International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 1 047 – 053

An Enhanced Initialization Method to Find an Initial Center for K-modes Clustering

S. Saranya
PG Scholar
Department of Computer Science and Engineering
Kongu Engineering College
Perundurai, Tamil nadu, India.

Dr.P.Jayanthi
Associate Professor
Department of Computer Science and Engineering
Kongu Engineering College
Perundurai, Tamil nadu, India.

Abstract **-** Data mining is a technique which extracts the information from the large amount of data. To group the objects having similar characteristics, clustering method is used. K-means clustering algorithm is very efficient for large data sets deals with numerical quantities however it not works well for real world data sets which contain categorical values for most of the attributes. K-modes algorithm is used in the place of K-means algorithm. In the existing system, the initialization of K- modes clustering from the view of outlier detection is considered. It avoids that various initial cluster centers come from the same cluster. To overcome the above said limitation, it uses Initial_Distance and Initial_Entropy algorithms which use a new weightage formula to calculate the degree of outlierness of each object. K-modes algorithm can guarantee that the chosen initial cluster centers are not outliers. To improve the performance further, a new modified distance metric -weighted matching distance is used to calculate the distance between two objects during the process of initialization. As well as, one of the data pre-processing methods is used to improve the quality of data. Experiments are carried out on several data sets from UCI repository and the results demonstrated the effectiveness of the initialization method in the proposed algorithm.

Keywords – *K-modes clustering; Outlier Detection; Initial cluster center; Initial_Distance; Initial_Entropy.*

*****

I. INTRODUCTION

Cluster analysis or clustering is grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. The clustering algorithms can be broadly classified into five types: hierarchical clustering, partitional clustering, density-based clustering, grid-based clustering and model-based clustering. K-means is one of the partitional clustering algorithms, it is very efficient for large data sets deals with numerical quantities, but it could not be efficiently used for real world data sets which contain categorical values for most of the attributes. To avoid this, [10, 11] proposed the K-modes algorithm which is used in place of K-means algorithm. In K- modes algorithm, (1) a simple matching dissimilarity measure for categorical objects is used;(2) calculation of modes instead of means of clusters is used;(3) It updates the modes using a frequency-based method in the clustering process to minimize the clustering cost function.

In general, the performance of K-modes algorithm is faster when to compare to K-means algorithm because it needs less iterations to converge [11]. It should be noted that

K-modes algorithm uses the same clustering process as like K-means algorithm except only clustering cost function[18]. K-means algorithm is sensitive to selection of initial cluster center and it produced the undesirable cluster structures [5]. Similarly, K-modes algorithm is also sensitive to the selection of initial cluster centres [14], but it provides as good initial cluster centres.

In this paper, initialization of K-modes clustering from the view of outlier detection is considered [14]. The main idea behind this concept is that outliers should not be selected as initial cluster centres and it avoids that various initial cluster centres come from the same cluster. In K-modes clustering, to solve the initialization problem, first propose a modified initialization algorithm (called Initial_Distance) via modified distance-based outlier detection technique. Second, to overcome the problems of distance-based technique, further present a modified Initial_Entropy based outlier detection technique within the framework of rough sets [17].

The above said algorithms calculate the degree of outlierness to avoid that outliers are selected as initial cluster centers.In addition, it avoids that various initial cluster centres come from the same cluster by calculating the distances between candidate centres and all currently existing initial centres. During the process of initialization, adopt a new modified distance metric—weighted matching distance metric to calculate the distance between two objects, which can obtain better results when compared to simple distance metric[14] for categorical attributes. Moreover, counting sort-based method is used to reduce the time complexities of algorithms Initial_Distance and Initial_Entropy by compute the partition of universe U induced by a given indiscernibility relation. Further to

improve the performance, one of the data pre-processing methods is used to handle the missing values in the data sets.

II. RELATED WORK

In past years, various initialization methods have been proposed to initialize cluster centres for K-modes clustering [3, 7, 9, 11, 19, 20]. S.S. Khan et al. proposed the random initialization method for K-modes clustering to initialize cluster centers by performing multiple clustering based on the categorical attribute. The K-modes algorithm must be executed several times to obtain desirable clustering results. Hence, it is necessary to design a non-random initialization algorithm for K-modes clustering to cluster categorical data.

