I.J. Information Technology and Computer Science, **2016, 3, 51-57**
Published Online March 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2016.03.06

Extension of K-Modes Algorithm for Generating Clusters Automatically

Anupama Chadha[*]
[*]Faculty of Computer Applications, MRIU, Faridabad, India
E-mail: anupamaluthra@gmail.com

Suresh Kumar[**]
[**]Faculty of Engineering and Technology, MRIU, Faridabad, India
E-mail: suresh.fet@mriu.edu.in

Abstract—K-Modes is an eminent algorithm for clustering data set with categorical attributes. This algorithm is famous for its simplicity and speed. The K-Modes is an extension of the K-Means algorithm for categorical data. Since K-Modes is used for categorical data so 'Simple Matching Dissimilarity' measure is used instead of Euclidean distance and the 'Modes' of clusters are used instead of 'Means'. However, one major limitation of this algorithm is dependency on prior input of number of clusters K, and sometimes it becomes practically impossible to correctly estimate the optimum number of clusters in advance. In this paper we have proposed an algorithm which will overcome this limitation while maintaining the simplicity of K-Modes algorithm.

Index Terms—Clustering, K-Modes clustering, Dependency, Prior input, Number of clusters.

I. INTRODUCTION

Clustering is the method of segregating the objects into groups such that the objects in a certain cluster have high degree of similarity with each other than the objects in the other clusters. Clustering process gained boom with the advent of various algorithms for numerical data. K-Means is one of the most commonly used partition based algorithm for numerical data. For categorical data however, various other algorithms are available. K-Modes which is an extension of K Means is one of such popular algorithms.

K-Modes algorithm is an extension of K-Means but with the following differences. First is that a Simple matching dissimilarity function suitable for categorical data is used instead of Euclidean distance. Secondly, Modes are used to represent centroids instead of Mean values and finally a frequency based method is used to find centroids in each iteration of the algorithm.

Also, as K-Modes is an extension of K-Means algorithm, the limitations of K-Means are also thus carried forward to K-Modes. One of them is the dependency on initial centroids to improve the accuracy of the clusters. A lot of research has been done, and some is still underway to address this.

Further, another major limitation applicable to both K-Means and K-Modes is to input the number of clusters at the very beginning based on anticipation. This sometimes may be off the mark, and hence requires an in-depth prior knowledge of the domain, and good prior idea of expected results. This restricts the algorithm to be used by experienced domain experts only. In this paper, attempt has been made to remove this need of prior anticipation and input of number of clusters, and still achieving good results with logical and optimum clusters generated automatically.

Work is also in progress to remove the limitation of providing to number of clusters as an input for mixed data sets and to find a Unified Metric for that works for both numerical and categorical data. Some work in these areas has been discussed in Literature Survey section. We are also working on an algorithm to remove the constraint of providing K as an input for mixed datasets using the metric as proposed in [1].

The paper is organized as follows: Section II covers Literature Survey, Section III introduces the proposed algorithm, Section IV gives the illustrative example of the algorithm using a small dataset, Section Analyses the results on real datasets and finally Section VI concludes the paper.

II. LITERATURE SURVEY

The Literature Survey is divided into three sections. The first section discusses the existing ways to select initial centroids to improve the accuracy of the clusters in K-Modes algorithm, second section discusses the ways to find an appropriate dissimilarity measure for the data- set containing both numerical and categorical data. The third section discusses the work already been done to remove the dependency on specifying the number of clusters, K in the algorithms dealing with categorical attributes.

A. Selection of Initial Centroids in K-Modes Algorithm

In the literature, many possibilities are provided to improve the accuracy of the clusters by improving the selection of initial centroids. Some of the work is discussed below:

Huang (1997) was the first one to suggest K-Modes, as an extension of the K-Means paradigm to categorical domains. He introduced new dissimilarity measures to deal with categorical objects, use of modes instead of means, and a frequency based method to be used to update modes in the clustering.

Again in next year (Huang, 1998) proposed the K-Prototype algorithm, which integrates the dissimilarity measure in the K-Means and K-Modes algorithms for clustering objects having mixed numeric and categorical attributes.

