Fuzzy Clustering Ensemble with Selection of Number of Clusters

Taoying Li, Yan Chen
Transportation Management Collage, Dalian Maritime University, Dalian, 116026, China
Email: ytaoli@126.com

***Abstract*— Existing clustering ensemble algorithms for partitioning data need to know the generating process of clustering members clearly and most of them are not suitable to categorical data. In order to partition categorical data conveniently, at same time broaden the application of clustering ensemble, a fuzzy clustering ensemble algorithm was proposed in this paper, which not only can be used to classify categorical data, but also be used to combine results of multi clustering for numerical data or mixed categorical and numerical data. The proposed algorithm firstly made use of relationship degree between different attributes to prune part of attributes. Next, took the distribution of clustering members into account, Descartes subset and relationship degree between any two different objects were used for establishing the relationships between objects, which were under unsupervised circumstances and could get the minimum value of objective function of clustering and obtain corresponding optimal partitions. Then, choose the number of clusters satisfying the difference and differential rate of objective function local maximum as the optimal number of clusters and its corresponding partitions are optimal clustering. Finally, the proposed algorithm was applied in Synthesis dataset, Fellow-Small dataset, Zoo dataset, and results show the algorithm is effective and feasible.**

***Index Terms*— fuzzy clustering; clustering ensemble; relationship degree; *k*-means clustering**

I. INTRODUCTION

Clustering is to group data points into several clusters and makes the intra-cluster similarity maximized and the inter-cluster similarity minimized [1], [2], [3]. Various types of clustering methods have been proposed and developed in [4]. Most clustering algorithms are used to partition numerical data whose inherent geometric properties can be exploited naturally to define distance functions between data points. However, some practical data is categorical and their values cannot be naturally ordered as numerical values. An example of categorical attribute given in [3] is shape whose values include circle, rectangle, ellipse, etc. K-means algorithm is a classical clustering algorithm, and it is well known for its efficiency in clustering large data sets [4], [5], [6], [7], however, it is usually used for numerical data and becomes useless for partitioning categorical data or mixed data. Due to the special properties of categorical attributes, the clustering of categorical data seems more complicated than that of numerical data. Many algorithms have been proposed for clustering categorical data in recent years [8]. K-modes algorithm and k-prototypes algorithm advanced by Huang spread the k-means algorithm for partitioning categorical data and mixed data [9], [10]. Many existing algorithms for partitioning categorical data and mixed data tend to be instable, random and poor accuracy [11].

Recently, ensemble learning technology has been widely used in many areas, such as biology feature recognition, Computer-aided medical diagnosis, text recognition, web information filtering, analysis of seismic waves, etc. Early research on ensemble learning technology focus on supervised learning, and the research on unsupervised learning, equaling to clustering ensemble, is in the initial stage [12], [13]. Clustering ensemble is to integrate some results of some existing clustering partitions for higher quality and better robust, and Topchy proved that the ensemble method can be better than anyone clustering algorithm in [14], [15].

The FCESNC (Fuzzy clustering ensemble with selection of number of clusters) algorithm is proposed in this paper and it can be used to partition categorical data or mixed data. FCESNC makes use of the relationship degree of any two different attributes to prune some attributes that are comparatively less influencing the results of clustering, which reduces the calculation amount. Then, according to the distribution of cluster members, *m* Descartes subset and relationship degree between objects were used for establishing the relationships between objects under unsupervised circumstance and got the minimum value of objective function of clustering. Then select the number of clusters while the difference and differential rate of objective function are local maximum, which will obtain the optimal number of clusters and is crucial to get the results of clustering.

This paper is organized as follows. Section 2 presents review on related work. In section 3, the FCESNC algorithm was proposed, and gives the steps and analysis the complexity. Instances are given in Section 4 and Section 5 concludes the paper.

II. RELATED WORK

This work was supported by the Doctoral Fund of Ministry of Education of China (No. 200801510001) and the Key Project of Chinese Ministry of Education (No. 209030).

doi:10.4304/jcp.5.7.1112-1119

Clustering ensemble/clustering combination were proposed by A. Strehl and J. Ghost in [2]. The Recently, the study mainly focused on two aspects, one is how to produce efficient clustering memberships and another is how to design the objective function for merging clustering membership [13], and the different clustering ensemble algorithms are given in [16] as shown in Figure 1.