Huang [4, 11] proposed the two non-random initialization methods for *K*-modes clustering. The first method which selects the first *K* objects from the dataset as initial cluster centres, and the second method which assigns the most frequent categories of data that are equally to Kinitial cluster centres. However, both of these two methods have some problems.

Sun et al. [19] proposed an iterative initial point's refinement algorithm given by Bradley and Fayyad [4] for K-modes clustering. Experiments are demonstrated that it leads to higher precision results and much more reliable than the random selection method without refinement.

Barbara et al. [3] proposed a heuristic algorithm COOLCAT and it found the K most 'dissimilar' objects from the data set by maximizing the minimum pairwise entropy of the chosen points, and used the K objects as initial cluster center

Z. Y. He [9] proposed two initialization methods for K-modes clustering based on the farthest-point heuristic. The first method is basic farthest-point heuristic (BFPH) and the second method is new farthest-point heuristic (NFPH). Experiments show that proposed initialization method obtain a better clustering accuracy than random selection initialization method for K-modes clustering.

Wu et.al.[20] proposed a density based initialization method for *K*-modes clustering.F. Y. Cao et al [7] proposed a novel initialization method for categorical data to select initial cluster centers by considering the distance between objects and the density of each object.. This method is superior to random initialization method and can be applied to large data.

Bai et al [6] proposed an initialization method for K-modes clustering, which can select initial cluster centers as well as the number of clusters. It improves the performance and scalability of proposed method for real data sets. The experimental results are demonstrated that the initialization method is very effective and can be applied to large data sets for k-modes clustering.

T. Bai et al. [14]proposed a global K-modes algorithm for clustering categorical data. The algorithm randomly selects Kini initial centers, where K is the predefined number of clusters, and it eliminates the redundant cluster centers by using an iterative optimization process.

Feng Jiang [14] proposed two different types of initialization algorithm for K-modes clustering are used, where the first is modified distance- based outlier detection technique and the second is modified partition entropy – based outlier detection technique. The above mentioned algorithms avoids that various initial cluster centres come from the same cluster and guarantee that the chosen initial cluster centres are not an outliers.

As above said in this paper, the initialization of K-modes clustering from the view of outlier detection is considered [1, 12, 13, 14]. Outlier detection was first discussed in statistics. In order to solve the problems of statistics-based outlier Detection technique, E.M. Knorr, R.T. Ng [14] proposed the distance-based outlier detection technique, which calculates distance between each object's to its neighbors for finding outliers.As consider the two different tasks in data mining, clustering and outlier detection have a some close relationship. Clustering can be used to detect outliers, and it emerges clusters far away from other clusters.

III. PROPOSED METHOD

A. Data Pre-Processing

Data Pre-Processing is an important tasks in data mining process. The different data Pre-Processing methods are Data Cleaning, Data Integration, Data Transformation and Data Reduction. In this paper, Data Cleaning method is used to handle the noisy, Incomplete and inconsistent of data. This method cleans the data by filling in missing values, Smoothing Noisy data, identifying and removing outliers and also resolving inconsistencies. Here, filling missing values method is used. This can be done by ignoring the tuples containing missing values, fill the missing values manually, use a constant value to fill in the missing values and use some probable values to fill missingvalues ,etc. Hence it improves the performance and quality of data.

B. New-Weighted Matching Distance

The existing methods used simple matching distance metric [14] to measure the distance between two objects for categorical attributes. It has some limitations. To overcome the limitations of simple matching distance metric, a weight matching distance is used [14].To improve the better performance of clustering results, a new

weighted matching distance metric is proposed in this paper, which is used to calculate the distance between two objects for categorical attributes at the time of initialization. Let Sig (att) represents the partition entropy-based significance of attributes [14].The new proposed weight of an attribute a in IS defined as:

If Sig (att) =0, Weight (att) $= \frac{1}{2} \times \left(1 + \frac{|\text{Att}| + C_{\text{zero}}}{\sqrt{|\text{Att}| - C_{\text{zero}}}}\right)$ (1)

If Sig (att) >0, Weight (att) = 1+ Sig (att) (2)

From above eq (1), |Att| denotes the number of attributes in Att, and C_{zero} denotes the number of attributes in Att whose significance value is equal to zero.The weighted matching distance is calculated between two objects uand vis defined as follows:

wd(u,v)= $\sum_{att \in Att}$ weight(att) × δ_a(u,v) (3)

Where

$$\delta_a(u,v) = \begin{cases} 1, & \text{if } f(u,att) \neq f(v,att) \\ 0, & \text{otherwise} \end{cases}$$

In this paper, the new weighted matching distance metric is used to calculate the distance between two objects at the time of initialization. This paper concentrates only on the issue of initialization for K-modes clustering. In Future, we will consider the issue of K-modes clustering based on the weighted matching distance metric.

