A big data analytics method for assessing creditworthiness of SMEs

Fuzzy equifinality relationships analysis

Citation for published version:
Shi, B, Bai, C & Dong, Y 2024, 'A big data analytics method for assessing creditworthiness of SMEs: Fuzzy equifinality relationships analysis', *Annals of Operations Research*. https://doi.org/10.1007/s10479-024-06054-w

Digital Object Identifier (DOI):
10.1007/s10479-024-06054-w

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Annals of Operations Research

A Big Data Analytics Method for Assessing Creditworthiness of SMEs: Fuzzy Equifinality Relationships Analysis

Baofeng Shi

College of Economics & Management

Northwest A&F University

No.3 Taicheng Road, Yangling, Shaanxi, 712100, P.R. China

Tel: 86-15388615466

E-mail: shibaofeng@nwsuaf.edu.cn

Chunguang Bai* (***Corresponding Author***)

School of Management and Economics

University of Electronic Science and Technology of China

No. 2006, Xiyuan Avenue, Chengdu, 611731, P.R. China

Tel: 86-13664228458

E-mail: Cbai@uestc.edu.cn

Yizhe Dong

University of Edinburgh Business School

University of Edinburgh

29 Buccleuch Place, Edinburgh, EH8 9JS, UK

Tel: 44 (0) 131 651 1056

E-mail: yizhe.dong@ed.ac.uk

A Big Data Analytics Method for Assessing Creditworthiness of SMEs: Fuzzy Equifinality Relationships Analysis

Abstract: Nowadays, many financial institutions are beginning to use Big Data Analytics (BDA) to help them make better credit underwriting decisions, especially for small and medium-sized enterprises (SMEs) with limited financial histories and other information. The various complexities and the equifinality problem of Big Data make it difficult to apply traditionalstatistical techniques to creditworthiness evaluation, or credit scoring. In this study, we extend the existing research in the field of creditworthiness assessment and propose a novel approach based on neighborhood rough sets (NRSs), to evaluate and investigate the complexities and fuzzy equifinality relationships in the presence of Big Data. We utilize a real SME loan dataset from a Chinese commercial bank to generate interval number rules that provide insight into the fuzzy equifinality relationships between borrowers' demographic information, company financial ratios, loan characteristics, other non-financial information, local macroeconomic indicators and rated creditworthiness level. In addition, the interval number rules are used to predict creditworthiness levels based on test data and the accuracy of the prediction is found to be 75.44%. One of the major advantages of using the proposed BDA approach is that it helps us to reduce complexity and identify equivalence relationships when using Big Data to assess the creditworthiness of SMEs. This study also provides important implications for practices in financial institutions and SMEs.

Keywords: Big Data Analytics; Credit Risk; Equifinality; Rough Set; Small and Medium-sized Enterprises

Acknowledgments: The research was supported by the the Major Program of National Social Science Foundation of China (No. 23&ZD175), the National Natural Science Foundation of China (Nos. 72173096, 72072021, 71873103)and the Credit Rating and Loan Pricing Project for Small Enterprise of the Bank of Dalian (No. 2012-01). Shi acknowledges financial support from the Tang Scholar Program of Northwest A&F University (No. 2021-04).

1 Introduction

Assessing and predicting credit risk is one of the most important topics for financial institutions and regulators (Medina-Olivares *et al.*, 2022; Lu *et al.*, 2022). The process of credit risk assessment is complex and unstructured. Many researchers and practitioners have developed statistical models for transforming relevant data into numerical measures so as to discriminate "good" from "bad" loan applications (Thomas *et al.*, 2002). Historically, traditional statistical techniques such as multivariate discriminate analysis (Lessmann *et al.*, 2015; D'Amato and Mastrolia, 2022), probit and logistic regressions (Altman, 1998) and multidimensional scaling (Molinero *et al.*, 1996) have been the most widely used methods for constructing rating systems. The more recent developments in credit scoring are oriented towards adapting artificial intelligence techniques, including neural networks (Artem and Stefan, 2017; Abedin *et al.*, 2018; Fombellida *et al.*, 2019), case-based reasoning (Cao and Zhai, 2022), gradient boosting (He *et al.*, 2018), support vector machines (SVM) (Yao *et al.*, 2017; Abedin *et al.*, 2019) and Bayes networks (Xia *et al.*, 2017; Billie, 2020). Although large efforts have been made to construct a credit scoring system, modeling credit risk for small and medium-sized enterprises (SMEs) is more difficult than for large companies because of insufficient or unverifiable information (Calabrese *et al.*, 2019; Bagale *et al.*, 2023; Li and He, 2023).

SMEs represent about 90% of businesses and more than 50% of employment worldwide according to the World Bank[1]. SMEs are generally considered the backbone of the economy in many countries. For example, by the end of 2021, SMEs in China had contributed more than 50% of tax revenue, more than 60% of GDP, more than 70% of technological innovation and more than 80% of urban labor employment, playing an important role in promoting economic and employment growth, and stimulating innovation vitality (Guangming Tech, 2022). Similarly, at the start of 2019, SMEs accounted for three-fifths of the employment and around half of the turnover in the UK private sector. Due to a justified tendency for prudential lending and the lack of elaborate SME default and creditworthiness prediction models, SMEs still face challenges in accessing finance from banks, which is a very important factor constraining the

[1] https://www.worldbank.org/en/topic/smefinance.

growth of the economy. One of the immediate economic consequences of the coronavirus pandemic is the sudden lack of liquidity affecting small and medium-sized businesses. Gaining access to the financial support they need to survive and develop, and having a better understanding of their credit risk drivers, has become even more important during (and will continue to be after) the coronavirus pandemic. Several studies have attempted to model credit risk for SMEs using loan data from different countries and regions, such as the United States (Altman & Sabato, 2007), Italy (Angilella & Mazzu, 2015), European countries (Pederzoli *et al.*, 2016), Brazil (Fernandes & Artes, 2016), China (Li *et al.*, 2016) and the United Kingdom (Calabrese *et al.*, 2019). However, there is still limited knowledge of the characteristics that determine SMEs' default behavior, particularly for those in emerging markets.

The digital age and new technologies have fostered a data explosion which is transforming the credit scoring industry. Financial institutions have begun to explore an "all data is credit data" approach, combining conventional credit information with thousands of data points collected or mined from various resources. Although more data are now available than ever before, in the absence of adequate approaches, financial institutions often face a myriad of challenges regarding how to extract and analyze information from Big Data, and make the smartest credit-granting decisions possible.

Complexity and equivalence in Big Data make the traditional approaches of creditworthiness evaluation, such as statistical techniques, difficult to apply (Hooman *et al.*, 2016). The complexity of Big Data refers to the fact that creditworthiness evaluation involves analyzing borrowers' demographic characteristics, financial information, credit history, public records, environmental and regulatory information, and digital footprints, involving large volumes, high velocities and several varieties of data, and their interactions (Bai *et al.*, 2019a).

An equifinality relationship is present when a system can reach the same outcome from different initial conditions and by a variety of different paths (Gresov & Drazin, 1997; Bai *et al.*, 2019a). Equifinality relationships exist in many fields, such as society (Verleye, 2019), the economy (Dahms, 2019), information technology (De Guinea & Raymond, 2020), creditworthiness assessment and so on. With the accumulation of customer data, credit data is gradually presenting the characteristics of Big Data. It not only includes structured data such as the financial information of SMEs, but also unstructured data such as demographic characteristics, the status of legal disputes, tax records, external political, economic, social, technological, environmental and competitive factors, and digital footprints. The conditional factors are not only related to the predicted credit risk outcome (i.e. the creditworthiness level), but also to each other. In exploring the uncertainty and diversity in the relationship between the conditional factors and the outcome, fuzzy equifinality analysis is a powerful tool. Hence, creditworthiness assessment needs to consider the complexity and equivalence, based on the various interactions and interdependencies amongst that increased number of conditional factors (Bai & Sarkis, 2018).

In recent years, BDA has attracted significant attention from academia and practitioners (Einav & Levin, 2014; Baru, 2018; Wamba *et al.*, 2018; Óskarsdóttir *et al.*, 2019)2. It could be a powerful tool that could help financial institutions make better credit underwriting decisions and more accurately predict SMEs' creditworthiness (Kshetri, 2016). However, despite increasing efforts to use data-driven approaches with BDA to aid decision making, theory-driven research aimed at understanding the equifinality relationships between different types of conditional factors and the outcome is still limited. The equifinality relationships can effectively solve the problem of the interpretability of creditworthiness assessment results, which is one of the main challenges of BDA or artificial intelligence. Very few studies consider whether the combined effect of varying levels of conditional factors, when uniquely bundled together, results in differing or similar levels of performance outcomes in the context of Big Data. We need to build a new theory-driven approach to mining those equivalence relationships, and thereby explain some of the complex phenomena and unanswered questions within the context of Big Data. Therefore, in this study, we seek to contribute to the operational research and Big Data literature by proposing a novel BDA approach which not only aims to reduce complexity and find equifinality relationships in Big Data, but at the same time predict the creditworthiness of new SMEs through those identified fuzzy equifinality relationships (Bai *et al.*, 2021).

[2] BDA involves the use of advanced analytic approaches (or techniques) which, in the processing of data with the "4Vs: high volume, high velocity, high veracity and high variety", can help the user to extract meaningful and useful knowledge, and facilitate data-driven decision making (Wamba *et al*., 2017; Zhan & Tan, 2020).

