, identify small businesses with high default risk as early as possible, and

then make correct credit decisions, thus reducing the risk pressure of financial institutions.

The contribution of this paper includes the development of a new integrative methodology that combines AFCM, SMOTE-NC, and APRIORI. Another contribution is developing insights and relevance of complex relationships between the feature attributes and default status. More broadly, this study provides initial evidence and relationships for generalized defaulting association rules for financial institutions to mine the root cause of small enterprise defaulting. This study also addresses some of the methodology issues facing previous techniques applied to credit evaluation. For example, the proposed methodology addresses the Boolean of quantitative data, the mining of low-frequency rules and the exploration of the complex relationship between feature attributes and credit risk, all of which limit the application power of previous techniques.

Following the Introduction, the article consists of five sections: Literature Review, Research Design, Empirical Analyses, and the Conclusion.

### *Literature Review*

The difficult and expensive financing of small enterprises has been a focus of academic research. For a long time, scholars have believed that small enterprises are affected by credit rationing, and the borrowers of small enterprises are either refused loans or the loan amount is lower than their loan application amount or the loan interest rate is unbearable (Chai et al., 2019; Y. Zhang et al., 2021). This is mainly because financial institutions may face disproportionately high monitoring costs due to the opaque information of small enterprises and the relative scarcity of mortgage assets compared with large enterprises or high achieving enterprises (Rostamkalaei & Freel, 2016). In addition, Freel et al. (2012) found that in addition to the above possible reasons, a reason why the demand for credit of small businesses can't be met is that small businesses not applying because they may be reluctant to apply due to the prospect of being rejected is almost twice that of being rejected. One of the key measures to solve the credit rationing of small businesses is to identify the credit risk of small businesses, so as to ensure that the probability of default can be accurately estimated (Y. Sun et al., 2022).

As for credit risk identification, the existing literature mainly focuses on credit evaluation indicator screening, credit score solution and credit rating. Those are the ways to predict loan customers' default status. Indicator screening mainly adopts the filtering method, embedding method and hybrid method to reduce the indicators, and then uses the indicators retained after reduction to predict the default status of customers (Hou et al., 2014;

Sefidian & Daneshpour, 2019; J. Sun et al., 2015; L. Yu et al., 2021; C. Zhang & Hu, 2020; Zhou et al., 2021). The credit score is used to represent the level of customer credit risk, and its solution mainly includes measurement and statistics, artificial intelligence and goal optimization (Ashofteh & Bravo, 2021; Luo, 2020; Mancisidor et al., 2020; Xia et al., 2018). The credit rating division consists of four mainstream methods: default probability threshold, credit score range, customer number distribution and default loss rate division (Chai et al., 2019; Krink et al., 2008; Shi et al., 2020). To improve the prediction effect of loan customer default, Bagging, Boosting, SMOTE and other methods that can retain a small number of default sample information have also been widely used (Niu et al., 2020; Shen et al., 2021; J. Sun et al., 2018).

In terms of feature attribute mining, Karlan and Zinman (2010) found that the borrower with a higher loan amount is less likely to default. Pope and Sydnor (2012) found that the default probability of black borrowers was 36% higher than that of white borrowers by using the Prosper data and Logit regression empirical analysis. Hildebrand et al. (2017) found that the loan projects with group leaders bidding had a higher default probability than those without group leaders bidding by the credit transaction data of the Prosper platform from February 2007 to April 2008. Bai et al. (2019) used the combination of a fuzzy rough set and fuzzy C-means method to build a credit risk evaluation model. Using the loan data of 2,044 farmers, they concluded that the education level and skills of farmers are highly related to their credit quality. The above literature makes a systematic analysis of credit risk, but it has not yet mined the feature attributes strongly associated with credit risk to find out the root cause of credit risk. In particular, when there are few default samples and more non-default samples, it is more difficult to explore the causes of the default. Therefore, this paper designs a combination method to explore the root causes of small business default.

Association rule analysis is the key technology of big data analysis, which can explore the correlation or relevance between different transactions from a large amount of data (Luo, 2020). APRIORI was proposed by Agrawal et al. (1993) to mine association rules. Subsequently, various scholars optimized this method and applied it widely in various fields. Lazcorreta et al. (2008) used this algorithm to analyze the behavior of individual users in a Web information system, and helped users choose the best customized links. Xie et al. (2020) optimized this algorithm to analyze the evolutionary characteristics of regional traffic congestion, laid the foundation for formulating advanced regional traffic control strategies, and helped to alleviate traffic

congestion. Based on this algorithm, Luo (2020) dug into the association rules and association degree between poverty indexes and poverty degree, and studied in-depth the causes of poverty among residents. As the purpose of this study is to mine the correlation between each feature attribute of small enterprises and default status, it is consistent with the performance of the APRIORI algorithm in mining the correlation between data items, and the algorithm has an extensive practical basis in various fields. Therefore, this paper adopted this method to analyze the default feature attributes and non-default feature attributes of small enterprises.

## Research Design

### *Interval Division of Quantitative Indicators*

In practice, association rules are generally divided into Boolean association rules and quantitative association rules. Boolean association rules deal with qualitative indicators with discrete values (Kabir et al., 2017), and quantitative association rules are used to deal with quantitative indicators with continuous values (Alataş & Akin, 2006). For example, if the association rule antecedent  $X$  is “female” and the association rule subsequent  $Y$  is “Default,” this rule is a Boolean association rule. If the association rule antecedent  $X$  is “industry prosperity index = 102” and the association rule subsequent  $Y$  is “non-default,” this rule is a quantitative association rule. The mining of quantitative association rules generally needs to convert quantitative data into Boolean data processing, that is, it is generally processed in an interval way (Alataş & Akin, 2006).

To ensure that the quantitative indicator interval division of small enterprises meets the goal of “the smaller the distance within the class, the greater the distance between classes,” this paper uses Adaptive Fuzzy C-means (AFCM) algorithm to convert the quantitative indicator value into Boolean data. The characteristics of this method are as follows. By solving the adaptive function composed of the distance between classes and the distance within classes of different feature attributes of the same indicator, a small enterprise quantitative indicator interval division model with improved fuzzy C-means is constructed. Following this the cluster number corresponding to the maximum of the adaptive function, that is, the optimal interval number, is obtained. This reflects the interval division idea of “the smaller the distance within the class, the greater the distance between the classes” of different feature attributes of the same indicator, which makes up for the deficiency that the clustering number of the existing fuzzy C-means algorithm needs to be set subjectively. More specifically, the objective function, cluster number and conditions are given by (Y. Sun et al., 2022)

$$\min J(U, c_1, \dots, c_l) = \sum_{i=1}^l \sum_{j=1}^m (u_{ij})^n d^2(X_j, c_i) \quad (1)$$

Where  $J(U, c_1, \dots, c_l)$  is the sum of squares of weighted distances from each feature attribute value of a quantitative indicator to the cluster center of the group;  $l$  is the total number of clusters, ( $i = 2, \dots, l$ );  $m$  is the number of small enterprises;  $u_{ij} \in [0, 1]$  represents the membership degree of a quantitative indicator value  $X_j$  of the  $j$ -th small enterprise belonging to class  $i$ ;  $n \in [1, \infty)$  is the weighted index of membership degree;  $c_i$  is the cluster center of  $i$ -th class, and the cluster center vector is  $C = \{c_i\}$ ;  $d(X_j, c_i)$  represents the Euclidean distance from the indicator value  $X_j$  to the cluster center  $c_i$ . Equations 2 and 3 are the necessary conditions for solving the minimum value of Equation 1, and  $K$  is the number of iterations.