C. Initial_Distance

In this section,present the initialization algorithm Initial_Distancefor K-modes clustering.In Initial_Distance algorithm, it is used to calculate the degree of outliers of each object to avoid initial cluster centre come from same cluster and calculating the distance between candidate initial cluster centres and all currently existing initial cluster centre to avoid different initial cluster centre come from the same cluster. However,for any object u∈U, this method gives only a binary classification of i.e., uis or is not an outlier [14]. In many cases, it is more useful to assign uas a degree of being an outlier. Therefore, introduce a concept called 'distance outlier factor (DOF)', which can quantify the degree of outlierness of a given object [12, 13].Given information table IS = (U, A,V,f), for any u∈U, the distance outlier factor of object uin IS defined as:

$$\text{DOF}(u) = \frac{|\{v \in U : wd(u,v) > d\}|}{|U|} \quad (4)$$

Where wd(u,v) denotes the weighted matching distance between objects uand v.In above eq (4), d isa given parameter.It should be taken as mid value within the range of 0 to$\sum_{att \in Att}$ weight(att). To calculate the possibility of each candidate cluster centre being as initial cluster centre,the following two factors are used together,

1. The degree of outlierness of each candidate initial cluster centre;
2. The distance between candidate initial cluster centre and all currently existing initial centres.

Consider a candidate centre v∈U-C, the possibility of v being a distance-based initial cluster centre is defined as:

$$P\_DIC(v) = \frac{\sum_{m=1}^{r} wd(v,c_m)}{r} \cdot \sqrt{DOF(v)} + \frac{\sum_{m=1}^{r} wd(v,c_m)}{r \times (1+\sqrt{DOF(v)})} \quad (5)$$

Where wd(v,c_m) is the weighted matching distance between v and c_m,C1={c_1,c_2 … … ,c_r) is the set of all currently existinginitial cluster centres,DOF(u) is the distance outlier factor,r is the number of currently existing initial cluster centre.

D. Initial_Entropy

The Initial_Distance algorithm based on distance-based outlier detection technique is very effective to detect outliers for non-parametric technique. Hence it is not feasible fordealing with large data sets containing categorical data because ofit's high time complexity. The distance-based method is used to detect outliers that it requires to select an appropriate distance metric to calculate a distance based outliers.Finding suitable distancemetric is difficultto doformany practical tasks and it may involve too many trials. Moreover, the quality of Initial_Distance algorithm based on distance-based technique is highly depending on the dis Parameter.To avoid the problems of distance-based technique, presentthe modified initialization algorithm calledInitial_Entropy for K-modes by using partition entropy based

Given:
S=(U,A,V,F),whereU={$u_1, u_2 \ldots . u_k$},A={$a_1, a_2 \ldots a_l$); K and dparameters,where K is the number of clusters and d isthe threshold value for distance based outliers.
1.Initialize C=0;
2.Calcualte the partition of attribute based on countingsort;
3.Calculate the partition entropy PE(Att) of U/IND(Att);
4.For each 1≤j≤ l,calculate the partition,partitionentropy of attribute U/IND(Att-{att_j}) and significanceSig(att_j) of attribute att_j;
5.Calculate the weight(a_{ttj}) of attribute att_j, where1≤j≤ l;
6.for m=1 to k do
7. for n=1 to k do
8. Calculate the weighted matching distance wd(u_m,v_n) between two objects
9. end
10.Calculate the distance outlier factor DOF(u_m);
11. end
12.sort domain U in ascending order based on DOF(u);
13.Select the first object v from U as the first initial cluster center;
14. while |C1| < K do
15. for m=1 to k do
16. if x_m not belongs to C1 then
17. for n=1 to |C1| do
18. calculate the weighted matching distance wd(u_m,c_m) between two objects u_m andc_m. where m is the current existing initial cluster centre i.e.,c_m ∈ C1.
19. end
20. Calculate the possibilityP_DIC(u_m) of object to select an next initial cluster center;
21. end;
22. end;
23 select q∈ U-C1,such thatP_DIC(q)=Max({ P_DIC(v):v∈U-C1}) and assignC1=C1U{q};
24.end
25. return C.