Rough set theory (RST) is a valuable tool that could help deal with the complexity and equifinality that occur in BDA research (Pawlak, 1982; Rajesh, 2022). The various rules that are the main results of RST could help identify potentially useful knowledge through BDA to address the complexity and equifinality concerns (Bai & Sarkis, 2018; Bai *et al.*, 2019b). However, the larger number of rules and crisp number rules are a formidable limitation when it comes to reducing complexity and finding equifinality in Big Data. Hence, we first introduce the neighborhood rough set (NRS) to reduce the complexity of SMEs' conditional factors and eliminate redundant information (Hu *et al.*, 2008). We then develop a novel approach based on the NRS to arrive at interval number rules that can be used to identify the various fuzzy equifinality relationships between the conditional factors (e.g. demographic characteristics, financial structure and performance information, non-financial information and economic environment factors) and creditworthiness levels. To the best of our knowledge, this is the first time that NRSs have been developed for investigating fuzzy equifinality relationships so as to predict the creditworthiness of SMEs. These interval number rules will help financial institutions and governmental agencies to predict the creditworthiness of SMEs in a complex and uncertain environment. To test the proposed approach, we use a dataset of SME loans granted by a Chinese commercial bank to identify fuzzy equifinality relationships regarding the creditworthiness of SMEs. This credit risk assessment approach based on Big Data can provide valuable insights for policy makers and practitioners regarding the management of broader information, such as non-financial information and social context.

The remainder of the paper is organized as follows. In Section 2, we conduct a literature review on credit risk assessment and prediction for SMEs, and highlight the contributions of our study. In Section 3, we develop a novel approach based on NRSs. In Section 4, the proposed approach is applied to real loan data to evaluate SMEs' creditworthiness from an equifinality perspective, and the experimental results are presented. Section 5 offers a discussion on the advantages of our proposed method. Section 6 highlights the theoretical, methodological and practical implications, based on the research results. The final section concludes, with limitations and future research directions.

2 Literature review

SMEs play an important role in the promotion of economic development and the building of modern economic systems (Altman and Sabato, 2007; Bai *et al.*, 2021). However, since the global financial crisis, financial intermediaries become more risk-averse and limit their credit exposures by avoiding credits to small businesses with limited available credit history, which in turn significantly increases the financing difficulty of SMEs and restricts their growth and development (Cowling *et al.*, 2012). Such problems are more pronounced after the outbreak of the coronavirus pandemic. One of the effective ways to address the financing problems of SMEs is to develop appropriate methods for financial intermediaries to accurately assess and predict the credit risk of SMEs when they make lending decisions. Therefore, financial institutions around the world have developed various credit risk assessment models to facilitate their lending decision marking. For example, to assess the creditworthiness of their business clients. many banks in the US employ Z-score and logistic regression models (Altman *et al.*, 2007), some Chinese banks, including Industrial and Commercial Bank of China (Sun *et al.*, 2022), Postal Savings Bank of China (Shi *et al.*, 2020) and Dalian Bank (Chai *et al.*, 2019), adopt the credit rating approach, and some banks in Italy apply multivariate linear discriminant analysis (MLDA) (D'Amato and Mastrolia, 2022). These models are mainly aimed at assessing the credit risks of large companies with greater available credit history and may not be very suitable for assessing the creditworthiness of SMEs as they tend to be informal, and have less available information.

Therefore, scholars have started to make efforts to address the key challenges in assessing SME credit risk from different perspectives. Early studies focused on the difficulties of financing SMEs, investigating the key factors of their profitability and credit risk (Martinez-Sola *et al.*, 2014). However, incomplete financial information (Van *et al.*, 2012), short operating histories, poorer performance, and high external environmental sensitivity, all of which result in higher information asymmetries and risk, and restrict most SMEs from obtaining financing through financial institutions. Recent literature has begun to focus on how to build methods for assessing and predicting the credit risk or creditworthiness of SMEs with limited information, so as to help banks to make wise lending decisions and minimize their losses. Some scholars use traditional statistical methods, such as probit and logistic regressions, multivariate discriminate analysis, survival analysis, spatial analysis, multi-objective optimization, and multidimensional scaling, to assess the credit risk of SMEs. For example, Altman *et al.*, (2007) and Ciampi (2015) use a logistic regression model for predicting SMEs' credit risk. Glennon and Nigro (2005) employ a discrete-time hazard procedure, a survival analysis technique, to take the effect of change in economic conditions over time into account in the SME credit risk assessment. Fernandes and Artes (2016) and Medina-Olivares (2022) develop credit risk models for small businesses by incorporating spatial dependence among firms. Kou *et al.* (2021) propose a credit risk prediction model based on a two-stage multi-objective feature selection that optimizes the number of features and improves model classification performance.

With the arrival of the fourth industrial revolution, many new technologies and approaches have emerged, such as Big Data, artificial intelligence, etc (Chen & Zhang, 2014). Scholars attempt to provide a better prediction on SME credit risk and default by using machine learning approaches. For example, Fantazzini and Figini (2009) propose a random survival forests model for modeling the credit risk of SMEs and find that their approach has a better prediction performance than the classical logit model. Kim *et al.* (2020) develop a deep learning model to support credit risk assessment decisions for SMEs, and demonstrate the superiority of deep learning over machine learning and rule-based benchmarks. Lu *et al.* (2022) propose a credit risk feature selection approach that integrates the binary opposite whale optimization algorithm (BOWOA) and the Kolmogorov–Smirnov (KS) statistic to improve the predictive performance of SMEs' credit risk. Zhao and Li (2022) use the SVM combined with back propagation neural network to predict the credit risk of SMEs. However, these models still can't fully address the SME finance problem which we discussed previously. SME information problem data are often insufficient for reliable assessment and prediction.

As data is growing exponentially in recent years, BDA has been attracting increasing attention and is considered as an efficient strategy for achieving high-performance prediction.

Liu *et al.* (2019) propose a Big-Data-driven credit assessment framework for SMEs, highlighting the combination of financial and non-financial data, including Big Data from businesses, governments, social media and networks. Stevenson *et al.* (2021) employ a deep learning approach to predict SME loan defaults based on both the text and structured data. They show that the text data influences credit default predictions and suggest that the textual loan assessment can be a new strategy for mitigating the limited data availability of SMEs. Wang *et al.* (2020) establish a credit assessment indicator system by incorporating the "online" specific indicators of e-commerce platforms. Then a nonlinear LS-SVM model combining with BDA is constructed to evaluate the credit risk of SMEs in the automobile manufacturing industry. These studies are mainly aimed at internet, e-Business and digital platform financing (Guo *et al.*, 2023), because it is easier to obtain relevant data in such situations.

Most of these existing studies focus on the improvement in the prediction accuracy of credit risk using some new or modified risk assessment models or multi-source heterogeneous data, while largely ignoring the process of the assessments. Our approach is different from theirs in the following three aspects. First, although we are in the era of Industry 4.0, few studies have used BDA for credit risk assessment based on large-scale and high-dimensional datasets. Moreover, most studies ignore the interpretability of the results and the identification of key indicators for credit assessment. We pay more attention to reducing the complexity of the data so as to focus on key determinants of credit default risk. By identifying key indicators, financial institutions can more easily collect relevant data at the lowest cost to complete their credit assessments. Second, although our ultimate goal is the creditworthiness assessment of SMEs, we also pay attention to the interpretability of the assessment results through equivalence relationships. With the transparency achieved through interpretability, financial institutions and enterprises can truly understand the reasons for their assessment results, improving access to finance for SMEs. Third, we build a set of fuzzy rules to help financial institutions assess the creditworthiness of SMEs. Although some rule-based methods are already used in credit assessment (Bai *et al.*, 2021), the rules are too strict. Therefore, a large number of rules are inapplicable for the interpretability of the assessment results. All in all, we need to build a new theory-driven approach to reduce the dimensionality of the data and mine the equivalence relationships, thereby predicting the creditworthiness of new SMEs through the identified fuzzy equifinality relationships and explaining some of the complex phenomena involved.

3 Methodology

In this section, we develop a multistep, novel approach which can be used to investigate the fuzzy equifinality relationships in the Big Data context, using NRSs (Appendix 1 provides detailed information about NRSs).

Step 1: Develop an adjusted neighborhood decision system

The Big Data generated from various types of information include both qualitative and quantitative data. We first need to preprocess these collected data and integrate them into a two-dimensional neighborhood decision system, *NDS*. Let *NDS* = 〈*U*, *C*, *D*, *V*, *W*〉 be a decision table or decision system, where *U* is a non-empty set of finite objects, usually called the universe. *C* is a non-empty finite set of conditional factors for the objects. *D* is a set of decision factors for the objects. *V* and *W* are the set of conditional and decision factor values. v_{xa} and w_{xd} respectively denote the values of conditional factor *a* and decision factor *d* for object *x*.

Second, we need to normalize those data for consistency, such that the values of each factor use similar scales. Some categorical data, such as the industry category, have no fixed order and do not need to be normalized. To complete this normalization of the values, we rely on expression (1) for the standardization values (Lu *et al.*, 2022)

$$\tilde{v}_{ij} = \frac{v_{ij} - v_j^{\min}}{v_j^{\max} - v_j^{\min}} \tag{1}$$

where $\tilde{v}_{ij}$ is the normalized value for object *i*, conditional factor *j*; $v_j^{\max}$ is the maximum value of factor *j*; $v_j^{\min}$ is the minimum value of factor *j*. This normalization will adjust all the standardization values (v_{ij}) such that they are normalized, with $0 \leqslant \tilde{v}_{ij} \leqslant 1$.