$$c_i = \frac{\sum_{j=1}^m (u_{ij})^n X_j}{\sum_{j=1}^m (u_{ij})^n} \quad (2)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{n-1}}} \quad (3)$$

For quantitative indicator  $X$ , under the conditions of meeting  $u_{ij} \in [0, 1]$  and  $\sum_{i=1}^l u_{ij} = 1$ , by solving the sum of the distance between the indicator value  $X_j$  of small enterprises in the class and the cluster center  $c_i$ , we can find out the clustering result of the quantitative indicator of small enterprises with the smallest distance in the class, so as to realize the Boolean data transformation of the quantitative indicator. The premise of getting cluster results is that we need to determine the cluster number  $l$  first. Then, we can solve it when the maximum value of  $L(l)$  is reached,

$$L(l) = \frac{\sum_{i=1}^l \sum_{j=1}^m (u_{ij})^n \|c_i - \bar{X}\|^2 / (l-1)}{\sum_{i=1}^l \sum_{j=1}^m (u_{ij})^n \|X_j - c_i\|^2 / (m-l)} \quad (4)$$

Where  $L(l)$  is the adaptive function value of cluster number  $l$ ,  $2 < l < m$ ,  $\bar{X} = \frac{\sum_{i=1}^l \sum_{j=1}^m u_{ij}^m X_j}{m}$  as the center vector of quantitative indicator  $X$ . The molecule  $\sum_{i=1}^l \sum_{j=1}^m (u_{ij})^n \|c_i - \bar{X}\|^2 / (l-1)$  of Equation 4 is the distance between different feature attribute classes of the same indicator. The greater the distance from the cluster

center  $c_i$  to the indicator center vector  $\bar{X}$ , the more significant difference between classes, that is, the difference between feature attributes of different classes is more obvious. The denominator  $\sum_{i=1}^l \sum_{j=1}^m (u_{ij})^n \|X_j - c_i\|^2 / (m-l)$  of Equation 4 is the intraclass distance of the feature attributes of the same indicator.

We can explain Equation 4 as follows. The smaller the distance from the indicator value  $X_j$  to the cluster center  $c_i$ , the smaller the difference between different feature attributes, that is, the closer these feature attributes are. The larger the numerator, the smaller the denominator and the larger the score value, indicating that the interval division results of different feature attributes of the same indicator can better reflect the standard of "the smaller the distance within the class and the greater the distance between classes." When the maximum value of  $L(l)$  is reached, the optimal cluster number can then be obtained.

We need to calculate the clustering center vector  $C$  and membership matrix  $U$  through continuous iteration until the difference between the iteration result of step  $K+1$  and that of step  $K$  is less than the threshold  $\varepsilon$ . At this time, the iteration is terminated to realize the interval division of quantitative indicators. Therefore, the iteration termination condition is

$$\|C^{(K+1)} - C^{(K)}\| < \varepsilon \quad (5)$$

Furthermore, the constraint conditions to ensure the maximum of Equation 4 are given,

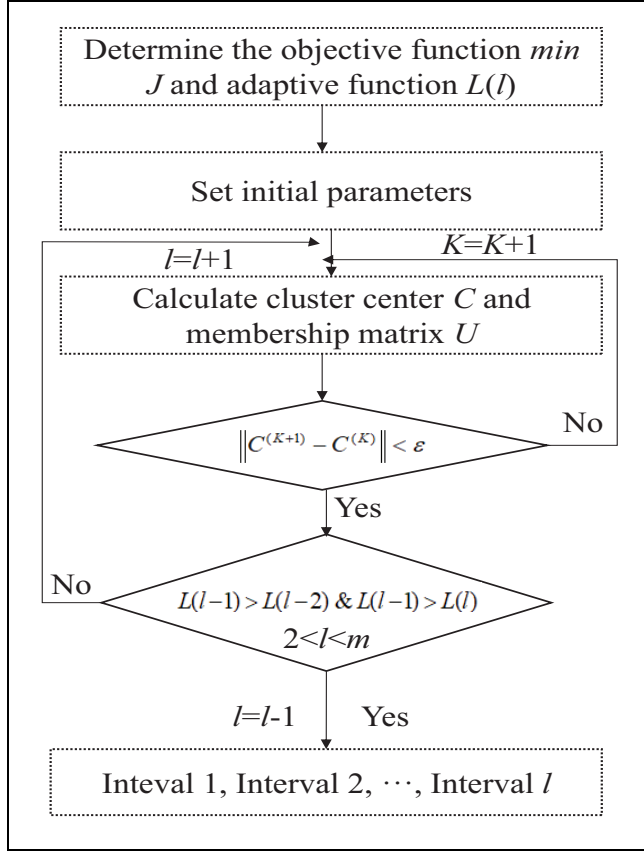
$$L(l-1) > L(l-2) \text{ and } L(l-1) > L(l) \quad (6)$$

The value  $L(l-1)$  of cluster number  $l-1$  is not only greater than the value  $L(l-2)$  corresponding to cluster number  $l-2$ , but also greater than the value  $L(l)$  corresponding to cluster number  $l$ , ensuring that  $L(l-1)$  is the maximum value; at that point the clustering number  $l-1$  is the optimal clusters of the quantitative indicator interval, and the clustering process ends.

To facilitate intuitive understanding, the implementation flow of the Adaptive Fuzzy C-means algorithm is shown in Figure 1.

### Association Rule Mining of Small Enterprise

**Default Sample Expansion Based on SMOTE-NC.** Due to the unbalanced characteristics of more nondefault samples and fewer default samples in credit evaluation data, it is easy to focus on the non-default high-frequency rules and fail to effectively mine the important low-frequency rules of default in association rule mining (Mahdi et al., 2022). Therefore, before mining association rules, it is necessary to expand the low-frequency default samples



**Figure 1.** Principle of adaptive fuzzy C-means interval division.

according to the unbalanced distribution characteristics of small enterprise data, to make up for the impact of data imbalance on low-frequency event association rules mining. In this paper, a default sample expansion model is constructed by Synthetic Minority Oversampling Technique-Nominal Continuous (SMOTE-NC). For the quantitative indicator  $S_{i,quantitative}^{new}$  in the newly synthesized default sample, it can be expressed as (Chawla et al., 2002)

$$S_{i,quantitative}^{new} = S_{i,quantitative} + \lambda(S_{i,quantitative}^l - S_{i,quantitative}) \quad (7)$$

Where  $S_{i,quantitative}$  represents the quantitative indicator set of defaulting small enterprises and  $S_{i,quantitative}^l$  is a qualitative indicator set,  $\lambda \in [0, 1]$  is a random number.  $S_{i,quantitative}^{new}$  is the indicator value with the highest frequency among the  $K$ -nearest neighbor of the default sample  $S_{i,quantitative}^l$ . For the qualitative indicator in the new default sample

$$S_i^{new} = S_{i,quantitative}^{new} \cup S_{i,qualitative}^{new} \quad (8)$$

According to the distribution characteristics of unbalanced data of default and non-default of small enterprise

data, the small enterprise default sample expansion model of SMOTE-NC is established to expand the samples of low-frequency default small enterprises, so as to balance the default and non-default samples in the data set and avoid the impact of data imbalance on the mining of association rules of low-frequency events. The difference between this method and the classical data expansion method SMOTE is that SMOTE cannot directly expand the data of qualitative indicators. It needs to digitize the qualitative indicators and then expand them according to Equation 8; SMOTE-NC does not need to numerate the qualitative indicator, it only needs to count the frequency of the feature attributes of qualitative indicators. This meets the data requirements of feature attribute mining in this paper, and makes up for the fact that of SMOTE cannot directly expand the qualitative indicators and the expanded qualitative indicator data classification is not clear. For example, if the feature attribute “male” contained in the “gender” indicator is assigned as 1 and “female” is assigned as 0.8, the newly generated sample of SMOTE may be 0.89 through interpolation, so it is difficult to determine which attribute the value of 0.89 should be.

