Fuzzy clustering analysis for the loan audit short texts

Lu Han
Central University of Finance and Economics

Zhidong Liu
Central University of Finance and Economics

Jipeng Qiang (jpqiang@yzu.edu.cn)
Yangzhou University

Zhuangyi Zhang
Central University of Finance and Economics

Research Article

Keywords: Fuzzy clustering analysis, Lexical analysis, Minimum spanning tree, Loan audit short texts

Posted Date: March 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2734237/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Knowledge and Information Systems on July 25th, 2023. See the published version at https://doi.org/10.1007/s10115-023-01943-1.

Abstract

In China, post loan management is usually executed in the form of visit survey by credit man. Through quarterly visit survey, a large number of loan audit short texts are collected, which contain valuable information for evaluating the credit status small and micro enterprises. However, there is still lack of methods for analyzing this kind of short texts. This paper proposes a method for processing these loan audit short texts called Fuzzy Clustering Analysis (FCA). This method firstly transforms short texts into a fuzzy matrix through lexical analysis; Then, the similarity between records is calculated based on each fuzzy matrix, and an association graph is constructed with the similarity. Finally, Prim minimum spanning tree is used to extract clusters based on different α cuts. Experiments with actual data from a commercial bank in China have revealed that FCA yields suitable clustering results when handling loan audit briefs. Moreover, it exhibits superior performance compared to BRICH, Kmean, and FCM..

1. Introduction

According to regulations, China's commercial banks are obligated to conduct quarterly investigations on the borrowers of each retail loan, so as to timely know the financial situation of the borrowers and avoid loan default to some extent. Generally, the borrowers of retail loans are mainly small median enterprise (SME) owners who lack standardized financial statements, resulting in untimely financial data updates. Therefore, in practice, commercial banks usually let credit salesmen to visit and investigate these SMEs at regular intervals, and require them to record the investigation information in the form of short texts. This has formed a large number of loan audit short texts. However, there is no standard specification for loan audit short text, resulting in significant differences in their contents and scopes. In addition, these briefs differ greatly from traditional financial data in terms of data form and descriptions. These briefs not only contain valuable information about the borrower's credit risk, but also have good timeliness. Meanwhile, different credit men deal with these records in a different manner. Experienced personnel can detect abnormal signals of SME operation in advance, and take necessary measures to prevent the occurrence of default. However, this work is currently done manually, leading to low efficiency and high costs. Additionally, there is no uniform specification on how to deal with these records. Therefore, there is a great need for developing methods to deal with this type of data to conduct post-loan management.

The widespread use of financial services has drawn the attention of researchers towards developing credit risk management models (Moscato, Picariello, & Sperlí, 2021). However, most of the credit risk management models focus on credit scoring and mainly uses customer financial data, primarily for pre-loan management (Dastile, Celik, & Potsane, 2020). Current short text analysis predominantly focuses on topic extraction (Rashid, Shah, & Irtaza, 2019; Cao, Xu, Yin, & Pan, 2022), emotional analysis (Yadollahi, Shahraki, & Zaiane, 2017; Fan, Zhao, Wen, Xu, & Chang, 2017; Žitnik, Blagus, & Bajec, 2022) and other aspects. Most loan audit short texts are objective records of the credit men’s visit, making it challenging to conduct emotional analysis as these texts lack positive and negative descriptions. Additionally, the subjects of these records mainly focus on SME's operation status, customer comments, the boss, or other managers, etc., which differ significantly from most topic modeling’s targets, such as news, social media comments, etc. For these loan audit short texts, there is no need to analyze their topics or sentiments. Instead, we must focus on identifying signals of default hidden behind these records..

To address the issue for finding hidden default signals within loan audit short texts that lack a clear target, we have employed clustering algorithms. Classical clustering algorithms include K-means based on partition, BIRCH based on hierarchy, and so on. Although K-means is simple to implement, it is quite sensitive to initial settings. The BIRCH method can dynamically cluster and handle noisy data, but is generally restricted to spherical clusters and is sensitive to the order of input data. Moreover, Either Kmeans or BRICH requires each data belongs solely to one cluster, which is problematic when dealing with a large number of clustering with unclear boundaries, especially those involving objects records.

To overcome this issue, Ruspini innovatively introduced fuzzy theory into clustering in 1969. Fuzzy clustering enhances traditional clustering by assigning a degree of membership to each cluster for each datum, thereby improving classification outcomes. As such, fuzzy clustering is less susceptible to initialization and possesses remarkable noise resistance. One of the most widely fuzzy clustering algorithms is the FCM (fuzzy c-means) algorithm. However, FCM has a significant drawback in that it is not adaptable to adding new records, as it requires reconstruction and retraining of the model.

Recently, scholars have been focusing on improving fuzzy clustering algorithms based on fuzzy similarity and fuzzy relationships. However, these algorithms alsoface the challenge of updating with new records, limiting their applications (Wang, Wang, & Wang, 2022). Here, we propose a new clustering method, which is called as fuzzy clustering analysis (FCA). FCA leverages lexical analysis to convert short texts into fuzzy matrices, calculates fuzzy similarity based on these matrices, and ultimately derives clusters based on similarity graphs using a minimum spanning tree. FCA is particularly well-suited for processing text records, being insensitive to initial settings and offering more convenience in update samples without requiring retraining of the model. Through the experiments on the loan audit short texts of a city commercial bank in China, we can find that FCA not only enables clustering the credit records to any degree according to the management regulations, but also outperforms other traditional clustering algorithms.

This paper is organized as follows. Section 1 introduces the background of this research. Section 2 summarizes recent works related to this study. Section 3 puts forwards the objectives. Section 4 illustrates the FCA method in detail. Section 5 describes the experimental results of FCA and compares the results with other cluster algorithms. Section 6 gives a discussion of FCA. Section 7 summarizes the whole study and gives a few suggestions for future work.

2. Related Work

We mainly study on the problem of credit risk management. Based on our research objectives and data characteristics, we summarize the credit risk assessment methods and the short text processing methods.

Generally, credit risk management usually uses credit score to quantify the borrower's default risk. Credit scoring is usually modeled on the borrower's financial statements by machine learning or some supervised learning methods (Silva, Pereira, & Magalhães, 2022). Credit score is divided into application score and behavior score. The application score uses the pre-loan digital demographic information, e.g., the number of dependents, time at current address, time at current employment, etc. The behavior score is mainly based on the loan repayment history (Kozodoi, Jacob, & Lessmann, 2022). Traditionally, financial institutions use a logistic regression to score borrowers (Altman, 2018). Nowadays, sophisticated machine learning models can be found that can replace the logistic regression model. In spite of high accuracies from machine learning (ML) models; ML models are generally unable to explain their predictions (Gunnarsson, Vanden Broucke, Baesens, Óskarsdóttir, & Lemahieu, 2021). Many researchers are committed to comparing the effects of different classification algorithms in credit scoring, the work can be seen from Louzada, Ara, & Fernandes, 2016; S. et al., 2019; Jiang, Lu, Wang, & Ding, 2023. While conducting credit analysis on traditional financial data, scholars gradually begin to pay attention to evaluate of credit risk hidden behind text data, a few works can be seen in Wang, Jiang, Zhao, & Ding, 2020; Stevenson, Mues, & Bravo, 2021; Acheampong & Elshandidy, 2021; Yang, Yuan, & Lau, 2022. These works not only provide some technical methods for processing text to manage credit risk, but also emphasize the irreplaceable value of using text in credit risk management.

With the development of natural language processing, the text data analysis methods have increasingly attracted researchers' attention (Dong et al., 2022). According to the objects, natural language processing is often divided into several tasks, such as lexical analysis, syntactic analysis, and semantic analysis (Erdem et al., 2022). According to the applications, natural language processing is commonly used in topic modeling with social media texts (Chen, Zhang, Liu, Ye, & Lin, 2019; Rashid, Shah, & Irtaza, 2019; Choudhary, Aggarwal, Subbian, & Reddy, 2022; Feng et al., 2022 Srivastava, Singh, Rana, & Kumar, 2022), sentiment analysis in decision-making and prediction (Shi, Zhu, Li, Gao, & Zheng, 2019; Wang, Niu, & Yu, 2020; Ahmed, Chen, & Li, 2020; Alekseev et al., 2021; Consoli, Barbaglia, & Manzan, 2022), and of course, some tasks such as abstract generation, human-computer dialogue, and language translation also need natural language analysis, however, these tasks are quite different from our objectives, so we will not discuss these topics here.

In the analysis of short texts, it is generally necessary to follow the three steps: domain dictionary construction, text representation and transformation, and model learning. From the literatures (Xu, Liu, & Araki, 2015; Zhou, Yu, & Hu, 2017; Han, Zhang, Yang, Shen, & Zhang, 2018; Xu et al., 2019; Ahmed, Chen, & Li, 2020; Alekseev et al., 2021; Chauhan & Shah, 2021; Gul, Räbiger, & Saygın, 2022), we can find that whether it is topic modeling or sentiment analysis, for short text processing, it is necessary to construct domain dictionary. In the work of Han, Zhang, Yang, Shen, & Zhang, 2018, they introduce mutual information from manual labeling to assign terms with Part-Of-Speech tags in the lexicon, then train their labels in a classifier, and finally integrate a completed lexicon-based sentiment analysis framework for sentiment learning. In the work of Xu et al., 2019, they focused on Chinese dictionaries by expanding the scope of vocabulary to distinguish the basic sentient words, the field sentient words, and the polymeric sentient words. Gul, Räbiger, & Saygın, 2022 assume that the documents are related to a certain topic, then through extracting the relevant concepts from a collection of unstructured textual documents, one can infer the context information for the concepts. For the text representation and transformation, most of the existing researches are to encode text into specific vectors based on dictionaries. Sinoara, Camacho-Collados, Rossi, Navigli, & Rezende, 2019 represent document collections based on embedded representations of words and word senses, they bring together the power of word sense disambiguation and the semantic richness of word and use word-sense embedded vectors to construct representations of document collections. Rahimi & Homayounpour, 2020 propose three different text representation methods with the term-document matrix and document-topic matrix, and apply tensor factorization to utilize the power of both matrices, then using these matrices, one can conduct text clustering for better performance. Wu, Zhao, & Li, 2020 utilize vectors to represent the similarity relationship between texts. Yu, 2020 constructs lemma vectors based on word frequency features, with which the frequency features of the word frequency and the frequency of the inverse frequency document are fused to construct the feature matrix, then the feature matrix is decomposed and transformed by the singular value to obtain a semantic space, which is performing semantic space transformation to obtain semantic vectors for text classification. Song, Gao, Yu, Zhang, & Zhou, 2021 propose a joint auto encoder to represent case text embedding representation, which consider the statistical features and content features of case texts together. Tang, Li, Li, Zhao, & Li, 2020 analyze a classic unsupervised term weighting method and three typical supervised term weighting methods in depth to illustrate how to represent test texts that can offset the drawback confronting TF-IDF. The task of text model learning can also be divided into supervised learning and unsupervised learning. If text can be converted into numerical matrices, many classical machine learning algorithms can be used for analysis (Erdem et al., 2022). Some interesting works can be found in Liu et al., 2019; Jung & Lee, 2020; Li, 2021; Cheerkoot-Jalim & Khedo, 2021.