Step 2: Identify the core conditional factor set

Initially, let $Atr = \varnothing$, which is the core conditional factor set (reduct). We will select some conditional factors a_k with information significance in the core conditional factor set *Atr* through the following sub-steps:

Sub-step 1: Calculate the distances between each of the objects on a conditional factor a_j, as shown in expression (2):

$$\Delta_j(x_i, x_k) = |\tilde{v}_{ij} - \tilde{v}_{kj}| \tag{2}$$

Sub-step 2: Compute the conditional neighborhood members $\delta_j(x_i)$ for each object x_i based on a given neighborhood size δ and a conditional factor a_j. The $\delta_j(x_i)$ is defined as the indistinguishable objects set with an object x_i for the similarity value of a conditional factor a_j under distance threshold δ:

$$\delta_j(x_i) = \{x_k \mid x_k \in U, \Delta_j(x_i, x_k) \le \delta\} \tag{3}$$

Sub-step 3: Compute the decision neighborhood members $\delta_j^D(x_i)$ for each object x_i based on a conditional factor a_j, decision factor *D* and given neighborhood size δ. $\delta_B^D(x_i)$ is defined as the indistinguishable objects set with an object x_i for the similarity value of a conditional factor a_j under distance threshold δ and the same value of decision factor *D*:

$$\delta_j^D(x_i) = \{x_k \mid x_k \in U, \Delta_j(x_i, x_k) \le \delta \; and \; D(x_i) = D(x_k)\} \tag{4}$$

Sub-step 4: Determine the lower NRS $POS_j^k(D)$ based on the inclusion measures and the selected inclusion threshold value *k* using expression (5):

$$POS_j^k(D) = \{x_i \mid I(\delta_j(x_i), D(x_i)) \ge k, x_i \in U\} \tag{5}$$

where the measure of inclusion is $I(\delta_j(x_i), D(x_i)) = \dfrac{|\delta_j^D(x_i)|}{|\delta_j(x_i)|}$, $where\ \delta_j(x_i) \ne \varnothing.$ |*| is the cardinality of a set *.

Sub-step 5: Determine the information significance $Sig(a_j, D)$ of a conditional factor

a_j. The information significance is the amount of influence a conditional factor has on a decision outcome (*D*). To calculate the information significance of a conditional attribute a_j with respect to decision level *D*, we use expression (6):

$$Sig(a_j, D) = \frac{|POS_j^k(D)|}{|U|} \tag{6}$$

Sub-step 6: Select the significance conditional factors and update the core conditional factor set *Atr*. We select the conditional factor a_k into core conditional factor set *Atr* if it satisfies $SIG(a_k, D) > \varepsilon$; then $Atr \bigcup a_k -> Atr$. ε is a positive infinitesimal real number used to control the convergence of *Atr*.

Step 3: Identify fuzzy equifinality relationships (interval number rules)

In this step, we develop a series of interval number rules to discern the fuzzy equifinality relationships based on the reduct *Atr* determined in Step 2. Rough set approaches regard any decision table as a set of generalized decision crisp number rules of the form

$$\wedge(g,v) \to \vee(d,w), \text{ where } g \in Atr, d \in D, v \in V_g, w \in W_D, D \notin Atr \tag{7}$$

where $\wedge(g,v)$ is called the condition or premise, and *v* represents the value of the reduct factor *g* and belongs to the set *V*. The outcome or conclusion is represented by $\vee(D,w)$, where *w* represents the value of outcome factor *D* and belongs to the set *W*. $\wedge$ and $\vee$ are Boolean notation for “and” and “or”, respectively.

The basic idea of interval number rules is to divide the multidimensional data space clearly using those fuzzy rules based on a fixed reliability degree. In other words, the proportion of objects in the interval area of rule division belonging to one decision level is greater than a fixed reliability degree. We set this fixed reliability to 0.9.

Sub-step 1: Calculate the distances between each of the objects in the reduct *Atr* as shown in expression (8):

$$\Delta_{Atr}(x_i, x_k) = \max_{j \in Atr} |\tilde{v}_{ij} - \tilde{v}_{kj}| \tag{8}$$

Sub-step 2: Compute the measure of inclusion $I(\delta_{Atr}(x_i), D(x_i))$ for each object x_i

using sub-steps 2 to 4 of Step 2 based on reduct set *Atr*.

Sub-step 3: Identify various initial rules. First, identify the interval number rules with the object x_i such that the measure $I(\delta_{Atr}(x_i), D(x_i))$ of inclusion (also called the fixed reliability degree) is greater than 0.9:

$$\wedge(\underline{r}_{ij} = \tilde{v}_{ij} - \delta, \overline{r}_{ij} = \tilde{v}_{ij} + \delta) \rightarrow \vee(D(x_i)) \tag{9}$$

Second, identify the interval number rules with the object x_i such that the measure $I(\delta_{Atr}(x_i), D(x_i))$ of inclusion is smaller than 0.9:

$$\begin{gathered}\wedge(\underline{r}_{it}, \overline{r}_{it}) = \begin{cases} (\frac{\tilde{v}_{it} + \tilde{v}_{kt}}{2}, \tilde{v}_{it} + \delta) & \tilde{v}_{kt} \leq \tilde{v}_{it} \\ (\tilde{v}_{it} - \delta, \frac{\tilde{v}_{it} + \tilde{v}_{kt}}{2}) & \tilde{v}_{kt} \geq \tilde{v}_{it} \end{cases} \rightarrow \vee(D(x_i)) \\ for \begin{cases} x_k \in \delta(x_i) \;\; D(x_i) \neq D(x_k) \\ t \in \max(\tilde{v}_{kt} - \tilde{v}_{it}) \end{cases}\end{gathered} \tag{10}$$

where $x_k \in \delta(x_i)$ is a member of the neighborhood of object x_i based on similarity or indistinguishability relationships, but having a different creditworthiness level. The main purpose of $(\frac{\tilde{v}_{it} + \tilde{v}_{kt}}{2}, \tilde{v}_{it} + \delta)$ $(\tilde{v}_{it} - \delta, \frac{\tilde{v}_{it} + \tilde{v}_{kt}}{2})$ is to narrow the scope of the interval number rules, so that the reliability degree of the objects within the rules is greater than the threshold 0.9. $t \in \max(\tilde{v}_{kt} - \tilde{v}_{it})$ is used to identify the biggest gap factor between objects x_i and x_k.

Sub-step 3: Reduce the number of rules. Integrate the interval number rules in a neighborhood such that they have the same decision value:

$$\begin{gathered}\wedge(\min(\underline{r}_{ij}, \underline{r}_{kj}), \max(\overline{r}_{ij}, \overline{r}_{kj})) \rightarrow \vee(D(x_i)) \\ for \;\; x_k \in \delta(x_i) \;\; D(x_i) = D(x_k) \;\; I(\delta_{Atr}(x_k), D(x_k)) \geq 0.9\end{gathered} \tag{11}$$

If $I(\delta_{Atr}(r_i^*), D(x_i)) \geq 0.9$, then retain the new integrated rule r_i^* and delete other rules $r_k \;\; for \;\; x_k \in \delta(x_i) \;\; D(x_i) = D(x_k) \;\; I(\delta_{Atr}(x_k), D(x_k)) \geq 0.9$. $\delta_{Atr}(r_i^*)$ represents a set of objects within an interval number rule r_i area.

If $I(\delta_{Atr}(r_i), D(x_i)) < 0.9$, do not update the interval number rule r_i, and retain other rules r_k.

Step 4: Use interval number rules to predict the outcome

In this step, the decision value of a new object is predicted using the interval number rules developed in Step 3.

Sub-step 1: Predict the decision value of a new object x_p which meets rule r_i:

$$D(x_p) = \{D(r_i) \mid (\underline{r_{ij}} \leq \tilde{v}_{pj} \leq \overline{r_{ij}}) \ \forall j \in Atr\} \tag{12}$$

Sub-step 2: Predict the decision value of a new object x_p which does not meet any rules. Then, calculate the distance between the object x_p and each rule:

$$\Delta_{Atr}(r_i, x_p) = \max_{j \in Atr} | r_{ij} - \tilde{v}_{pj} | \tag{13}$$

where $\underline{r_{ij}} \leq \tilde{v}_{pj} \leq \overline{r_{ij}}$, $|r_{ij} - \tilde{v}_{pj}|$=0; $\tilde{v}_{pj} \leq \underline{r_{ij}}$, $|r_{ij} - \tilde{v}_{pj}|$=$\underline{r_{ij}}$-$\tilde{v}_{pj}$; $\tilde{v}_{pj} \geq \overline{r_{ij}}$, $|r_{ij} - \tilde{v}_{pj}|$=$\tilde{v}_{pj}$-$\overline{r_{ij}}$.

Sub-step 3: Compute the neighborhood members $\delta_{Atr}(x_p)$ for each rule r_i based on a given neighborhood size δ and core conditional factors *Atr*:

$$\delta_{Atr}(x_p) = \{r_i \mid r_i \in R, \Delta_{Atr}(r_i, x_p) \leq \delta\} \tag{14}$$

Sub-step 4: Calculate the satisfaction degree of the predicted object for each rule r_i that belongs to the neighborhood members $\delta_{Atr}(x_p)$:

$$u_{pi} = \left[\sum_{k \in \delta(x_p)} \left(\frac{\Delta_{Atr}(r_i, x_p)}{\Delta_{Atr}(r_k, x_p)}\right)^2 \right]^{-1} \tag{15}$$

where u_{pi} indicates the degree of association or membership function of object x_p with rule r_i.

Sub-step 5: Summarize the satisfaction degree of each decision value for a new object x_p, and then use the maximum satisfaction degree to determine the decision value of object x_p:

$$D(x_p) = \{w \mid w = \max(\sum_{D(r) \in W_1} u_{pr}, \ldots \sum_{D(r) \in W_w} u_{pr})\} \tag{16}$$

4 Experimental analysis

4.1 Real credit dataset

For our experiments, we use a dataset composed of loans granted to SMEs by a Chinese city commercial bank between 1998 and 2013. We apply the definition of SMEs proposed by the Ministry of Industry and Information Technology of the People's Republic of China[3]. The bank was established in March 1998. By the end of 2022, it had ten branches (167 business outlets) in ten cities, with more than 5,000 employees and total assets of 472.1 billion Chinese yuan (approximately 66.5 billion US dollars based on 7.1 yuan equaling approximately one US dollar). The bank aims to serve local economic development and offer financing services to SMEs. The characteristics of the dataset are as follows:

- We initially obtain more than 5,000 loan observations from the bank, then exclude loans for which more than 30% of the variable values are missing. The final sample consists of 3,111 loans granted to SMEs. Each loan is described by 83 variables, such as company financial ratios, company and owner demographic descriptors, loan characteristics, payment and public records, other non-financial information, local macroeconomic indicators and the creditworthiness level as rated by the bank.
- The dataset has an average loan value of 3.5 million yuan with a mean term of eight months. The maximum loan amount is 60 million yuan and the minimum loan amount is 1,000 yuan. The maximum loan term is sixty months (i.e. five years), and the minimum is just one month. The sample is rather diversified and distributed across 12 industries. The characteristics of the dataset and related variables used in evaluating SME creditworthiness are given below in Table 1 and detailed definitions of the variables are presented in Appendix 2.