*Mining Default Feature Attributes of Small Enterprises Based on APRIORI.* In order to mine the default feature attributes of small enterprises, it is necessary to establish association rules between the feature attributes of small enterprises and their corresponding default status. Support and Confidence are two measurement criteria that can accurately describe the correlation between variables (Mahdi et al., 2022). The degree of support  $f(X \rightarrow Y)$  refers to the probability of simultaneous occurrence of the preceding item  $X$  (i.e., the feature attribute contained in the evaluation indicator) of the association rule and the subsequent item  $Y$  (i.e., default status) of the association rule (Hong et al., 2020),

$$f(X \rightarrow Y) = \text{count}(XY) / \text{count}(T) \quad (9)$$

Where  $T$  represents the number of small loan enterprises,  $\text{count}^*$  represents the frequency of corresponding events, and  $\text{count}(XY)$  represents the frequency of simultaneous occurrence of feature attribute  $X$  and default status  $Y$ . Equation 9 depicts the regular relationship between feature attribute  $X$  contained in the small enterprise credit evaluation indicator and default status  $Y$ . The greater the support  $f(X \rightarrow Y)$ , the stronger the correlation between feature attribute  $X$  and default status  $Y$ .

Confidence represents the percentage of default status  $Y$  of small enterprises with feature attribute  $X$ , expressed as a conditional probability (Hong et al., 2020),

$$g(X \rightarrow Y) = \text{count}(XY) / \text{count}(X) \quad (10)$$

The greater the confidence  $g(X \rightarrow Y)$ , the greater the probability of default status  $Y$  of small enterprises with feature attribute  $X$ .

It is necessary to select an appropriate association rule mining algorithm to mine default feature attributes by using support and credibility. Due to its easy understanding, low data requirements and strong interpretability of recognition rules, the APRIORI algorithm is widely used in mining the association rule in the medical field, engineering and other industries (Kavšek & Lavrač, 2006; Tatavarthi & Sambasiva, 2017; H. Yu et al., 2011). The core of APRIORI is to find the frequent itemset  $I$  where its support  $f$  is not less than the minimum threshold  $\delta$  in the database through iteration, and then construct the association rule set  $Z$  where the confidence  $g$  is not lower than the minimum confidence threshold  $\zeta$  (Mahdi et al., 2022). The implementation process is as follows.

**Build a Frequent Itemset  $I$ .**  $f(X_i^b \rightarrow Y)$  is the support between  $X_i^b$  and  $Y$  obtained by Equation 10,  $\delta$  is the minimum support threshold. Then,

$$I = \{(X_i^b, Y) | f(X_i^b \rightarrow Y) \geq \delta\}, \exists X_i^b \rightarrow Y \quad (11)$$

Where  $X_i^b$  as the  $b$ -th feature attribute ( $b = 1, 2, \dots, l$ ) of the  $i$ -th indicator, and  $l$  is the number of feature attributes of the  $i$ -th indicator, ( $i = 1, 2, \dots, n$ );  $n$  is the number of indicators, and  $Y$  is the default status of small enterprises ( $Y = 1$  for defaulting small enterprises,  $Y = 0$  for non-defaulting small enterprises). Equation 11 indicates that when the  $b$ -th feature attribute  $X_i^b$  and default status  $Y$  occur at the same time, the support  $f(X_i^b \rightarrow Y)$  is greater than or equal to the threshold  $\delta$ .  $X_i^b \rightarrow Y$  belongs to frequent itemset  $I$ . It shows that the frequency of the feature attribute  $X_i^b$  and default status  $Y$  is very high. That is, the feature attribute  $X_i^b$  is highly related to the default status  $Y$ .

**Build a Strong Association Rule Set  $Z$ .** On the premise that  $X_i^b \rightarrow Y$  belongs to frequent itemset  $I$ , when the confidence is greater than or equal to the confidence threshold  $\zeta$ ,  $X_i^b \rightarrow Y$  can be retained in the strong association rule set  $Z$ . Then,

$$Z = \{(X_i^b, Y) | g(X_i^b \rightarrow Y) \geq \zeta, (X_i^b, Y) \in I\}, \exists X_i^b \rightarrow Y \quad (12)$$

Equation 12 indicates that there is a strong correlation between the feature attributes  $X_i^b$  and the default status  $Y$ , and the reliability is high.

**Determine the Feature Attributes  $X_{i, \max}^b$  Most Relevant to the Default Status of Small Enterprises.** Suppose the feature attribute  $X_{i, \max}^b$  contained in the  $i$ -th indicator has the

strongest correlation with the default status  $Y$  in the strong association rule set  $Z$ , then

$$\max f(X_{i, \max}^b \rightarrow Y), \forall (X_i^b, Y) \in Z \quad (13)$$

## Empirical Analyses

### Sample Selection and Data Source

This paper uses 1,231 small business loans from 1994 to 2012 of a city commercial bank in China as an empirical sample. According to the Classification Standard of Small and Medium-sized Enterprises issued by the Ministry of Industry and Information Technology of the People's Republic of China, the data was collected from 11 industries, such as wholesale, retail, construction, transportation, leasing and business services. At the same time, the data covers nine provincial regions: Beijing, Hebei, Henan, Jiangxi, Liaoning, Shanghai, Sichuan, Tianjin and Chongqing. Among all of the loans, only 35 had defaulted.

Based on the high-frequency credit evaluation indicators of small enterprises of authoritative institutions such as Moody's, S&P, Fitch, China Construction Bank, combined with the availability of data, this paper has established a small enterprise credit evaluation indicator system composed of 10 standard levels, such as small enterprise solvency, profitability, operation ability, growth ability, external macro conditions, internal non-financial factors and enterprise credit. There are 81 indicators contained in this data such as asset liability ratio, return on total assets and industry prosperity indicator. On this basis, the indicators reflecting information redundancy are eliminated through partial correlation analysis, and the indicators that can significantly distinguish the default status of small enterprises are selected by Probit regression. A small enterprise credit rating indicator system composed of 23 indicators such as " $X_1$  Quick ratio" and " $X_{14}$  years of employment related industries" was constructed, as shown in Table 1. Descriptive statistics of 23 indicators are shown in Tables 2 and 3 shows the scores of different feature attributes of quantitative indicators. The system includes 13 quantitative indicators and 10 qualitative indicators, and the corresponding AUC is as high as 98.62% (Chai et al., 2019). The feature attributes and corresponding coding of 23 indicators are shown in Table 4.