The drawback of these algorithms is that the relative frequencies of attribute values is not taken into account in a cluster centroid. This results in allocation of less similar objects in a cluster.

Sun, Zhu, Chen (2002) introduced an initialization method based on Bradley's iterative initial-point refinement algorithm (Bradley &Fayyad, 1998) to the K-Modes clustering,

The limitation of the algorithm is that many parameters have to be asserted in advance. Also high time cost.

Barbar ́a, Couto, Li (2002) suggested COOLCAT algorithm to deal with clustering of data streams. This algorithm is based on notion of entropy.

The limitation is the dependency on an input parameter m that represents the size of the smallest cluster.

San, Huynh, Nakamori (2004) proposed a new notion of cluster centers and dissimilarity measure. The experimental results showed betterness of the algorithm over the K-Modes algorithm.

The limitation of this algorithm is that its results have been compared with the K-Modes algorithm proposed by Huang (1998) but not with the other algorithms.

Cao, Liang, Bai (2009) proposed a initialization method for categorical data. The distance between objects was calculated based on the frequency of attribute values. The density of an object was defined as the total distance between an object and all objects from data. Due to high time cost of calculating the densities of all objects, the process was limited to a sub-sample dataset.

The drawback lies in the fact that since the sub-sample is selected randomly, so the single clustering result cannot be guaranteed.

Khan, Ahmad (2013)introduced an algorithm in which some objects which are very similar to each other have same cluster membership irrespective of the choice of initial cluster centers. Also it was based on the idea that some of the attributes inthe dataset whose number of attribute values are less than or equal toK have higher discriminatory power and were called Prominent attributes. These prominent attributes were used to generate clusters.

The limitation of this algorithm is that the accuracy of the clusters produced is not better than some of the other proposed algorithms.

B. Dissimilarity Metric

Also ample amount of work is carried out in developing dissimilarity measure to deal with the datasets containing categorical data and mixed data. Some of the work is discussed below:

He, Deng, Xu (2005) proposed a dissimilarity measure based on the similarity between a data object and cluster mode. The authors observed that this similarity is directly proportional to the sum of relative frequencies of the common values in the mode.

The proposed algorithm carried forward the same weakness as in K-Modes of choosing the initial modes randomly.

Ahmad, Dey (2007) proposed a dissimilarity measure based on the distance between two attribute values of same attribute. The authors found out the fact that similarity of two attribute values is dependent on their relationship with other attributes.

Since in this method distance is calculated for each attribute, it is not suitable for noisy and high dimensional datasets.

Also (Ahmad, Dey, 2007) proposed a cost function based on the one proposed by Huang. The proposed cost function added weight wt for numeric attributes computed from the given dataset. In the process, all numeric attributes were normalized and discretized to do the calculations.

The results obtained by this algorithm can be improved further by improving discretizing methods for numeric valued attributes.

Ng, Li, Huang, He (2007) proposed a new dissimilarity measure in which the modes of clusters were updated in each iteration. This method utilized some theorems proposed by authors themselves.

The algorithm with the new similarity measure requires more computational time than the original K-Modes algorithm.

Lee, Lee, Park (2009) suggested a new measure called DVD (Domain Value Dissimilarity). The information about distribution of data correlated to each categorical value was used to define the dissimilarity measure.

The drawback lies in the fact that this algorithm requires more computation time and also the memory as compared to original K-Modes algorithm.

Ienco, Pensa, Xu (2011) proposedamethodcalled DILCA (Distance Learning for Categorical Attributes). The distance between two values of a categorical attribute Ai was determined by the way in which the value of the other attributes Ajwere distributed in the dataset.

The drawback lies in the fact that the performance of this algorithm depends on some input parameter.

Desai, Singh, Pudi (2011) suggested the method DISC (Data-Intensive Similarity Measure for Categorical Data). This measure didn't require any domain knowledge to understand the data- set.

The algorithm is iterative and it requires feedback from a classifier for more accurate results.

