Rough K-modes Clustering Algorithm Based on Entropy

Qi Duan, You Long Yang, and Yang Li

Abstract—**Cluster analysis is an important technique used in data mining. Categorical data clustering has received a great deal of attention in recent years. Some existing algorithms for clustering categorical data do not consider the importance of attributes for clustering, thereby reducing the efficiency of clustering analysis and limiting its application. In this paper, we propose a novel rough k-modes clustering algorithm based on entropy. First, we integrated the knowledge of information entropy to define a new dissimilarity measure that takes into account the importance of attributes for clustering and improves the quality of clustering. Then, applying the theory of rough set analysis, we used upper and lower approximation to deal with uncertain clusters, which allowed us to offer an improved solution for uncertainty analysis. Finally, our experimental results demonstrated that our proposed algorithm performed better than other conventional clustering algorithms in terms of clustering accuracy, purity, and F1-measure.**

Index Terms—**categorical data, clustering algorithm, dissimilarity measure information entropy, rough set, uncertainty analysis, upper and lower approximation.**

I. Introduction

CLUSTERING analysis [1] is one of the most popular data analysis tools in the field of data mining. A cluster is a set of data objects that are similar to other objects in the same cluster, but dissimilar to objects in other clusters [2]. The clustering method is widely used in different fields, such as pattern recognition [3], trend analysis [4], social media [5], medical systems [6], customer segmentation [7], and cloud computing [8].

In recent years, categorical data clustering has drawn a great deal of attention. Some research papers have discussed the problems related to categorical data [9], [10], and various categorical data clustering methods have been proposed, as in [11], [12], [13], [14], [15], [16], [17], [18], [19], [20].

Huang proposed the k-modes [12] algorithm and fuzzy k-modes [13] by extending the standard k-means [11] algorithm and fuzzy k-means with a simple matching method for categorical data. The k-modes algorithm substituted modes for means, using a based-frequency method to update modes in order to get the minimum cost function. However, this approach did not consider fully the dissimilarity between two values of the same attributes.

Prasad et al. [14] proposed a new modeling strategy called collaborative fuzzy rules generation. This method obtained a better effect and reduced the root mean square error value. Ng et al. [15] used a dissimilarity measure based on relative frequency that improved the precision of clustering effectively. However, it assumed that each object makes the same contribution to the cluster mode value of the objects in the cluster. Guha et al. [16] presented the ROCK (Robust Clustering using links) algorithm based on a hierarchical approach that defined a similarity threshold and used the concept of shared (common) neighbors to improve clustering. This algorithm had the advantage of breaking the traditional method based on a phase metric between two points. Instead, ROCK obtained the distance between two points by using the concept of shared neighbors, making it suitable for large data sets.

Ganti et al. [17] proposed a new hierarchy-based clustering algorithm named CACTUS. This summarization-based algorithm defined the concept of using three evaluation attribute values. CACTUS had the advantage of speed, scanning just two times for the number of data sets, which reduces scanning times and makes the algorithm suitable for large data sets. The COOLCAT algorithm was proposed by Barbar et al. [18]. This algorithm used information entropy to choose the initial cluster mode, making it suitable for large-scale data. However, the clustering criterion function was unable to reflect internal similarities between attribute values of different objects. Brendall et al. [19] presented a neighbor propagation algorithm that combined messages with similar matrices, and defined two types of information exchange between the underlying cluster modes. This algorithm had significant results for use with face images and text clustering. Cao et al. [20] presented a new dissimilarity measure for the k-modes clustering algorithm that demonstrated the importance of attributes for clustering. This algorithm could be used effectively with large data sets.

In our research, we proposed a novel entropy-based rough k-modes (ER-k-modes) clustering algorithm. We developed a new dissimilarity measure that took into account the significance of each attribute and the distance between categorical values in order to evaluate the dissimilarity between data object and mode. In this way, the algorithm was able to provide a better analysis by using the upper and lower approximation of the rough set clusters when the data are noisy, inaccurate, or incomplete. The experimental results on UCI datasets showed that our approach was both reasonable and effective.

The remainder of this paper is organized as follows. In section 2, we introduce rough set theory and the k-modes clustering algorithm. We also introduce information entropy. We present our method in section 3. In section 4, the experimental results demonstrate the advantage of algorithm. In Section 5, we present our conclusions and recommendations for future research.