(Insert Table 1 about here)

4.2 Experimental setup and results

[3] http://www.gov.cn/zwgk/2011-07/04/content_1898747.htm.

We apply the aforementioned proposed methodology to our loan dataset to generate interval number rules that provide insight into the fuzzy equifinality relationships between SMEs' demographic information, financial ratios, other non-financial information, macroeconomic indicators and creditworthiness level.

Step 1: Develop adjusted neighborhood decision system

We initially collected nearly 500,000 pieces of data through financial institutions, the National Bureau of Statistics of China, the Internet and other sources. We preprocess all of the collected data and integrate them into a two-dimensional neighborhood decision system, *NDS* = $\langle U, C, D, V, W\rangle$. $U = \{x_i \mid 1, \mathrm{K}, 3111\}$ is a set of 3,111 SMEs. $C = \{c_j \mid 1, \mathrm{K}, 83\}$ contains 83 conditional factors for each SME, as defined in Appendix 2. The conditional factors include four categories: loan characteristics, financial ratios, non-financial information and macroeconomic indicators. $D=\{d\}$ is a decision factor (creditworthiness level) for each SME. A number of experts were working alongside loan officers in the financial institution to evaluate SMEs and determine their creditworthiness levels, or *D*[4]. *D* takes one of four levels (1=very low creditworthiness, 2=low creditworthiness, 3=high creditworthiness and 4=very high creditworthiness). Next, we normalize the data using expression (1) to make sure the data for each input lie within the same interval. Thus, all the factor values (v_{ij}) are normalized to within [0, 1].

Step 2: Identify the core conditional factor set

The normalized data are used to reduce the conditional factors by means of expressions (2)-(7) to those that will be used to generate the core conditional factor set *Atr* of SME loan characteristics, financial ratios, non-financial information and macroeconomic indicators. We set the neighborhood size δ = 0.1 to control neighborhood members, the inclusion threshold value k=0.6 to determine the lower NRS and ε=0.01 to control the convergence of *Atr*. The final core conditional factor set is *Atr*={*Principal, Quick asset ratio, Capitalization rate, Superquick asset ratio, Long-term asset suitability rate, Net profit divided by total operating costs and expenses, Total profit divided by total operating costs and expenses, Inventory*

[4] Inputs from 17 experts at the China Banking Regulatory Commission (CBRC), the Postal Savings Bank of China (PSBC) and the city branch of the CBRC were utilized to determine the creditworthiness level.

turnover velocity, Velocity of fixed assets, Velocity of equity, Loans from bank divided by company's total bank loans, Education background, Automobile and real estate, Monthly family income}. The information significance level of each core conditional factor is presented in Table 2.

(Insert Table 2 about here)

Step 3: Identifying fuzzy equifinality relationships

We identify a series of interval number rules to discern the fuzzy equifinality relationships between SMEs' core conditional factors and creditworthiness levels. A total of 2,058 interval number rules are generated. Table 3 presents some of the generated interval number rules. For example, the first row in Table 3 shows: ($Atr1$ = [0, 0.102]) $\wedge$ ($Atr2$ = [0.582, 0.782]) $\wedge$ ($Atr3$ = [0, 0.1]) $\wedge$ … $\wedge$ ($Atr14$ = [0, 0.121]) → (*Creditworthiness* =4). This interval number rule indicates that, if an SME's *Principal* factor ($Atr1$) belongs to [0, 0.102], *Quick asset ratio* factor ($Atr2$) belongs to [0.582, 0.782], …, and *Monthly family income* belongs to [0, 0.121], then this SME's creditworthiness will be very high.

(Insert Table 3 about here)

Step 4: Use intervals to predict outcome

We randomly select 11% of the new SMEs from the whole sample and use the interval number rule to make predictions of their creditworthiness. The creditworthiness levels of 342 new SMEs are predicted through the 2,058 interval number rules obtained in Step 3. Let us take the first new SME as an example to show how the creditworthiness of a SME is predicted. The first new SME does not satisfy any rules, but falls into the neighborhood of 122 rules. We use expressions (14)-(17) to predict the creditworthiness level y_1 of SME, and find that $\sum_{D(r)=0.25} u_{pr} = 3.73\%$, $\sum_{D(r)=0.5} u_{pr} = 48.99\%$, $\sum_{D(r)=0.75} u_{pr} = 32.17\%$ and $\sum_{D(r)=1} u_{pr} = 15.11\%$. Then, we predict that the creditworthiness level y_1 of SME is equal to 2 (i.e. it has a low creditworthiness level).

The creditworthiness levels of the other 341 new SMEs in the predictive set are determined by the same process. The distribution of the predicted accuracy for each creditworthiness level is shown in Table 4. The overall predicted accuracy is 75.44%, reflecting the proportion of times out of 342 that the predicted creditworthiness level was the same as the expert-rated creditworthiness level.

(Insert Table 4 about here)

5 Discussion

When compared to traditional statistical techniques, the complexity of the relationships (interactions among factors), absence of parametric requirements and many equifinality relationships are all advantages of this methodology.

5.1 Fuzzy equifinality relationships in Big Data

Big Data makes us more aware that our world is diverse. A system or object can reach the same outcome from different initial conditions and by a variety of different paths, which is termed equifinality (Gresov & Drazin, 1997). Hence, identifying equifinality relationships is a major challenge in the Big Data context, especially given the uncertainty and complexity of the relationship between the conditional factors and outcomes (Bai *et al.*, 2016). We introduce a novel approach that is a viable solution for identifying those equivalence relationships by means of interval number rules. The generated intervals provide three insights into the equifinality relationships between the creditworthiness and various conditional factors of SMEs.

First, these rules show a number of different paths by which SMEs can reach the same creditworthiness level from various initial conditional factors. In all, 2,058 interval number rules are generated. Our findings have three main implications. (1) Our approach can help SMEs to come up with their own development strategies or plans to improve their creditworthiness based on their own resources. For example, these rules can help determine the locus of investment and effort required to boost specific factors that will help an SME achieve creditworthiness improvements. This provides more possibilities for the development of enterprises. (2) It can also be used as a reference for financial institutions, to specify relevant processes and policies for the evaluation of SMEs' creditworthiness levels. For example, it indicates the kind of data a financial institution should collect and rely on. Financial institutions should not only focus on the values of the factors, but also on their combination or configuration. (3) It would be helpful for researchers and analysts to study the equivalence relationships between SMEs' conditional factors, and their creditworthiness, an equivalence relationship being a type of non-linear relationship that can reveal the effects of different combinations or configurations of conditional factors on a decision factor.

Second, these interval number rules show the complexity and uncertainty of equifinality in Big Data. In the real world, complex and uncertain situations are a constraint we have to face when evaluating SMEs' creditworthiness. Traditional rules of rough sets are based on crisp data. It is difficult for SMEs to fully meet the requirements of the rules, which limits the scope of their application. Hence, interval number rules form one way of capturing the uncertainty present in the fuzzy equifinality relationships, which helps to tackle financial exclusion.

Third, these rules can be used to predict the creditworthiness levels of SMEs and make lending decisions about new applications. The predicted accuracy is shown in Table 4. The most consistent result is 88.89% for SMEs with a creditworthiness level of 3 (i.e. high creditworthiness). One reason for the high accuracy is that most of the rules apply to a creditworthiness level of 3 (there are 446 rules for very low or level 1 creditworthiness, 286 for low or level 2, 702 for high or level 3, and 624 for very high or level 4). From Table 4, it can be observed that, when more objects generate more rules, a higher predictive accuracy rate results. Another reason for the high accuracy is the diversity of rules for creditworthiness level 4. These results also imply that we need to collect data with a greater number and diversity of objects to generate more rules and thereby achieve a more consistent predictive result.

5.2 Complexity in Big Data

Complexity is another major challenge of BDA. Vachon and Klassen (2002) suggest that the complexity includes numerousness, interrelation and systems unpredictability. Numerousness refers to the “4Vs” of information, namely simultaneously having a large volume, high velocity, high veracity and several varieties of data. In our case, there is a large number of conditional attributes. In order to achieve effective creditworthiness evaluation, financial institutions should focus on the core conditional factors that play a significant role in the prediction and determination of creditworthiness. Our approach reduces the original conditional factor set from 83 factors to 14 core factors. This core conditional factor set has the following advantages. (1) Financial institutions can, effectively and at a low cost, collect data and assess the creditworthiness of SMEs. (2) SMEs can focus on key conditional factors to achieve low-cost and efficient creditworthiness enhancement, so as to obtain financing more easily. (3) This smaller set of core factors can improve the accuracy of the prediction and enable the avoidance of the influence of noisy information on the prediction results. (4) It can help identify equifinality relationships more clearly if there is a reduced rule set. (5) It also helps to reduce the computational complexity of the BDA approach to creditworthiness prediction. The number of rules we generate is smaller than the number of data objects, but the traditional rough set generates about a hundred rules for every object, which significantly increases the computational complexity of predicting the creditworthiness level. When new objects are introduced, our approach can quickly generate new rules based on the differences between the new objects and the old objects. However, with traditional rough sets, one would need to recalculate the rules based on the differences among all the objects, which would require a lot of additional computation.