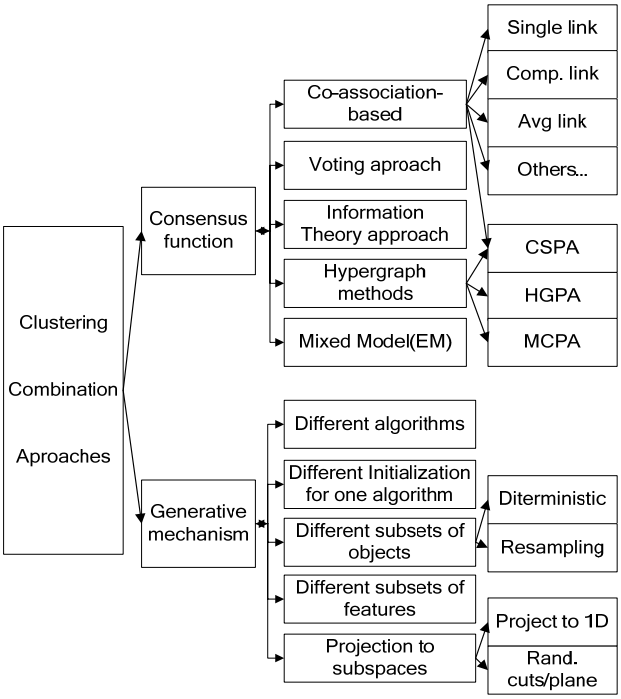

Figure1. Taxonomy of different approaches to clustering combination; top side: different approaches how to obtain the diversity in clustering; bottom side: different consensus function to find the clustering ensemble.

A. Producing Clustering Memberships

K-means algorithm is also a classical method for producing membership, which randomly selects the initial cores and number of clusters, then training particular times, and then obtains the memberships. The algorithm lows complexity and is suitable to small data. However, k-means algorithm won't be suitable to partition multi-dimension data or complex data.

Reference [17] uses random sampling to produce sub-dataset, which is thought to reflect the structure of dataset and real distribution. Reference [18] made use of Bootstrap to establish memberships. Reference [19] mainly researched on combining multi-weak clustering, which is similar to clustering ensemble. Reference [20] adopted random projection for high Dimensional Data Clustering to produce memberships. Reference [21] used adaptive clustering ensemble, the memberships of which were produced on a particular sequence and k-means algorithm was used to partition dataset every time, and authors used one parameter to verify the probability that one point was abstracted. Reference [22] focused on the difference degree of memberships and the bigger difference degree, the better the clustering.

B. Designing the Objective Function

Reference [2] proposed three different methods for ensemble memberships of clustering, which adopted co-association matrix, hyper graph. Reference [23], [24] used EA (Evidence Accumulation) method to get the final results using co-association matrix. The WSnnG (Weighed Shared nearest neighbors Graph) was advanced in [25], which made co-association matrix to sparse and then get the final results. Reference [26] proposed HBGF (Hybrid Bipartite Graph Formulation) method, which produced a bipartite graph, if a point belongs to one cluster, then used an edge between the point and the core of the cluster. Then the graph theory was used to get the results. Reference [27] proposed a algorithm called JSDCC (Jensen-Shannon Divergence based Clustering Combination), which abstracted a joint distribution model from co-association matrix, next computed the value of Jensen-Shannon between data point and model, and then decided which cluster the point should belong, another algorithm took the results of clustering to be the original features of data points, which mapped the memberships to data points, and the used Kerouac method to complete clustering.

In the process of clustering ensemble, above all, we should generate m clustering memberships of dataset X and then merge the results of these m clustering memberships according to the mutual function. The research on the design of mutual function is a hot topic in clustering ensemble. All of CEMC (cluster ensemble based mixed attribute cluster) in [20], and the mixed method of CDC (Categorical data clustering) algorithm and CE (cluster ensemble) in [3] partition categorical data by drawing on the idea of clustering ensemble algorithm. However, these algorithms need to know the generating methods of the clustering memberships. The calculation process with a larger amount of data is so complex that will not be able to run, and greatly affect the efficiency of the algorithm while only know the results and do not understand the generation of cluster memberships.

III. FCESNC ALGORITHM

Assume there are n data points $X=\{x_1,x_2,\ldots,x_n\}$ with m attributes, and the ith attribute has k_i different values and its weight is w_i. We can get a partition on the attribute as (1).

$$R_i = \{C_{i,1}, C_{i,2}, \cdots C_{i,k_m}\}, 1 \le i \le m. \quad (1)$$

Here, R_i is the partition on the ith attribute that denotes results of the ith clustering, $C_{i,j}$ is the jth cluster of the ith partition, and $\sum_{i=1}^{m} w_i = 1$.

By inspiring from [3], we know that partition on each attribute can be regarded as a clustering membership, and then obtain m clustering memberships $R=\{R_1,R_2,\ldots,R_m\}$, and there are k_ith clusters in R_i.