3. Research Objectives

In this research, we introduced Fuzzy Clustering Analysis (FCA) for loan audit short texts. The main objectives of our research are as follows:

- The first main objective of this research is to propose an applicable intelligent approach that can be directly deal with loan audit short texts. In order to deal with the work more quickly, accurately and objectively, and reduce manual labor, we carry out named entities recognition algorithm to find the basic entities and attributes, then use fuzzy membership to code documents into matrixes, finally we conduct fuzzy clustering analysis with Prim minimum spanning trees, which is more effective.
- The second objective is to solve the incremental clustering problem with increasing samples. Traditional clustering algorithms, such as KNN, KMeans and BIRCH, are greatly affected by the initial settings. When there are a certain number of new samples, the results of the model often need to be retrained. The FCA method proposed in this paper can effectively ensure the stability of clustering results. When new samples come, it is not necessary to retrain the model, and it can be simply performed through incremental calculation.

- The final objective is to evaluate and interpret the results of text clustering, so as to discover hidden knowledge that can lead risk control in practical business.

4. Fuzzy Clustering Analysis

In this section, we illustrate fuzzy clustering analysis (FCA), which is used for discovering hidden knowledge from loan audit short texts. Here, FCA is mainly divided into three steps. Firstly, lexical analysis is carried out, which is mainly to find the key tokens. Secondly, the fuzzy matrix is generated, in which the main purpose is to transform the unstructured short text into standard numerical fuzzy matrix, fuzzy matrix is based on fuzzy numbers which is defined as “degree of truth or determination” instead of Boolean logic “1 or 0”. The final step is fuzzy clustering, the step will assign each cluster with $\alpha - cut$ degree, which is to control the similarity degree of clusters. The FCA steps can be seen in the following.

4.1 Lexical analysis

Most of the short credit audit documents come from the quarterly visit and report of credit salesmen, so there is a certain amount of noise, moreover, it is quite different from the general thesaurus in terms of word frequency distribution. Meanwhile, similar to other Chinese lexical analysis, theses short texts also need to deal with word segmentation, punctuations, words variations, special characters, etc. We apply these steps to preprocess the documents (Han, Rajasekar, & Li, 2022). Here, we randomly select 500 documents to construct bag of words.

4.1.1 Segmentation and annotation

First, we use Baidu (https://ai.baidu.com/tech/nlp_basic/lexical) to conduct tagging according to syntactical functions and morphological features, then every short text can change into labeled segmented words, a few of the labeled data can be seen in Fig. 1. We remove almost the function words, such as articles, prepositions, conjunctions, auxiliary words, and interjections. Meanwhile because of the lack of changes, we also deleted a few notional words, such as pronoun (Li & Han, 2023). Finally, the left words are used in analyzing bag of words.

4.1.2 Bag of words

In this step, we use the bag of words for extracting features from the labeled data, which is the-state-of-art method for calculation of entities’ occurrences in document collection (Ferreira, Lins, Simske, Freitas, & Riss, 2016). In the bag of words, documents are represented by multiset of words, which is called a bag. Let the document be noted as , and the entities be noted as , then after finding the frequencies of entities in the document, Eq. (1) can be used to calculate the frequencies of entities in documents.

$$f_{i,j} = \sum_{k}^{m_{i,j}} m_{k,j}$$

1

Where, $m_{i,j}$ is the number of the occurrences of the considered term t_i in the document. The range of and is dependent on terms occurrences in documents which are changed according to text corpus.

In each document, entities are sorted by their occurrences, which are converted into a vector with the occurrences. In our work, this step is conducted through Weka. Then we count the frequency of entities, furthermore, according to Zipf's law (Takahashi & Tanaka-Ishii, 2019), we know that few words are responsible for the largest proportion of a written text, we only select entities whose cumulative frequency accounts for 60% as the entities of this study (Wang, Lin, & Han, 2023). Part of the entities after calculating the bag of words can be seen in Fig. 2. The frequency and cumulative frequency of the entities can be seen in Fig. 3. From Fig. 3, we choose the words from ‘enterprise’ to ‘tax’ as entities.

4.2 Fuzzy matrixes calculating

Here we focus on all the descriptions and behaviors of these target entities which are the only three verbs and adjectives words around the entity. For example, “boss” in these documents has actions and descriptions such as be contacted, be not contacted, be at work, study outside, be on travel, good, frugal, hardworking, etc., which can be defined as attributes and noted as x_i, the numbers of attributes with every entity can be seen in Table 1.

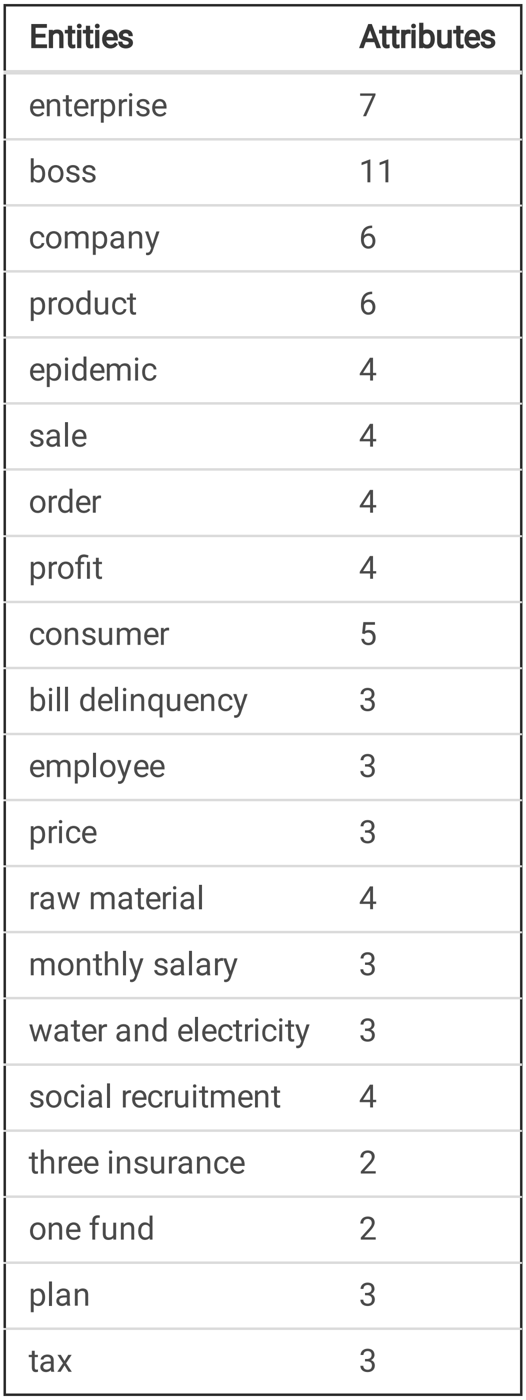
Then, we calculate the membership function with any attribute using term probability as Eq. (2).

$$\underset{\sim}{A}(x_i) = \frac{f_{i,j}(x_i)}{\sum_i f_j(x_i)}$$

2

where $\sum_i f_j(x_i)$ represents the total frequency of all attributes with entity j, and $f_{i,j}(x_i)$ represents the frequency of attribute x_i in one document, then the short loan document can be defined as a fuzzy set $\underset{\sim}{A} = \frac{\underset{\sim}{A}(x_1)}{x_1} + \frac{\underset{\sim}{A}(x_2)}{x_2} + \cdots + \frac{\underset{\sim}{A}(x_n)}{x_n}$, where x_i represents an attribute, and $\underset{\sim}{A}(x_i)$ represents membership. After all entities have done the same process in attributes learning, each loan audit short texts can be characterized by a fuzzy vector. Table 2 shows the fuzzy matrixes of the five records.

Table 1
Total number of attributes



Entities	Attributes
enterprise	7
boss	11
company	6
product	6
epidemic	4
sale	4
order	4
profit	4
consumer	5
bill delinquency	3
employee	3
price	3
raw material	4
monthly salary	3
water and electricity	3
social recruitment	4
three insurance	2
one fund	2
plan	3
tax	3

Table 2
Fuzzy matrix of the five records

	D1	D2	D3	D4	D5
enterprise	3/7	2/7	0	0	0
boss	1/11	0	3/11	2/11	1/11
company	0	0	1/6	1/6	0
product	0	0	0	0	1/6
epidemic	0	0	0	0	1/4
sale	1/4	0	0	0	0
order	0	0	0	0	1/2
profit	0	0	0	0	1/2
consumer	0	0	1/5	0	0
bill delinquency	0	0	0	1/3	0
employee	1/3	0	0	0	0
price	0	0	0	0	1/3
raw material	0	0	0	0	1/4
monthly salary	0	2/3	0	1/3	0
water and electricity	0	1/3	0	0	0
social recruitment	0	0	0	1/2	0
three insurance	0	0	0	0	0
one fund	0	0	0	1	0
plan	0	0	0	0	1/3
tax	0	0	0	0	0

4.3 Fuzzy clustering

4.3.1 Minimum spanning trees

Firstly, calculating the similarity between samples. there are several methods to calculate the distance, such as Eq. (3) to Eq. (5), which is named as Euclidean distance, Hamming distance and Chebyshev distance. And the distance graph of 10 samples can be seen in Fig. 4.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m} \left(\underset{\sim}{A}(x_{ik}) - \underset{\sim}{A}(x_{jk})\right)^2}$$

3

$$d(x_i, x_j) = \sum_{k=1}^{m} \left|\underset{\sim}{A}(x_{ik}) - \underset{\sim}{A}(x_{jk})\right|$$

4

$$d(x_i, x_j) = \bigvee_{k=1}^{m} \left|\underset{\sim}{A}(x_{ik}) - \underset{\sim}{A}(x_{jk})\right|$$

5

Then, we can calculate the similarity with two samples through Eq. (6).

$$s(x_i, x_j) = \frac{1}{d(x_i, x_j)}$$

6

Secondly, construct minimum spanning tree. We can get an undirected graph uses samples as vertices and the similarity between samples as the weight of edges. The degree of the vertex can be calculated by $\sum r_{ij}$. Then then minimum spanning tree can be got through Prim algorithm (Pop, 2020). The results with 10 samples can be seen in Fig. 5.