Algorithm 1.Initial_Distance

outlier detection technique within the framework of rough set. This algorithm is used to find an initial cluster center for K-modes clustering. In Initial_Entropy, for any candidate initial cluster center y, to calculate the possibility of y being an initial cluster center by using two factors which is used in Initial_Distance algorithm. The following two factors are: (1) the degree of outlierness of y; (2) the distance between candidate center y and each current existing initial cluster center.

IV. EXPERIMENTAL ANALYSIS

To evaluate the performance of algorithms Initial_Distance and Initial_Entropy, the results of the two algorithms are compared with Feng Jiang initialization method on different categorical datasets. In the Soybean data set, already it reaches the value 1 in existing paper [14].so rest of the data sets are considered for experiments. The different performance metrics are used to evaluate the quality of clustering results and compared with different initialization methods.The following five categorical data sets were used to test Initial_Distance and Initial_Entropy. The data sets used in this paper are listed below:

1. Zoo data set.
2. Breast Cancer Wisconsin
3. Mushroom data set.
4. Lung Cancer data set.
5. Congressional Voting Records data set.

The different properties of five data sets are taken from the UCI Machine Learning Repository [2] (Bache K and Lichman M, 2013) as shown in Table 1.

The Initial_Distance and Initial_Entropy algorithms were compared with the Feng Jiang. In the experiments, first used Initial_Distance and Initial_Entropy to select an K initial cluster centres for each of five data set given in table 1.second, Feng Jiang proposed initialization methods is used to select an K initial cluster center. Third, obtained a corresponding clustering result by using K-modes algorithm proposed by Huang [11] for each of five data set of initial cluster centres generated by a specific initialization method. Finally, the obtained K-modes clustering results are compared with each other for five data sets for each initialization method.

This paper is implemented the Initial_Distance and Initial_Entropy algorithms in mat lab. Experiments were conducted on a Intel core i3 machine with 4 GB RAM and 1TB hard disk, running on the windows XP operating system.

To measure the performance of clustering results, adopted the performance metrics used by Wu et al. [14].For a given data set D and clustering algorithm A, assume that data set D contains K classes which is denoted by $Cl_1,\ldots Cl_K$, and the clustering algorithm A partitions data set D into K clusters which is denoted by $Cl'_1,\ldots\ldots Cl'_K$.For each of objects $1\leq i \leq K$, let $m_i$represents the number of objects which is correctly assigned to the class Cl_i(i.e., $m_i = |Cl_i \cap Cl'_i|$);n_i represents the number ofobjects which is incorrectly assigned to the class Cl_i(i.e., $n_i = |Cl'_i| - m_i$.t_i represents the number ofobjects which is incorrectly rejected from class Cl_i(i.e., $t_i = |Cl_i| - m_i$.The performance

ofclustering results generated by algorithm Initial_Distance and Initial_Entropy was measured by the following three metrics: Precision(PR), Recall(RE) and Accuracy (AC)[14], where

Given:IS=(U,A,V,F),where$U=\{x_1, x_2 \ldots . x_n\}$,$A=\{a_1, a_2 \ldots a_m\}$; K and dis parameters,where K is the number of clusters and dis is the threshold value for distance based outliers.
1. Initialize C=0;
2. Calculate the weight (a_j) of attribute a_j, where 1≤j≤ m, by using steps 1-4 of Algorithm 1.
3. Construct the weight-based sequence of attribute S=<$a'_1 \ldots . a'_m$>and the weight-based sequence of attribute subsets AS=<$A_1 \ldots A_m$> in IS.
4. for j=1 to m do
5. calculate the weight W(A_j) of attribute subset A_j;
6. calculate the partition U/IND({a_j}) and U/IND({A_j}) ,partition entropy PE({a_j}) and PE({A_j}) of attributes and attribute subsets.
7. for i=1 to n do
8. calculate the partition entropy of attribute for each object PE_{x_i}({a_j}) and each attribute subsetsPE_{x_i}({A_j})
9. calculate the significances of attribute $sig_{\{a_j\}}$ (x_i) and attribute subset $sig_{\{A_j\}}(x_i)$ of object x_i
10. end
11.end
12.calculate the partition entropy outlier factor PEOF(x) of object x;
13. sort domain U in ascending order based on PEOF(x);
14. Remaining steps 13-25 follows as in Algorithm1.