A. Definitions

Definition1: m subset of Descartes $C^m(j_1, j_2, \cdots, j_m)$ is an intersection among all m

clusters, which come from *m* clustering memberships and any two do not come from the same and the number of elements is more than one as (2).

$$C^m(j_1, j_2, \cdots, j_m) = \overbrace{C_{1,j_1} \cap C_{2,j_2} \cap \cdots \cap C_{m,j_m}}^{m} \quad (2)$$

Here, $1 \le j_i \le k_i, i = 1,2,\cdots,m$, $|C^m(j_1, j_2, \cdots, j_m)| > 1$, C_{i,j_i} is the j_ith cluster of ith clustering membership.

In order to present the research figuratively, the relationship degree between objects is introduced.

Definition2: relationship degree between objects x_l and x_h is denoted as (3).

$$R(x_l, x_h) = \begin{cases} \sum_{i=t_1}^{t_s} w_i & if, \alpha \\ \min(R(x_l, x_p), R(x_p, x_h)) \times \frac{|A \cap B|}{\min\{|A|, |B|\}} & else if, \beta \\ 0 & else \end{cases} \quad (3)$$

Here, w_i is the weight of the ith attribute, and $\sum_{i=1}^{m} w_i = 1$,

$\alpha = \exists A, x_l \in A, x_h \in A, and A = C^s(j_{t_1}, j_{t_2}, \cdots, j_{t_s})$

$\beta = \exists A, B, x_l \in A, x_h \in B, x_p \in A \cap B$.

In order to illustrate the definition mentioned above, an example can be showed as follows.

TABLE I EXAMPLE

object	Attributes	
	A_1	A_2
x_1	a	c
x_2	b	c
x_3	a	d
x_4	a	c

There are three objects with two attributes in Table I, given the weights of attributes are 0.4 and 0.6.

We can get the partition R_1 on A_1 and R_2 on A_2 are

R_1= {{x_1, x_3, x_4}, {x_2}};

R_2= {{x_1, x_2, x_4}, {x_3}}.

The m subset of Descartes $C^m(j_1, j_2, \cdots, j_m)$ is $C^2(j_1, j_2)$ ={x_1, x_4}.

The relationship degree matrix can be shown as Table II.

TABLE II RELATIONSHIP DEGREE MATRIX

	x_1	x_2	x_3	x_4
x_1	1	0.6	0.4	1
x_2	0.6	1	0.27	0.6
x_3	0.4	0.27	1	0.4
x_4	1	0.6	0.4	1

B. Algorithm for Pruning Redundant Attributes

In practice, there are many data with a lot of attributes, which are interrelated to each other. The greater the possibility of one attribute decided by attributes, the higher the opportunity that it a redundant attribute. Redundant attributes reduce the efficiency of the algorithm and possibly affect the accuracy. Thereby, we should prune redundant attributes before adopting learning algorithms.

The relationship degree between attributes is used for pruning redundant attributes in this paper.

Definition3: Relationship degree of attributes as and at is denoted as (4).

$$R(a_s, a_t) = \frac{\sum_{i_t}^{k_t} \sum_{i_s}^{k_s} |C_{s,i_s} \cap C_{t,i_t}|}{n} \quad (4)$$

Here, if $|C_{s,i_s} \cap C_{t,i_t}| = 1$, because of the definition 1 $|C^m(j_1, j_2, \cdots, j_m)| > 1$, $|C_{s,i_s} \cap C_{t,i_t}| = 0$.

First, we should give the minimum relationship degree between attributes $R_{\min}$, and then we can get a half-angle matrix according to Definition3. If the relationship degree between two attributes is large than $R_{\min}$, one can be pruned. If many the relationship degrees between an attribute and other attributes are large than $R_{\min}$, the attribute should be reserved and others should be pruned.

C. Objective Function

We give the objective function as (5). Let k fixed, the partition satisfying F minimum is the optimal partition.

$$F(C,T) = \min \frac{k\sqrt{\sum_{l=1}^{k} \sum_{j=1}^{n} \tau_{lj}[1 - R(c_l, x_j)]}}{\sum_{l=1}^{k} \left[1 - \frac{1}{n} \sum_{j=1}^{n} R(c_l, x_j)\tau_{lj}\right]} \quad (5)$$

here $\tau_{lj} = \{0,1\}$ is the membership degree of the jth object belonging to the lth cluster, c_l is center of the lth cluster, x_j is the jth object, m is the number of clusters and the strength of the incentive for clustering on more dimensions is controlled by the condition of the parameter $\gamma > 0$, $D_{lj} = [1 - R(c_l, x_j)]$ is the distance between the jth object and core of cluster that it belongs, $\overline{D_l} = \left[1 - R(c_l, \bar{x})\right]$ is the average distance among distances of objects in lth cluster to their core.