Definition 1

Suppose that $A : X \to [0, 1]$ is a fuzzy subset of . Then, for any $\alpha \in [0, 1]$, the $\alpha - cut$ are crisp sets as $A^{\alpha} = \{x|(x \in X) \wedge (A(x) \geqslant \alpha)\}$.

Thirdly, we can use the $\alpha - cut$ to find the clusters in the tree. Because similarity is used as edges to associate vertices, $\alpha - cut$ can get different connectivity subtrees at different thresholds. One subtree is a cluster.

Finally, the vertex with the maximum degree in each cluster is added with other samples, and the distance is recalculated to generate a subtree until all samples are included in a corresponding cluster.

The whole process can be seen in Algorithm 1.

Algorithm 1 Fuzzy clustering analysis

Inputs: fuzzy matrix X, a

Outputs: clusters {yi}

```
Step 1: Calculate distance r_ij using Eq. (3)-(5)
Step 2: Generate minimum spanning tree
Prim()
{MST = {s};
while (1) {V = not in the Vertex list;
if (V all in the list ) break;
Add MST list: dist[V] = 0;
for ( V' nearest vertex W )
if ( E (V,W) < dist[W] )
{dist[W] = E (V,W) ;
parent[W] = V;}
}
Step 3: Select edges in the minimum spanning tree as cluster {yi} with similarity is greater than a
Step 4: Add the node with the maximum degree in a cluster to new X, and repeat step 1 until all nodes are in {yi}
```

4.3.2 Optimal $\alpha - cut$

If there are m samples and k attributes, then one sample can be noted as $x_i = (x_{i1}, x_{i2}, \cdots x_{ik})$, where $i = 1, 2, \cdots m$. Then we can calculate the mean of sample, which is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. With one $\alpha - cut$, it can get clusters, and let the jth cluster have the number of samples as n_j, and every sample in the jth cluster be noted as $\bar{x}^{(j)} = (\bar{x}_1{}^{(j)}, \bar{x}_2{}^{(j)}, \cdots, \bar{x}_k{}^{(j)})$, then we can get F statistics as Eq. (7).

$$F = \frac{\sum_{j=1}^{r} n_j ||\bar{x}^{(j)} - \bar{x}||^2/(r-1)}{\sum_{j=1}^{r}\sum_{i=1}^{n_j} ||x^{(j)} - \bar{x}^{(j)}||^2/(n-r)}$$

7

Where $||\bar{x}^{(j)} - \bar{x}|| = \sqrt{\sum_{k=1}^{m} (\bar{x}_k^{(j)} - \bar{x}_k)^2}$. It can be seen that the numerator of F statistics represents the distance between clusters, and the denominator of F statistics represents the distance between samples within the cluster. Thus, the larger the value of F statistic is, the better the clustering has done. Therefore, when we conduct the experiments, we can set the threshold of F statistics to find the optimal $\alpha - cut$ through $F > F_{0.05}(r-1, n-r)$.

5. Experimental Results

In this section, we illustrate FCA with experiments of loan audit short texts. The data of this study is from the credit audit data of small and micro enterprises (SMEs) of a city commercial bank in China in 2020. There are 5866 SMEs that have the loan records. According to regulations, credit business auditors need to at least visit these SMEs every month and record the relevant information in time. There are 11,2632 records and 22 basic variables, of which 4 variables are the daily records of the survey of credit salesmen. Here we put our focus on the 4 variables- operation status, manager status, abnormal conditions, others. The 4 variables are all short texts. We deleted the missing records and duplicated records, and randomly selects 2 records in each SME, then we get the initial training sample with 1,1732 records. The work flow of the experiments can be seen in Fig. 6.

We will introduce the results of the experiments in two parts: in the first part we explore FCA clustering results in details, and in the other part we compare the FCA clusters with KNN and Kmean using text indicators such as IDF, Entropy and Purity (Moscato, Picariello, & Sperlí, 2021) to make a general conclusion.

5.1 Results of FCA clustering

Here we will explore the whole process of FCA clustering. Firstly, we explore the general results with the document samples under different level of $\alpha - cut$; secondly, we compare every result with F statistics to find the optimal a-cut; finally, we use the optimal $\alpha - cut$ to analyze the whole sample and get the clusters.

The lexical analysis is conducted by Baidu (https://ai.baidu.com/tech/nlp_basic/lexical), and using the bag of word as we proposed in section 4.1. And the fuzzy matrix calculating and fuzzy clustering is conducted by Matlab 2018b.

5.1.1 Numbers of minimum spanning trees under different $\alpha - cut$

We use Eq. (3) to Eq. (5) to calculate the distance between different samples. As α changes, the number of minimum spanning trees changes which is shown in Fig. 7.

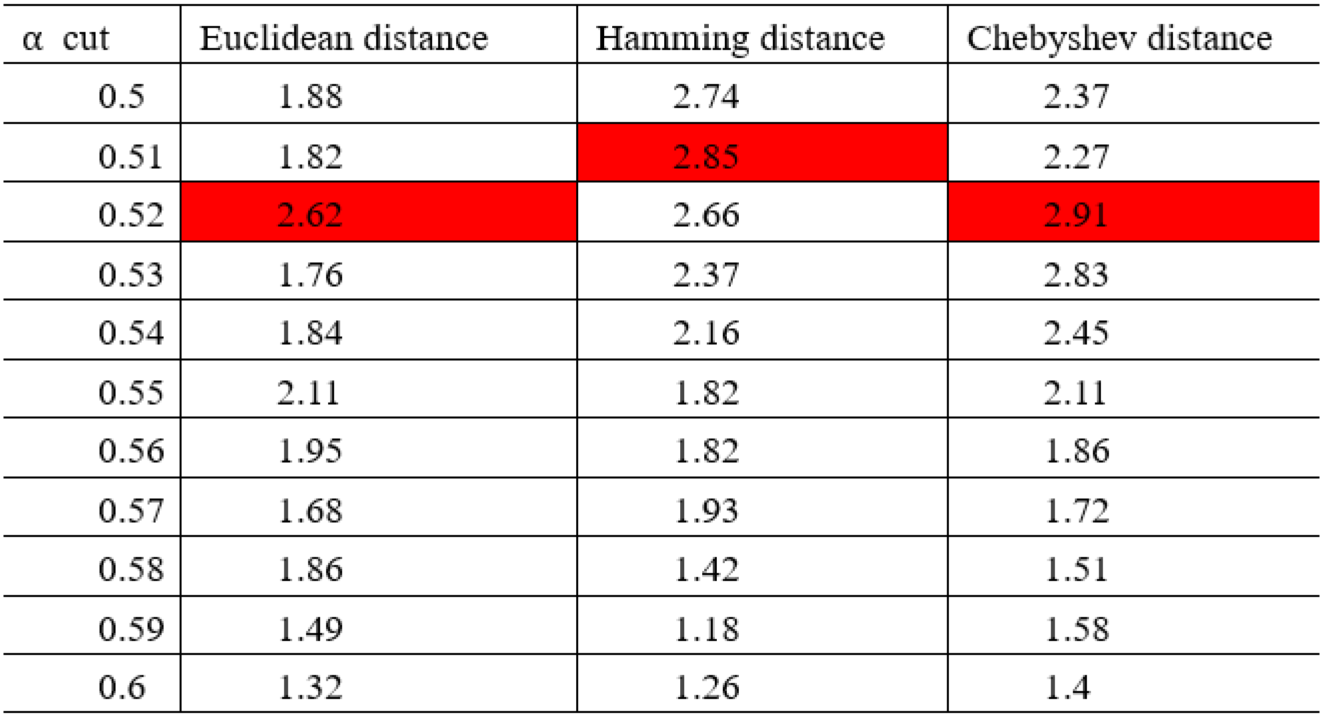
It can be seen from Fig. 7, no matter what distance definition method is used when α is less than 0.5, the number of minimum spanning trees tends to be quite stable, and it is near 0; and when α is greater than 0.6, the number of minimum spanning trees begins explosive growth. Therefore, when α is greater than

0.5 and less than 0.6, the number of minimum spanning trees has a relatively ideal result for model training. The corresponding numbers of minimum spanning trees are 83 and 973, respectively.

5.1.2 Optimal α cut

Here we start with α as 0.5, then we calculate the F statistics for each result of 0.01 increments in α, and the upper limit is 0.6. The results can be seen in Table 3.

Table 3. F statistics with different **α** cut using different distance



α cut	Euclidean distance	Hamming distance	Chebyshev distance
0.5	1.88	2.74	2.37
0.51	1.82	2.85	2.27
0.52	2.62	2.66	2.91
0.53	1.76	2.37	2.83
0.54	1.84	2.16	2.45
0.55	2.11	1.82	2.11
0.56	1.95	1.82	1.86
0.57	1.68	1.93	1.72
0.58	1.86	1.42	1.51
0.59	1.49	1.18	1.58
0.6	1.32	1.26	1.4

It can be seen from Table 3 that both under Euclidean distance and Chebyshev distance, the optimal α cut is 0.52, while under Hamming distance, the optimal α cut is 0.51. At the same time, combining with the discussion of the number of minimum spanning trees in 5.1.1, we can see that no matter what distance calculation is used, the numbers of the minimum spanning trees increase rapidly with the increase of α. In the range of 0.5–0.52, the number of minimum spanning trees generated by Chebyshev distance is the smallest, that means it can get the least clusters. Therefore, we set the optimal α cut as 0.52 and the distance calculation as Chebyshev distance as Eq. (5). Finally, we trained all records using the optimal α cut and Chebyshev distance, and got 276 minimum spanning trees, that is, 276 clusters.

5.2 Comparing results with other clustering algorithms

In order to compare with other clustering algorithms, we take BRICH, Kmeans and Fuzzy C-means clustering (FCM) into experiments. We use the same fuzzy matrices like Table 2, and treat every entity as a variable (Khan & Lohani, 2022).