II. Related works

In this section, we review the techniques central to our work. In Section 2.1, we provide the basic concepts of

Manuscript received May 23, 2016; revised October 10, 2016.

Q. Duan, is with the Department of Mathematics and statistics, Xidian University , Xi'an, Shaanxi 710071, China e-mail:duanqixd@163.com.

Y. Yang and Y. Li are with the Department of Mathematics and statistics, Xidian University, Xi'an, Shaanxi.

rough set theory such as categorical information systems, upper approximation, and lower approximation. In Section 2.2, we review the basic methodology for using the k-modes clustering algorithm.

A. Rough set

Rough set theory [21] was introduced by Pawlak in 1982 as a new mathematical technique that could deal with vague, incomplete, or uncertain knowledge. Its main idea is exported decision-making of problem and classification rules by knowledge reduction under the premise of the same ability of classification. At present, rough set theory has been applied successfully to machine learning, decision analysis, procedure control, pattern recognition, data mining, and other fields. Rough set theory has been studied widely in research concerning categorical data, including [22], [23], [24], [25], [26], [27], [28].

Let $S = (U, A, V, f)$ be a quaternary information system. U is a nonempty finite set of objects, called the universe. A is a nonempty finite set of attributes.

Let $V = \bigcup\{V_a | a \in A\}$, where V_a is the domain of attribute a. Define $f : U \times A \longrightarrow V$, called an information function, for any $a \in A$, and $x \in U$, $f(x, a) \in V_a$.

$S = (U, A, V, f)$ also called $S = (U, A)$.

For any subset $B \subseteq A$, $x, y \in U$, the indiscernibility relation to B is defined as:

$IND(B) = \{(x, y) \in U \times U | a \in B, f(x, a) = f(y, a)\}$.

The equivalence class is:

$[x]_B = \{y \mid \forall y \in U, (x, y) \in IND(B)\}$.

Then, we describe the upper and lower approximation of rough set theory:

Set $X \subseteq U$, $B \subseteq A$, the upper approximation and lower approximation of X with respect to B can be defined as follow:

$\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\}$,

$\overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \emptyset\}$.

$BN_B(X) = \overline{B}(X) - \underline{B}(X)$ is called the boundary region of X .

B. k-modes algorithm

The k-modes clustering algorithm is the conventional algorithm for analyzing categorical data. It uses a simple matching dissimilarity measure to calculate the cluster mode and allocate each object to the nearest cluster, then recalculate the mode of each cluster with the based-frequency method.

Let $U = \{x_1, x_2, \dots, x_n\}$ be a set of n categorical objects, where each object $x_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}(1 \leq i \leq n)$ is described by m categorical attributes $A_1, A_2, \dots, A_m$.

Set c_i is the mode of cluster C_i, and each C_i is composed of n_i objects, where $C_i = \{v_1, v_2, \dots, v_{n_i}\}$, and then to the cluster mode c_i takes the value of the highest frequency in each category attribute.

Given any two objects x_i and x_j, the categorical attribute distance is defined as:

$$d(x_i, x_j) = \sum_{i=1}^{m} \delta(x_{il}, x_{jl}) \tag{1}$$

$$\delta(x_{il}, x_{jl}) = \begin{cases} 1, & x_{il} \neq x_{jl}; \\ 0, & x_{il} = x_{jl}. \end{cases} \tag{2}$$

Huang presented the objective function for the k-modes clustering algorithm, defined as:

$$F(W, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} \omega_{il} d(x_i, x_j) \tag{3}$$

$\omega_{il} \in \{0, 1\}$,$1 \leq l \leq k$, $1 \leq i \leq n$, $\sum_{i=1}^{m} \omega_{il}(x_{il}, x_{jl}) = 1$, $1 \leq i \leq n$,$0 < \sum_{i=1}^{m} \omega_{il}(x_{il}, x_{jl}) < n$,$1 \leq i \leq n$.

Where W is a $n \times k \in \{0, 1\}$ matrix, and Z is a $k \times m$ matrix containing k cluster centers. $\omega_{il} = 1$ denotes the ith categorical object to the lth cluster.

III. Proposed rough k-modes clustering algorithm based on entropy

In this section, we present our new entropy-based rough k-modes (ER-k-modes) clustering algorithm. First, we defined a new dissimilarity measure based on entropy, which we explain in Section 3.1 and illustrate using an example. In Section 3.2, we introduce our rough k-modes clustering algorithm based on entropy and provide details of its development.