Proof: If all objects are in true clusters, which equals to that the inside distance among objects in a cluster is minimum and distance among objects in different clusters is maximum. Therefore, the inside distance $\sqrt{\sum_{l=1}^{k} \sum_{j=1}^{n} \tau_{lj}[1 - R(c_l, x_j)]}$ minimum and the average outside distance $\frac{1}{k} \sum_{l=1}^{k} \left[1 - \frac{1}{n} \sum_{j=1}^{n} R(c_l, x_j)\tau_{lj}\right]$ maximum, so

$\frac{k\sqrt{\sum_{l=1}^{k}\sum_{j=1}^{n}\tau_{lj}[1-R(c_l,x_j)]}}{\sum_{l=1}^{k}\left[1-\frac{1}{n}\sum_{j=1}^{n}R(c_l,x_j)\tau_{lj}\right]}$ is the minimum, so F will be minimum.

Let T fixed, and we can get C according to (6).

$$c_{li}=\frac{\sum_{j=1}^{n}\tau_{lj}x_{ji}}{\sum_{j=1}^{n}\tau_{lj}}(l=1,\cdots,k,i=1,\cdots,m) \quad (6)$$

Let C fixed, we can get T according to (7).

$$\tau_{lj}=\begin{cases}1 & 1-R(c_l,x_j)\le 1-\underset{r=1}{\overset{k}{Sup}}(c_r,x_j)\\ 0 & else\end{cases} \quad (7)$$

According to (5), we know that the smaller of the distance between objects in the same cluster, the larger the relationship degree between them, and the smaller the F, and the larger the distance between objects from different clusters, the smaller the F. Therefore, if k fixed, the formula (5) satisfies the aim of clustering that the distance between objects in a cluster is as small as possible and that in different clusters is as large as possible.

If k is unknown, we can obtain its value according (6) and (7). From Reference [28], we know that the optimal number of clusters $k\le\sqrt{n}$, so the scope of k value is $[1,\sqrt{n}]$.

The difference of objective function $\Delta F(k)$ is denoted as (8).

$$\Delta F(k)=\left|F(k)-F(k+1)\right| \quad (8)$$

The differential rate of objective function $\hat{F}(k)$ is denoted as (9).

$$\hat{F}(k)=\frac{\Delta F(k)}{F(k)} \quad (8)$$

The k satisfying both of $\Delta F(k)$ and $\hat{F}(k)$ local maximum is the optimal value.

D. Processes of FCESNC

FCESNC algorithm can be shown as follows:

1. Init parameters, objects X, minimum relationship degree between attributes $R_{\min}$, number of clusters k, weight $W=\{w_i\}$, the maximum iteration number Z.
2. Prune redundant attributes according to $R_{\min}$.
3. Search for $m\to 2$ subsets of Descartes.
4. Let iteration number z=1, search for k sets with maximum subsets of Descartes, and choose one object from each of them as initial k cores c_l.
5. Compute the membership degree τ_{lj} of x_j belonging to the lth cluster.
6. Compute the value of objective function F.
7. If z+1>Z or the cores fixed, stop the algorithm, else go to 8.

Compute c_{li}, and search for the nearest object to c_{li} from X as the new core, and let $z=z+1$, go to 5.

IV. INSTANCES

The proposed algorithm was firstly used for partitioning Synthesis dataset, and then for classifying practical Yellow-small dataset and Zoo dataset from UCI dataset in [29]. Because categorical data is difficult to be denoted in Figure, we can just show the value of objective in Figure.

A. Synthesis dataset

There are 30 practical points needing to be classified, and their proper partitions are {{1-10}, {11-20}, {21-30}}, where number j denotes the jth point. TABLE III gives results of 9 different clustering, which can be results of partitioning categorical data, numerical data or mixed data.

TABLE III ERROR RATE OF DIFFERENT ALGORITHM

Serial Number	Clustering Memberships
1	{1-30}
2	{{1-17,20,25},{18-19,21-24,26-30}}
3	{{1-6,8-9,13,21},{7,11-12,14,16-18,20,25},{10,15,19,22-24,26-30}}
4	{{1-9,12,},{10-11,13-17,23-24},{18-22,25-30}}
5	{{1-7,9-11},{8,12-18,20,25},{19,21-24,26-30}}
6	{{1-8,10-11},{9,12-18,25},{19-24,26-30}}
7	{{1-7,9-11},{8,12-17,20},{18-19,21-30}}
8	{{1-11},{12-17},{19,21-24,26-30},{18,20,25}}
9	{{1-7,9-11},{12-17},{19,21-24,26-30}{8},{18,20,25}}

Then we know that there will be 9 clustering memberships, and assume that all memberships have same weight because we don't know which result of membership is best, we just think that their same probabilities are equal to others, so we can ignore weights of memberships. At the same time, we know that 1th membership just has one cluster, and we can prune it from memberships, which are similar to prune attributes from dataset, or we can reserve it in the memberships, which don't influence the results.