### Analysis of Mining Results of Non-defaulting and Defaulting Feature Attributes of Small Enterprises

**Interval Division Results of Quantitative Indicators.** Taking the interval division of the quantitative indicator " $X_1$  Quick ratio" as an example, this paper explains the interval

**Table 1.** Original Data of Credit Evaluation Indicators of Small Enterprises.

Indicator type	Indicator	Non-defaulting small business number ( $Y_i=0$ )			Defaulting small business number ( $Y_i=1$ )		
		XQY 001	...	XQY 1196	XQY 1197	...	XQY 1231
Quantitative indicator	$X_1$ Quick ratio	0.0007	...	1.1422	6.0000	...	8.7981
	$X_2$ Overdrive ratio	0.3956	...	0.6670	0.1767	...	0.0742
	$X_3$ Return on total assets	0.029	...	0.002	0.020	...	0.000
	...	...	...	...	...	...	...
	$X_{13}$ Mortgage and pledge score	0.65	...	0.00	0.10	...	0.00
Qualitative indicator	$X_{14}$ Years of employment in related industries	Working years $\geq 8$ years	...	Working years $\geq 8$ years	Working years $\geq 8$ years	...	2 years $\leq$ working years $< 5$ years
	$X_{15}$ Patent status	The enterprise has no patent	...	Missing data	The enterprise has no patent	...	Missing data
	$X_{16}$ Product sales scope	Domestic products	...	Product export	other	...	Missing data
	...	...	...	...	...	...	...
	$X_{23}$ Number of contract breaches between enterprises	Contract breach between enterprises for 0 time	...	Contract breach between enterprises for 0 time	Contract breach between enterprises for 3 times or more	...	Contract breach between enterprises for 3 times or more

division process of the quantitative indicator. Let the initial cluster number  $l^0 = 2$ , stop threshold  $\varepsilon = 1E - 5$ , fuzzy number  $\omega = 2$ , the initial membership matrix  $U^0$  is a 0 matrix of  $t * m$ , and the initial cluster center  $C^0$  is a 0 matrix of  $1 * t$  (Zeng et al., 2014). The membership

matrix  $U^K = \begin{pmatrix} 0.0005 & \dots & 0.1336 \\ \vdots & \vdots & \vdots \\ 0.9965 & \dots & 0.0050 \end{pmatrix}_{4 \times 1231}$ . The

clustering center  $c^K$ , interval division of indicator “ $X_1$  Quick ratio” and adaptive function value  $L(4)$  can be obtained by using the original data of indicator “ $X_1$  Quick ratio” in Table 2 in combination with Equations 1–6, as shown in Table 4. The interval division results in column 3 of Table 5 are feature coded, and the results are shown in column 4 of Table 6.

Similarly, according to the above Section 3.1(4) adaptive fuzzy C-means AFCM interval division steps, the interval division of the remaining 12 quantitative indicators in Table 2 can be completed.

**Mining Results of Default Feature Attributes**

*Default Feature Attributes of Small Enterprises Mining Based on SMOTENC-APRIORI.* Firstly, according to the SMOTENC default sample expansion method (See section 3.2.1), the original data of 1,231 small enterprises (1,196 non-

defaulting small enterprises and 35 defaulting small enterprises) are expanded to form 2,392 small enterprise expansion data (1,196 non-defaulting small enterprises and 1,196 defaulting small enterprises). Then, the extended data is used to encode the feature attributes, and the quantitative indicator data is transformed into Boolean data.

Secondly, after continuous iterative attempts, the minimum support threshold is set to  $\delta = 0.05$  and the minimum confidence threshold is set to  $\zeta = 0.50$ . Using the original data and extended data encoded by feature attributes, combined with Section 3.2.2 above, the feature attributes of small enterprises strongly associated with default ( $Y_i = 1$ ) and non-default ( $Y_i = 0$ ) are mined respectively. The results are shown in Tables 6 and 7.

*Analysis of Feature Attributes of Small Enterprises Strongly Associated With Non-Default ( $Y_i=0$ ).*

Using the expanded data of 2,392 small enterprises, one non-defaulting feature is mined for each indicator, as shown in Table 6. Take “ $X_1$  quick ratio” as an example: the strong association relation between the feature attribute of “ $X_1$  quick ratio” and non-default is  $X_1^2 = [0.8668, 3.7894)$ , indicating that small loan enterprises with “quick ratio” in the range of  $[0.8668, 3.7894)$  are less likely to default than other small enterprises. At the same time, the support of the



**Table 2.** Descriptive Statistics.

Indicator	Obv	Minimum	Maximum	Mean	Std. dev	Median	Mode
X <sub>1</sub> Quick ratio	1,231	0.000	8.800	1.734	2.080	1.110	0.001
X <sub>2</sub> Overdrive ratio	1,231	0.000	5.630	0.888	0.926	0.651	0.169
X <sub>3</sub> Return on total assets	1,231	0.000	0.770	0.078	0.102	0.042	0.000
X <sub>4</sub> Operating profit margin	1,231	0.000	0.650	0.177	0.179	0.120	0.002
X <sub>5</sub> Gross profit margin	1,231	0.000	0.560	0.133	0.163	0.070	0.002
X <sub>6</sub> Net cash flow from operating activities	1,231	-96,200,000	96,860,000	630,735.	15,181,880	0.000	0.000
X <sub>7</sub> Turnover speed of accounts receivable	1,231	0.050	335.030	17.130	49.304	4.150	0.048
X <sub>8</sub> Growth rate of total assets	1,231	-1.000	2.990	0.137	0.655	0.026	-1.000
X <sub>9</sub> Industry prosperity index	1,231	65.710	165.100	138.914	12.982	143.600	153.200
X <sub>10</sub> Consumer price index	1,231	0.000	121.400	102.395	3.458	102.700	102.700
X <sub>11</sub> Per capita disposable income of urban residents	1,231	3,058	36,230	19,333.467	4,817.255	19,014	21,293
X <sub>12</sub> Total value of the legal representative automobile and real estate	1,231	0.000	22,910,000	3,450,460	4,367,938	1,690,000	0.000
X <sub>13</sub> Mortgage and pledge score	1,231	0.000	1.000	0.584	0.302	0.700	0.700
X <sub>14</sub> Years of employment in related industries	1,231	1.000	4.000	3.125	1.154	4.000	4.000
X <sub>15</sub> Patent status	1,231	1.000	2.000	1.070	0.255	1.000	1.000
X <sub>16</sub> Product sales scope	1,231	1.000	3.000	1.725	0.539	2.000	2.000
X <sub>17</sub> Loan default record of the legal representative	1,231	1.000	4.000	3.133	1.234	4.000	4.000
X <sub>18</sub> Credit card records of legal representative	1,231	1.000	2.000	1.630	0.483	2.000	2.000
X <sub>19</sub> Residential status	1,231	1.000	5.000	3.410	1.937	5.000	5.000
X <sub>20</sub> Local residence of the company's legal representative	1,231	1.000	3.000	2.320	0.886	3.000	3.000
X <sub>21</sub> Time of holding the post	1,231	1.000	4.000	2.151	1.088	2.000	1.000
X <sub>22</sub> Corporate credit in recent 3 years	1,231	1.000	4.000	3.257	1.237	4.000	4.000
X <sub>23</sub> Number of contract breaches between enterprises	1,231	1.000	4.000	3.257	1.277	4.000	4.000
Default status	1,231	1.000	4.000	3.125	1.154	4.000	4.000

association rule  $f(X_1^2 \rightarrow Y_i = 0) = 0.46$ , indicates that 46% of the 2,392 small enterprises have an “X<sub>1</sub> quick ratio” in the range of [0.8668, 3.7894] and there is no default. Further analysis shows the confidence of the association rule  $g(X_1^2 \rightarrow Y_i = 0) = 0.5319$ , indicating that small enterprises with the quick ratio in the range of [0.8668, 3.7894] have a 53.19% probability that they will not default. The specific description of non-defaulting feature attributes of 23 indicators is shown in column 2 of Table 8.