A. Dissimilarity measure based on entropy

The k-modes algorithm considers that each object makes the same contribution to the cluster mode. However, in real life, different categorical attributes have diverse effects on the clustering result. For example, when providing an analysis of customer information, the customer's phone number, name, and other data of this type are categorical attributes that are useless for clustering computing. A traditional decision tree algorithm introduces the concept of entropy to determine the importance of each classification attribute, then redefines the dissimilarity between objects according to their importance. The dissimilarity of categorical attributes based on entropy is defined as follows:

Set U is a set of n categorical objects, which are divided into k-different clusters, say $C_i(i = 1, 2, \dots, k)$, each C_i contains n_i objects. The entropy of U is defined as,

$$E(U) = -\sum_{i=1}^{k} p_i \log_2(p_i) \tag{4}$$

p_i is the probability of ith cluster C_i of U, where $p_i = \frac{n_i}{n}$. V_a is a set of different attribute values, U_v is the subset of the value v of attribute A, namely, $U_v = \{u \in U \mid A(u) = v\}$, the entropy of U_v is $E(U_v)$, and the expect entropy of attribute A is defined as,

$$E(U, A) = \sum_{v \in V_a} \frac{|U_v|}{|U|} E(U_v) \tag{5}$$

The information gain of U is defined as,

$$Gain(U, A) = E(U) - E(U, A) \tag{6}$$

Now, we introduced a new dissimilarity measure by using $d^{Entropy}(x_i, x_j)$ in Definition 1.

Definition 1. Given any two objects x_i and x_j, the distance measure can be defined as follows:

TABLE I
ARTIFICIAL DATA USED AS AN EXAMPLE.

Objects	a_1	a_2	a_3	a_4
x_1	A	A	B	B
x_2	A	B	A	B
x_3	A	B	B	B
c_1	A	B	B	B
x_4	B	A	B	B
x_5	B	B	A	B
x_6	A	B	A	C
c_2	B	B	A	B
x_7	B	D	A	B
x_8	A	D	B	B
x_9	B	B	A	B
c_3	B	D	A	B

$$d^{Entropy}(x_i, x_j) = \sum_{l=1}^{m} \omega_l \phi(x_{il}, x_{jl}) \tag{7}$$

$$\phi(x_{il}, x_{jl}) = \begin{cases} 1, & x_{il} \neq x_{jl}; \\ 0, & x_{il} = x_{jl}. \end{cases} \tag{8}$$

$$\omega_l = \frac{Gain(U, A)}{E(U)_A} \tag{9}$$

where

$$E(U)_A = \sum_{v \in V_a} \frac{|U_v|}{|U|} \log_2 \frac{|U_v|}{|U|} \tag{10}$$

Let us consider the following example that illustrates the limitations of the simple matching dissimilarity measure.

Example. Table 1 shows the artificial data set of nine objects: $U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\}$, where c_l, c_2, c_3 denote the cluster modes. Each object has four attributes values $A = \{a_1, a_2, a_3, a_4\}$. According to Eq.(1): $d(c_1, x_2) = 0+0+1+0 = 1$; $d(c_2, x_2) = 1+0+0+0 = 1$; $d(c_3, x_2) = 1+1+0+0 = 2$.

This means that we cannot determine the cluster to which x_2 should be assigned c_1 or c_2. We solved this problem by using the Definition 1 distance measure: $d^{Entropy}(c_1, x_2) = 0.0734$, $d^{Entropy}(c_2, x_2) = 0.3823$, $d^{Entropy}(c_3, x_2) = 0.8604$,

Then the object x_2 can be determinately assigned into the c_1.

B. Rough k-modes clustering algorithm based on entropy

In the next step of our work, we applied rough set theory to deal with uncertainty problems and propose a suitable objective function for determining the accuracy of clustering results. The goal of rough set theory is to provide a method for classifying uncertain or incomplete knowledge using the concepts of upper and lower approximate regions of a set. Based on this idea, we differentiated between the boundary point and lower approximation point in the cluster and endowed them with different weights. The cluster mode was able to show fully the distribution of data points in the cluster, and its accuracy could be improved. Our clustering algorithm, based on rough set theory, followed these three principles:

1. An object can belong only to a cluster of the lower approximation.

2. If the object belongs to the lower approximation of a cluster, it also belongs to the upper approximation of the cluster.