Next, we can get the objective function $F(k)$, difference $\Delta F(k)$ and differential rate $\hat{F}(k)$ of objective function according to processes of FCESNC, which can be shown as Figure.2, Figure.3 and Figure.4.

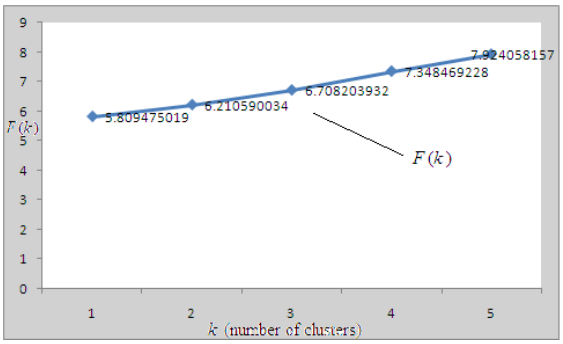

Figure 2. Objective function $F(k)$

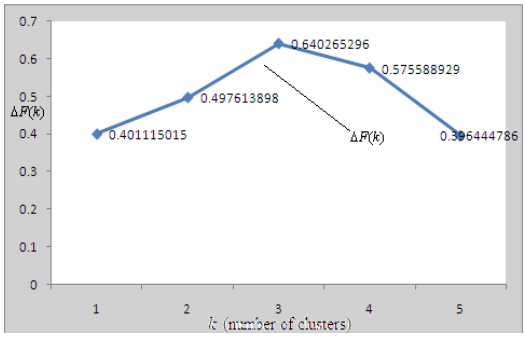

Figure 3. Difference $\Delta F(k)$ of objective function

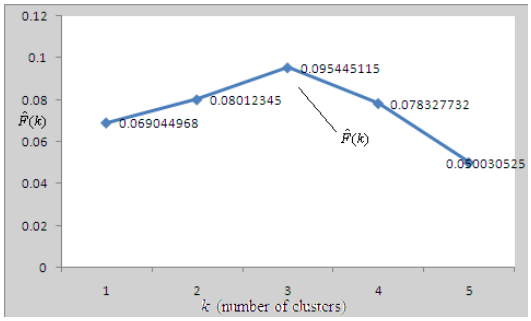

Figure 4. Differential rate $\hat{F}(k)$ of objective function

From Figure 3 and Figure 4, we know that the optimal k for this instance is 3, which is correct. For k=3, we get the result is {{1-11}, {12-18, 20, 25}, {19, 21-24, 26-30}} and its error rate is 0.1, which is better than any membership with k=3.

B. Yellow-Small Dataset

The yellow-small dataset includes 20 objects with 4 attributes, whose weight W={0.2,0.2,0.3,0.3}, according to FCESNC algorithm, we get results as TABLE IV.

TABLE IV RELATIONSHIP DEGREE MATRIX

Value of k	$F(k)$	$\Delta F(k)$	$\hat{F}(k)$
1	4.59	0.04	0.01
2	4.63	0.74	0.16
3	3.89	0.56	0.14
4	3.33	0	0

The results can also be shown in Figure 5, Figure 6, Figure 7.