To verify the reliability of the feature attribute mining method established in this paper, original data and expanded data are used to find the feature attributes strongly associated with non-default ( $Y_i = 0$ ), as shown in Table 6. By comparing the similarity of association rules between the expanded data and the original data, it is not difficult to find that 20 of the 23 association rules extracted from the extended data mining are the same as the association rules of the original data mining. The feature attributes of the two dataset mining have a

**Table 3.** Scoring Results of Qualitative Indicators.

Indicator	Feature attributes	Score
X <sub>14</sub> Years of employment in related industries	0 < working years <2 years, or the data is missing	1
	2 years ≤ working years <5 years	2
	5 years ≤ working years <8 years	2
	Working years ≥ 8 years	4
...	...	...
X <sub>15</sub> Patent status	The enterprise has no patent or data is missing	1
	The enterprise has less than 5 patents in the same industry	2
	The enterprise has 5 or more patents in the same industry	3
X <sub>23</sub> Number of contract breaches between enterprises	Contract defaults for 3 times or more, or data is missing	1
	...	...
	Contract breach between enterprises for 0 times	4
	...	...

**Table 4.** Indicator Feature Attributes and Their Codes.

Indicator type	Indicator name	Feature attribute	Feature attribute coding
Quantitative indicator	X <sub>1</sub> Quick ratio	[0.0007, 0.8668)	X <sub>1</sub> <sup>1</sup>
		[0.8668, 3.7894)	X <sub>1</sub> <sup>2</sup>
		[3.7894, 7.3765)	X <sub>1</sub> <sup>3</sup>
		[7.3765, 8.7981]	X <sub>1</sub> <sup>4</sup>
...	...	...	...
Qualitative indicator	X <sub>13</sub> Mortgage and pledge score	[0.0000, 0.0700)	X <sub>13</sub> <sup>1</sup>
		...	...
		[0.8299, 0.9417)	X <sub>13</sub> <sup>11</sup>
		[0.9417, 1.0000)	X <sub>13</sub> <sup>12</sup>
Qualitative indicator	X <sub>14</sub> Years of employment in related industries	0 < working years <2 years, or the data is missing	X <sub>14</sub> <sup>1</sup>
		2 years ≤ working years <5 years	X <sub>14</sub> <sup>2</sup>
		5 years ≤ working years <8 years	X <sub>14</sub> <sup>3</sup>
		Working years ≥ 8 years	X <sub>14</sub> <sup>4</sup>
...	...	...	...
Qualitative indicator	X <sub>23</sub> Number of contract breaches between enterprises	Contract breach between enterprises for 3 times or more, or lack of data	X <sub>23</sub> <sup>1</sup>
		Contract breach between enterprises for 2 times	X <sub>23</sub> <sup>2</sup>
		Contract breach between enterprises for 1 time	X <sub>23</sub> <sup>3</sup>
		Contract breach between enterprises for 0 times	X <sub>23</sub> <sup>4</sup>

**Table 5.** Interval Division Results of “X<sub>1</sub> Quick Ratio.”

(1) No.	(2) Cluster center c <sup>k</sup>	(3) Interval division	(4) Feature attribute coding
1	0.2292	[0.0007, 0.8668)	X <sub>1</sub> <sup>1</sup>
2	1.5022	[0.8668, 3.7894)	X <sub>1</sub> <sup>2</sup>
3	5.8757	[3.7894, 7.0200)	X <sub>1</sub> <sup>3</sup>
4	8.6242	[7.3765, 8.7981)	X <sub>1</sub> <sup>4</sup>
Adaptive function L(4)=(5998.3605, 6710.5752, 9076.5517, 6489.5129)			

coincidence degree of 87% (≈ 20/23), which shows that the data expanded by SMOTE-NC is highly consistent with the original data mining results. The feature mining model established in this paper can be applied and popularized in credit risk identification.

*Analysis of the Feature Attributes of Small Enterprises Strongly Associated With Default (Y<sub>i</sub>=1).* Using the expanded data of 2,392 small enterprises, one defaulting feature is mined for each indicator, as shown in Table 7. Take “X<sub>1</sub> Quick ratio” as an example; the strong association

**Table 6.** Mining Results of Feature Attributes of Small Enterprises Strongly Associated With Non-Default ( $\gamma_i = 0$ ).

No.	Association rule mining based on extended data				Are the association rules of extended data and original data mining the same				Association rule mining based on original data			
	Association rule antecedent X	Association rule consequent Y	Support f	Confidence g	Yes	Association rule antecedent X	Association rule consequent Y	Support f	Confidence g	Association rule antecedent X	Association rule consequent Y	Support f
1	X <sub>1</sub> <sup>2</sup>	0	0.46	0.5319	Yes	X <sub>1</sub> <sup>2</sup>	0	0.48	0.9832			
2	X <sub>2</sub> <sup>4</sup>	0	0.1	0.6802	Yes	X <sub>2</sub> <sup>4</sup>	0	0.14	0.9825			
3	X <sub>3</sub> <sup>2</sup>	0	0.22	0.5802	Yes	X <sub>3</sub> <sup>2</sup>	0	0.25	0.9775			
4	X <sub>4</sub> <sup>4</sup>	0	0.06	0.8377	Yes	X <sub>4</sub> <sup>4</sup>	0	0.10	1.0000			
5	X <sub>5</sub> <sup>2</sup>	0	0.07	0.7421	Yes	X <sub>5</sub> <sup>2</sup>	0	0.10	0.9916			
6	X <sub>6</sub> <sup>4</sup>	0	0.56	0.6587	Yes	X <sub>6</sub> <sup>4</sup>	0	0.72	0.9865			
7	X <sub>7</sub> <sup>2</sup>	0	0.24	0.6834	Yes	X <sub>7</sub> <sup>2</sup>	0	0.33	0.9850			
8	X <sub>8</sub> <sup>7</sup>	0	0.11	0.5148	Yes	X <sub>8</sub> <sup>8</sup>	0	0.10	0.9829			
9	X <sub>9</sub> <sup>5</sup>	0	0.22	0.8874	Yes	X <sub>9</sub> <sup>4</sup>	0	0.54	0.9881			
10	X <sub>10</sub> <sup>2</sup>	0	0.53	0.6210	Yes	X <sub>10</sub> <sup>2</sup>	0	0.65	0.9838			
11	X <sub>11</sub> <sup>5</sup>	0	0.21	0.5375	Yes	X <sub>11</sub> <sup>5</sup>	0	0.23	0.9750			
12	X <sub>12</sub> <sup>2</sup>	0	0.57	0.5856	Yes	X <sub>12</sub> <sup>2</sup>	0	0.38	0.9936			
13	X <sub>13</sub> <sup>9</sup>	0	0.18	0.7824	Yes	X <sub>13</sub> <sup>9</sup>	0	0.28	0.9826			
14	X <sub>14</sub> <sup>4</sup>	0	0.57	0.5077	Yes	X <sub>14</sub> <sup>4</sup>	0	0.57	0.9759			
15	X <sub>15</sub> <sup>2</sup>	0	0.06	0.6045	Yes	X <sub>15</sub> <sup>2</sup>	0	0.92	0.9735			
16	X <sub>16</sub> <sup>2</sup>	0	0.35	0.9348	Yes	X <sub>16</sub> <sup>2</sup>	0	0.63	0.9923			
17	X <sub>17</sub> <sup>4</sup>	0	0.40	0.7737	Yes	X <sub>17</sub> <sup>4</sup>	0	0.61	0.9893			
18	X <sub>18</sub> <sup>2</sup>	0	0.35	0.9276	Yes	X <sub>18</sub> <sup>2</sup>	0	0.63	0.9923			
19	X <sub>19</sub> <sup>5</sup>	0	0.40	0.7547	Yes	X <sub>19</sub> <sup>5</sup>	0	0.59	0.9862			
20	X <sub>20</sub> <sup>3</sup>	0	0.33	0.9401	Yes	X <sub>20</sub> <sup>3</sup>	0	0.60	0.9919			
21	X <sub>21</sub> <sup>2</sup>	0	0.19	0.7155	Yes	X <sub>21</sub> <sup>2</sup>	0	0.27	0.9761			
22	X <sub>22</sub> <sup>4</sup>	0	0.42	0.8578	Yes	X <sub>22</sub> <sup>4</sup>	0	0.71	0.9897			
23	X <sub>23</sub> <sup>4</sup>	0	0.55	0.6757	Yes	X <sub>23</sub> <sup>4</sup>	0	0.74	0.9857			

