Acta Technica *62* (2017), No. 4A, 405–412

Entropy-based improved fuzzy clustering algorithm for credit rating of banks

Xiaohui Wang[2]

Abstract. When handling clustering problems, fuzzy clustering algorithms divide data in a "soft" manner. Owing to this feature, fuzzy clustering has been widely applied to many business scenarios of data mining. However, the initialization of parameters in fuzzy clustering needs to be optimized. In existing studies, the number of clusters is generally determined empirically. The advantage of fuzzy clustering algorithms is not fully exploited this way, and their computational stability is thus questioned. In this study, information entropy is introduced to improve the parameter initialization of fuzzy clustering algorithms. By applying the improved algorithm to the credit rating process of banks, its feasibility and accuracy were validated. The customer rating result of the improved fuzzy clustering algorithm is in good accordance with the rating given by the bank based on its business practice.

Key words. Fuzzy clustering, credit rating, Information entropy.

1. Introduction

Data mining means to explore and analyze large quantities of data in order to discover meaningful patterns and rules (Berry and Linoff, 1997). With the higher and higher demands for data processing in recent years, clustering remains as the key and fundamental concept and algorithm in data mining and has always been a hot topic in the area [1], despite the development of various new algorithms. Clustering algorithms are widely used in a variety of fields, including business intelligence (BI), marketing, text analysis, image processing, pattern recognition, and so on. However, direct application of general clustering algorithms rarely produces satisfying computation results in many business scenarios. This is also an important reason for the attention received by the studies of clustering algorithms [2].

General clustering methods are based on classic set theory. The studied samples

[1]Acknowledgement - The author acknowledges the Scientific Research Projects of Colleges and Universities in Shandong Province (Grant: J15WB18).

[2]Workshop 1 - School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan, China; e-mail: gkwxh123@163.com

are clearly divided by their properties. Therefore, there are clear boundaries among the resultant clusters. A common example can be used to elaborate the drawback of this feature. Assuming that a person is classified as poor, ordinary, or rich based on his/her income, corresponding to the range of [0, 30k], [30k, 70k], or [70k, 120k]. Although a person making 70k and a person making 120k are both classified as rich, the big difference between them cannot be reflected [3]. This problem is solved by the application of fuzzy clustering methods.

Fuzzy clustering has significant advantage in solving "grey" problems with preferences, such as credit evaluation and risk control. Although a large number of data mining algorithms are available, it is not easy to merge different algorithms. On the one hand, it is limited to the specific business scenario; on the other hand, each algorithm has special requirements on data type. Based on comparative experiments, we decided to apply information entropy to improve fuzzy clustering. As placecountry-regionChina's financial market continues to grow and improve and the policies for interest rates continue to loosen, the competition for customer resources has been growing more intense among banks, small loan companies, investment management firms, and other financial institutions. The pressure of the bank in the screening of customer credit rating and risk assessment of loan business gradually increase. The current economic situation influences the credit business; most middle and small enterprises in placecountry-regionChina urgently need assistance through financing and loans. In the meantime, the increased availability of additional financing leads to greater uncertainty in the credit business. Apparently, the development of outstanding and stable data mining algorithms is of practical importance.

2. Basic Principles

2.1. Fuzzy clustering

Clustering is the process of distinguishing and classifying items according to certain requirements and rules in order to re-divide a dataset into several categories, or clusters. In this process, the data points in a particular cluster should be as similar as possible, whereas the data points in different clusters should be as different as possible. Conventional clustering analysis methods often rigorously divide data samples, and each object to be identified is categorized strictly. Such categorization produces obvious boundaries between the clusters. From the perspective of fuzzy clustering, the degree of membership of an object is either 1 or chmetcnvTCSC0NumberType1NegativeFalseHasSpaceTrueSourceValue0UnitNamein0 in such conventional clustering. However, in reality, this is often not the case: for example, in analyzing the bank credit business. If customer credit is categorized and classified by using common clustering algorithms, then a portion of potential customers would be missed. In addition, the risk of a certain business cannot be correctly evaluated. Clearly, this is counterintuitive to maximizing the profits of financial institutions.