Interrelation refers to the interactions or interconnectivity among the data. In our case, it is shown that the conditional factors will affect a decision factor, and also each other. The following points can be made. (1) Since the conditional factors affect a decision factor, we can use this to determine the factors that hold important information, so as to retain those important factors and delete the unimportant ones. Using NRSs, one loan characteristic, ten financial ratios and three non-financial information factors for SMEs are deemed to be the core conditional factors based on information significance. This shows that the main reference for experts in financial institutions when rating SMEs is their financial indicators. The presence of three non-financial factors indicates that financial institutions should also collect a wide variety of information from different sources. (2) Regarding the interaction of the conditional factors, we mainly use rules to identify the impact of the combination of these conditional factors on the decision factor. 2,058 interval numbers are determined. This result suggests that a complex non-linear relationship exists between the conditional factors and the creditworthiness of SMEs.

System unpredictability refers to the fact that the results of a system are always uncertain, with insufficient information, information uncertainty, dynamic decision parameters and changing decision boundaries. In our case, system unpredictability may exist since similar SMEs receive different creditworthiness evaluations. The intervals can be used to address the unpredictability of the complex system as they can guide decision makers and researchers to focus only on rules that provide strong evidence of correlation or causation with respect to the creditworthiness outcomes. Thus, for this Big Data example, interval number rules have been described that help us more fully understand system unpredictability.

Identifying core variables refers to simplifying the data structure, when performing a credit assessment, from the huge amount and high dimensionality of information. Our method can reduce the dimensionality of credit data effectively and identify core variables for SMEs, and, as mentioned above, it reduces the original conditional factor set from 83 factors to 14 core factors. The core factors are mainly distributed among the categories of loan characteristics, financial ratios and non-financial information. “Monthly family income”, “Automobile and real estate” and “Education background” are the three most important factors for credit assessment, and they are just personal information about the owners. Therefore, the personal characteristics of the owners of SMEs are critical to the operation of enterprises. Financial information factors make up the largest category, with nine factors. Therefore, we need to systematically evaluate the possibility of default based on different financial perspectives. Macroeconomic factors are not found among the core factors, which is mainly because they have the same impact on all SMEs and cannot be effectively avoided individually, so they are not used as creditworthiness assessment factors.

5.3 Comparative analysis

In this section, we conduct a comparative analysis to validate the effectiveness of our approach. The first comparison method is developed by Bai *et al.* (2019a) which is a knowledge rules methodology based on fuzzy RST and fuzzy C-means clustering. The method was used to evaluate and investigate the complex relationships between farmers' characteristics, macro-environmental factors and farmers' creditworthiness levels. We replicate their method to predict the creditworthiness levels of the same 342 SMEs studied earlier, based on the same original objects, conditional attributes and decision attributes. The predicted results for each creditworthiness level are reported in Table 5. In addition, we also apply a hybrid decision tree model (Farid *et al.*, 2014) and a neural network model based on the genetic algorithm (Örkcü and Bal, 2011) on our dataset for the purpose of comparison. The prediction accuracies across four different methods are summarized in Table 6.

(Insert Table 5 about here)

(Insert Table 6 about here)

The overall prediction accuracies of Bai *et al.*'s (2019a), Farid *et al.*'s (2014) and Örkcü & Bal's (2011) methods show that only 61.11%, 44.74% and 48.83% of the predicted creditworthiness levels are the same as the expert-rated creditworthiness levels. The prediction accuracies of their methods are much lower than that of our approach (75.44%), verifying the effectiveness of our method. More specifically, compared with Bai *et al.*'s (2019a) method, the prediction accuracies of the Bai *et al.* (2019a) for 'bad' borrowers (i.e. creditworthiness levels 1 and 2) are slightly better than those of our approach, but the performance of that method is much worse than ours in predicting 'good' borrowers (i.e. levels 3 and 4). As shown in Table 5, there are a significant proportion of 'good' borrowers that are misclassified as 'bad' borrowers. Therefore, if lending decisions are made based on Bai *et al.* (2019a) method, some high-quality borrowers are likely to be rejected for loans, which in turn leads to a negative effect on the profitability of the bank. Moreover, compared to the 2,058 rules of our approach, Bai *et al.* (2019a) generate 15,985 rules which significantly increase the complexity of the calculation. Furthermore, in Table 6, we find that Farid *et al.* (2014) and Örkcü & Bal (2011) methods produce a bad prediction performance on borrowers with lower credit ratings, which could significantly increase the rate of loan defaults and lead to a large capital loss.

6 Research implications

6.1 Theoretical implications

This study offers some theoretical and methodological implications for using BDA to understand borrowers' default behaviors and help financial institutions to make sound credit decisions. First, with the rapid development of Internet economy and finance, using BDA for credit assessment has become a trend. In this era of Big Data, knowing how to quickly collect data and accurately identify factors that are strongly related to SMEs' creditworthiness from high-dimensional datasets is the key for SMEs to ease financing constraints and/or access credit at better terms (Kshetri, 2016). There are two points to make here: (1) BDA expands the sources of information that can be used for assessing credit risk. The traditional data used for credit assessments are mainly related to lenders' financial borrowing and repayment behaviors. The information is mainly obtained from a credit bureau and a borrower's application. However, the traditional credit system disadvantages small and micro enterprises that generally don't have healthy cash flow and enough information in their credit files. With new-age technology, a huge amount of non-financial and unstructured data, such as public and property records, transaction data, mobile data, voice recordings and social profile data, can be obtained through various channels such as the Internet, call centers, branches, and many other open sources. Alternative data can be supplemental for those borrowers who have insufficient traditional credit information for creditworthiness assessment. Combining alternative data with traditional data for credit scoring can help lenders make more informed credit decisions among a wider number of clients and accelerate financial inclusion. (2) BDA expands factors that are considered in credit quality assessment. Through data mining in Big Data, more hidden patterns and correlations can be uncovered, and some potential factors that strongly affect borrowers' creditworthiness can be identified. For example, based on BDA, we find that "Monthly family income", "Automobile and real estate" and "Education background" are the three most important factors that should be considered for credit assessment. By mainly focusing on these important and useful factors, the costs of data collection and calculations can be significantly reduced.

Second, when the prediction results of BDA have an impact on the development of SMEs, an interpretable conclusion will become crucial. The equivalence relationships can offer financial institutions an explanation of the credit rating/ranking of SMEs, so that they can understand why an SME has obtained a given rating and provide operable guidance for SMEs to help them improve their own credit ratings. It can be seen that interpretability is a necessity for protecting the rights and interests of SMEs, and an important indicator for financial institutions that operate with integrity, transparency and fairness. In sum, by using appropriate sources of credit information and factor selection methods, the BDA, as we proposed, can sharpen the accuracy of credit scoring models, reduce the computational complexity and provide useful insights into SMEs' creditworthiness and development.

6.2 Methodological implications

From a methodological point of view, RST is a useful, intelligent, mathematical tool that can be used in BDA to provide valuable insights into credit risk evaluation (Wang *et al.*, 2017; Bai *et al.*, 2019a). The major applications of rough sets include attribute or object reductions to address the problem of large factors and large volume objects, and the identification of complex relationships (non-linear relationships) among attributes that cannot be identified by other techniques such as regression or fuzzy systems. Attribute or object selections and reductions play an important role in credit risk evaluation (Abedin *et al*., 2019). Moreover, a rough set is a simple tool that does not require parametric assumptions or additional information about the data, such as a priori distribution characteristics or the possible values used in fuzzy set theory (Bai *et al.*, 2016). Rules generated from rough sets can also be used to address the predictability of complex systems, especially in the Big Data context.

However, rough sets also have some limitations when it comes to BDA, in that they cannot deal effectively with continuous numeric data. Although the NRS, as an extension of the rough set, has greater flexibility in dealing with such data (Bai & Sarkis, 2014), it is hard to identify fuzzy rules to address complex and uncertain relationships. Traditional crisp numbers, meanwhile, require strict classification or well-defined boundaries between objects, which is not suitable for direct application to Big Data.

In this study, we develop a novel approach, based on the flexible measures of distance and inclusion in the NRS, to identify interval number rules that are more applicable than the rules derived from RST. Most artificial intelligence methods address the uncertainty of credit risk evaluation from a probabilistic perspective (Abedin *et al.*, 2018), while this article adopts interval number rules to address this issue. Those interval number rules will allow for flexibility and ambiguity (fuzziness) that can be used to evaluate the fuzzy equifinality relationships and will also overcome the limitation that rough sets can only recognize crisp number rules (Bai & Sarkis, 2018). This approach can mitigate various credit system complexity and equifinality problems. Hence, we use this novel approach to investigate and evaluate the fuzzy equifinality relationships between SMEs' loan characteristics, financial information, non-financial information, external economic and competitive factors, and creditworthiness levels. To the best of our knowledge, this is the first time that NRSs have been developed to investigate fuzzy equifinality relationships so as to predict the creditworthiness of SMEs.

6.3 Practical implications

This study provides an effective and useful credit assessment tool based on BDA for financial institutions. By adopting this method, lenders can have a more accurate assessment on the creditworthiness of SMEs, which can reduce losses caused by loan defaults and improve the soundness and safety of financial institutions (Lu *et al.*, 2022). With the development of e-commerce, there is a need for a wider range of SMEs to conduct credit assessments, and accuracy is the core of this task (Guo *et al.*, 2023). In particular, as illustrated in Table 6, our method provides an excellent prediction performance for SMEs with a high level of creditworthiness, which can considerably facilitate lending institutions to identify good borrowers and significantly increase revenue and profitability. Moreover, credit data generally has a very high dimensionality and contains a lot of noise. Financial institutions can adopt our proposed method to reduce the dimensionality of credit data, so as to reduce the costs of data collection, calculation and assessment. Furthermore, our method can identify the equivalence relationships and provide interpretable results for credit analysis. These equivalence relations can not only help financial institutions understand the reasons behind financing decisions, but also create credit risk governance knowledge, which can be applied to the collection of credit data, credit risk monitoring, etc.