**Table 7.** Mining Results of Feature Attributes of Small Enterprise Strongly Associated With Default ( $Y_i = 1$ ).

Association rule mining based on extended data					
No.	Association rule antecedent $X$	$\rightarrow$	Association rule subsequent $Y$	Support $f$	Confidence $g$
1	V13	$\rightarrow$		0.22	0.7452
2	V21	$\rightarrow$		0.25	0.7620
3	V31	$\rightarrow$		0.6	0.6231
4	V41	$\rightarrow$		0.19	0.6644
5	V51	$\rightarrow$		0.46	0.6471
6	V61	$\rightarrow$		0.1	0.9620
7	V71	$\rightarrow$		0.66	0.6363
8	V86	$\rightarrow$		0.29	0.5163
9	V93	$\rightarrow$		0.27	0.6797
10	V101	$\rightarrow$		0.47	0.6364
11	V111	$\rightarrow$		0.25	0.9652
12	V121	$\rightarrow$		0.41	0.6367
...	...	...	...	...	...
22	V221	$\rightarrow$		0.33	0.6916
23	V231	$\rightarrow$		0.33	0.6474

relation between the feature attribute of “ $X_1$  Quick ratio” default is  $X_1^3 = [3.7894, 7.3765]$ . Therefore small enterprises with a quick ratio in the range of  $[3.7894, 7.3765]$  are more likely to default than other small enterprises. At the same time, the support of the association rule  $f(V12 \rightarrow Y_i = 1) = 0.22$ , indicates that 22% of the 2,392 small enterprises have a “quick ratio” of  $[3.7894, 7.3765]$  and defaulted. The confidence of the association rule  $g(V12 \rightarrow Y_i = 1) = 0.7452$ , indicates that when the quick ratio is in the range of  $[3.7894, 7.3765]$ , small enterprises have a 74.52% probability of default. The specific description of defaulting feature attributes of the 23 indicators is shown in column 3 of Table 8.

The 23 feature attributes strongly associated with default are the attributes that financial institutions should pay attention to when granting credit. In particular, the following five feature attributes deserve more attention. The return on total assets of the enterprise is in the range  $[0, 0.0344]$ , and the turnover speed of accounts receivable is in the range  $[0.0484, 3.9500]$ . The enterprise has less than five patents in the same industry, and the product sales scope is unclear or not in two ways (domestic sales and export sales). Legal representative (or person in charge of the enterprise) personal credit card has default record or missing data. And the legal person has held the position for less than 2 years. Their probability of simultaneous occurrence with enterprise default (i.e., support  $f$ ) exceeds 60%. It shows that small enterprises with these five feature attributes have a great possibility of default.

**Conclusion**

Given the unbalanced characteristics of “more non-default samples and fewer default samples” in credit

evaluation, it is easy to focus on non-default rules in association rule mining and not effectively mine default rules. This paper used SMOTE-NC to generate defaulting small enterprises. It used the APRIORI algorithm to mine small enterprise default feature attributes, and made an empirical analysis with 1,231 small enterprise credit data.

Based on the division of different feature attributes of the same indicator, the model can find out the feature attributes strongly associated with default ( $Y_i = 1$ ) and non-default ( $Y_i = 0$ ). It is suggested that financial institutions should pay attention to the 23 feature attributes strongly associated with default excavated in this paper. Five attributes in particular should be looked at: the return on total assets of small enterprises is in the range  $[0.0000, 0.0344]$ ; the turnover speed of accounts receivable of small enterprises is in the range  $[0.0484, 3.9500]$ ; the enterprise has less than five patents in the same industry; the product sales scope of small enterprises is unclear, or it is not in the two ways of domestic sales and export sales; the personal credit card of the legal representative (or the person in charge of the enterprise) of a small enterprise has a default record or missing data; and the legal person has held the position for less than 2 years. The probability of small enterprises with these five feature attributes and default at the same time is high, that is, the support  $f$  exceeds 60%, which shows those small enterprises have a strong default risk. In addition, among the 23 non-defaulting feature attributes of SMOTE-NC extended data mining, 20 feature attributes are the same as 20 of the 23 feature attributes mined from the original data (the coincidence degree is 87%), indicating that the extended data is highly consistent with the original data mining results. The feature

**Table 8.** The Description of Non-Defaulting and Defaulting Feature Attributes of Small Enterprises for Extended Data.

(1) Indicator name	(2) Non-defaulting feature attributes	(3) Defaulting feature attributes
X <sub>1</sub> Quick ratio	X <sub>1</sub> <sup>2</sup> =[0.8668, 3.7894)	X <sub>1</sub> <sup>3</sup> =[3.7894, 7.3765)
X <sub>2</sub> Overdrive ratio	X <sub>2</sub> <sup>4</sup> =[0.5222, 0.6878)	X <sub>2</sub> <sup>1</sup> =[0.000, 0.136)
X <sub>3</sub> Return on total assets	X <sub>3</sub> <sup>2</sup> =[0.0344, 0.1022)	X <sub>3</sub> <sup>1</sup> =[0.0000, 0.0344)
X <sub>4</sub> Operating profit margin	X <sub>4</sub> <sup>4</sup> =[0.0552, 0.0756)	X <sub>4</sub> <sup>1</sup> =[0.0016, 0.0159)
X <sub>5</sub> Gross profit margin	X <sub>5</sub> <sup>2</sup> =[0.0296, 0.0520)	X <sub>5</sub> <sup>1</sup> =[0.0016, 0.0296)
X <sub>6</sub> Net cash flow from operating activities	X <sub>6</sub> <sup>4</sup> =[-3586873.83, 1780504.17)	X <sub>6</sub> <sup>1</sup> =[-96199357.9424, -55592466.3700)
X <sub>7</sub> Turnover speed of accounts receivable	X <sub>7</sub> <sup>2</sup> =[3.95, 13.56)	X <sub>7</sub> <sup>1</sup> =[0.0484, 3.9500)
X <sub>8</sub> Growth rate of total assets	X <sub>8</sub> <sup>7</sup> =[0.0571, 0.1509)	X <sub>8</sub> <sup>6</sup> =[-0.0304, 0.0571)
X <sub>9</sub> Industry prosperity index	X <sub>9</sub> <sup>5</sup> =[142.40, 150.40)	X <sub>9</sub> <sup>3</sup> =[123.70, 132.00)
X <sub>10</sub> Consumer price index	X <sub>10</sub> <sup>2</sup> =[102.00, 121.40)	X <sub>10</sub> <sup>1</sup> =[0, 102)
X <sub>11</sub> Per capita disposable income of urban residents	X <sub>11</sub> <sup>5</sup> =[18423.08, 20541.00)	X <sub>11</sub> <sup>1</sup> =[3058.00, 11994.38)
X <sub>12</sub> Total value of the legal representative automobile and real estate	X <sub>12</sub> <sup>2</sup> =[4000000, 15000000)	X <sub>12</sub> <sup>1</sup> =[0, 4000000)
X <sub>13</sub> Mortgage and pledge score	X <sub>13</sub> <sup>9</sup> =[0.6850, 0.7449)	X <sub>13</sub> <sup>2</sup> =[0.07, 0.19)
X <sub>14</sub> Years of employment in related industries	X <sub>14</sub> <sup>4</sup> ="Years of employment ≥ 8 years"	X <sub>14</sub> <sup>2</sup> ="2 years ≤ Years of employment < 5 year"
X <sub>15</sub> Patent status	X <sub>15</sub> <sup>2</sup> ="The enterprise has less than 5 patents in the same industry"	X <sub>15</sub> <sup>1</sup> ="The enterprise has no patents or data missing"
X <sub>16</sub> Product sales scope	X <sub>16</sub> <sup>2</sup> ="Product domestic sales"	X <sub>16</sub> <sup>1</sup> ="Other, or data missing"
X <sub>17</sub> Loan default record of the legal representative	X <sub>17</sub> <sup>4</sup> ="With loan record and no default record"	X <sub>17</sub> <sup>1</sup> ="Default record, unsettled, or data missing"
...	...	...
X <sub>23</sub> Number of contract breaches between enterprises	X <sub>23</sub> <sup>4</sup> ="Contract breach between enterprises for 0 times"	X <sub>23</sub> <sup>1</sup> ="Contract defaults for 3 times or more, or data is missing"