The BRICH results of 500 records using Euclidean distance are shown in Fig. 8. Similar to that in FCA, if we take the optimal Euclidean distance as 2.6 with BRICH, we can get 193 clusters with the whole records. Moreover, if we take the clusters as 276, then BRICH returns the optimal Euclidean distance as 1.82.

Both the Kmeans and FCM need to specify the initial settings with number of clusters, so we explore the results with the clusters from 100 to 300 using Kmeans and FCM, and here we use R in Eq. (8) as the indicator to choose the optimal result. If there are m clusters and x_i is one record, then x_m is the mean of the cluster, and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the mean of the whole sample. From Eq. (8), we know that the less the R is, the better the result will be. The results of Kmeans and FCM can be seen in Fig. 9. From Fig. 9, we can find that the optimal number of clusters with Kmeans is 140, and the optimal number of clusters with FCM is 200.

$$D_{out} = \sqrt{\sum_{i=1}^{m} (x_m - \bar{x})^2}$$

$$D_{in} = \sqrt{\sum_{i=1}^{m} \sum_{p \in C_i} (p - x_m)^2}$$

$$R = \frac{D_{out}}{D_{in}}$$

8

There are different clustering results due to different algorithm settings. In order to further compare the comprehensive performance of these four methods, we explore the performance of clustering, which is measured using Entropy. Always Entropy is the measure of clusters coherence, which can be calculated by Eq. (9). In Eq. (9), L is the number of entities, C_{kl} is a number of entities with category l, c_k is the total amount of documents in k clusters, D is the number of records. Lower entropy indicates better performance. We compare the four clustering methods based on entropy from 100 to 300 clusters, as shown in Fig. 10. From Fig. 10, we can see that FCA with 240 clusters has the smallest entropy, which is the optimal setting. At the same time, in the comparison of these methods, it can be seen that entropy of kmeans is relatively high, so this method may not be suitable for analyzing these text records.

In short, through this series of comparative experiments, we can find that FCA has a better performance on these 11,2632 loan audit short text, and the optimal setting is to take α cut as 0.52, use the Chebyshev distance, and finally it will obtain about 240 clusters. In addition, BRICH, Kmeans and FCM need to reconstruct and retrain the model when meeting with new records, while FCA only needs to process new records separately and update the whole results. Therefore, from this perspective, FCA is also more suitable to deal with this issue.

6. Discussion

Although FCA has a good performance in the analyzing loan audit short text, this method still has the following shortages.

Firstly, the study only analyzes these loan audit short text records from a city bank in one year, and the results show that it is to divide these records into a number of clusters. Whether each cluster is reasonable or not still needs discussions. At the same time, according to the results, these loan audit text records eventually generated a large number of subcategories, though this is the optimal setting according to the indicators in our experiments, too many subcategories also bring a lot of difficulties to the actual management. Therefore, it is worth studying in the future to optimize the method’s setting. After all, in the FCA, any number of categories can be obtained by adjusting the threshold of α cut. There must be a trade off on the number of categories and management requirements.

Secondly, the analysis with Chinese short text in this study needs further exploration. Although this paper proposes a method to transform short text into fuzzy matrix, it is still quite rough. First of all, this method is based on a recognition dictionary which is constructed by us with 500 samples. Second, this method only focuses on the three content words before and after the entity, which will loss a number of attributes. Finally, this method cannot depict the whole semantic sentence with the fuzzy numerical matrix. In the fuzzy matrix, the fuzzy number only represent the number of attributes, which loss some key information of their positions Therefore, there will need more improvement on the text processing in further.

Lastly, in terms of the FCA parameter setting, the current α cut is obtained by comparing experimental results, so it requires a large number of repeated experiments, which seriously affects the efficiency and deployment of the method. In the future, if we can theoretically analyze the functional relationship of α cut, the number of clusters and entropy or other evaluation indicators, and get the method to theoretically optimize the parameter setting, it will be a significant improvement.

7. Conclusion

The current pressing issue in China's commercial banks is finding a solution to analyze loan audit short texts. In this paper, we propose a method that transforms text records into standard fuzzy matrix, which are then used to generate clusters through the minimum spanning tree. Through the comparison experiments with BRICH, Kmeans and FCM, we find that not only does FCA proved better clustering results, but it also has the capability to process incremental data with greater efficiency.

However, the text processing utilized in this study remains rather rough. Furthermore, the number of clusters obtained in this study is far more than the number of categories concerned by the current management norms (the credit management of China's commercial banks implements five-level classification). Therefore, it will be useful to explore the text processing technology, and optimize the FCA setting in combination with the management practice.

Declarations

Acknowledgements

The work was supported by the National Natural Science Foundation of China (Grant No. 72101279).

References

1. Acheampong A, Elshandidy T (2021) Does soft information determine credit risk? Text-based evidence from European banks. JOURNAL OF INTERNATIONAL FINANCIAL MARKETS INSTITUTIONS & MONEY, 75. doi: 10.1016/j.intfin.2021.101303
2. Ahmed M, Chen Q, Li ZH (2020) Constructing domain-dependent sentiment dictionary for sentiment analysis. Neural Comput Appl 32(18):14719–14732. 10.1007/s00521-020-04824-8
3. Alekseev, V., Egorov, E., Vorontsov, K., Goncharov, A., Nurumov, K.,… Buldybayev,T. (2021). TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation. DATA & KNOWLEDGE ENGINEERING, 135.doi: 10.1016/j.datak.2021.101921
4. Altman EI (2018) A fifty-year retrospective on credit risk models, the Altman Z-score family of models and their applications to financial markets and managerial strategies. J CREDIT RISK 14(4):1–34. 10.21314/JCR.2018.243
5. Cao J, Xu X, Yin X, Pan B (2022) A risky large group emergency decision-making method based on topic sentiment analysis. Expert Syst Appl 195:116527. https://doi.org/10.1016/j.eswa.2022.116527
6. Chauhan U, Shah A (2021) Topic Modeling Using Latent Dirichlet allocation: A Survey. ACM-CSUR 54(7). 10.1145/3462478
7. Chen Y, Zhang H, Liu R, Ye Z, Lin J (2019) Experimental explorations on short text topic mining between LDA and NMF based Schemes. Knowl Based Syst 163:1–13. https://doi.org/10.1016/j.knosys.2018.08.011
8. Cheerkoot-Jalim S, Khedo KK (2021) A systematic review of text mining approaches applied to various application areas in the biomedical domain. J Knowl Manage 25(3):642–668. 10.1108/JKM-09-2019-0524
9. Choudhary N, Aggarwal CC, Subbian K, Reddy CK (2022) Self-supervised Short-text Modeling through Auxiliary Context Generation. ACM Trans Intell Syst Technol 13(3):51. 10.1145/3511712
10. Consoli S, Barbaglia L, Manzan S (2022) Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. Knowl Based Syst 247:108781. https://doi.org/10.1016/j.knosys.2022.108781
11. Dastile X, Celik T, Potsane M (2020) Statistical and machine learning models in credit scoring: A systematic literature survey. Appl Soft Comput 91:106263. https://doi.org/10.1016/j.asoc.2020.106263
12. Dong, C., Li, Y., Gong, H., Chen, M., Li, J., Shen, Y.,… Yang, M. (2022). A Survey of Natural Language Generation. ACM Comput. Surv., 55(8). doi: 10.1145/3554727
13. Erdem, E., Kuyu, M., Yagcioglu, S., Frank, A., Parcalabescu, L., Plank, B.,… Korvel,G. U. A. Z. (2022). Neural Natural Language Generation: A Survey on Multilinguality,Multimodality, Controllability and

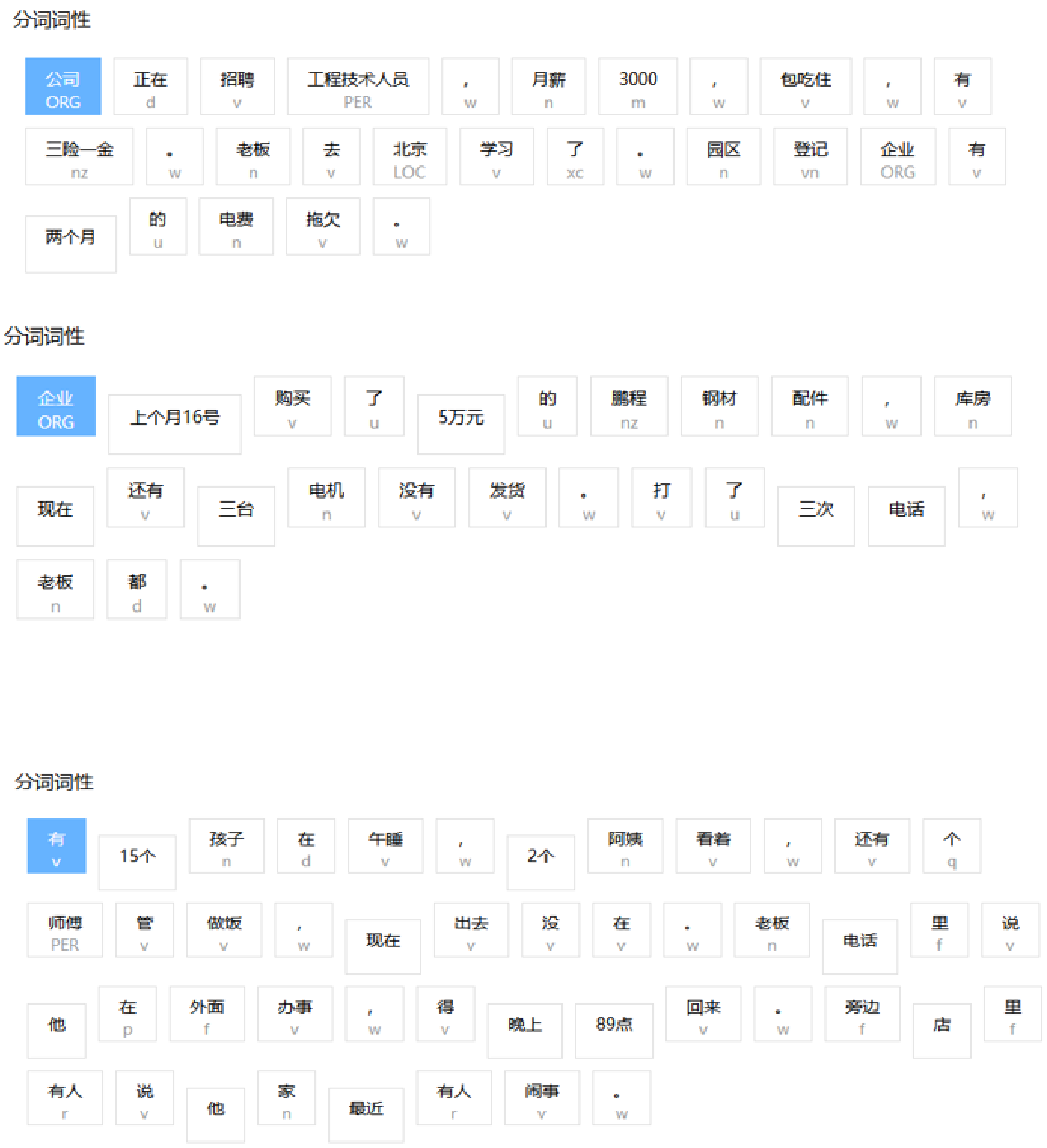
Learning. J. Artif. Int. Res., 73. doi: 10.1613/jair.1.12918