3. If the object does not belong to the lower approximation of any clusters, it belongs to the upper approximation of two, or more than two, clusters.

According to the above features, we divided the points that might belong to one cluster or another cluster into the clusters upper approximation. The determining points belong to a clusters lower approximation. We divided the data points into the clusters upper approximation when the dissimilarity between a data point and one cluster mode was close to the dissimilarity between the point and another cluster mode. If such points did not exist, then the data points were divided into the lower approximation of the cluster. We could find the boundary points of the cluster by upper and lower approximation. Obviously, if the weight of the boundary point was smaller, the the distribution of the clusters mode was more reasonable.

Therefore, we introduced w_{lw}(the weight of the lower approximation) and w_{bn}(the weight of the boundary point), where $w_{lw} + w_{bn} = 1$, stop threshold ε.

Then, we proposed the following function to calculate the cluster modes, where the dissimilarity measure adopted the measure given by Definition 1. Subsequently, we introduced the objective function of Definition 2 and process of algorithm.

Definition 2. The objective function was defined by Eq.(11).

where c_i is the mode of cluster C_i, $1 \leq i \leq n$; $\underline{B}(C_l)$ is the lower approximation of cluster C_l; $\overline{B}(C_l)$ is the upper approximation of cluster C_l; $BN(C_l) = [\overline{B}(C_l) - \underline{B}(C_l)]$ is the boundary region of cluster C_l; w_{lw} is the importance of the lower approximation; and ω_{bn} is the importance of the boundary region; $w_{lw} + w_{bn} = 1$.

Our rough k-modes clustering algorithm based on entropy is described as follows:

Input: Data set U; Cluster number k; Under the approximate weight w_{lw}, boundary weights ω_{bn}; Rough clustering threshold η; Stop threshold ε.

Output: Clustering result $\{C_1, C_2, \ldots, C_k\}$.

Step 1. Randomly select k objects as the initial cluster rough mode.

Step 2. According to Definition 1, compute the distance between arbitrary object x_j and the rough mode c_i.

Step 3. If $|d_{min}^{Entropy}(x_j, c_j) - d_{premin}^{Entropy}(x_j, c_k)| < \eta$, then x_j belongs to upper approximation of mode c_j,c_k, otherwise x_j belongs to lower approximation of mode c_j.

Step 4. Calculate the cluster rough mode based on the frequency method.

Step 5. Use Definition 2 to calculate ℓ_{ER},if $|\ell_{ER} - \ell_{ER}^{pre}| < \varepsilon$ stops. Otherwise, return to step 2.

IV. EXPERIENCE

We utilized MATLAB to perform an experimental program and analyze the effectiveness of the algorithm. Experimental data came from the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets.html). Data set descriptions are shown in Table 2.

$$\ell_{ER} = \begin{cases} \omega_{lw} \times \sum\limits_{l=1}^{k} \sum_{x_i \in \underline{B}(C_l)} d^{Entropy} + \omega_{bn} \times \sum\limits_{l=1}^{k} \sum_{x_i \in BN(C_l)} d^{Entropy}, if \underline{B}(C_l) \neq \emptyset, BN(C_l) \neq \emptyset; \\ \omega_{lw} \times \sum\limits_{l=1}^{k} \sum_{x_i \in \underline{B}(C_l)} d^{Entropy}, if \underline{B}(C_l) \neq \emptyset, BN(C_l) = \emptyset; \\ \omega_{bn} \times \sum\limits_{l=1}^{k} \sum_{x_i \in BN(C_l)} d^{Entropy}, if \underline{B}(C_l) = \emptyset, BN(C_l) \neq \emptyset; \end{cases} \tag{11}$$

TABLE II
SUMMARY OF THE REAL DATA SETS'CHARACTERISTICS.

Data set	Objects	Attributes	Classes
Soybean	47	35	4
Zoo	101	17	7
Breast-cancer	699	10	2
Mushroom	8124	23	2

TABLE III
THE AC FROM THE THREE DIFFERENCE ALGORITHM ON FOUR DATA SETS.