Cao, Liang, Li, Bai, Dang (2012) suggested a new dissimilarity measure based on theidea of biological and genetic taxonomy and rough membership function. The accuracy of the clusters produced is more that the K-Modes with Ng's dissimilarity measure and Huang's dissimilarity measure.

The runtime of K-Modes with this measure is more than that of the K-Modes with Huang's measure.

C. Removing the Dependency on Inputting K

In this section we will be discussing two of the research papers dealing with this limitation of inputting the value of K.

San, Huynh (2004) proposed an algorithm which used a regularization parameter to control the number of clusters inthe clustering process. A suitable value of regularization parameter was chosen to find the most stable clustering results.

The major limitation of the proposed algorithm is that an input parameter representing the initial cluster centers is required.

Cheung, Jia (2013) presented penalized competitive learning algorithm that required some initial value of k which should not be less than the original value of k.

The drawback lies in the fact that it requires some initial value of k which should not be less than the original value of k. The resulting clusters are more accurate than the original K-Modes and K-Modes with Ng's dissimilarity metric (Ng's k-modes). But this algorithm too has much computation involved in it, as in the clustering algorithm proposed in [12].

III. Proposed Algorithm

As discussed in section 2.3, the algorithms proposed to overcome the limitation of inputting the value of K for categorical data require either a range of K, or some other input parameter which indirectly computes the value of K.In the proposed algorithm we have extended K-Modes algorithm where no initial direct or indirect input in needed.

Modified K-Modes algorithm without inputting the value of K

In the algorithm, the significance of attributes proposed in [1] is utilized. The significance of the attributes is calculated to create partitions initially and is not utilized in clustering. Significance of an attribute defines the importance of that attribute in the dataset. Those attributes, which display a good separation of co-occurrence of values into different groups, play a more significant role in clustering of data elements. In other words, an attribute plays a significant role in clustering, provided any pair of its attribute values are well separated against all attributes i.e. have an overall high value of d(x,y), for all pairs of x and y [1].To find centroids, the frequency based method proposed in [15] is used. The distance between an object and the centroid is calculated using the distance measure proposed in [19].

Input: dataset of n objects with m categorical attributes

Output: clusters or groups distributing the objects in the given dataset

Steps:

1. Find the most significant attribute from the given m attributes using the way proposed in [1].
2. Create initial clusters with attribute values with maximum distance in most significant attribute in different clusters.
3. Find the centroids of the clusters using the algorithm as proposed in [15] as stated below:

   a. Calculate the frequencies of all categories for all attributes and store them in a category array in descending order of frequency for every cluster. Here, $c_{i,j}$ denotes category iof attribute j and f $(c_{i,j})\geq$ $f(c_{i+1,j})$ where $f(c_{i,j})$ is the frequency of category $c_{i,j}$.
   b. Assign the most frequent categories in the two category arrays as centroids.

4. Find the distance of every object from its centroid in all the clusters as proposed in [19]. This distance is calculated as:

Distance(x,q) is the distance between object x and centroid q and is calculated as

$$d(x,q) = \sum_{j=1}^{m}(1 - f_{x_j}) \quad (1)$$

where $x=(x_1,x_2,\ldots\ldots x_m)$, $q=(q_1,q_2,\ldots..q_m)$ and $f_{xj}$is the relative frequency of category x_j within c.

The minimum of these distance values (other than 0) is represented as d.

5. Find the outliers in the initial clusters according to the objective function as stated below:

If Distance(x, q) <=d then x not an outlier.

6. Calculate the new centroids of the clusters as discussed in step 3.
7. Find the number of attributes in which records have matching values in all the clusters. Maximum of these values is represented as n. Take out all the outliers in all of the clusters having values matching in n-1 attributes.
8. If outliers with values matching in n-1 attributes at different positions then

   a. Put those outliers back in cluster whose distance from the centroid is less as compared to others.
   b. Else put all the outliers with values matching in n-1 attributes back in the corresponding clusters.

9. Calculate the distance of every outlier from the new cluster centroids and find the outliers not satisfying the objective function in step 5.
10. Let B= $\{Y_1,Y_2,\ldots..Yp)$ be the set of outliers obtained in step 9.