14. Fan F, Zhao WX, Wen J, Xu G, Chang EY (2017) Mining collective knowledge: inferring functional labels from online review for business. Knowl Inf Syst 53(3):723–747. 10.1007/s10115-017-1050-4
15. Feng, J., Zhang, Z., Ding, C., Rao, Y., Xie, H.,… Wang, F. L. (2022). Context reinforced neural topic modeling over short texts. Information Sciences, 607, 79–91. doi: https://doi.org/10.1016/j.ins.2022.05.098
16. Ferreira R, Lins RD, Simske SJ, Freitas F, Riss M (2016) Assessing sentence similarity through lexical, syntactic and semantic analysis. Comput Speech Lang 39:1–28. https://doi.org/10.1016/j.csl.2016.01.003
17. Gunnarsson BR, Broucke V, Baesens S, Óskarsdóttir B, M., Lemahieu W (2021) Deep learning for credit scoring: Do or don't? Eur J Oper Res 295(1):292–305. https://doi.org/10.1016/j.ejor.2021.03.006
18. Han L, Rajasekar A, Li S (2022) An evidence-based credit evaluation ensemble framework for online retail SMEs. Knowl Inf Syst 64(6):1603–1623. 10.1007/s10115-022-01682-9
19. Han HY, Zhang JP, Yang J, Shen YR, Zhang YS (2018) Generate domain-specific sentiment lexicon for review sentiment analysis. MULTIMEDIA TOOLS AND APPLICATIONS 77(16):21265–21280. 10.1007/s11042-017-5529-5
20. Jiang C, Lu W, Wang Z, Ding Y (2023) Benchmarking state-of-the-art imbalanced data learning approaches for credit scoring. Expert Syst Appl 213:118878. https://doi.org/10.1016/j.eswa.2022.118878
21. Jung H, Lee BG (2020) Research trends in text mining: Semantic network and main path analysis of selected journals. EXPERT SYSTEMS WITH APPLICATIONS, 162. doi: 10.1016/j.eswa.2020.113851
22. Khan MS, Lohani QMD (2022) Topological analysis of intuitionistic fuzzy distance measures with applications in classification and clustering. Eng Appl Artif Intell 116:105415. https://doi.org/10.1016/j.engappai.2022.105415
23. Kozodoi N, Jacob J, Lessmann S (2022) Fairness in credit scoring: Assessment, implementation and profit implications. Eur J Oper Res 297(3):1083–1094. https://doi.org/10.1016/j.ejor.2021.06.023
24. Li S, Han L (2023) A Two-Stage NER Method for Online-Sale Comments. Springer Nat Singap. 10.1007/978-981-19-2768-3_26
25. Li M (2021) Capturing the Risk Signals for a Specific Emerging Technology: An Integrated Framework of Text Mining. IEEE Trans Eng Manage 68(5):1245–1258. 10.1109/TEM.2019.2930335
26. Liu, S., Wang, X., Collins, C., Dou, W., Ouyang, F., El-Assady, M.,… Keim, D. A. (2019).Bridging Text Visualization and Mining: A Task-Driven Survey. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, 25(7), 2482–2504. doi: 10.1109/TVCG.2018.2834341
27. Louzada F, Ara A, Fernandes GB (2016) Classification methods applied to credit scoring: Systematic review and overall comparison. Surv Oper Res Manage Sci 21(2):117–134. https://doi.org/10.1016/j.sorms.2016.10.001

28. Moscato V, Picariello A, Sperlí G (2021) A benchmark of machine learning approaches for credit score prediction. Expert Syst Appl 165:113986. https://doi.org/10.1016/j.eswa.2020.113986
29. Pop PC (2020) The generalized minimum spanning tree problem: An overview of formulations, solution procedures and latest advances. Eur J Oper Res 283(1):1–15. https://doi.org/10.1016/j.ejor.2019.05.017
30. Rahimi Z, Homayounpour MM (2020) Tens-embedding: A Tensor-based document embedding method. Expert Syst Appl 162. 10.1016/j.eswa.2020.113770
31. Rashid J, Shah SMA, Irtaza A (2019) Fuzzy topic modeling approach for text mining over short text. Inf Process Manag 56(6):102060. https://doi.org/10.1016/j.ipm.2019.102060
32. Ruspini EH (1969) A new approach to clustering. Inf Control 15(1):22–
33. S., M., Z., A., Y., T., R., H., M., S. H.,… H., Z. (2019). An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. IEEE Access,7, 93010–93022. doi: 10.1109/ACCESS.2019.2927266
34. Silva DMB, Pereira GHA, Magalhães TM (2022) A class of categorization methods for credit scoring models. Eur J Oper Res 296(1):323–331. https://doi.org/10.1016/j.ejor.2021.04.029
35. Sinoara RA, Camacho-Collados J, Rossi RG, Navigli R, Rezende SO (2019) Knowledge-enhanced document embeddings for text classification. Knowl Based Syst 163:955–971. 10.1016/j.knosys.2018.10.026
36. Shi Y, Zhu LY, Li W, Gao K, Zheng YC (2019) Survey on Classic and Latest Textual Sentiment Analysis Articles and Techniques, vol 18. INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY & DECISION MAKING, pp 1243–1287. 410.1142/S0219622019300015
37. Song R, Gao S, Yu Z, Zhang Y, Zhou G (2021) Case2vec: joint variational autoencoder for case text embedding representation. Int J Mach Learn Cybernet 12(9):2517–2528. 10.1007/s13042-021-01335-3
38. Srivastava R, Singh P, Rana KPS, Kumar V (2022) A topic modeled unsupervised approach to single document extractive text summarization. Knowl Based Syst 246:108636. https://doi.org/10.1016/j.knosys.2022.108636
39. Stevenson M, Mues C, Bravo C (2021) The value of text for small business default prediction: A Deep Learning approach. Eur J Oper Res 295(2):758–771. 10.1016/j.ejor.2021.03.008
40. Takahashi S, Tanaka-Ishii K (2019) Evaluating Computational Language Models with Scaling Properties of Natural Language. Comput Linguist 45(3):481–513. 10.1162/coli_a_00355
41. Tang Z, Li W, Li Y, Zhao W, Li S (2020) Several alternative term weighting methods for text representation and classification. Knowl Based Syst 207. 10.1016/j.knosys.2020.106399
42. Wang J, Lin J, Han L (2023) Word2vec Fuzzy Clustering Algorithm and Its Application in Credit Evaluation. Springer Nat Singap. 10.1007/978-981-19-2768-3_56
43. Wang Z, Jiang C, Zhao H, Ding Y (2020) Mining Semantic Soft Factors for Credit Risk Evaluation in Peer-to-Peer Lending. J Manage Inform Syst 37(1):282–308. 10.1080/07421222.2019.1705513

44. Wang L, Niu JW, Yu S (2020) SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis. IEEE Trans Knowl Data Eng 32(10):2026–2039. 10.1109/TKDE.2019.2913641
45. Wang HY, Wang J, Wang G (2022) A survey of fuzzy clustering validity evaluation methods. Inf Sci 618:270–297. 10.1016/j.ins.2022.11.010
46. Wu Y, Zhao S, Li W (2020) Phrase2Vec: Phrase embedding based on parsing. Inf Sci 517:100–127. 10.1016/j.ins.2019.12.031
47. Xu J, Liu J, Araki K (2015) A Hybrid Topic Model for Multi-Document Summarization, vol E98D. IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS, pp 1089–1094. 510.1587/transinf.2014EDP7229
48. Xu, G. X., Yu, Z. H., Yao, H. S., Li, F., Meng, Y. T.,… Wu, X. (2019). Chinese Text Sentiment Analysis Based on Extended Sentiment Dictionary. IEEE ACCESS, 7, 43749–43762.doi: 10.1109/ACCESS.2019.2907772
49. Yadollahi A, Shahraki AG, Zaiane OR (2017) Current State of Text Sentiment Analysis from Opinion to Emotion Mining. ACM Comput Surv 50(2):25. 10.1145/3057270
50. Yang K, Yuan H, Lau RYK (2022) PsyCredit: An interpretable deep learning-based credit assessment approach facilitated by psychometric natural language processing. Expert Syst Appl 198:116847. https://doi.org/10.1016/j.eswa.2022.116847
51. Yu H (2020) Bibliographic automatic classification algorithm based on semantic space transformation. MULTIMEDIA TOOLS AND APPLICATIONS 79(13–14):9283–9297. 10.1007/s11042-019-7400-3
52. Zhou H, Yu H, Hu R (2017) Topic evolution based on the probabilistic topic model: a review. Front Comput Sci 11(5):786–802. 10.1007/s11704-016-5442-5
53. Žitnik S, Blagus N, Bajec M (2022) Target-level sentiment analysis for news articles. Knowl Based Syst 249:108939. https://doi.org/10.1016/j.knosys.2022.108939

Figures



分词词性

公司	ORG
正在	d
招聘	v
工程技术人员	PER
，	w
月薪	n
3000	m
，	w
包吃住	v
，	w
有	v
三险一金	nz
。	w
老板	n
去	v
北京	LOC
学习	v
了	xc
。	w
园区	n
登记	vn
企业	ORG
有	v
两个月	
的	u
电费	n
拖欠	v
。	w