2.2. FCM algorithm

The Fuzzy C-Means (*FCM*) clustering algorithm was proposed by Bezdek in 1973. It is a classification-based clustering algorithm, and the primary premise is to maximize the similarity among objects in the same cluster while minimizing the similarity among objects in different clusters[7−9].

Consider a dataset $X = \{x_1, x_2, \ldots, , x_n\}$, where each sample is composed of s attributes. Fuzzy clustering means dividing samples into c classes. $V = \{v_1, v_2, \cdots v_n\}$ is c cluster centers. In fuzzy clustering, a sample does not strictly belong to a certain class, rather a degree of membership is assigned to each sample[7].

Let u_{ik} indicate the degree to which the k^{-th} sample belongs to the i^{-th} class: $0 \leq u_{ik} \leq 1$ and $\sum_{i=1}^{c} u_{ik} = 1$. The target function is defined as $J(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m (d_{ik})^2$, where $d_{ik} = \|x_k - v_i\|$.

Clearly, $J(U,V)$ indicates the quadratic sum of the weighted distance between the samples and the cluster center, and the weight is the m^{-th} power of the degree of membership of sample x_k to the $^{i-th}$ class. By computing the minimum of the target function, the algorithm can be further detailed as follows:

(1) Predefine c, m, and the degree of membership matrix U^0, and the number of iterations $\mathrm{l} = 0$.

(2) Calculate the cluster center V as:

$$v_i^{(l)} = \sum_{k=1}^{N} (u_{ik}^{(l)})^m x_k \Big/ \sum_{k=1}^{N} (u_{ik}^{(l)})^m (i = 1, 2, \cdots, c), (1 < m).$$

(3) Update the degree of membership matrix U as

$$u_{ik}^{(l+1)} = 1 \Big/ \sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}} \forall i, \forall k,$$

where $d_{ik} = \|x_k - v_i\|$ is the Euclidean distance from the k^{-th} sequence to the i^{-th} center.

For given $\varepsilon > 0$, an iterative calculation should be performed to the given initial value until $\max\{\left|u_{ik}^{l} - u_{ik}^{l-1}\right|\} < \varepsilon$. If the condition is not met, let $l = l + 1$, and go to step (2).

3. Algorithm improvement

3.1. Information entropy-based fuzzy clustering

In 1948, C.E. Shannon introduced the concept of entropy into information theory for the first time. The following equation shows the measurement of the quantity of information using information entropy[10]:

$$H(x) = -\sum_{i=1}^{r} P(a_i) \log P(a_i)$$

where a_i indicates the various symbols at the information source end, and $H(x)$ is the statistical mean of the overall quantity of information at the source.

Generally, the number of clusters c is determined empirically before the computation; however, the value of c directly influences the convergence. The application of information entropy greatly improves this limitation of fuzzy clustering. A value range for c is first determined$[c_{\min}, c_{\max}]$. During the iterative computation of the FCM algorithm, a degree of membership matrix $U^{(c)}$ is produced for each value of c. Then, the entropy $H_c(u)$ of each matrix is calculated using the information entropy formula, and the c value that generates the smallest entropy value is the final number of clusters.

Furthermore, $H_c(u) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij} \log u_{ij}$ is the total entropy value of a degree of membership matrix. During the iteration, $c_{\max} - c_{\min}$ entropy values can be obtained, and the number of clusters corresponding to the smallest entropy value is the final number of clusters.

3.2. Algorithm

The information entropy-based improved fuzzy clustering algorithm is detailed as follows:

(1)Set a maximum $c_{\max}$ and a minimum $c_{\min}$ number of clusters with a threshold of ε. The initial c value is set as $c_{\min} - 1$.

(2)Initialize the parameters of the *FCM* algorithm: the degree of membership matrix $U^{(0)}$, $l = 0$, and $c = c + 1$.

(3)Calculate the cluster center V, and update the degree of membership matrix U, $l = l + 1$.

(4)If $\max\{|u_{ik}^{l} - u_{ik}^{l-1}|\} < \varepsilon$, terminate the algorithm and obtain a cluster number c. Otherwise, repeat step (3).

(5)Calculate $H_c(u)$, and compare the resulting $H_c(u)$ values from different values of c, and identify the smallest c value.