11. Repeat until (B==Φ)

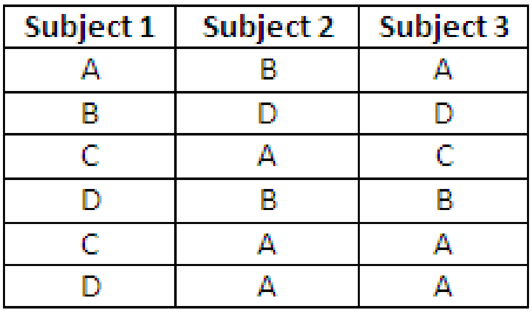
   a. Assume the set B as a new dataset.
   b. Perform steps 1 to 10

IV. Illustrative Example

The proposed algorithm is explained using a small dataset with categorical attributes representing the academic record of students in three subjects.

The dataset contains the academic details of the students in terms of Grades A, B, C and D in three subjects. The proposed Clustering algorithm will group the students based on their academic performance.

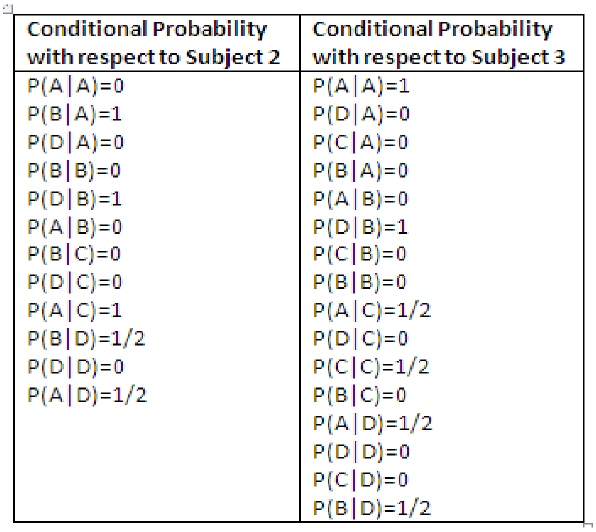
Table 1. Academic Record of Students.



Subject 1	Subject 2	Subject 3
A	B	A
B	D	D
C	A	C
D	B	B
C	A	A
D	A	A

1. In order to find the most significant attribute, calculate the conditional probabilities and the distance between various values of the attributes as given in Table 2 and Table 3.

Table 2.Probability Table for Subject1.

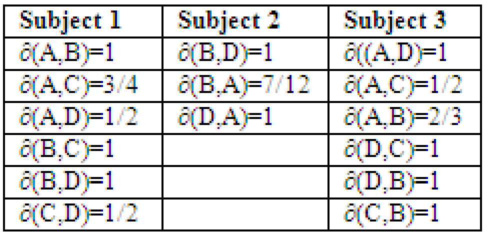


Conditional Probability with respect to Subject 2	Conditional Probability with respect to Subject 3
P(A|A)=0	P(A|A)=1
P(B|A)=1	P(D|A)=0
P(D|A)=0	P(C|A)=0
P(B|B)=0	P(B|A)=0
P(D|B)=1	P(A|B)=0
P(A|B)=0	P(D|B)=1
P(B|C)=0	P(C|B)=0
P(D|C)=0	P(B|B)=0
P(A|C)=1	P(A|C)=1/2
P(B|D)=1/2	P(D|C)=0
P(D|D)=0	P(C|C)=1/2
P(A|D)=1/2	P(B|C)=0
	P(A|D)=1/2
	P(D|D)=0
	P(C|D)=0
	P(B|D)=1/2

Similarly the conditional probabilities of Subject2 and Subject3 can be calculated.

The distance between various values of the attributes is calculated using these conditional probabilities as shown in Table 3.

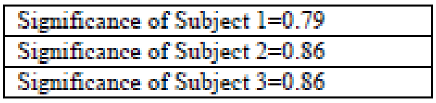
Table 3. Distance between Various Values of the Attributes of Table 1.