mining model established in this paper is accurate and reliable and can be used as a reference for credit risk identification of commercial banks.

The financing of small enterprises has always been a focus of attention for the Chinese government. In 2021, the amount of loans to small enterprises reached 72.11 trillion yuan, an increase of 28.056% compared to 2020 (CBIRC, 2020, 2021). Recently, the People's Bank of China, China Banking and Insurance Regulatory Commission and other departments issued the Notice on Further Strengthening the Support of Deferred Repayment of Principal and Interest for Small and Micro Enterprises Loans, which indicates that for loans to small and micro enterprises (including individual industrial and commercial households and small and micro enterprise owners' operating loans) due in the fourth quarter of 2022, the repayment date can be extended to June 30, 2023 at the longest in principle (The Central Peoples Government of the Peoples Republic of China [CPGPRC], 2022). The increase in the loan amount for small enterprises and the extension of the repayment period undoubtedly increase the default risk

of small enterprises as explained undertaken by financial institutions. Mining the default feature attributes of small enterprises in this paper can help financial institutions identify default small businesses in advance, and enter loan intervention measures in advance, so as to avoid adverse effects on their sustainable development caused by the increased of credit risk of small enterprises.

This study is based on the expansion data of default samples to mine the default feature attributes and non-defaulting feature attributes of small enterprises, which may increase the calculation cost of the model. We would like to continue to improve the feature attribute mining model. This study uses the data of 1,231 small businesses in a commercial bank in China to analyze their default feature attributes and non-default feature attributes. It is hoped that with the improvement of data collection, further tests can be conducted on the robustness of the design model in this paper. In addition, due to the high degree of correlation between the feature attributes and the default status of small enterprises, we will further explore the design of a credit risk evaluation model for small enterprises based on the

feature attributes, in order to overcome the problem of low recognition of default risk of small enterprises caused by the weak correlation between traditional indicators and default status.

## Appendix I. Interval Division Steps of AFCM Algorithm

**Step 1.** Set the initial cluster number  $l^0$ , stop threshold  $\varepsilon$ , fuzzy number  $\omega$ , initial membership matrix  $U^0$ , initial cluster center  $C^0$ , and the adaptive function of cluster number  $l = 1$  as  $L(1)$ .

**Step 2.** Calculate the cluster center  $C^K$ ,  $C^{K+1}$  and membership matrix  $U^K$  according to Equations 1–3.

**Step 3.** Judge whether the stop threshold of Equation 5 is satisfied. If it is satisfied, enter *Step 4*, otherwise, return to *Step 2* and assign a value  $K = K + 1$ .

**Step 4.** Judge whether the constraints of Equation 6 are met. If so, end the iteration, otherwise return to *Step 3* and assign a value  $l = l + 1$ .

## Appendix 2. A Case of Feature Attribute Mining

For ease of understanding, this paper takes a dataset composed of six non-defaulting and four defaulting enterprises as an example to illustrate the feature attribute mining process, as shown in Appendix table 1. Set that indicator X includes three feature attributes  $X^1$ ,  $X^2$ , and  $X^3$ , and the minimum support threshold  $\delta = 1/10$ , minimum confidence threshold  $\zeta = 1/5$ .

**Step 1.** Calculate the support and confidence. it is easy to get from Equation 10, and the support  $f(X^1 \rightarrow 1) = \text{count}(X^1 \& Y = 1) / \text{count}(T) = 3/10$ ,  $f(X^2 \rightarrow 1) = 1/10$  and  $f(X^3 \rightarrow 1) = 0$  corresponding to the three feature attributes  $X^1$ ,  $X^2$  and  $X^3$ ; the confidence  $g(X^1 \rightarrow 1) = \text{count}(X^1 \& Y = 1) / \text{count}(X^1) = 3/5$ ,  $g(X^2 \rightarrow 1) = 1/4$  and  $g(X^3 \rightarrow 1) = 0$ .

**Step 2.** Build frequent itemset  $I$ . Because  $f(X^1 \rightarrow 1) = 3/10 > \delta = 1/10$  and  $f(X^2 \rightarrow 1) = 1/10 = \delta$ , therefore, the frequent itemset  $I$  is composed of  $X^1 \rightarrow 1$  and  $X^2 \rightarrow 1$ , that is,  $I = \{(X^1, 1), (X^2, 1)\}$ .

**Step 3.** Constructing strong association rule set  $Z$ . In frequent itemset  $I$ , because  $g(X^1 \rightarrow 1) = 3/5 > \zeta = 1/5$  and  $g(X^2 \rightarrow 1) = 1/4 > \zeta = 1/5$ , therefore, the strong association rule set  $Z$  is composed of  $X^1 \rightarrow 1$  and  $X^2 \rightarrow 1$ , that is  $Z = \{(X^1, 1), (X^2, 1)\}$ .

**Step 4.** Mining the feature attribute  $X_{r, \max}^b$  with the greatest correlation with default. In the strong association rule set  $Z$ , because  $f(X^1 \rightarrow 1) = 3/10 > f(X^2 \rightarrow 1) = 1/10$ , the feature attribute most related to default is  $X^1$ .

**Appendix Table I.** Example of Correspondence Between Feature Attributes and Default Status for Small Enterprises.

Enterprise number No.	Feature attribute of indicator X	Default status $Y_i$
1	$X^1$	1
2	$X^1$	1
3	$X^1$	1
4	$X^1$	0
5	$X^1$	0
6	$X^2$	1
7	$X^2$	0
8	$X^2$	0
9	$X^2$	0
10	$X^3$	0


## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the National Natural Science Foundation of China (Nos. 72173096, 71873103, 71503199), the Key Project of National Natural Science Foundation of China (No. 71731003), the Social Science Foundation of Shaanxi Province (No. 2018D51). Shi acknowledges financial support from the Tang Scholar Program of Northwest A&F University. Chai acknowledges financial support from the Graduate Science and Technology Innovation Project of College of Economics & Management, Northwest A&F University (JGKC2021-02).

## ORCID iD

Baofeng Shi  <https://orcid.org/0000-0003-1244-5886>

## References

- Agrawal, R., Imieliński, T., & Swami, A. (1993). *Mining association between sets of items in massive database* [Conference session]. International proceedings of the ACM-SIGMOD international conference on management of data (pp. 207–216).
- Alataş, B., & Akin, E. (2006). An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules. *Soft Computing*, 10(3), 230–237.
- Ashofteh, A., & Bravo, J. M. (2021). A conservative approach for online credit scoring. *Expert Systems with Applications*, 176, 114835.
- Bai, C., Shi, B., Liu, F., & Sarkis, J. (2019). Banking credit worthiness: Evaluating the complex relationships. *Omega*, 83, 26–38.