分词词性

企业	ORG
上个月16号	
购买	v
了	u
5万元	
的	u
鹏程	nz
钢材	n
配件	n
，	w
库房	n
现在	
还有	v
三台	
电机	n
没有	v
发货	v
。	w
打	v
了	u
三次	
电话	
，	w
老板	n
都	d
。	w

分词词性

有	v
15个	
孩子	n
在	d
午睡	v
，	w
2个	
阿姨	n
看着	v
，	w
还有	v
个	q
师傅	PER
管	v
做饭	v
，	w
现在	
出去	v
没	v
在	v
。	w
老板	n
电话	
里	f
说	v
他	
在	p
外面	f
办事	v
，	w
得	v
晚上	
89点	
回来	v
。	w
旁边	f
店	
里	f
有人	r
说	v
他	
家	n
最近	
有人	r
闹事	v
。	w

Figure 1

A few samples of tagging with the loan audit short texts

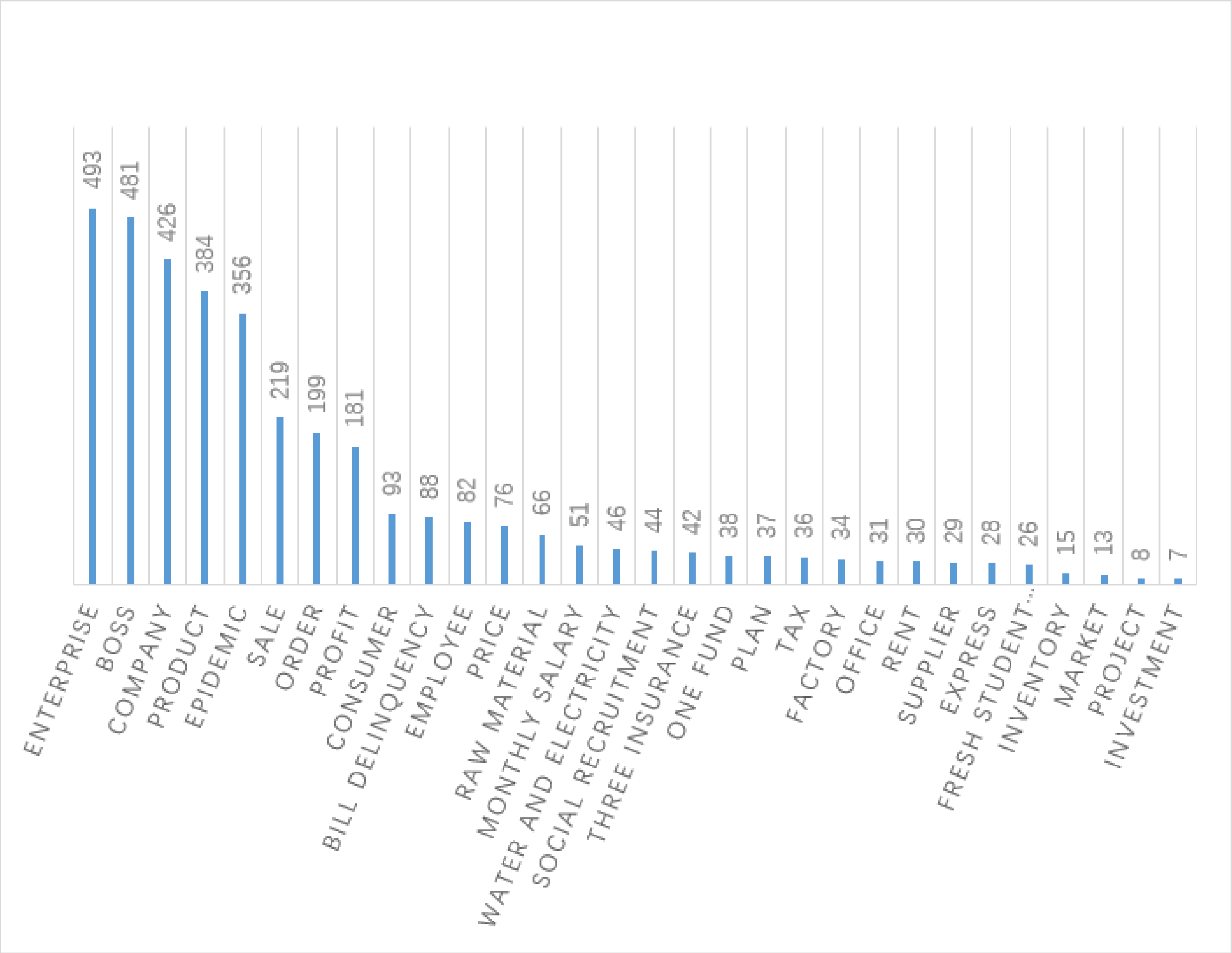