Subject 1	Subject 2	Subject 3
∂(A,B)=1	∂(B,D)=1	∂((A,D)=1
∂(A,C)=3/4	∂(B,A)=7/12	∂(A,C)=1/2
∂(A,D)=1/2	∂(D,A)=1	∂(A,B)=2/3
∂(B,C)=1		∂(D,C)=1
∂(B,D)=1		∂(D,B)=1
∂(C,D)=1/2		∂(C,B)=1

The significance of an attribute is calculated by taking the average of all the distance values as shown in Table 4.

Table 4.The Significance of the Attributes of Table 1.

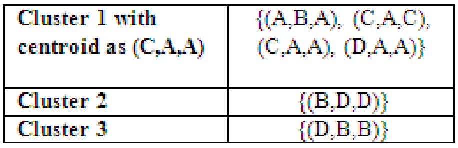


Significance of Subject 1=0.79
Significance of Subject 2=0.86
Significance of Subject 3=0.86

The attributes Subject2 and Subject3 have same significance value as shown in Table 4.So any attribute can be chosen as the most significant attribute. We have chosen Subject3 as most significant attribute. The maximum distance in attribute Subject3 is between values {A, D}, {D, C}, {D, B} and {C, B} which is 1. Create clusters for Table1 such that these set of values are not in the same cluster as shown in Table 5.

2. Three clusters are created initially such that the values {A, D}, {D, C}, {D, B} and {C, B} are not in the same cluster for attribute Subject3.
3. Initial clusters with their centroids are shown in table 5.

Table 5.Initial Clusters with their Centroids.



Cluster 1 with centroid as (C,A,A)	{(A,B,A), (C,A,C), (C,A,A), (D,A,A)}
Cluster 2	{(B,D,D)}
Cluster 3	{(D,B,B)}

4. In cluster1 the distance of the object {D, A, A) is minimum and is 0.5. Further cluster2 and cluster3 contain only on object so distance cannot be calculated. Therefore d=0.5.
5. Obtain the new cluster1 after removing the outliers according to the objective function mentioned in step5 of the algorithm. There will be no change in cluster2 and cluster3 as there is only one object in these clusters.

The outliers in Cluster1 are: {(A, B, A), (C, A, C)}

6. Find the new centroid of Cluster1 after removing the outliers.

Cluster1 after removing the outliers: {(C, A, A), (D, A, A)}.

New centroid of cluster1

Cluster 1: (C, A, A)

7. Here n=2, as cluster1 has records matching in 2 attributes. The outliers (A, B, A) and (C, A, C) can both be put back in cluster1 as (A, B, A) has matching value with other records in cluster1 at attribute subject3, (C, A, C) has matching value at attribute subject2.Cluster2 and cluster3 will remain same as they contain only one object.
8. Only (C, A, C) is put back in cluster1 as its distance from the centroid (C, A, A) is less as compared to (A, B, A).
9. New Cluster1:{(C,A,A), (D,A,A), (C,A,C)}

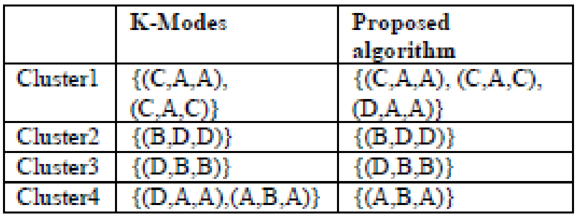
New Centroid of Cluster 1: (C, A, A)

Try to adjust the outlier (A, B, A) in Cluster1, Cluster2 and Cluster3 by finding the distance from their centroids. (A, B, A) does not satisfy the objective function mentioned in step 5 of the algorithm for any of the cluster.

10. As the outlier (A,B,A) does not satisfy the objective function for any of the clusters, so set B ={(A,B,A)}

Since only one record is left in set B, create a new cluster and put this record, otherwise the steps 1 to10 are repeated. Finally, four clusters are obtained. The clusters obtained are compared with the clusters obtained using K-Modes algorithm using software Rapid Miner as shown in Table 6.