This study also provides a low-cost credit improvement tool for SMEs. We have identified many equivalent relationships that SMEs can use as a guide to improve their creditworthiness. These equivalence relationships can not only help SMEs understand the reasons for their current scores, but also provide them with multiple paths for improving their creditworthiness. Based on these equivalent paths and their own situations, SMEs can efficiently and effectively choose appropriate investment strategies to improve their creditworthiness. In addition, the method provided in this paper significantly reduces the dimensionality of credit data, from the original 83 variables to the 14 core variables, without prejudicing the predictive performance. This will help SMEs to focus on those core variables with a low cost, and thereby improve their creditworthiness and access to finance (Van Caneghem and Van Campenhout, 2012).

Finally, our study also generates policy implications for the government and policy makers. In the context of Industry 4.0, BDA brings new opportunities for the fast and accurate assessment and prediction of the creditworthiness of SMEs. Therefore, the government should support and encourage financial institutions to apply BDA and other advanced methods for risk assessment. The findings from our study also can provide insights for government policies on improving access to finance for SMEs and promoting the growth and development of the sector.

7 Conclusion and future work

In the past decade, researchers and practitioners have begun to pay attention to the creditworthiness of SMEs in the context of Big Data. It has been shown that much of the financial industry is collecting large amounts of data from a variety of sources. Hence, using Big Data in creditworthiness calculations is a challenge which needs to be studied. Equifinality relationships are a universal and important problem in Big Data but have not been thoroughly investigated. Hence, we develop a novel approach based on NRSs to identify equifinality relationships between different types of conditional factors, and the outcome of creditworthiness.

We apply the proposed method to a real SME loan dataset. We find a core set of conditional factors for SMEs' creditworthiness, and identify fuzzy equivalence relationships between various borrowers' demographic information, financial ratios, loan characteristics, other non-financial information, local macroeconomic indicators and creditworthiness levels. We predict the creditworthiness levels from the test data based on the identified fuzzy equifinality relationships. The results show good accuracy in the prediction of SMEs' creditworthiness levels. This investigation is one of the first to provide insights into fuzzy equifinality relationships based on Big Data and the results can be used to advance decision making. Our approach can easily deal with the "4Vs" and real-time changes in Big Data, and provides a powerful tool that can help both traditional financial institutions and fintech companies to make better credit underwriting decisions and more accurate predictions about the creditworthiness of their potential customers.

Like any other study, ours has some limitations that also leave room for further investigation. First, although we reduce the number of rules to a large extent, we increase the complexity of the rules with many conditional factors. It would be helpful to find some rules in a short form to help SMEs better understand these equivalence relations. Second, the data distribution of each factor is uneven, making it inappropriate to use the same distance threshold to determine the neighborhood members in this approach. Further work could use the fuzzy C-means method to determine the neighborhood members according to the distribution of the data.

Conflict of interest statement

The authors declare that there are no conflict of interests.

References

Abedin, M. Z., Chi, G., Colombage, S., & Moula, F. E. (2018). Credit default prediction using a support vector machine and a probabilistic neural network. *Journal of Credit Risk, Forthcoming, 14(2),* 1-27.

Abedin, M. Z., & Guotai, C. (2019). An optimized support vector machine intelligent technique using optimized feature selection methods: Evidence from Chinese credit approval data. *Journal of Risk Model Validation, 13(2),* 1-46.

Altman, E. (1998). The important and subtlety of credit rating migration. *Journal of Banking and Finance*, *22*, 1231-1247.

Altman, E., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the U.S. market. *Abacus*, *43*(3), 332-357.

Angilella, S., & Mazzù, S. (2015). The financing of innovative SMEs: A multicriteria credit rating model. *European Journal of Operational Research*, *244*(2), 540-554.

Artem, B., & Stefan, L. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications*, *86*, 42-53.

Bagale, G. S., Vandadi, V. R., *et al.* (2023). Small and medium-sized enterprises' contribution in digital technology. *Annals of Operations Research*, *326(1),* 3-4. DOI: 10.1007/s10479-021-04235-5.

Bai, C. G., & Sarkis, J. (2014). Determining and applying sustainable supplier key performance indicators. *Supply Chain Management: An International Journal*, *19*(3), 275-291.

Bai, C. G., & Sarkis, J. (2018). Honoring complexity in sustainable supply chain research: A rough set theoretic approach. *Production Planning and Control, 29*(16), 1367-1384.

Bai, C. G., Dhavale, D., & Sarkis, J. (2016). Complex investment decisions using rough set and fuzzy C-means: An example of investment in green supply chains. *European Journal of Operational Research*, *248*(2), 507-521.

Bai, C. G., Shi, B. F., Liu F, & Sarkis, J. (2019a). Banking credit worthiness: Evaluating the complex relationships. *Omega*, *83*, 26-38.

Bai, C., Govindan, K., Satir, A., & Yan, H. (2019b). A novel fuzzy reference-neighborhood rough set approach for green supplier development practices. *Annals of Operations Research*, in press, DOI: 10.1007/s10479-019-03456-z.

Bai, C., Kusi-Sarpong, S., Khan, S. A., & Vazquez-Brust, D. (2021). Sustainable buyer–supplier relationship capability development: a relational framework and visualization methodology. *Annals of Operations Research*, *304*, 1-34.

Baru, C. (2018). How to deliver translational data-science benefits to science and society. *Nature, 561*, 464.

Billie, A. (2020). Using Bayesian networks to perform reject inference. *Expert Systems with Application*, *137*, 349-356.

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.

Cao, Y., & Zhai, J. (2022). A survey of AI in finance. *Journal of Chinese Economic and Business Studies*, 22 (2), 125-137.

Chai, N. N., Wu, B., Yang, W. W., Shi, B. F. (2019). A multicriteria approach for modeling small enterprise credit rating: Evidence from China. *Emerging Markets Finance and Trade*, *55*(11): 2523-2543.

Chen, C. L. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314-347.

Ciampi, F. (2015). Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms. *Journal of Business Research*, *68*(5), 1012-1025.

Cowling, M., Liu, W., & Ledger, A. (2012). Small business financing in the UK before and during the current financial crisis. *International Small Business Journal*, 30(7), 778-800.

Dahms, S. (2019). Foreign-owned subsidiary knowledge sourcing: The role of location and expatriates. *Journal of Business Research*, *105*, 178-188.

D'Amato, A., & Mastrolia, E. (2022). Linear discriminant analysis and logistic regression for default probability prediction: the case of an Italian local bank. *International Journal of Managerial and Financial Accounting, 14*(4), 323-343.

De Guinea, O. A., & Raymond, L. (2020). Enabling innovation in the face of uncertainty through IT ambidexterity: A fuzzy set qualitative comparative analysis of industrial service SMEs. *International Journal of Information Management, 50*, 244-260.

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, *346*(6210), 1243089.

Fantazzini, D., & Figini, S. (2009). Random survival forests models for SME credit risk measurement. *Methodology and computing in applied probability*, 11(1), 29-45.

Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., Strachan, R. (2014). Hybrid decision tree and nave bayes classifiers for multi-class classification tasks. *Expert Systems with Applications, 41*(4), 1937-1946.

Fernandes, G. B., & Artes, R. (2016). Spatial dependence in credit risk and its improvement in credit scoring. *European Journal of Operational Research*, 249(2), 517-524.

Fombellida, J., Martin-Rubio, I., Romera-Zarza, A., & Andina, D. (2019). KLN, A new biological koniocortex based unsupervised neural network: Competitive results on credit scoring. *Natural Computing*, *18*(2), 265-273.

Glennon, D., & Nigro, P. (2005). Measuring the default risk of small business loans: A survival analysis approach. *Journal of Money, Credit and Banking*, 923-947.

Gresov, C., & Drazin, R. (1997). Equifinality: Functional equivalence in organization design. *Academy of Management Review, 22*(2), 403-428.

Guangming Tech. (2022). Enhance core competitiveness of SMEs and foster more and more 'small giants'. Guangming Technology Website. Aviliable at: 15 December 2022. https://tech.gmw.cn/2022-03/08/content_35571859.htm.

Guo, J., Jia, F., Yan, F., & Chen, L. (2023). E-commerce supply chain finance for SMEs: the role of green innovation. *International Journal of Logistics Research and Applications,* 1-20.

He, H. L., Zhang, W. Y., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications*, *98*, 105-117.

Hooman, A., Marthandan, G., Yusoff, W. F., Omid, M., & Karamizadeh, S. (2016). Statistical and data mining methods in credit scoring. *The Journal of Developing Areas*, *50*(5), 371-381.

Hu, Q., Yu, D., Liu, J., & Wu, C. (2008). Neighborhood rough set based heterogeneous feature subset selection. *Information Sciences*, *178*(18), 3577-3594.

Kim, A., Yang, Y., *et al.* (2020). Can deep learning predict risky retail investors? A case study in financial risk behavior forecasting. *European Journal of Operational Research*, *283*(1), 217-234.

Kou, G., Yu, H., *et al.* (2021). Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. *Decision Support Systems, 140*, 113429.

Kshetri, N. (2016). Big data's role in expanding access to financial services in China. *International Journal of Information Management,* 36(3), 297-308.

Lappas, P. Z., & Yannacopoulos, A, N. (2021). A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. *Applied Soft Computing, 107*, 107391.

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

Li, W. X., & He, T. (2023). Banking structure and government policies regarding SMEs financing. *Journal of Chinese Economic and Business Studies*, 21(3), 387-402.

Li, K., Niskanen, J., Kolehmainen, M., & Nishanen, M. (2016). Financial innovation: Credit default hybrid model for SME lending. *Expert Systems with Applications*, *61*, 343-355.

Lin, C., Liu S., & Wei L. (2022). Banking and innovation: a review. *Journal of Chinese Economic and Business Studies*, Published: 03 Nov 2022, DOI: 10.1080/14765284.2022.2127397.

Liu, Y. D., Song, Y. N., *et al.* (2019). Big-data-driven model construction and empirical analysis of SMEs credit assessment in China. *Procedia Computer Science, 147*, 613-619.