(6)If $c > c_{\max}$, c is the final number of clusters. Otherwise, go to step (2).

(7)Using the c value determined by the previous steps, compute the final clustering result using the FCM algorithm.

4. Case study

When rating the credit levels of corporations, banks and other financial institutions often investigate and collect data concerning the state of business, profitability, assets, and other economic indicators of the corporations. An institution would then select customers that are of high quality and focus on those customers. A great deal of attention would be paid to potential customers, whereas the low-quality customers with higher risks would be sifted out.

In this study, the data of multiple large corporations' business which was collected by a commercial bank in their daily business interactions was used to test the proposed algorithm. The data consists of 124 records: each is composed of six attributes, including total asset, sales revenue, total profit, net profit, asset-liability

ratio, and credit rating. Each corporation's credit rating was assigned by the bank based on the customer's state of business in recent years. The following table details the data

Table 1. Customer data (unit: 10K)

Client ID	Total assets	Sales revenue	Total profit	Net profit	Asset/liability rate (%)	Credit rate
001	55631.00	78732.00	11713.00	8784.00	27.16	5
002	41237.00	194751.00	1689.00	1267.00	53.03	5
.....	.....	.....	.....	.....	.....	.....
123	307086.00	239908.00	27639.00	20729.00	40.62	7+
124	111083.00	104032.00	11489.00	8617.00	39.86	9

Based on existing credit ratings of customers, the accuracy of the fuzzy clustering algorithm can be evaluated. The F-measure was adopted, as shown by the following equation[11]:

$$u_{ik}^{(l+1)} = 1 \Big/ \sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}} \forall i, \forall k,$$

where P is the precision, R is the recall rate, and β is the weight to adjust between precision and recall rate. *F-measure* is a common evaluation standard used in the information retrieval field to determine retrieval precision. It considers both positive and negative effects during evaluation, thus is gradually introduced into evaluation of classification methods. By making the parameter $\beta = 1$, the most commonly used *F1-measure* is applied.

The fuzzy clustering algorithm that has been improved by the introduction of information entropy, first selects a range for the number of clusters: the maximum number is $c_{\max} = 9$, and the minimum is$c_{\min} = 2$. Pre-computation revealed that when $c = 3$, the information entropy value is 14.32, which is the smallest. The following table shows the entropy values of the degree of membership matrices that result from different c values.

Table 2. Entropy values corresponding to different numbers of clusters

Number of clusters	2	3	4	5	6	7	8
Information entropy	16.03	14.32	18.96	25.70	30.91	42.41	50.76

After determining the number of clusters $c = 3$, the clustering result of the data sample set can be computed using the FCM algorithm. The following figure is the scatter diagram of the data set and the three cluster centers, and it is plotted based on the first two data items: the total assets and sales profit.

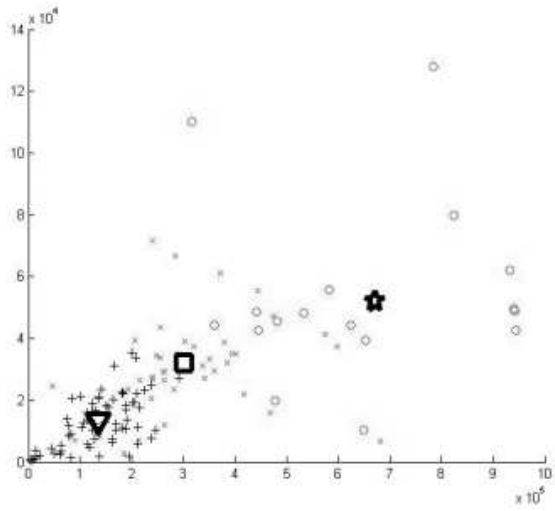

Fig. 1. Scatter diagram and cluster centers of the data set (cluster 1: '+', Cluster 2: 'o', Cluster 3: '*')

In the data set, the credit rating for each customer ranges from 1 to 9. To compare the clustering accuracy, the credit ratings were divided into three classes. Customers with credit ratings of 1 to 4 are classified as low-credit level, those with credit ratings of 5 to 7 are classified as medium-credit level, and finally, those with credit ratings of 8 to 9 are classified as high-credit level. In addition, three values (3, 6, and 9) were chosen to label the three clustered data sets. This way, after calculating the expected credit level of each customer and with the aid of the degree of membership matrix, the clustering result can be evaluated by referencing those three credit levels.