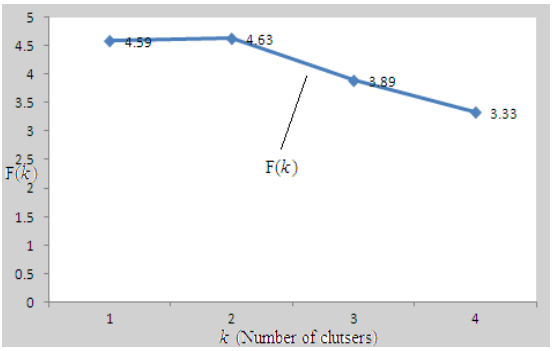

Figure 5. The value of $F(k)$ with different k

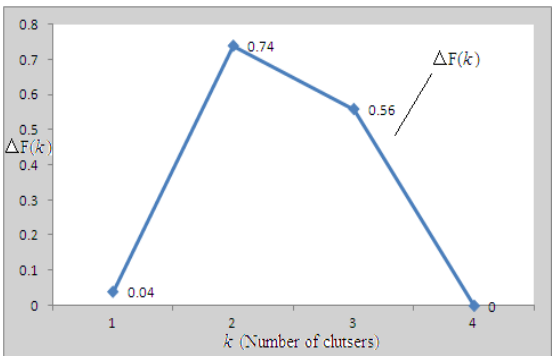

Figure 6. The value of $\Delta F(k)$ with different k

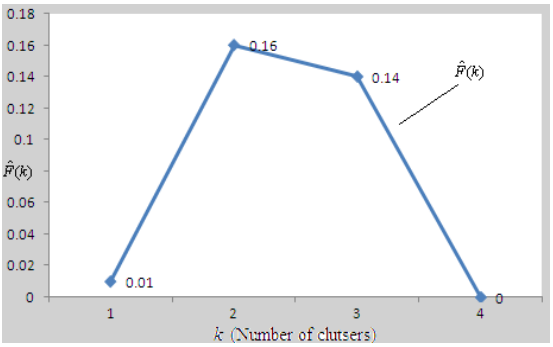

Figure 7. The value of $\hat{F}(k)$ with different k

Then we can obtain the optimal k according to (8) and (9). Because there is only one local maximum of difference and differential rate of objective function, so optimal k =2, and the error rate of its results of clustering is 0.05, which is fully consistent to the goal.

C. Zoo Dataset

Zoo dataset includes 101 objects with 16 attributes, which are related to each other. For example, Mammals are hair, blood, and breast-feeding animals, attributes of being Mammals and related to being hair, we can prune the attribute hair.

In Zoo dataset, we let R_{min}= 0.8, and finally we get 8 attributes for clustering, which are {4, 5, 7, 9, 10, 11, 13, 16}, the Arabic numerals are the labeling of attributes. Let w_i =0.125 for these 8 attributes.

Then, according to the processes of FCESNC, we can get the objective function $F(k)$, difference $\Delta F(k)$ and differential rate $\hat{F}(k)$ of objective function as shown in Figure.8, Figure.9 and Figure.10.

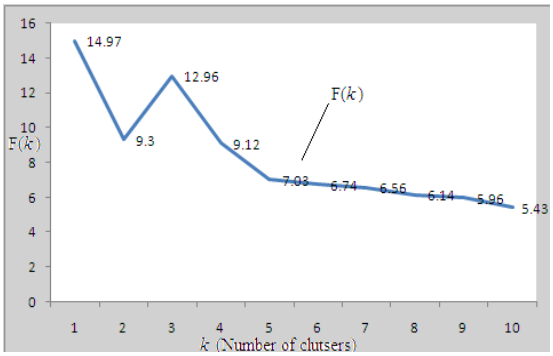

Figure 8. Objective function $F(k)$

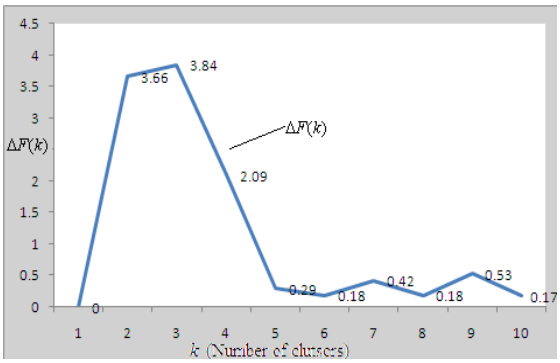

Figure 9. Difference $\Delta F(k)$ of objective function

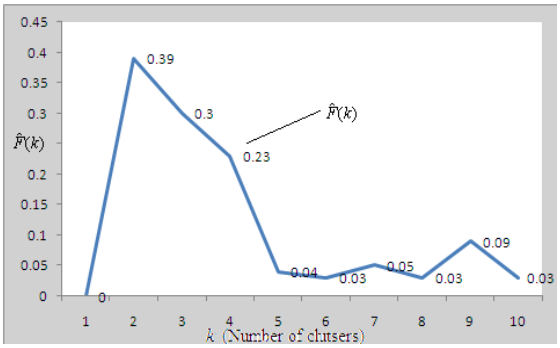

Figure 10. Differential rate $\hat{F}(k)$ of objective function

Figure.8 denotes the change of value of objective function $F(k)$ with the change of k.

Figure.9 and Figure.10 show that change of difference $\Delta F(k)$ and differential rate $\hat{F}(k)$ of objective function with the change of k. At the same time, we know that difference and differential rate of objective function are local maximum while k=2,7,9, and corresponding values are:

$$\Delta F(2) = 3.66, \Delta F(7) = 0.42, \Delta F(9) = 0.53$$

$$\hat{F}(2) = 0.39, \hat{F}(7) = 0.06, \hat{F}(9) = 0.09$$

So results while k=2, 7, 9 are relatively better than others. Comparing with the goal, the results with k=7 are shown in Table V.