Lu, Y., Yang, L., Shi, B. F., Li, J. X., Abedin, M. Z. (2022). A novel framework of credit risk feature selection for SMEs in Industry 4.0. *Annals of Operations Research*, Published: 25 July 2022. DOI: 10.1007/s10479-022-04849-3.

Medina-Olivares, V., Calabrese, R., *et al.* (2022). Spatial dependence in microfinance credit default. *International Journal of Forecasting*, *38*(3), 1071-1085.

Molinero, M. C., Gomez, A. P., & Cinca, S. C. (1996). A multivariate study of Spanish bond ratings. *Omega*, *24*(4), 451-462.

Örkcü, H. H., & Bal, H. (2011). Comparing performances of backpropagation and genetic algorithms in the data classification. *Expert Systems with Applications, 38*(4), 3703-3709.

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., Baesens, B. (2019). The value of Big Data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, *74*, 26-39.

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences*, *11*(5), 341-356.

Pederzoli, C., Thoma, G., & Torricelli, C. (2013). Modelling credit risk for innovative SMEs: The role of innovation measures. *Journal of Financial Service Research*, *44*, 111-129.

Rajesh, R. (2022). Sustainability performance predictions in supply chains: grey and rough set theoretical approaches. *Annals of Operations Research*, 310(1), 171-200.

Shi, B. F., Chi, G. T., & Li, W. P. (2020). Exploring the mismatch between credit ratings and loss-given-default: A credit risk approach. *Economic Modelling, 85*, 420-428.

Stevenson, M., Mues, C., & Bravo, C. (2021). The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research*, 295(2), 758-771.

Sun, Y., Chai, N., *et al.* (2022). Assessing and predicting small industrial enterprises' credit ratings: A fuzzy decision making approach. *International Journal of Forecasting*, *38*(3): 1158-1172.

Thomas, L. C., Crook, J. N., & Edelman, D. B. (2002). *Credit scoring and its applications*. Philadelphia: SIAM.

Vachon, S., & Klassen, R. D. (2002). An exploratory investigation of the effects of supply chain complexity on delivery performance. *IEEE Transactions on Engineering Management*, *49*(3), 218-230.

Van Caneghem, T., & Van Campenhout, G. (2012). Quantity and quality of information and SME financial structure. *Small Business Economics,* 39, 341-358.

Verleye, K. (2019). Designing, writing-up and reviewing case study research: An equifinality perspective. *Journal of Service Management*, *30*(5), 549-576.

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, *70*, 356-365.

Wamba, S. F., Gunasekaran, A., Dubey, R., & Ngai, E. W. (2018). Big data analytics in operations and supply chain management. *Annals of Operations Research*, *270*, 1-4.

Wang, F. T., Ding, L. H., Yu, H. X., & Zhao, Y. J. (2020). Big data analytics on enterprise credit risk evaluation of e-Business platform. *Information Systems and E-Business Management, 18*(3), 311-350.

Wang, H., Xu, Z., & Pedrycz, W. (2017). An overview on the roles of fuzzy set techniques in big data processing: Trends, challenges and opportunities. *Knowledge-Based Systems*, *118*, 15-30.

Xia, Y. F., Liu, C. Z., Li, Y. Y., & Liu, N. N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225-241.

Yao, X., Crook, J., & Andreeva, G. (2017). Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research*, *263*(2), 679-689.

Zhan, Y., Tan, K. H. (2020). An analytic infrastructure for harvesting big data to enhance supply chain performance. *European Journal of Operational Research*, *281*(3), 559-574.

Zhang, Z.P., Chi, G.T., Colombage, S., Zhou. Y. (2022). Credit scoring model based on a novel group feature selection method: The case of Chinese small-sized manufacturing enterprises. *Journal of The Operational Research Society*, *73*(1), 122-138.

Zhao, J. F., Li, B. (2022). Credit risk assessment of small and medium-sized enterprises in supply chain finance based on SVM and BP neural network. *Neural Computing & Applications*, *34*(15), 12467-12478.

Tables and Appendices:

Table 1: Construction of dataset for SMEs

SMEs	Loan information		Financial ratios		Non-financial factors		Macroeconomic indicators		Creditworthiness level
	Principal (RMB)	…	Debt-to-assets ratio	…	Hierarchy of new product	…	GDP growth rate	…	
SME 1	1,200,000	…	0.640	…	No new product certificate	…	16.2	…	1
SME 2	13,000,000	…	0.949	…	New product certificate issued by social organizations	…	16.2	…	1
SME 3	5,000,000	…	0.643	…	No new product certificate	…	16.2	…	3
SME 4	1,300,000	…	0.624	…	No new product certificate	…	16.2	…	1
SME 5	6,000,000	…	0.334	…	No new product certificate	…	16.2	…	2
SME 6	1,500,000	…	0.523	…	No new product certificate	…	16.2	…	3
SME 7	3,000,000	…	0.526	…	No new product certificate	…	16.2	…	3
SME 8	5,000,000	…	0.523	…	New product certificate issued by provincial government department	…	16.2	…	3
SME 9	900,000	…	0.040	…	No new product certificate	…	16.2	…	4
SME 10	1,800,000	…	0.182	…	No new product certificate	…	16.2	…	3
…	…	…	…	…	…	…	…	…	…
SME 3109	1,590,000	…	0.556	…	No new product certificate	…	10.3	…	1
SME 3110	1,627,000	…	0.556	…	No new product certificate	…	10.3	…	1
SME 3111	1,578,000	…	0.556	…	No new product certificate	…	10.3	…	1

Table 2: Information significance values of core conditional factor set

Atr	Core conditional factor	Information significance
*Atr*1	Principal	0.017
*Atr*2	Quick asset ratio	0.017
*Atr*3	Capitalization rate	0.016
*Atr*4	Superquick asset ratio	0.020
*Atr*5	Long-term asset suitability rate	0.032
*Atr*6	Net profit divided by total operating costs and expenses	0.018
*Atr*7	Total profit divided by total operating costs and expenses	0.034
*Atr*8	Inventory turnover velocity	0.017
*Atr*9	Velocity of fixed assets	0.022
*Atr*10	Velocity of equity	0.022
*Atr*11	The loan from the bank divided by the firm's total bank loans	0.015
*Atr*12	Education background	0.037
*Atr*13	Automobile and real estate	0.046
*Atr*14	Monthly family income	0.057

Table 3: Examples of interval number rules for creditworthiness level (*D*) = 4

*Atr*1	*Atr*2	*Atr*3	*Atr*4	*Atr*5	*Atr*6	*Atr*7	*Atr*8	*Atr*9	*Atr*10	*Atr*11	*Atr*12	*Atr*13	*Atr*14	*D*
[0, 0.102]	[0.582, 0.782]	[0, 0.1]	[0, 0.1]	[0, 0.1]	[0, 0.1]	[0,0.1]	[0,0.105]	[0,0.1]	[0,0.1]	[0.9,1]	[0.9,1]	[0,0.144]	[0,0.121]	4
[0, 0.1]	[0.599,0.799]	[0.9, 1]	[0.451, 0.651]	[0, 0.113]	[0.369, 0.569]	[0.729,0.929]	[0,0.1]	[0,0.102]	[0,0.118]	[0.433,0.633]	[0.9,1]	[0.118,0.318]	[0.254,0.454]	4
[0, 0.101]	[0.061, 0.261]	[0.9, 1]	[0.152, 0.352]	[0.429, 0.629]	[0, 0.106]	[0.372,0.572]	[0,0.104]	[0,0.184]	[0,0.143]	[0.658,0.858]	[0.3,0.5]	[0.336,0.536]	[0.608,0.808]	4
[0, 0.102]	[0.045, 0.245]	[0.9, 1]	[0.128, 0.328]	[0.036, 0.236]	[0, 0.176]	[0.418,0.618]	[0,0.105]	[0,0.163]	[0.025,0.225]	[0,0.194]	[0.8,1]	[0,0.187]	[0.254,0.454]	4
[0, 0.111]	[0.003, 0.203]	[0.886, 1]	[0, 0.146]	[0, 0.105]	[0, 0.1]	[0,0.1]	[0,0.1]	[0,0.1]	[0,0.1]	[0.9,1]	[0.9,1]	[0,0.1]	[0,0.142]	4
[0, 0.123]	[0.198, 0.398]	[0.9, 1]	[0.319, 0.519]	[0, 0.113]	[0, 0.163]	[0.412,0.612]	[0,0.103]	[0,0.101]	[0,0.13]	[0.025,0.225]	[0.9,1]	[0.118,0.318]	[0.75,0.95]	4
[0, 0.111]	[0.042, 0.242]	[0.9, 1]	[0.106, 0.306]	[0.798, 0.998]	[0.013, 0.213]	[0.427,0.627]	[0,0.107]	[0.317,0.517]	[0.025,0.225]	[0.041,0.241]	[0.9,1]	[0.336,0.536]	[0,0.1]	4

Table 4: Distribution of predicted accuracy

Credit level		Predicted				Number	Accuracy	Number of rules
		1	2	3	4			
Actual	1	82	4	10	6	102	80.39%	446
	2	0	14	18	6	38	36.84%	286
	3	4	3	96	5	108	88.89%	702
	4	14	1	13	66	94	70.21%	624
Sum		100	22	137	83	342	75.44%	2058

Table 5: Distribution of predicted accuracy, using the knowledge rules methodology developed by Bai *et al.* (2019a)

Creditworthiness level		Predicted				Number	Accuracy	Number of rules
		1	2	3	4			
Experts' ratings	1	87	6	5	4	102	85.29%	2495
	2	3	16	15	4	38	42.11%	3624
	3	10	14	58	16	108	62.96%	6875
	4	17	27	2	48	94	51.06%	2991
Sum		117	63	90	72	342	61.11%	15985

Table 6: Prediction accuracies of our method and three other methods

Creditworthiness level		Number	Our method	Bai *et al.*'s (2019a) method	Farid *et al.*'s (2014) method	Örkcü & Bal's (2011) method
Experts' ratings	1	102	80.39%	85.29%	40.20%	64.71%
	2	38	36.84%	42.11%	23.68%	31.58%
	3	108	88.89%	62.96%	37.04%	22.22%
	4	94	70.21%	51.06%	67.02%	69.15%
overall		342	75.44%	61.11%	44.74%	48.83%

Appendix 1. Background on neighborhood rough sets (NRSs)

The definitions of NRSs presented in this paper are based on the developments of Hu *et al.* (2008).