Algorithm 2. Initial_Entropy

TABLE 1

Properties of the five UCI data Sets.

Properties	Zoo	Breast	Mushroom	Lung	Voting
No of Classes	7	2	2	3	2
No of Instances	101	699	8124	32	435
No of Attributes	16	9	22	56	16
Missing Values	No	Yes	Yes	Yes	Yes

$$PR = \frac{\sum_{i=1}^{K} \frac{m_i}{m_i+n_i}}{K} \; ; \; RE = \frac{\sum_{i=1}^{K} \frac{m_i}{m_i+t_i}}{K} \; ; \quad AC = \frac{\sum_{i=1}^{K} m_i}{|D|} \qquad (6)$$

The above metrics are calculated and its results are compared with each above mentioned initialization method for six data sets.

V. RESULT AND DISCUSSION

From the Table 2-10 and Fig 1-5, the performance of the K-modes clustering results are shown respectively on the Zoo, Breast, Mushroom, Lung, and voting data sets. Already the Soybean data set reached the value 1in existing paper[14].so, Calculate the performance metrics :'AC' ,'PR','RE' for rest of the data set listed in the Table 1.

TABLE 2

Confusion Matrix for Initial_Distance on Zoo data set

Class	a	b	c	d	e	f	g
a	39	0	2	0	0	0	0
b	0	18	0	2	0	0	0
c	1	0	0	3	1	0	0
d	0	0	0	13	0	0	0
e	0	0	0	0	3	1	0
f	0	0	0	0	0	7	0
g	0	0	0	0	0	3	8

TABLE 3

Confusion Matrix for Initial_Entropy on Zoo data set

Class	a	b	c	d	e	f	g
a	37	2	0	2	0	0	0
b	0	20	0	0	0	0	0
c	1	0	1	2	1	0	0
d	0	0	0	13	0	0	0
e	0	0	0	0	4	0	0
f	0	0	0	0	0	8	0
g	0	0	0	0	0	1	9