TABLE V RESULTS OF FCESNC WHILE K=7

ith Cluster	Objects in Each Cluster
1	{ 1-2,4-7,10-11,18,20,23,29-30,32-33,36-37, 45-46,48-51,55-56,64-71, 75-76,85,94-95,97,99}
2	{3,8-9,13-14,19,35,39, 61-63,74,77,81,83,87,93}
3	{12,17,21-22,24,28,34,38,44, 57-60,72,79-80,84,88,91,96,101}
4	{25,31,40-41,43,52,82,89,98,100}
5	{26-27,42,53,90,92}
6	{73}
7	{15-16,47,54,78,86}

In TABLE V, the Arabic numerals in sets are the labeling of objects, and we know that 9 objects are partitioned into wrong clusters and the error rate of clustering is 0.089. Comparing to other algorithms in [3], the results were shown in TABLE VI.

TABLE VI ERROR RATE OF DIFFERENT ALGORITHM

Algorithms	Error Rate
Squeezer	0.190
GAClust	0.210
ccdByEnsemble	0.234
FCESNC	0.089

From TABLE VI, we know that the error rate of FCESNC is much smaller than other algorithms, and the proposed FCESNC algorithm only depends on the distribution of partitions based on the categorical data, and can obtain the final results under unsupervised circumstances.

From the instances above, we know that the proposed FCESNC algorithm can be used to partition categorical data and combine memberships from multi clustering that are suitable to categorical data, numerical data and mixed categorical data and numerical data.

V. CONCLUSIONS

Presently, clustering mainly focuses on numerical data and is relatively mature then categorical data and mixed numerical data, and the clustering ensemble algorithm under supervised condition also is relatively more mature and widely used then under unsupervised condition. In order to broaden the fields of clustering, a great deal of attention was attracted and the unsupervised clustering ensemble becomes one of hot topics in clustering.

We proposed the FCESNC (Fuzzy clustering ensemble algorithm for partitioning categorical data) algorithm in this paper, which can be used for both partitioning categorical data and mixed numerical and categorical data, also can be applied to joint results of multi clustering. FCESNC algorithm makes use of the relationship degree of attributes for pruning a part of attributes with close relationship to each other. Then, according to the distribution of clustering memberships, m Descartes subset and relationship degree between objects were used for establishing the relationships between objects under unsupervised circumstances and got the minimum value of objective function of clustering. Then we obtained the optimal value of number of clusters by making the

difference and differential rate of objective function local maximum and its corresponding partitions were optimal clustering relatively.

Finally, the proposed FCESNC algorithm was applied in classifying Synthesis dataset, Yellow-small dataset and Zoo dataset, and results show it is effective and feasible for partitioning categorical data and combining results of multi clustering no matter the data points have categorical value, numerical value or mixed value.

ACKNOWLEDGMENT

This work was supported by the Doctoral Fund of Ministry of Education of China (No. 200801510001) and the Key Project of Chinese Ministry of Education (No. 209030).

REFERENCES

[1] H.Y. Wang, H.B. Lu, and Z.P. Li. "Study of high dimensional data clustering based on cluster ensembles algorithm", *Electronic Measurement Technology*, Chengdu: South West Jiaotong University Press, vol.31, no.4, pp.41-45, 2008. (In Chinese)

[2] A. Strehl and J.Ghosh. "Cluster ensembles a knowledge reuse framework for combining multiple partitions". *Journal of Machine Learning Research*, Cambridge: MIT. Press, vol.3, no.3, pp. 583-617, 2003.

[3] Z.Y. He, X. F. Xu, and S.C. Deng, "A cluster ensemble method for clustering categorical data", *Information Fusion*, pp.143–151, June 2005,

[4] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*. Prentice Hall, 1988.

[5] Y. Zhang, A.W. Fu, C.H. Cai, and P.A. Heng, "Clustering categorical data", *Proc. of ICDE'00*, pp. 305–305, 2000.

[6] M.R. Anderberg, Cluster Analysis for Applications. *Academic*, 1973.

[7] G.H. Ball and D.J. Hall, "A Clustering Technique for Summarizing Multivariate Data", *Behavioral Science*, vol. 12, pp. 153-155, 1967.

[8] J.B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proc. Fifth Symp. Math. Statistics and Probability*, vol. 1, AD 669871, pp. 281-297, 1967.

[9] Z. X. Huang. "Extensions to the k-means algorithm for clustering large datasets with categorical values". *Data Mining and Knowledge Discovery*, vol.2, no.1, pp. 283-304, 1998.

[10] Z. X. Huang. "Clustering large data sets with mixed numeric and categorical values", *Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Singapore: World Scientific, pp. 21-34, 1997.