The following table presents a comparison between the clustering result of a conventional clustering algorithm and that of the improved clustering algorithm by using the Purity method.

Table 3. Clustering results of two clustering algorithms

	K-means	FCM	Bank
1~4	15	20	24
5~7	38	44	45
8,9	32	40	55
$F1-measure$	81.33%	90.26%	

As the clustering results demonstrate, the combination of the fuzzy clustering algorithm and the degree of membership matrix resulted in a credit rating accuracy of 0.9026, which is much higher than the 0.8133 achieved by the conventional clustering algorithm. The FCM algorithm tends to produce more credit levels of medium during

classification. For example, customer 018 is classified to the class 1˜4 by K-means algorithm. Using the FCM algorithm, the customer'ss credit rating has degrees of membership to the three levels of 0.5217, 0.3514, and 0.1269, which leads to a credit rating of 4.8158. Thus, this customer is classified to medium-credit level, and will be further developed as an important customer. Apparently, FCM algorithm processes ratings of such customers from a more realistic perspective than K-means algorithm.

Therefore, a greater number of potential valuable customers (or customers with risk) can be discovered by referencing the customer credit levels obtained via the improved fuzzy clustering algorithm. Compared to common clustering algorithms, the proposed algorithm produces more realistic credit ratings as well as greatly contributes to the credit business of banks.

5. Conclusions

This paper discusses the possibility of applying information entropy-based improved fuzzy clustering algorithms to the corporate customer credit rating process of banks. The advantages of using fuzzy clustering to determine credit customer selection risks were analyzed. Using actual business data from a commercial bank, the algorithm was validated. Furthermore, analyses demonstrated that the "endogeneity" of the parameter initialization process of the proposed entropy-based improved fuzzy clustering algorithm makes the algorithm more reasonable and practical. Meanwhile, the application of fuzzy clustering adds a competitive edge to banks in the customer selection process. The clustering result is more objective as it shows the trends of the risk of different businesses.

Admittedly, however, despite that the fuzzy clustering algorithm classifies customers' credit levels more realistically, it tends to produce classifications that are concentrated in the middle, which means the differences between credit levels are reduced. Therefore, some information of important customers may be omitted when applying this algorithm to customer evaluation process of banks. This disadvantage of the algorithm shall be optimized and perfected in future studies.

References

[1] V. E. Castro: *Why so many clustering algorithms:a position paper.* SIGKDD Exploration News letter *4* (2002),65–75.

[2] J. Vesanto, E. Alhoniemi: *Clustering of the self-organizing map.* IEEE Trans *11* (2000), 586–600.

[3] M. Kaya, R. Alhajj: *Genetic algorithm based framework for mining fuzzy association rules.* Fuzzy Sets and Systems *152* (2005), 587–601.

[4] L. A. Zadeh: *Fuzzy Sets.* Information and Control (1965), No. 8, 338–353.

[5] P. Alam, D. Booth: *The use of fuzzy clustering algorithm and self-organizing neural networks for identifying potentially failing banks: an experimental study.* Expert Systems with Applications *18* (2000), 185–199.

[6] N. Grira, & M. E. Houle: *Best of Both: A Hybridized Centroid-Medoid Clustering Heuristic.* Proc. of the 24th International Conference on Machine Learning (2007).

[7] X. WANG, Y. WANG, L. WANG: *Improving fuzzy c-means clustering based on feature-weight learning.* Pattern Recognition Letters *25* (2004), 1123–1132.

[8] R. O. DUDA, P. E. HART, D. G. STORK: *Pattern Classification (2nd Edition).* Wiley-Interscience (2004).

[9] J. C. BEZDEK: *Pattern Recognition with Fuzzy objective Function Algorithm.* New York PIenum Press (1981).

[10] U. M. FAYYAD: *Data mining and knowledge discovery.* lEEE Expert (1996), 20–25.

[11] A. ROMBEL: *CRM Shifts to Data Mining to Keep Customers.* Global Finance *15* (2001), 97–98.

Received November 16, 2016