Data set	K-modes	Ng' k-modes	ER-k-modes
Soybean	0.6170	0.6809	0.7021
Zoo	0.6535	0.7327	0.7723
Breast-cancer	0.8927	0.9056	0.9113
Mushroom	0.7903	0.7976	0.8035

A. Evaluation index

This experiment with clustering accuracy, purity of clustering and F1-measure was able to reflect the effect of clustering:

$$AC = \sum_{i=1}^{k} \frac{a_i}{n} \tag{15}$$

$$PR = \frac{\sum\limits_{i=1}^{k} \frac{a_i}{a_i + b_i}}{k} \tag{16}$$

$$RE = \frac{\sum\limits_{i=1}^{k} \frac{a_i}{a_i + d_i}}{k} \tag{17}$$

where n denotes the number of objects in the data sets, k is the number of clusters, the clustering result is $C = \{C_1, \ldots, C_k\}$, a_i is the number of objects that that are correctly assigned to the ith cluster, b_i is the number of objects that that were erroneously assigned to the ith cluster, and d_i represents the number of objects that should have been assigned to ith cluster but were not assigned. The comprehensive evaluation index (F-Measure) is a weighted harmonic mean of accuracy and recall that integrates the two indicators of evaluation index to reflect the overall index. Its formula was,

$$F = \frac{(a^2 + 1)AC * RE}{a^2(AC + RE)} \tag{18}$$

The parameters of $a = 1$ are the most common form of $F1$:

$$F1 = 2\frac{AC * RE}{AC + RE} \tag{19}$$

TABLE IV
THE PR FROM THE THREE DIFFERENCE ALGORITHM ON FOUR DATA SETS.

Data set	K-modes	Ng' k-modes	ER-k-modes
Soybean	0.7872	0.7872	0.7872
Zoo	0.8317	0.8317	0.8812
Breast-cancer	0.8927	0.9056	0.9113
Mushroom	0.7903	0.7976	0.8035

TABLE V
THE F1-MEASURE FROM THE THREE DIFFERENCE ALGORITHM ON FOUR DATA SETS.

Data set	K-modes	Ng' k-modes	ER-k-modes
Soybean	0.6982	0.7161	0.7180
Zoo	0.7231	0.7745	0.7854
Breast-cancer	0.8361	0.8494	0.8559
Mushroom	0.6687	0.6775	0.6847

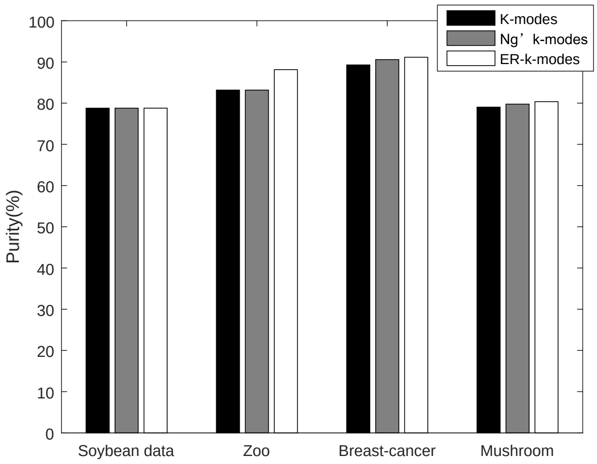

Fig. 1. Comparison of the AC of three algorithm on four data sets

Clustering accuracy (AC) shows the proportion of the objects correctly assigned out of all the objects in the cluster. Higher clustering accuracy means greater correctness. The calculation of purity (PR) is based on the number of objects correctly assigned to the cluster and the number of objects erroneously assigned to the cluster, which gives the average purity of clusters. Higher clustering purity means fewer errors of assigned to the objects.

Recall rate (RE) is to compute according to correctly assigned to the objects number of the cluster and should be given but not assigned to the objects number of the cluster, and get the averages of k right clusters that assigned to the cluster of objects. The higher that should be assigned to the recall rate is, the less that but not assigned to the number of