Table 6.Results of K-Modes and Proposed algorithm for Academic Dataset.



	K-Modes	Proposed algorithm
Cluster1	{(C,A,A), (C,A,C)}	{(C,A,A), (C,A,C), (D,A,A)}
Cluster2	{(B,D,D)}	{(B,D,D)}
Cluster3	{(D,B,B)}	{(D,B,B)}
Cluster4	{(D,A,A),(A,B,A)}	{(A,B,A)}

The results in Table 6 show that both the algorithms have tried to categorize the dataset with objects having similarity in grades in one of the three available subjects, though the results are slightly different.

V. Results and Discussion

To explore the practical application of the proposed algorithm, the algorithm is tested on a number of real world datasets of different sizes and dimensions namely Car Evaluation and Credit Approval.

(http://archive.ics.uci.edu/ml/).

We have compared the results of the proposed algorithm with results of original K-Modes, obtained using Rapid Miner (http://rapidminer.com/).

Initial value of K was supplied as predefined value for the known data.

Experiment on Car Evaluation data

Car Evaluation dataset consists of 6 categorical attributes and 1728 instances. In this dataset there are four pre-defined classes, approximately 97% of the instances fall into 2 classes and rest 7% fall in remaining 2 classes.

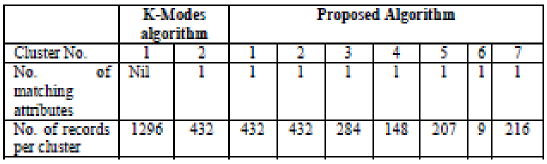
The results generated by the K-Modes algorithm in Rapid Miner and the proposed algorithm are shown in Table 7.

In results generated by the K-Modes algorithm, Cluster1 records are not matching in even one of the attributes.Cluster2 records are having similarity in one attribute named a1 with value ”high”.

The proposed algorithm generates 6 clusters on this dataset.

Cluster1 has records matching in one attribute named a1 with value “vhigh”. Cluster2 has records matching in attribute a4(value “2”).Cluster3 has records matching in attribute a1 with value “high”, Cluster4 has records matching in attribute a2 (value “high”), Cluster5 with matching value “med” for attribute a1, cluster6 with matching attribute as a1 with value “med” and a4 with value “4”, Cluster7 with matching value “low” for attribute a1. These results are summarized below:

Table 7.Results Generated by the K-Modes and the Proposed Algorithm for Car Evaluation Data.



	K-Modes algorithm		Proposed Algorithm						
Cluster No.	1	2	1	2	3	4	5	6	7
No. of matching attributes	Nil	1	1	1	1	1	1	1	1
No. of records per cluster	1296	432	432	432	284	148	207	9	216

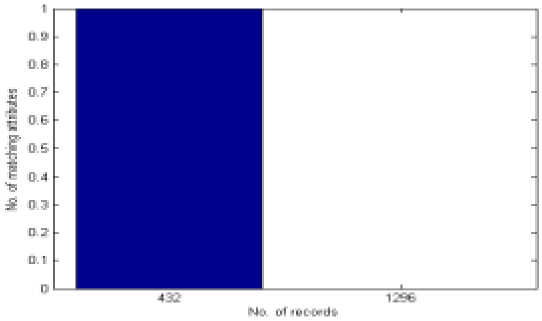


Fig.1. No. of Matching Attributes in Each Cluster using K-Modes.

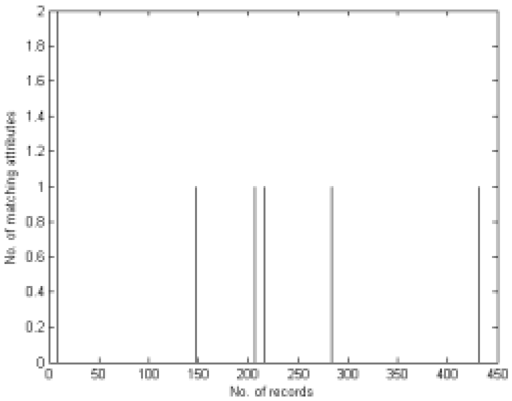


Fig.2. No. of Matching Attributes in Each Cluster using Proposed Algorithm.