Definition 1: Given an arbitrary $x_i \in U$ and $B \subseteq C$, the neighborhood $\delta_B(x_i)$ of x_i in factor set *B* is defined as

$$\delta_B(x_i) = \{x_k \mid x_k \in U, \Delta_B(x_i, x_k) \leq \delta\}, \qquad \text{(A1)}$$

where Δ is a distance function.

Definition 2: Let x_i and x_k denote two objects in *m*-dimensional space $A=\{a_1,a_2, \ldots, a_\mathrm{m}\}$. Then, an appropriate distance function between x_i and x_k is as follows:

$$\Delta_A^q(x_i, x_k) = (\sum_{j=1}^{m} | f(x_i, a_j) - f(x_k, a_j) |^q)^{1/q} \qquad \text{(A2)}$$

where $f(x,a_j)$ denotes the value of sample *x* in the *j*th attribute *a_j;* *q*=1 is defined as a Manhattan distance; *q*=2 is a Euclidean distance; and *q*=∞ is a Chebychev distance.

Definition 3: Let X_1, X_2, ..., X_N be the object subsets with decisions 1 to *N*, the lower and upper approximations of decision *D* with respect to attributes *B*, defined as

$$\underline{N_B}D=\bigcup_{d=1}^{N}\underline{N_B}X_d, \overline{N_B}D=\bigcup_{d=1}^{N}\overline{N_B}X_d \tag{A3}$$

where

$$\underline{N_B}X_d=\{x_i \mid \delta_B(x_i)\subseteq X_d, x_i\in U\}, \overline{N_B}X_d=\{x_i \mid \delta_B(x_i)\bigcap X_d\neq\varnothing, x_i\in U\}. \tag{A4}$$

Definition 4: Given two sets *A* and *B* in the universe *U*, we define *A*'s degree of inclusion in *B* as

$$I(A,B)=\frac{|A\bigcap B|}{|A|},\ where\ A\neq\varnothing \tag{A5}$$

where |*| is the cardinality of a set.

Definition 5: Given any subset $X\subseteq U$ and inclusion measure, the lower and upper sets of *X* are

$$\begin{aligned}\underline{N^k}X&=\{x_i \mid I(\delta(x_i),X)\geq k, x_i\in U\},\\ \overline{N^k}X&=\{x_i \mid I(\delta(x_i),X)\geq 1-k, x_i\in U\}.\end{aligned} \tag{A6}$$

where 0≤k≤1

Definition 6: The dependency degree of *D* (decision attribute) to a set *B* is defined as

$$\gamma_B(D)=\frac{|\underline{N_B}D|}{|U|} \tag{A7}$$

where $\gamma_B(D)$ reflects the ability of set *B* to approximate *D*. Clearly, $0\leq\gamma_B(D)\leq 1$. If $\gamma_B(D)$=1, we can say that *D* completely depends on *B*; otherwise, we say that D_γ depends on B.

Definition 7: The significance of a_j in *B* is defined as

$$Sig(a_j,B,D)=\gamma_B(D)-\gamma_{B-a_j}(D) \tag{A8}$$

If $Sig(a_j,B,D)=0$, we say a_j is superfluous, which means a_j is useless for *B* to approximate *D*.

Appendix 2. Characteristic attributes of SME credit evaluation

Category	Characteristic attribute	Explanation
Basic loan characteristics	*Principal*	The principal of the loan obtained by the SME
	Industry	The industry type of the SME, classified according to MIIT document No. 300 (MIIT, Ministry of Industry and Information Technology of PRC)
	Loan term	The length of time within which the SME has to repay
Financial ratios	*Debt-to-assets ratio*	Total liabilities divided by total assets
	Ratio of net cash flows	"Cash flow from operating activities" divided by current liabilities
	Quick asset ratio	Quick assets divided by current liabilities
	Liquidity ratio	Current assets divided by current liabilities
	Net cash flow -to- main business income ratio	"Cash flow from operating activities" divided by "prime operating revenue"
	EBIT -to- current liabilities ratio	"Earnings before Interest and Tax (EBIT)" divided by current liabilities
	Capitalization rate	"Long-term liabilities" divided by "Long-term liabilities plus owners' equity"
	Net cash flow -to- asset ratio	"Cash flow from operating activities" divided by total assets
	Equity ratio	-
	Superquick asset ratio	"Cash + short-dated securities + net receivables" divided by current liabilities
	Net cash flows from operating activities -to- net profit ratio	"Cash flow from operating activities" divided by net profit
	Net assets divided by the sum of short-term loan balance and long-term loan balance	-
	Capital immobilization ratio	"Total assets minus current assets" divided by "the average owners' equity"
	Cash ratio	"Cash + securities" divided by current liabilities
	Long-term asset suitability rate	"The sum of long-term liabilities and owners' equity" divided by "the sum of fixed assets and long-term investment"
	Total outstanding loans divided by total equity	"The owed debt principal" divided by total equity
	Total outstanding loans divided by total assets	"The owed debt principal" divided by total assets
	Net cash flows from operating activities -to- non-current liability ratio	"Cash flow from operating activities" divided by non-current liability
	Net cash flows from operating activities -to- "assets plus funds borrowed" ratio	"Cash flow from operating activities" divided by total assets after borrowing
	EBITDA divided by liabilities	-
	Return on equity	"Net profit" divided by "Owner's equity"
	Net cash flows from operating activities divided by sales revenue	-
	Net profits divided by sales revenue	-
	Return on total assets	-
	Operating profit margin	-

Cont. Appendix 2

Category	Characteristic attribute	Explanation
Financial ratios	*Net profit divided by total operating costs and expenses*	"Net profit" divided by (operating costs + selling expenses + management expenses + financial expenses)
	Gross profit rate	-
	Total profit divided by total operating costs and expenses	-
	EBITDA (earnings before interest, taxes, depreciation and amortization)	-
	EBITDA divided by total revenue	-
	Net profit	-
	Net cash flows from operating activities	-
	Total input of cash flows from operating activities	-
	Receivable turnover velocity	"Revenue from main operations" divided by [(original value of the beginning accounts receivable balance + original value of the balance of the accounts receivable at the end of the period) / 2]
	Inventory turnover velocity	"Revenue from main operations" divided by [(beginning inventory + ending inventory)/2]
	Total assets turnover velocity	"Net sales" divided by [(total assets at the beginning of the year + total assets at the end of the year) / 2]
	Velocity of liquid assets	"Total operating income" divided by "current assets"
	Velocity of fixed assets	"Total operating income" divided by "fixed assets"
	Velocity of equity	"Total operating income" divided by "equity"
	Working capital ratio	"Working capital" divided by "current assets"
	Rate of return on investment	"Investment income (after tax)" divided by "investment cost"
	Accounts payable turnover velocity	"Operating costs" divided by "accounts payable"
	Cash cycle	Inventory turnover speed + accounts receivable turnover speed - accounts payable turnover speed
	Revenue growth rate	-
	Profit growth rate	-
	Total assets growth rate	-
	Rate of capital accumulation	Growth rate of shareholders' equity
	Retained earnings growth rate	-
Non-financial information	*Firm age*	The number of years a business owner has worked in the industry
	Audit or not	Whether SME is audited annually by an auditing firm
	Hierarchy of new products	Whether the new products produced by the SME have passed certification
	Patent condition	Whether SME has any invention patent
	Business life	Length of time the business has been in operation
	Bank account condition	Whether SME has set up a basic account with the bank
	Sales scope	Whether products are sold abroad

Cont. Appendix 2

Category	Characteristic attribute	Explanation
Non-financial information	*Whether brand products*	Whether the main products are brand-name products
	Loans from bank divided by total loans	Loans from bank divided by total loans of SME
	Education background	Education background of firm's owner at the time the SME obtains the loan
	Default records of legal representatives	Number of credit loan defaults by firm's owner
	Credit history of legal representatives	Number of credit card defaults by firm's owner
	Marital status	Marital status of firm's owner at the time the SME obtains the loan
	Residence status	Whether firm's owner owns his (her) property
	The length of time for local residency	The length of time the person has resided locally
	Gender	Male or female
	Age	Age of owner at the time the SME obtains the loan
	Automobile and real estate	Total value of owner's automobiles and real estate
	Monthly family income	-
	Amount of time holding the position	-
	Type of registered capital	Capital registered or physical capital registered
	Enterprise credit over last three years	Number of defaults by the SME in the past three years
	Tax records	Whether SME pays tax on time
	Legal dispute number	Whether there is any legal dispute over the operation of the SME
	Lawful operation or not	Whether there are industrial or commercial commendations or penalties regarding the SME's operations
	Number of breaches of contract	Number of contract breaches among enterprises
	Credit score based on collateral	The converted score of the pledged goods submitted by the SME
Macroeconomic indicators	*Business cycle index*	The business sentiment index published by the Chinese Bureau of Statistics at the time the SME obtains the loan
	Urban residents per capita savings at the end of the year (Yuan)	Local per capita deposit balance at the time the SME applies for the loan
	GDP growth rate (%)	Local GDP growth rate at the time the SME applies for the loan
	CPI (Consumer Price Index)	Loacl CPI at the time the SME applies for the loan
	Urban citizens' per capita disposable income	-
	Engel coefficient	The Engel coefficient of the homeplace at the time the SME applies for the loan
Creditworthiness level		1-very low creditworthiness, 2-low creditworthiness, 3-high creditworthiness, 4-very high creditworthiness