TABLE 4

Confusion Matrix for Initial_Distance on Breast data set

Class	Benign	Malignant
Benign	455	3
Malignant	31	210

TABLE 5

Confusion Matrix for Initial_Entropy on Breast data set

Class	Benign	Malignant
Benign	456	2
Malignant	29	212

TABLE 6

Confusion Matrix for Initial_Distance on Mushroom data set

ISSN: 2321-8169
047 – 053

Class	Poisonous	Edible
Poisonous	3115	801
Edible	10	4198

TABLE 7
Confusion Matrix for Initial_Entropy on Mushroom data set

Class	Poisonous	Edible
Poisonous	3025	891
Edible	10	4198

TABLE 8
Confusion Matrixfor Initial_Distance on Lung data set

Class	a	b	c
a	5	4	0
b	0	13	0
c	1	0	9

TABLE 9
Confusion Matrix for Initial_Entropy on Lung data set

Class	a	b	c
a	8	1	0
b	2	11	0
c	1	4	5

TABLE 10
Confusion Matrix for Initial_Distance on Vote data set

Class	Republican	Democrat
Republican	161	7
Democrat	36	231

TABLE 10
Confusion Matrix for Initial_Distance on Vote data set

Class	Republican	Democrat
Republican	161	7
Democrat	36	231

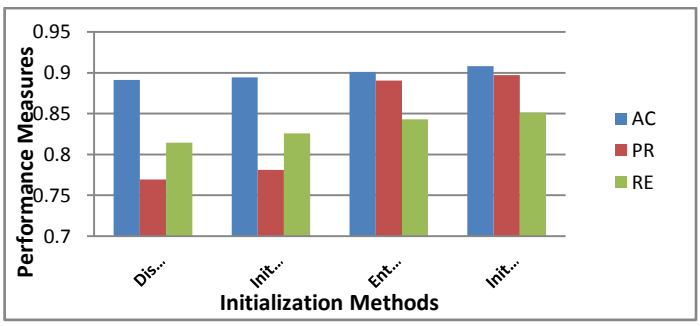


Fig. 1. Clustering results on the Zoo data set

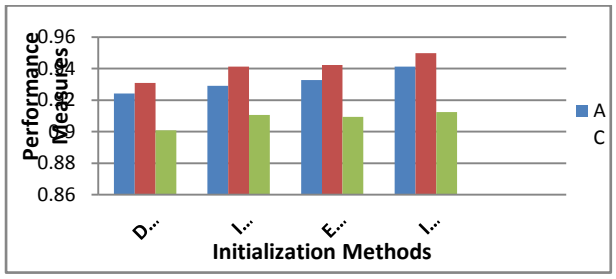


Fig. 2. Clustering results on the Breast data set

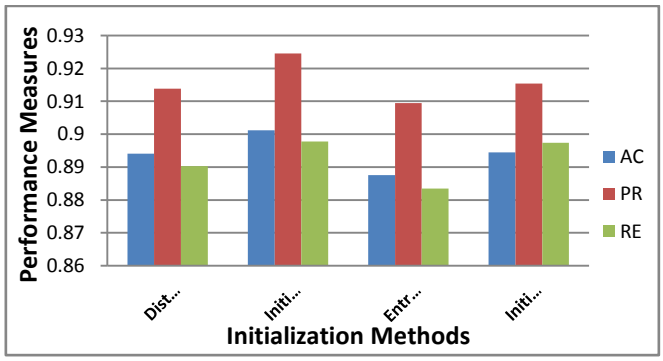
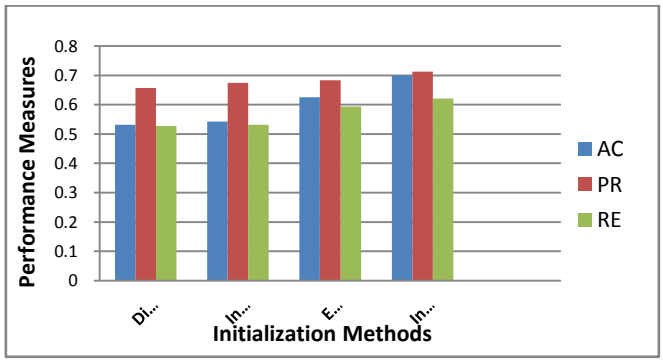


Fig.3. Clustering results on the Mushroom data set



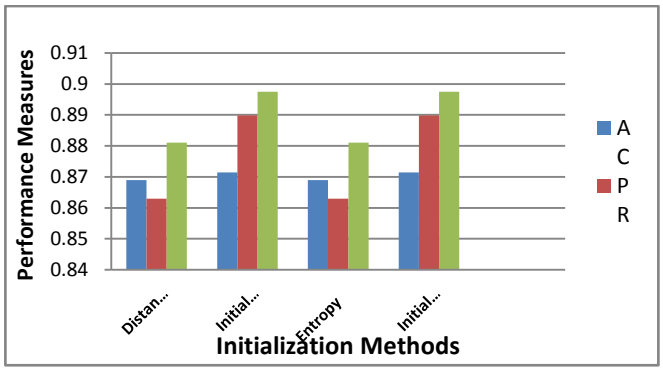


Fig.5. Clustering results on the Voting data set

VI. CONCLUSION

The selection of initial cluster center is the important tasks for K-modes clustering, since improper selection of initial cluster center may result in undesirable cluster

structures. In this paper, presented the Initial_Distance and Initial_Entropy algorithms to select an initial cluster centres in K-modes clustering combined with the detection of outliers. Consider a given object x, the above said algorithms is used to calculate the possibility of x being an initial cluster center based on following two factors: (1) the degree of outlierness of x; (2) the distance between the objects x and each existing initial cluster center, where the first factor can avoid that outliers are selected as initial cluster center and the second factor can avoid that various initial cluster center are come from same cluster. Hence the proposed method is tested with five categorical data sets and experimental results shown the effectiveness of clustering algorithm. In future, other outlier detection techniques may use to select an initial cluster center for K-modes clustering algorithm. Further, Fuzzy K-modes algorithm may apply to Initial_Distance and Initial_Entropy to improve the performance of clustering result.