As can be seen in Fig. 1And Fig. 2, all the clusters generated using proposed algorithm are similar in one attribute value whereas in original K Modes one of the two clusters generated contains records which have no similarity.

Experiment on Credit Approval data

This dataset has originally 15 mixed attributes (both categorical and numerical) and 690 instances out of which 37 instances contain missing values. In our experiment, we have removed 6 numerical attributes and instances with missing values. The resulting dataset contains only 655 instances. There are twopre defined classes. The results are described in Table 8.

Table 8. Results Generated by the K-Modes and the Proposed Algorithm for Credit Approval Data.

	K-Modes algorithm		Proposed Algorithm			
Cluster No.	1	2	1	2	3	4
No. of matching attributes	2	2	2	3	6	5
No. of records per cluster	507	158	498	156	7	4

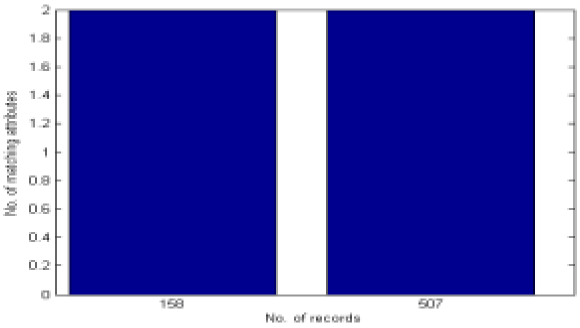


Fig.3. No. of Matching Attributes in Each Cluster using K-Modes.

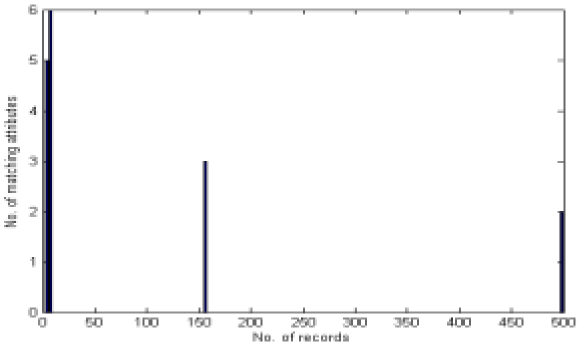


Fig.4. No of Matching Attributes in Each Cluster using Proposed Algorithm.

As depicted from Table 8, the clusters generated by K-Modes algorithm for K=2 have records matching at the most in two attributes. In the proposed algorithm, 4 clusters are generated with cluster 3 having records that match in 6 out of 9 attributes and represents the records that are most similar in the entire data- set. This is shown in Fig .3and Fig. 4 also. In Fig. 4, the first two dark lines represent number of matching records in cluster3 and cluster4.

VI. Conclusion and Future Work

In this paper we have proposed an algorithm based on K-Modes algorithm which creates appropriate number of clusters without need of prior input of number of clusters, K. The results obtained are found to be better than the original K-Modes in terms of the quality of clusters. The proposed algorithm has generated clusters such that the similarity of objects in a cluster in terms of number of attributes with matching values is maximum. This way we can obtain a cluster of objects form the given dataset with maximum similarity without providing the value of K initially. Though the time taken by the proposed algorithm is more as compared to the original K-Modes.

In Future Work, attempt will be made to reduce the computation time for clusters generation, especially with a large number of attributes. Also an algorithm for mixed datasets with no dependency on K as an input will be proposed.

References

[1] Ahmad, A., Dey, L. A K-Mean Clustering Algorithm for Mixed Numeric and Categorical Data. Data & Knowledge Engineering, 2007, 63: 503–527.

[2] Ahmad, A., Dey, L. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. Pattern Recognition Letters, 2007, 28 (1): 110–118.

[3] Bai, L., Liang, J., Dang, C., Cao, F. A cluster centers initialization method for clustering categorical data. Expert Systems with Applications, 2012, 39: 8022-8029.