[11] Y. Wang and L. Yang. "A clustering algorithm for mixed valued data based on aggregate function", *Journal of Dalian University of Technology*, Dalian: Dalian University of Technology, vol.46, no.3, pp. 446-448, 2006. (In Chinese)

[12] Y. Zhao, B. Li, X. Li, W. H. Liu, S. J. Ren. "Cluster ensemble method for databases with mixed numeric and categorical values", *J. Tsinghua Univ (Sci & Tech)*, Beijing: Tsinghua University, vol.46, no. 10, pp. 1673-1676, 2006. (In Chinese)

[13] L. Yang and W. Wang, "Clustering Ensemble Approaches: An Overview", *Application Research of Computers*, Chengdu: Application Research of Computers Press, vol.22, no.12, pp. 8-10, 2005. (In Chinese)

[14] Y. Yang, F. Jin, and M. Kamel. "Latest development of clustering ensemble", *Computer Engineering and Applications*, Beijing: Publishing House of Journal of Computer Engineering and Applications, vol.44, no.11, pp. 142- 144, 2008. (In Chinese).

[15] A. Topchy, A. K. Jain and W. Punch. "A mixture model for clustering ensembles". *Proceedings of the 4th SIAM International Conference on Data Mining*, Lake Buena Vista, Florida, 2004, pp. 379-390.

[16] B Minaei-Bidgoli, A Topchy and W F Punch. "A Comparison of Resampling Methods for Clustering Ensembles". *Proceedings of Intl. Conf. on Machine Learning, Models, Technologies and Applications*, pp. 939-945, 2004.

[17] B Minaei-Bidgoli, A Topch, and W F Punch. "Ensembles of Partitions via Data Resampling". *Proceedings International Conference on Information Technology, Coding and Computing*, vol.2, pp. 188-192, 2004.

[18] B Fischer and J M Buhmann. "Path-based Clustering for Grouping of Smooth Curves and Texture Segmentation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.25, no.4, pp. 513-518, 2003.

[19] A Topchy, A K Jain and W F Punch. "Combining Multiple Weak Clusterings". *Proceedings of the 3rd IEEE International Conference on Data Mining*, pp. 331-338, 2003.

[20]X Z Fern and C E Brodley. "Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach". *Proceedings of the 20th International Conference on Machine Learning*, pp. 186-193, 2003.

[21] A. Topchy, B. Minaei-Bidgol, A. K. Jain, et al. "Adaptive Clustering Ensembles". *Proceedings of the 17th International Conference on Pattern Recognition,* England, vol.1, pp. 272-275, 2004.

[22] L.I. Kuncheva, S T Hadjitodorov. "Using Diversity in Cluster Ensembles". *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pp.1214-1219, 2004.

[23] A Fred and A K Jain. "Data Clustering Using Evidence Accumulation". *Proceedings of the 16 th International Conference on Pattern Recognition*, vol. 4, pp. 276-280, 2002.

[24] A Fred and A K Jain. "Evidence Accumulation Clustering Based on the K-means Algorithm". *Proceedings of the International Workshops on Structural and Syntactic Pattern Recognition*, pp. 442-451, 2002.

[25] H Ayad and M Kamel. "Finding Natural Clusters Using Multi-Clusterer Combiner Based on Shared Nearest Neighbors". *Proceedings of the 4th International Workshop on Multiple Classifier Systems*, vol. 2709 of Lecture Notes in Computer Science, Sp ringer, pp. 166-175, 2003.

[26] X Z Fern and C E Brodley. "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning". *The 21st International Conference on Machine Learning*, 2004.

[27] H Ayad, O A Basir and M Kamel. "A Probabilistic Model Using Information Theoretic Measures for Cluster Ensembles". *Proceedings of the 5 th International Workshop on Multiple Classifier Systems*, vol. 3077 of Lecture Notes in Computer Science, Sp ringer, pp. 144-153, 2004.

[28] S. Yang, Y. Li, X. Hu, R. Pan, "Optimization Study on k-Value of K-means Algorithm", *Systems Engineering-theory & Practice*, Beijing: Systems Engineering Society of China, vol.2, pp. 97-101, 2006. (In Chinese)

[29] http://www.ics.uci.edu/~mlearn/databases.

Taoying Li was born in Anhui Province, China on September 1983. Taoying Li received the BE degree in information management and information system at Dalian Maritime University, Dalian, China, in 2005 She is currently working toward the Successive Postgraduate and Doctoral Progr- am in the Transportation Management College, Dalian Maritime University.

She had carried out some projects and published several papers. She majors in Management science and Engineering and her research interests include data mining, system engineering, and artificial intelligence.

Yan Chen was born in Liaoning Province, China on December 1952. Yan Chen received the BE degree in computer software at Dalian Maritime College, Dalian, China, in 1978, the MS degree in computer application at Dalian Maritime College in 1989 and the PhD degree in Management Science and Engineering at Dalian University of Technology in 2000.

She is a professor in Transportation Management College, Dalian Maritime University, and the dean of the Key Laboratory in Liaoning Province on Logistics Shipping Management System Engineering. She has published three books and published more than 100 papers. Her main research interests include data mining, system engineering, special data mining, decision-making support system and data warehouse.