Figure 2

Part of the entities after calculating the bag of words

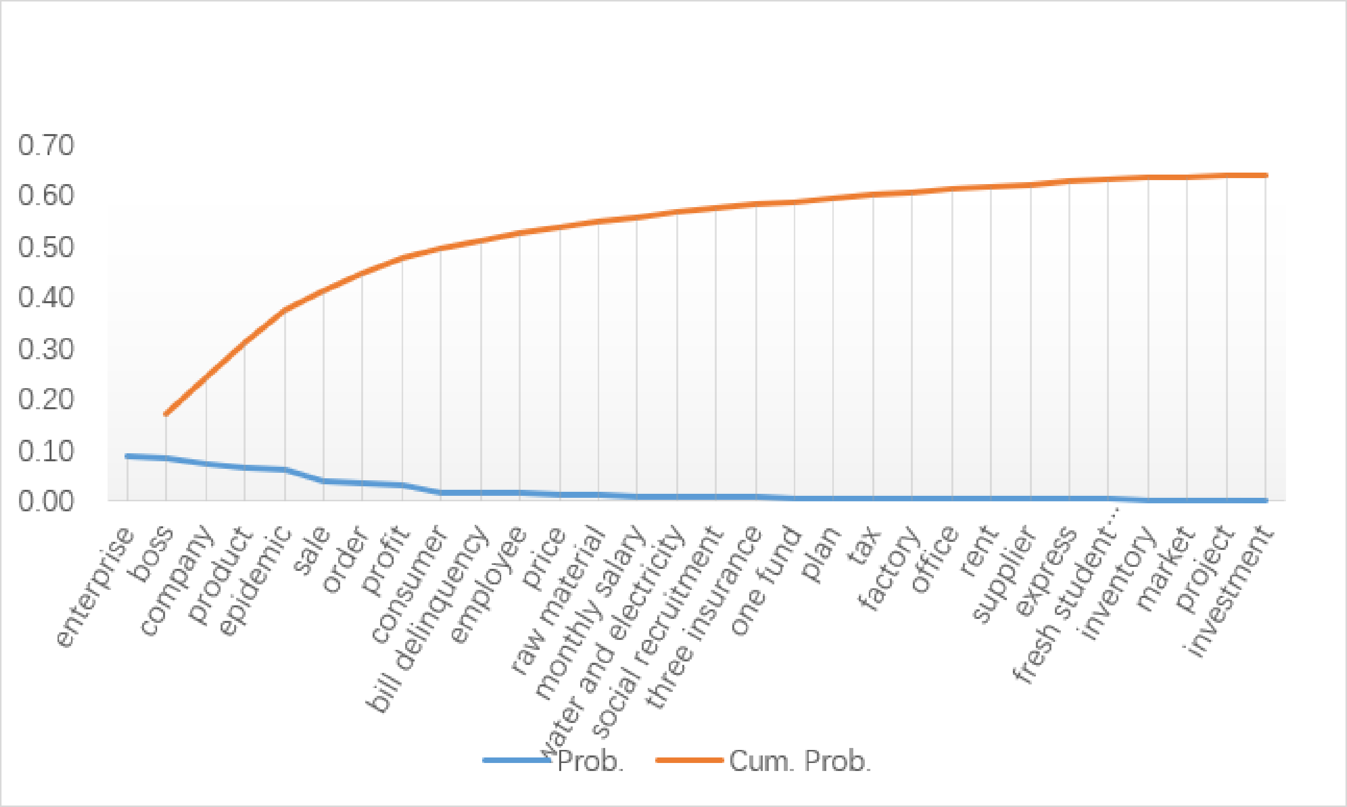


Figure 3

Frequency and cumulative frequency of the entities

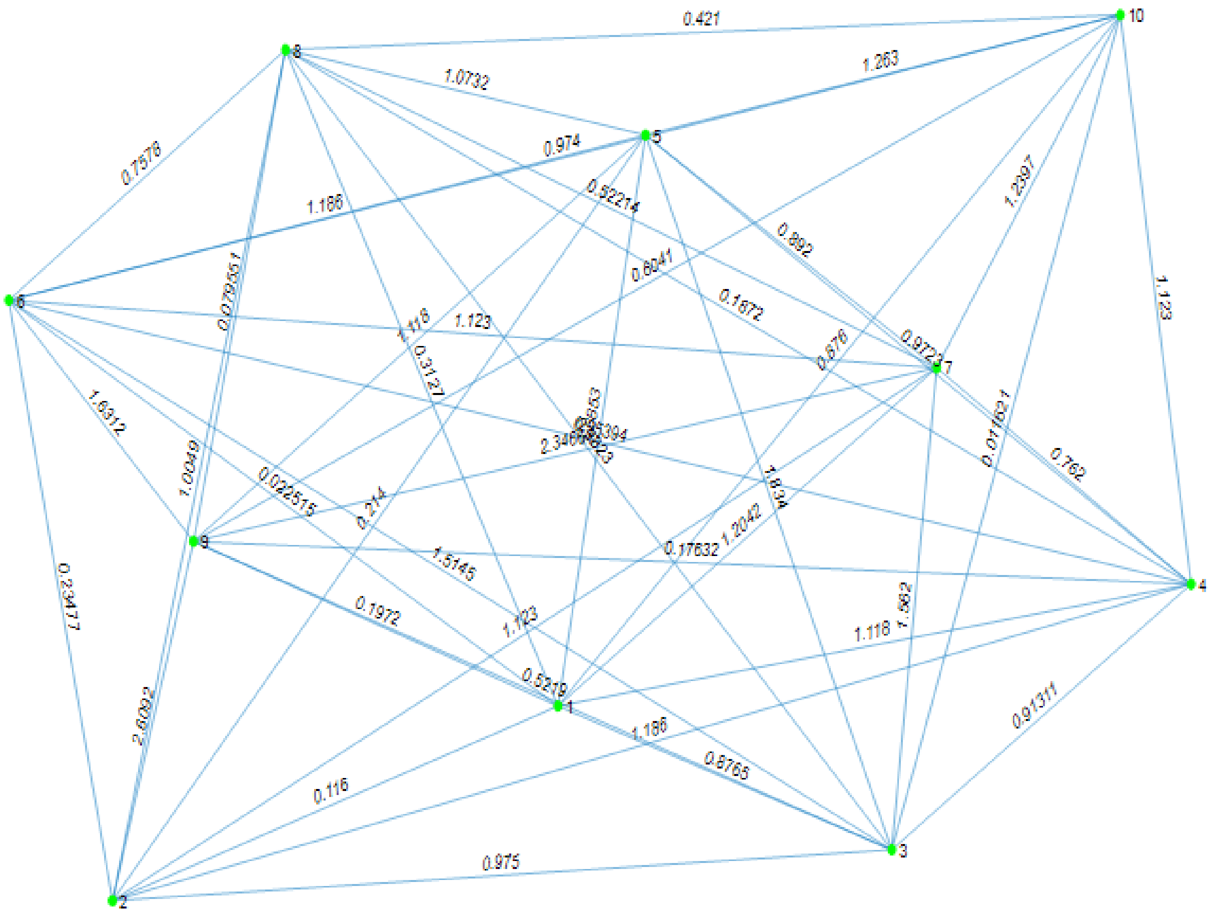


Figure 4

The distance graph of 10 samples

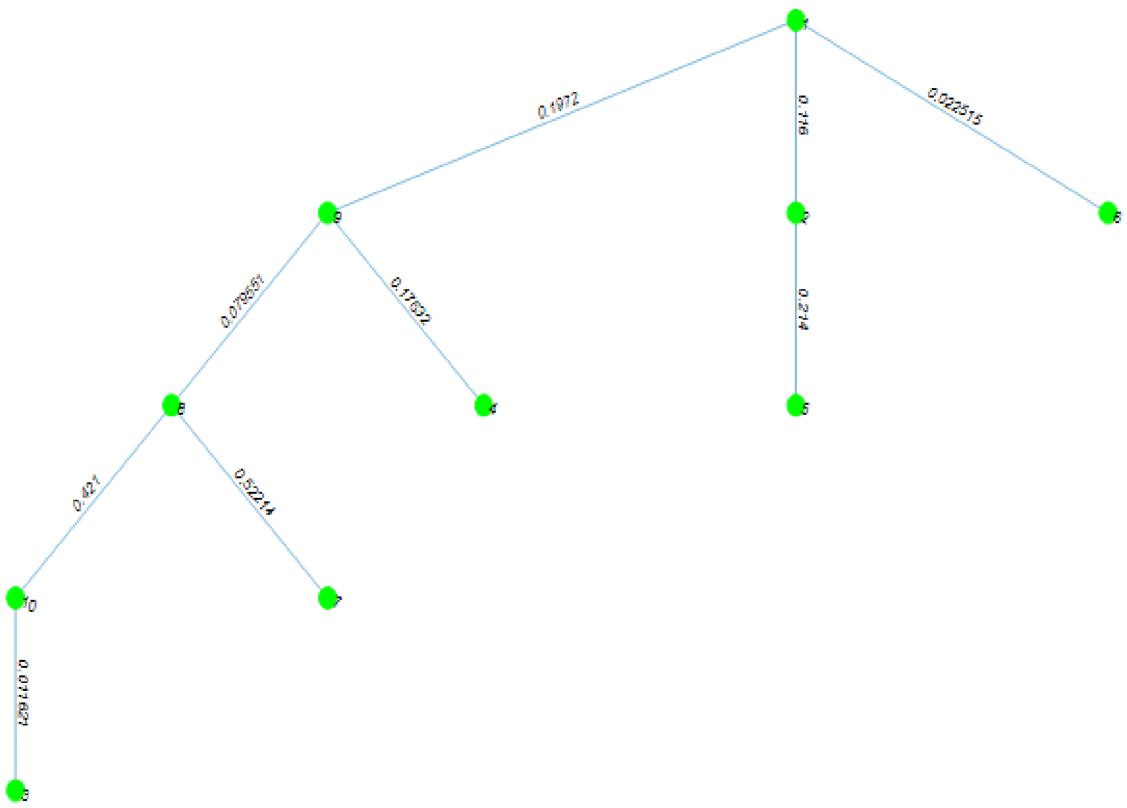


Figure 5

The minimum spanning tree of 10 samples

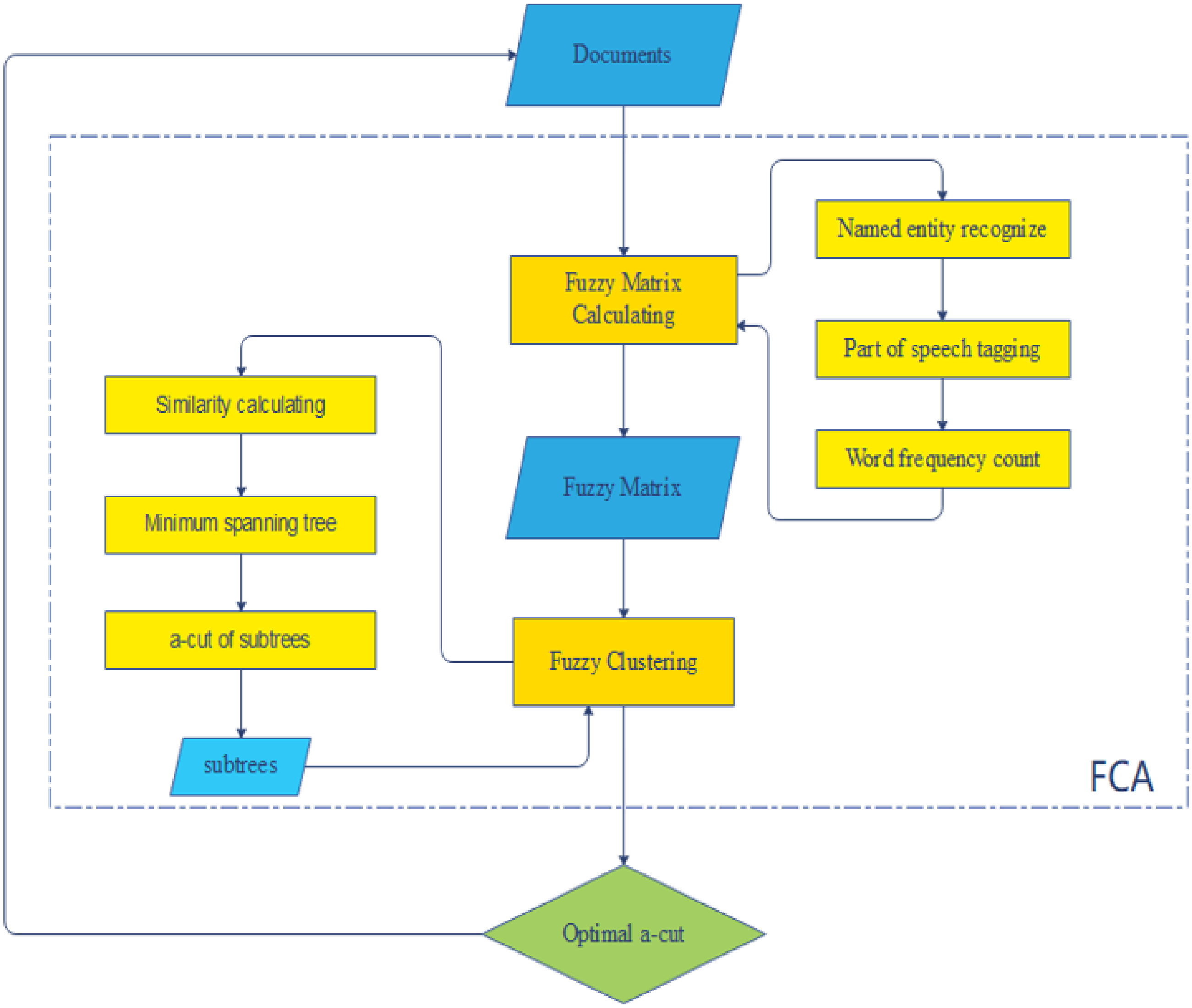


Figure 6

The work flow of FCA experiments