[4] Barbar ́a, D., Couto, J., Li, Y). COOLCAT: An entropy-based algorithm for categorical clustering. CIKM '02 Proceedings of the eleventh international conference on Information and knowledge management: 582-589.

[5] Basak, J., De, R., K., Pal, S., K. Unsupervised feature selection using a neuro-fuzzy approach. Pattern Recognition Letters, 1998, 19: 997–1006.

[6] Bradley, P., S., Fayyad, U., M. Refining initial points for k-means clustering. Proceedings of 15th international conference on machine learning (ICML98), 1998: 91–99.

[7] Cao F., Liang, J., Bai, L. A new initialization method for categorical data clustering. Expert Systems with Applications, 2009, 36: 10223-10228.

[8] Cao, F., Liang J., Li D., Bai, L., Dang, C. A dissimilarity measure for the k-Modes clustering algorithm. Knowledge-Based Systems, 2012, 26: 120–127.

[9] Cheung, Y., Jia, H. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. Pattern Recognition, 2013, 46: 2228–2238.

[10] Desai, A., Singh, H., Pudi, V. DISC: Data Intensive Similarity Measure for Categorical Data. Proceedings of Advances in Knowledge Discovery and Data Mining – 15thPacific Asia Conference, 2011, 6635: 469 – 481.

[11] Ienco, D., Pensa, R., G., Meo, R.From Context to Distance: Learning Dissimilarity for Categorical Data Clustering. ACM Transactions on Knowledge Discovery from Data, 2011, 0(0):1-22.

[12] H. Liao, M.K. Ng, “Categorical Data Clustering with Automatic Selection of Cluster Number”, Fuzzy Information and Engineering 1 (1), 2009: 5-25.

[13] He, Z., Deng, S., Xu, X. Improving K-Modes Algorithm Considering Frequencies of Attribute Values in Mode. Computational Intelligence and Security Lecture Notes in

Computer Science, 2005, 3801: 157-162.

[14] Huang, Z. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. In Proceeding SIGMOD workshop research issues on data mining and knowledge discovery, 1997: 1–8.

[15] Huang, Z. Extensions to the *k*-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery, 1998, 2: 283–304.

[16] Khan, S., S., Ahmad A. Cluster Center Initialization for Categorical Data Using Multiple Attribute Clustering. Expert Systems with Applications, 2013, 40(18): 7444–7456.

[17] Lee, J., Lee, Y., Park, M. Clustering with Domain Value Dissimilarity for Categorical Data, Advances in Data Mining. Applications and Theoretical Aspects, Lecture Notes in Computer Science, 2009, 5633: 310-324.

[18] Ng, M., K., Li, M., J., Huang, J., Z., He, Z. On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29 (3): 503-507.

[19] San, O., M., Huynh, V., Nakamori, V. An Alternative Extension of the *k*-Means Algorithm for Clustering Categorical Data. International Journal Appl. Math. Computer. Sci., 2004, 14(2): 241–247.

[20] Sun, Y., Zhu, Q., Chen, Z. An iterative initial-points refinement algorithm for categorical data clustering. Pattern Recognition Letters, 2002, 23: 875–884.

[21] Yeung, D., S., Wang, Y., S. Improving performance of similarity-based clustering by feature weight learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24 (4): 556–561.

[22] Http://archive.ics.uci.edu/ml/

[23] Http://rapidminer.com/

Authors' Profiles

Dr. Suresh Kumar is a Professor in Department of Computer Science and Engineering, Manav Rachna International University. He has 14 years of experience and his research interests include Big Data, Data Mining, Networking.

Ms. Anupama Chadha is a research scholar in Department of Computer Science and Engineering, Manav Rachna International University. Her area of interest include Data Mining, Software Engineering.

How to cite this paper: Anupama Chadha, Suresh Kumar,"Extension of K-Modes Algorithm for Generating Clusters Automatically", International Journal of Information Technology and Computer Science(IJITCS), Vol.8, No.3, pp.51-57, 2016. DOI: 10.5815/ijitcs.2016.03.06