- Belitski, M., Guenther, C., Kritikos, A. S., & Thurik, R. (2022). Economic effects of the COVID-19 pandemic on entrepreneurship and small businesses. *Small Business Economics*, 58(2), 593–609.
- Calabrese, R., Andreeva, G., & Ansell, J. (2019). Birds of a Feather" fail together: exploring the nature of dependency in SME defaults. *Risk Analysis*, 39(1), 71–84.
- Chai, N., Wu, B., Yang, W., & Shi, B. (2019). A multicriteria approach for modeling small enterprise credit rating: evidence from China. *Emerging Markets Finance and Trade*, 55(11), 2523–2543.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1), 321–357.
- China Banking and Insurance Regulatory Commission (CBIRC). (2019). Loans to inclusive small and micro enterprises of banking financial institutions (Legal person) <http://www.cbirc.gov.cn/cn/view/pages/ItemDetail.html?docId=890468&itemId=954&generaltype=0>.
- China Banking and Insurance Regulatory Commission (CBIRC). (2020). *Loans to Inclusive Small and Micro Enterprises of Banking Financial Institutions in 2020*. <http://www.cbirc.gov.cn/cn/view/pages/ItemDetail.html?docId=966736&itemId=954&generaltype=0>.
- China Banking and Insurance Regulatory Commission (CBIRC). (2021). *Loans to Inclusive Small and Micro Enterprises of Banking Financial Institutions in 2021*. <http://www.cbirc.gov.cn/cn/view/pages/ItemDetail.html?docId=1018528&itemId=954&generaltype=0>.
- Freel, M., Carter, S., Tagg, S., & Mason, C. (2012). The latent demand for bank debt: Characterizing “discouraged borrowers”. *Small Business Economics*, 38(4), 399–418.
- Hildebrand, T., Puri, M., & Rocholl, J. (2017). Adverse incentives in crowdfunding. *Management Science*, 63(3), 587–608.
- Hong, J., Tamakloe, R., & Park, D. (2020). Application of association rules mining algorithm for hazardous materials transportation crashes on expressway. *Accident Analysis and Prevention*, 142, 105497.
- Hou, C., Nie, F., Li, X., Yi, D., & Wu, Y. (2014). Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *IEEE Transactions on Cybernetics*, 44(6), 793–804.
- Kabir, M. M. J., Xu, S., Kang, B. H., & Zhao, Z. (2017). A new multiple seeds based genetic algorithm for discovering a set of interesting Boolean association rules. *Expert Systems with Applications*, 74, 55–69.
- Karlan, D., & Zinman, J. (2010). Expanding Credit Access: Using randomized supply decisions to estimate the impacts. *Review of Financial Studies*, 23(1), 433–464.
- Kavšek, B., & Lavrač, N. (2006). APRIORI-SD: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, 20(7), 543–583.
- Krink, T., Paterlini, S., & Resti, A. (2008). The optimal structure of PD buckets. *Journal of Banking & Finance*, 32(10), 2275–2286.
- Lazcorreta, E., Botella, F., & Fernández-Caballero, A. (2008). Towards personalized recommendation by two-step modified APRIORI data mining algorithm. *Expert Systems with Applications*, 35(3), 1422–1429.
- Luo, C. (2020). A comprehensive decision support approach for credit scoring. *Industrial Management & Data Systems*, 120(2), 280–290.
- Mahdi, M. A., Hosny, K. M., & Elhenawy, I. (2022). FR-Tree: A novel rare association rule for big data problem. *Expert Systems with Applications*, 187, 115898.
- Mancisidor, R. A., Kampffmeyer, M., Aas, K., & Jenssen, R. (2020). Deep generative models for reject inference in credit scoring. *Knowledge-Based Systems*, 196, 105758.
- Niu, K., Zhang, Z., Liu, Y., & Li, R. (2020). Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. *Information Sciences*, 536, 120–134.
- Nizaeva, M., & Coskun, A. (2019). Investigating the relationship between financial constraint and growth of SMEs in south Eastern Europe. *Sage Open*, 9, 1–15.
- Pope, D. G., & Sydnor, J. R. (2012). What’s in a picture?: Evidence of discrimination from prosper.com. *Journal of Human Resources*, 46(1), 53–92.
- Rostamkalaei, A., & Freel, M. (2016). The cost of growth: Small firms and the pricing of bank loans. *Small Business Economics*, 46(2), 255–272.
- Sefidian, A. M., & Daneshpour, N. (2019). Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model. *Expert Systems with Applications*, 115, 68–94.
- Shen, F., Zhao, X., Kou, G., & Alsaadi, F. E. (2021). A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Applied Soft Computing*, 98, 106852.
- Shi, B., Chi, G., & Li, W. (2020). Exploring the mismatch between credit ratings and loss-given-default: A credit risk approach. *Economic Modelling*, 85, 420–428.
- Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*, 425, 76–91.
- Sun, J., Lee, Y. C., Li, H., & Huang, Q. H. (2015). Combining B&B-based hybrid feature selection and the imbalance-oriented multiple-classifier ensemble for imbalanced credit risk assessment. *Technological and Economic Development of Economy*, 21(3), 351–378.
- Sun, Y., Chai, N., Dong, Y., & Shi, B. (2022). Assessing and predicting small industrial enterprises’ credit ratings: A fuzzy decision-making approach. *International Journal of Forecasting*, 38(3), 1158–1172.
- Tatavarthi, U. D., & Sambasiva, R. P. (2017). Applicability of APRIORI based association rules on medical data. *International Journal of Applied Engineering Research*, 12(20), 9451–9458.
- The Central People’s Government of the People’s Republic of China (CPGPRC). (2022). Notice on further strengthening the support for the delayed repayment of principal and interest of loans to small and micro enterprises. [http://www.gov.cn/zhengce/zhengceku/2022-11/14/content\\_5726949.htm](http://www.gov.cn/zhengce/zhengceku/2022-11/14/content_5726949.htm).
- Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*, 93, 182–199.
- Xie, D. F., Wang, M. H., & Zhao, X. M. (2020). A spatiotemporal APRIORI approach to capture dynamic associations of regional traffic congestion. *IEEE Access*, 8, 3695–3709.

- Yu, H., Wen, J., Wang, H., & Jun, L. (2011). An improved APRIORI algorithm based on the Boolean matrix and hadoop. *Procedia Engineering*, 15(1), 1827–1831.
- Yu, L., Yu, L., & Yu, K. (2021). A high-dimensionality-trait-driven learning paradigm for high dimensional credit classification. *Financial Innovation*, 7(1), 1–20.
- Zeng, R., Zhang, L., Xiao, Y., Mei, J., Zhou, B., Zhao, H., & Jia, J. (2014). An approach on fault detection in diesel engine by using symmetrical polar coordinates and image recognition. *Advances in Mechanical Engineering*, 6, 1–9.
- Zhang, C., & Hu, N. (2020). A new method for computing letter of credit risks. *Sage Open*, 10, 1–11.
- Zhang, Y., Xing, C., & Tripe, D. (2021). Redistribution of China's green credit policy among environment-friendly manufacturing firms of various sizes: Do banks value small and medium-sized enterprises? *International Journal of Environmental Research and Public Health*, 18(1), 33.
- Zhou, Y., Uddin, M. S., Habib, T., Chi, G., & Yuan, K. (2021). Feature selection in credit risk modeling: An international evidence. *Economic Research-Ekonomska Istrazivanja*, 34(1), 3064–3091.