REFERENCES

[1] Angiulli, S. Basta, S. Lodi, C. Sartori, Distributed Strategies for Mining Outliers in Large Data Sets, IEEE Transactions on Knowledgeand Data Engineering 25(7)(2013)1520–1532.

[2] K. Bache, M. Lichman, UCI machine learning repository, 2013, http://archive.ics.uci.edu/ml.

[3] E.Barbara, J. Couto, Y. Li, and COOLCAT: An entropy-based algorithm for categorical clustering, in: Proc. of the Eleventh Int. Conf. on Information and Knowledge Management, 2002, pp. 582–589.

[4] P.S. Bradley, U.M. Fayyad, Refining initial points for *K*-means clustering, in: Proc. of the 15th Int. Conf. on Machine Learning, Morgan Kaufmann,1998, pp.91–99.

[5] F.Y. Cao, J.Y. Liang, G. Jiang, An initialization method for the *K*-Means algorithm using neighborhood model, Computers and Mathematics with Applications 58 (3) (2009) 474–483.

[6] S. Chawla, A. Gionis, *K*-means-: A Unified Approach to Clustering and Outlier Detection, in: Proc. of the 13th SIAM Int. Conf. on Data Mining, Texas, USA, 2013, pp.189–197.

[7] F.Y. Cao, J.Y. Liang, L. Bai, A new initialization method for categorical data clustering, Expert Systems and Applications 36 (7) (2009) 10223–10228.

[8] A.Fred, A.K. Jain, Combining Multiple Clustering Using Evidence Accumulation, IEEE Transactions on Pattern Analysis and Machine Intelligence 27(6) (2005) 835–850.

[9] Z.Y.He, Farthest-point heuristic based initialization methods for *K*-modes clustering, 2006, CoRR, abs/cs/0610043.

[10] Z.X. Huang, A fast clustering algorithm to cluster very large categorical datasets in data mining, in: Proc. of the SIGMOD Workshop Research Issues on Data Mining and Knowledge Discovery, Canada,1997, pp.1–8.

[11] Z.X. Huang, Extensions to the *k*-means algorithm for clustering large datasets with categorical values, Data Mining and Knowledge Discovery 2(3) (1998) 283–304.

[12] E.Jiang, Y.F. Sui, C.G. Cao, A rough set approach to outlier detection, International Journal of General Systems 37(5)(2008)519–536.

[13] E.Jiang, Y.F. Sui, C.G. Cao, A hybrid approach to outlier detection based on boundary region, Pattern Recognition Letters 32(14)(2011)1860–1870.

[14] F Jiang, Guozhu Liu, Junwei Du, Yuefei Sui, Initialization of K-modes clustering using outlier detection techniques, Pattern Recognition Letters 332(2016)167–183.

[15] S.S. Khan, A. Ahmad, Computing initial points using density based multi scaled at a condensation for clustering categorical data, in: Proc. of the 2nd International Conference on Applied Artificial Intelligence (ICAAI03), Kolhapur, India, 2003.

[16] P. Mitra, C.A. Murthy, S.K. Pal, Density-based multi scaled at a condensation, IEEE Transactions on Pattern Analysis and Machine Intelligence 24(6) (2002) 734–747.

[17] Z. Pawlak, Rough Sets, International Journal of Computer and Information Sciences11 (5) (1982)341–356.

[18] Y.H.Qian, F.J.Li, J.Y.Liang, B.Liu, C.Y.Dang, Space structure and clustering of categorical data, IEEE Transactionson Neural Networks and Learning Systems (2015).Accepted.

[19] Y. Sun, Q.M. Zhu, Z.X. Chen, An iterative initial-points refinement algorithm for categorical data clustering, Pattern Recognition Letters 23(7)(2002)875–884.

[20] S.Wu, Q.S.Jiang, Z.X.Huang, A new initialization method for clustering categorical data, in: Proc. Of the 11th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data mining, in: Springer LNAI, vol. 4426, 2007, pp.972–980.