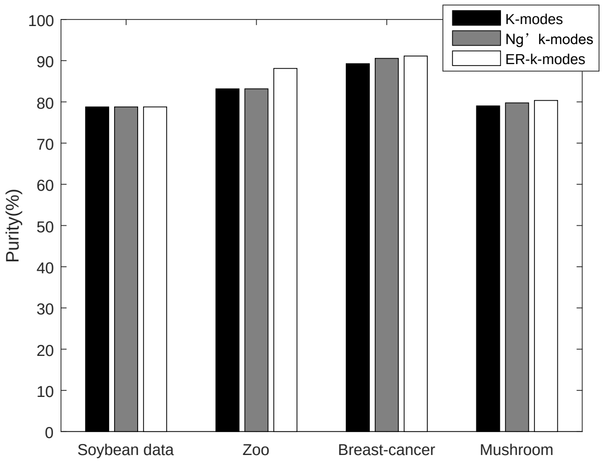

Fig. 2. Comparison of the PR of three algorithm on four data sets

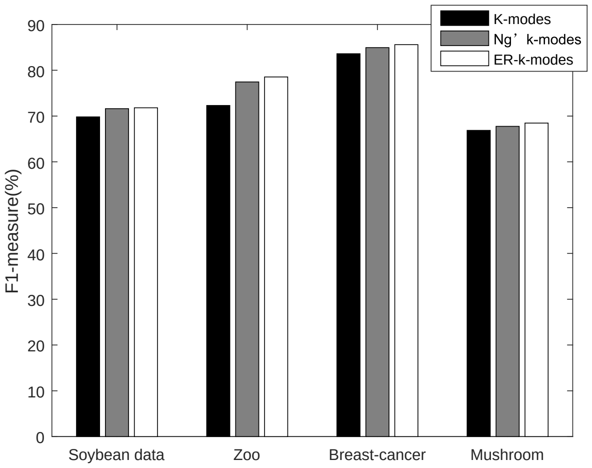

Fig. 3. Comparison of the F1-measure of three algorithm on four data sets

objects in the cluster is. The higher F1 is, the more effective the test method is.

B. Empirical results and analysis

Now we compared our method with k-modes and Ng'k-modes that handle categorical data algorithms. Some of the data sets had missing attribute values; therefore we replaced those values with special values in all of the data sets. In the Zoo data set, Mushroom data set, and Soybean data set have some missing attributes. In order to ensure the integrity of the data set and the accuracy of the results, their values are 0.

From Table 3, Table 4, and Table 5 it can be seen that our method proved superior to the original k-modes algorithm and Ng'k-modes algorithm in the four data sets. Table 3 shows the AC results of all four data sets. We list the clustering accuracy (AC) of k-modes, Ng'k-modes, and ER-k-modes on the Soybean data set as 0.6170, 0.6809, 0.7021, respectively. Table 4 displays the clustering purity (PR) results of all four data sets. The PR of k-modes, Ng'k-modes, and ER-k-modes on the Soybean data set have the same values. The PR of k-modes, Ng'k-modes, and ER-k-modes on the Zoo data set is shown as 0.8317,0.8317, 0.8812, respectively. Table 5 displays the four data sets with F1 -measure results. The F1-measure of k-modes, Ng'k-modes, and ER-k-modes on the Mushroom data set is shown as 0.6687, 0.6775, 0.6847, respectively.

As can be seen from Fig. 1, Fig. 2, and Fig. 3, accuracy, purity, and F1-measure of ER-k-modes were obviously better than the results for k-modes and Ng'k-modes. These results validated the algorithms effectiveness, where $w_{lw} = 0.95$, $\eta = 0.01$, $\varepsilon = 0.0001$.

In order to test the effects of the importance of lower approximation for the algorithm, we selected the Soybean data set to experiment and compare the F1-measure under the same condition by taking different weights. It can be seen from Fig.4 that the results region was between 0.80-1.00, which is a relatively optimal effect.

V. Conclusion and Future Work

In this paper, we proposed a new rough k-modes clustering algorithm based on entropy for categorical data. The algorithm combined the concept of information entropy with the power of rough set theory, which can process categorical data and improve the quality of the clustering while effectively solving the k-modes algorithm for uncertain objects. We introduced the idea of the upper and lower approximation of rough set and boundary sets to handle a border of uncertain data points in the process of clustering. The experimental results demonstrated that the proposed algorithm had higher accuracy and better convergence ability .