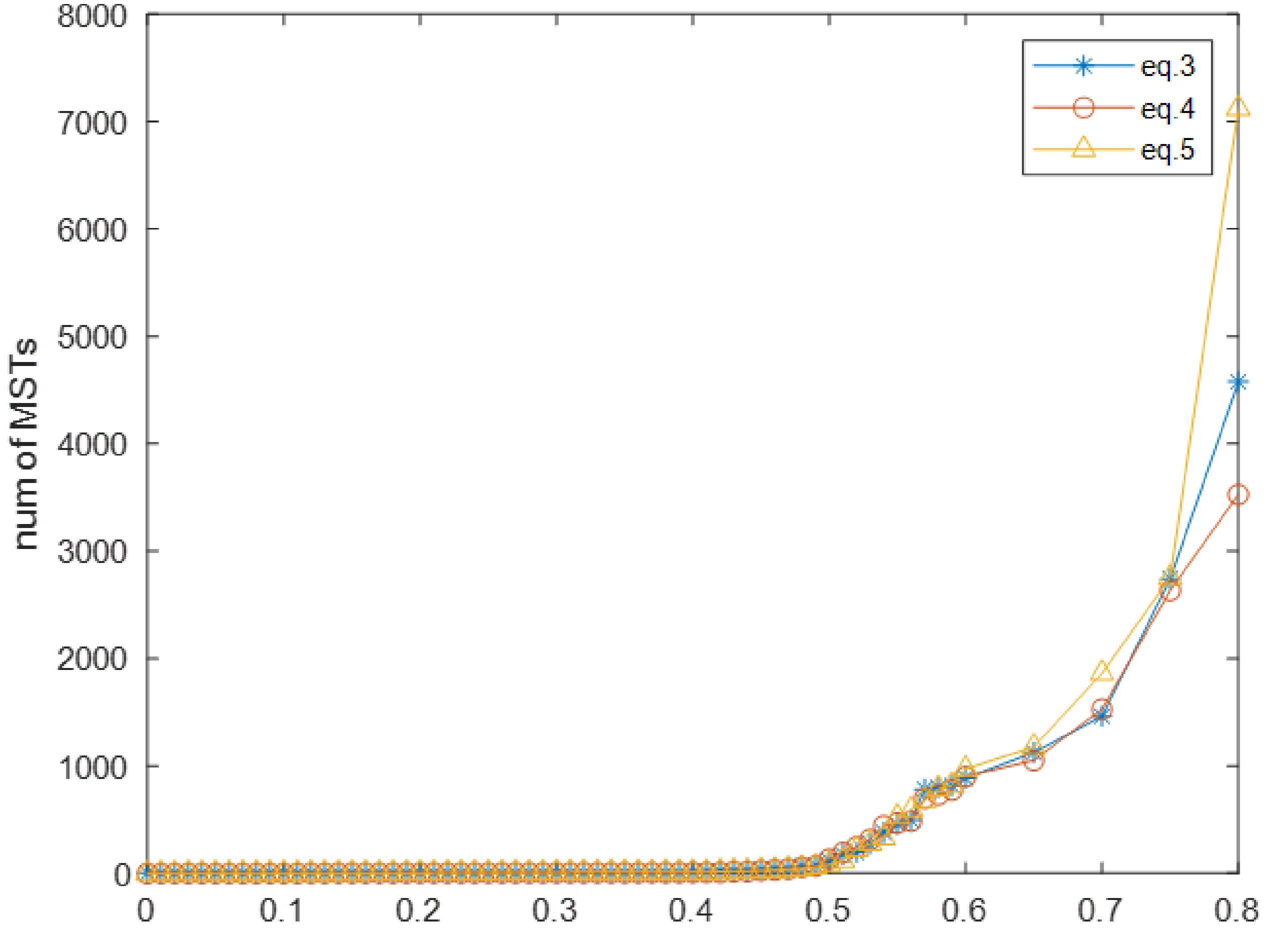


Figure 7

The number of minimum spanning trees with different a-cut

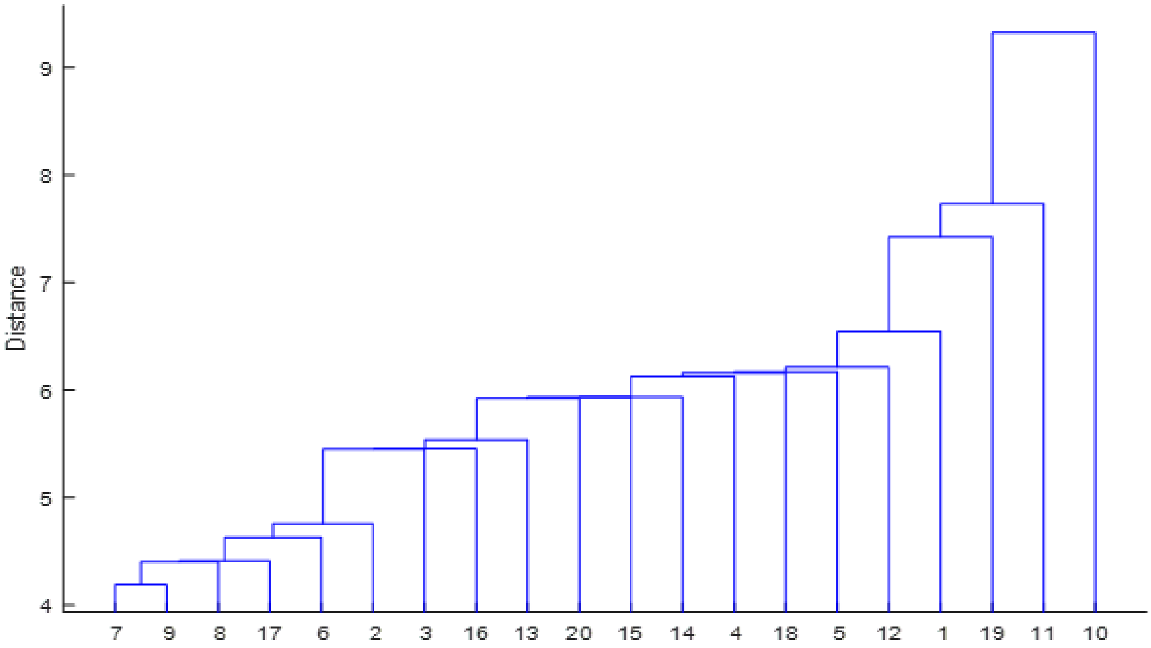


Figure 8

The BRICH results of 500 records using Euclidean distance

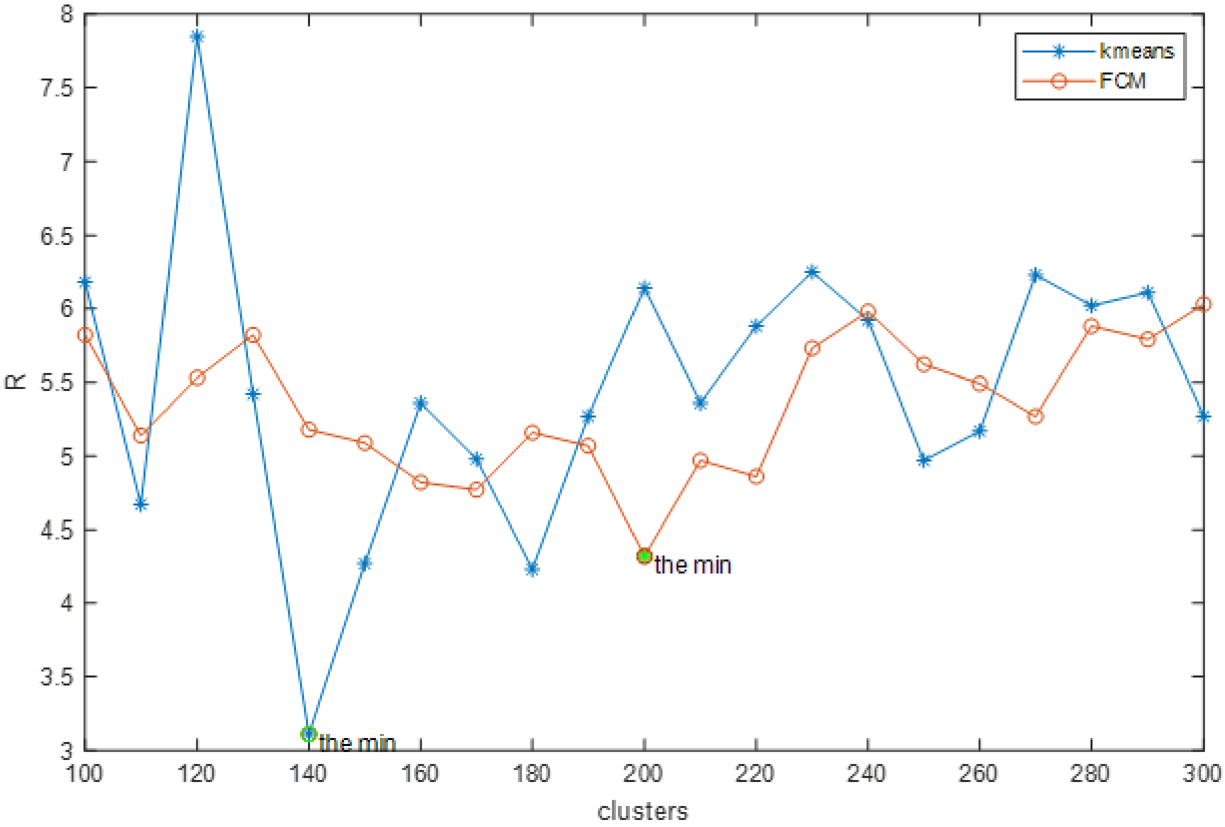


Figure 9

The results of Kmeans and FCM

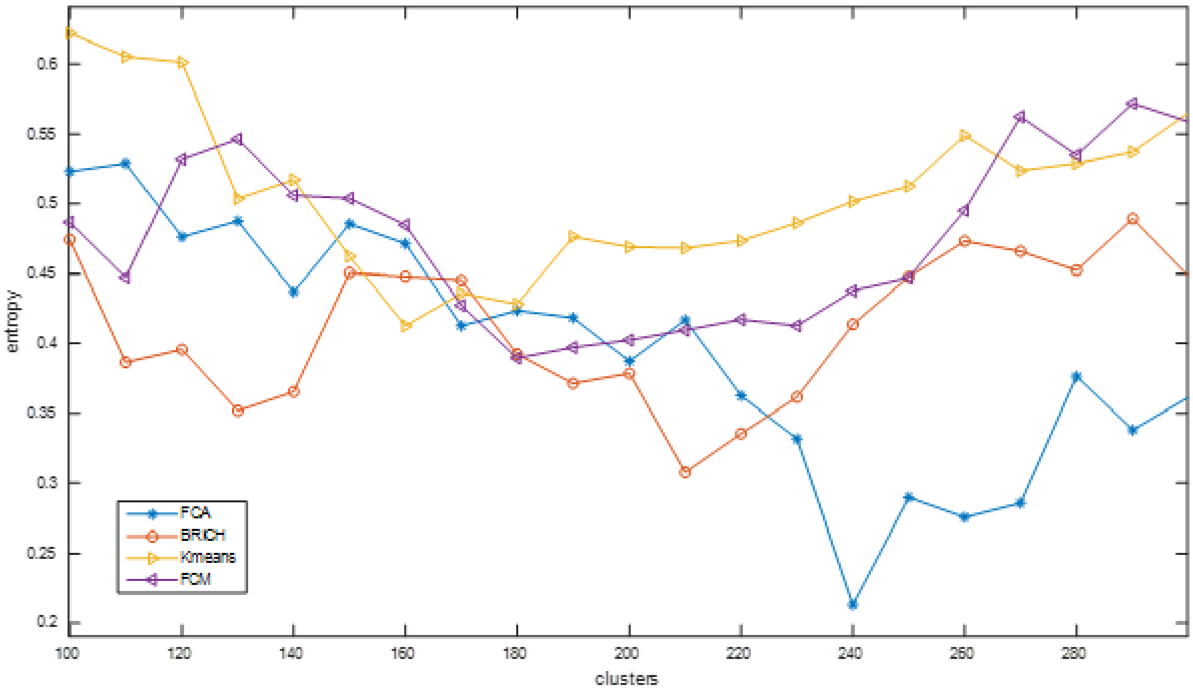


Figure 10

Entropy of four clustering methods under different clusters