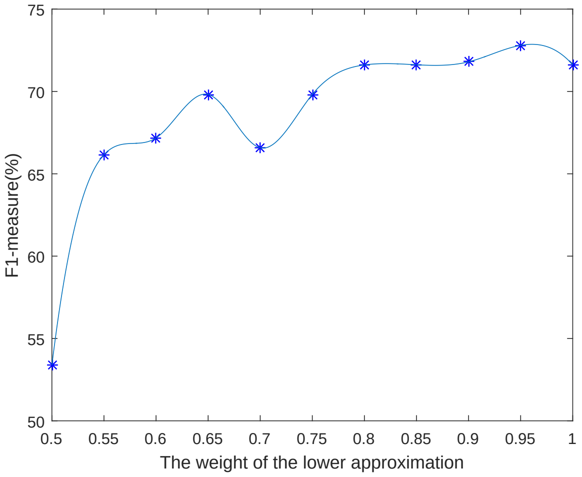

Fig. 4. The result from the variety of w_{lw} value

Clustering analysis is one of the important methods in data mining. We have demonstrated that our algorithm is worthy of further development for use with categorical data. While our current paper is limited to categorical data only, cluster analysis still has many content areas in need of further study.

(1) In the real world, its common to have mixed data. A future goal for research is to develop the best means for improving our proposed algorithm to make it suitable for mixed data processing.

(2) At present, many indirect methods of data collection lead to production of quantities of uncertain data. Research on uncertain data and clustering algorithms is a future challenge for clustering analysis.

(3) We are also challenged to study how to determine the

clustering number in the clustering process. Many clustering algorithms require the user to input the number of clusters into the clustering analysis, which not only increases the users burden, but also makes it difficult to control clustering quality.

Although there are still many problems that need to be solved for the advancement of clustering analysis, we predict that the continuing efforts of researchers and the demands for practical application will bring about significant development of clustering technology.

ACKNOWLEDGMENT

The authors would like to thank the editors and the anonymous reviewers for helpful comments and suggestions.

REFERENCES

[1] J. W. Han and M. Kamber, "Data mining concepts andconcepts and techniques, " San Francisco, USA Morgan Kaufmann, 2001.

[2] J. T. Tou and R. C. Gonzalez, "Pattern Recognition Principles," London, 1974.

[3] U.I. Saha, "Modified differential evolution based fuzzy clustering for pixel classification in remote sensing imagery," *Pattern Recognition*, vol. 42, no. 9, pp. 2135-2149, 2009.

[4] A. Popescul, G. W. Flake, S. Lawrence, L. H. Ungar and C. L. Giles, "Clustering and identifying temporal trends in document databases," in *Prococeedings IEEE Advances in Digital Libraries 2000 (ADL 2000), 2002*, pp. 173-182.

[5] T. Sakai, K. Tamura, and H. Kitakami, "Extracting Attractive Local-Area Topics in Georeferenced Documents using a New Density-based Spatial Clustering Algorithm," *IAENG International Journal of Computer Science*, vol. 41, no. 3, pp. 185-192, 2014.

[6] U. Maulik, "Medical image segmentation using genetic algorithms," *IEEE Transaction Information Technology BioMedicine*, vol. 13, no. 2, pp. 166-173, 2009.

[7] D. S. Boone and M. Roehm, "Retail segmentation using artificial neural networks," *International Journal of Research in Marketing*, vol. 19, no. 3, pp. 287-301, 2002.

[8] X. Zhong, G. Yang, L. Li and L. Zhong, "Clustering and correlation based collaborative filtering algorithm for cloud platform," *IAENG International Journal of Computer Science*, vol. 43, no. 1, pp. 108-114, 2016.

[9] E. Cesario, G. Manco and R. Ortale, "Top-down parameter-free clustering of highdimensional categorical data," *IEEE Transaction on Knowledge and Data Engineering*, vol. 19, no. 12, pp. 1607-1624, 2007 .

[10] H. L. Chen, K. T. Chuang and M. S. Chen, "On data labeling for clustering categorical data," *IEEE Transaction on Knowledge and Data Engineering*, vol. 20, no. 11, pp. 1458-1472, 2008.

[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1967*, pp. 281-297.

[12] Z. X. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283-304, 1998.

[13] Z. X. Huang and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data ," *IEEE Transaction on Fuzzy Systems*, vol. 7, no. 4, pp. 446-452, 1999.

[14] M. Prasad, D. L. Li, C. T. Lin, S. Prakash, J. Singh and S. Joshi, "Designing Mamdani-Type Fuzzy Reasoning for Visualizing Prediction Problems Based on Collaborative Fuzzy Clustering," *IAENG International Journal of Computer Science*, vol. 42, no. 4, pp. 404-411, 2015.

[15] M. K. Ng, M.J. Li, Z.X. Huang and Z.Y. He, "On the impact of dissimilarity measure in k-Modes clustering algorithm," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 503-507, 2007.

[16] S. Guha, R. Rastogi and K. Shim, "Rock:A robust clustering algorithm for categorical attributes," *Proceedings of the IEEE International Conference on Data Engineering 1999*, pp. 512-521.

[17] V. Ganti, J. E. Gekhre and R. Ramakrishnan, "CACTUS-Clustering categorical data using summaries," *Proceedings of 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1999*, pp. 73-83.

[18] D. Barbarg, J. Couto and Y. Li, "COOLCAT:An entropy-based algorithm for categorical clustering," *Proceedings of the l lth International Conference on Information and Knowledge Management 2002*, pp. 582-589.

[19] J. F. Brendan and D. Delbert, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972-976, 2007.

[20] F. Y. Cao, J. Y. Liang, D. Y. Li, L. Bai and C. Y. Dang, "A dissimilarity measure for the k-Modes clustering algorithm," *Knowledge-Based Systems*, vol. 26, pp. 120-127, 2012.

[21] Z. Pawlak, "Rough Sets: Theoretical Aspects of Resoning About Data," *Kluwer Academic*, MA, USA, 1992.

[22] H. M. Chen, T. R. Li, D. Ruan, J. H. Lin and C. X. Hu, "A rough-set-based incremental approach for updating approximations under dynamic maintenance environments," *IEEE Transaction on Knowledge and Data Engineering*, vol. 25, no. 2, pp. 274-284, 2013.

[23] Y. Y. Yao and S. K.M. Wong, "A decision theoretic framework for approximating concepts," *Internal Journal of Man-Machine Studies*, vol. 37, no. 6, pp. 793-809, 1992.

[24] Y. Y. Yao, "Probabilistic rough set approximations," *Internal Journal Approximation Reasoning*, vol. 49, no. 2, pp. 255-271, 2008.

[25] J. B. Zhang, T. R. Li and H. M. Chen, "Composite rough sets for dynamic data mining," *Information Sciences*, vol. 257, no. 5, pp. 81-100, 2014.

[26] H. M. Chen, T. R. Li, C. Luo, S. J. Horng and G. Y. Wang, "A rough set-based method for updating decision rules on attribute valuescoarsening and refining," *IEEE Transaction on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2886-2899, 2014.

[27] B. Li and J. Q. Shen, "The Initial Boundary Value Problem of the Generalized Wigner System," *Engineering Letters*, vol. 23, no. 2, pp. 110-114, 2015.

[28] J. Hu , T. R. Li , H. J. Wang, H. Fujita, "Hierarchical cluster ensemble model based on knowledge granulation," *Knowledge-Based Systems*, vol. 91, no. 5, pp. 179-188, 2016.

Qi Duan was born in Weinan, Shaan Province, China in 1990. She received her B.S. degree in the Department of Mathematics from Shangluo University in 2009, M.S. degree in the Department of Mathematics from in Xidian University in 2014. She is major in Probabilistic graphical, data analysis and its application.

Youlong Yang was born in Weinan, Shaanxi, China in 1967. He received his B.S. degree, M.S. degree in the Department of Mathematics from Shaanxi Normal University, in 1990 and 1993, respectively. And received the Ph.D. degree at Northwestern Polytechnical University in 2003. In 2006, he was out bounded post-doctoral mobile stations in Xidian University, and then was sent to the United States University of Rochester as a national public school student in 2007. His scientific interest is Probabilistic Graphical Models and Time Series.

Youlong is full professor, PhD supervisor in Xidian University and executive director of the Mathematical Association of Shaanxi Province. Prof. Yang has published more than 40 papers. His papers were published as the first-named author in journals such as Information Sciences, International Journal of Approximate Reasoning, Acta Application Mathematic, Control Theory and Applications. He has received one reward of Shaanxi Provincial Science and Technology and two Bureau departmental level research awards and so on.

Yang Li was born in Xi'an, Shaanxi, China in 1991. She received her B.S. degree from Baoji University of Arts and Sciences in 2010 and begin work for a M.S. degree in the Department of Mathematics at Xidian University in 2014. She is major in Statistical inference.