Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation

Shehroz S Khan
National University of Ireland Galway,
Department of Information Technology,
Galway, Republic of Ireland
s.khan1@nuigalway.ie

Dr. Shri Kant
Scientific Analysis Group,
Defence R&D Organization,
Metcalfe House, Delhi, India-110054
shrikant@scientist.com

Abstract

Clustering accuracy of partitional clustering algorithm for categorical data primarily depends upon the choice of initial data points (modes) to instigate the clustering process. Traditionally initial modes are chosen randomly. As a consequence of that, the clustering results cannot be generated and repeated consistently. In this paper we present an approach to compute initial modes for *K*-mode clustering algorithm to cluster categorical data sets. Here, we utilize the idea of Evidence Accumulation for combining the results of multiple clusterings. Initially, n $F-$ dimensional data is decomposed into a large number of compact clusters; the *K*-modes algorithm performs this decomposition, with several clusterings obtained by N random initializations of the *K*-modes algorithm. The modes thus obtained from every run of random initializations are stored in a Mode-Pool, P_N. The objective is to investigate the contribution of those data objects/patterns that are less vulnerable to the choice of random selection of modes and to choose the most diverse set of modes from the available Mode-Pool that can be utilized as initial modes for the *K*-mode clustering algorithm. Experimentally we found that by this method we get initial modes that are very similar to the actual/desired modes and gives consistent and better clustering results with less variance of clustering error than the traditional method of choosing random modes.

1 Introduction

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters [Guha et al, 1998]. The most distinct characteristics of clustering operation in data mining is that the data sets often contain both numeric and categorical attribute values. This requires the clustering algorithm to be capable of dealing with the complexity of the inter- and intra-relation of the data sets expressed in different types of the attributes, no matter numeric or categorical [Michalski et al, 1998]. The *K*-means algorithm [Jain and Dubes, 1988] is one of the most popular clustering algorithms because of its efficiency in clustering large data sets [Anderberg, 1973]. However, *K*-means clustering algorithm fails to handle data sets with categorical attributes because it minimizes the cost function that is numerically measured.

K-means does not guarantee unique clustering because we get different results with randomly chosen initial cluster centers [Sing-Tze Bow, 2002] and hence the results cannot be relied with confidence. The *K*-means algorithm gives better results only when the initial partitions are close to the final solution [Jain and Dubes, 1988]. Several attempts have been reported to generate *K*-prototype points that can be used as initial cluster centers. A recursive method for initializing the means by running *K* clustering problems is discussed by Duda and Hart [1973]. Bradley et al [1997] reported that the values of initial means along any one of the *m* coordinate axes are determined by selecting the *K* densest "bins" along that coordinate. Bradley and Fayyad [1998] proposed a procedure that refines the initial point to a point likely to be close to the modes of the joint probability density of the data. Mitra et al [2002] suggested a method to extract prototype points based on Density Based Multiscale Data Condensation. Khan and Ahmad [2004] presented an algorithm to compute initial cluster centers for *K*-means clustering algorithm. Their algorithm is based on two experimental observations that some of the patterns are very similar to each other and that is why they have same cluster membership irrespective to the choice of initial cluster centers. Also, an individual attribute may provide some information about initial cluster center. The initial cluster centers computed by using their methodology are found to be

very close to the desired cluster centers with improved and consistent clustering results.

Various clustering algorithms have been reported to cluster categorical data. Ralambondrainy [1995] presented an approach by using *K*-means algorithm to cluster categorical data. The approach is to convert multiple category attributes into binary attributes (using 0 and 1 to represent either a category absent or present) and treat the binary attributes as numeric in the *K*-means algorithm. Gower and Diday [1991] used a similarity coefficient and other dissimilarity measures to process data with categorical attributes. The *K*-mode clustering algorithm [Huang, 1997] extends the *K*-means paradigm to cluster categorical data by using a simple matching dissimilarity measure (hamming distance) for categorical objects and modes instead of *means* for clusters.

Most of the above mentioned algorithms for clustering categorical data require a random selection of initial data points in addition to apriori knowledge of number of clusters (*K*). This leads to the problem that clustering results are dependent on the selection of initial modes. Choosing different initial modes lead to different cluster structures and hence the clustering results cannot be repetitively generated. Furthermore, inappropriate choice of initial modes leads to undesirable clustering results. Machine learning practitioners find it difficult to count on such clustering results.

Zhexue Huang [1998] presented two methods of initialization for categorical data for *K*-mode clustering algorithm and showed that if diverse initial modes are chosen then it could lead to better clustering results. Sun et al [2002] proposed an iterative method based on initial points refinements algorithm for categorical data clustering to the setting up of the initial points so as to map the categorical data sets to clustering results that have better consistency rates. They applied Bradley and Fayyad's iterative initial point refinement algorithm [Bradley and Fayyad, 1998] to the *K*-modes clustering to improve the accuracy and repetition of clustering results. They used sub-sampling method to carry the clustering iteratively several times so that effect of skewed distributed data should not affect the final clustering results. Khan S.S. et al [Khan and Ahmad, 2003] presented an algorithm to compute initial modes using Density based Multiscale Data Condensation. They showed that by choosing initial modes this way consistent and efficient clustering results were achieved.

Kant et al [1994] presented an Automatic and Stable Clustering Algorithm for clustering numerical data. They showed stable clustering of data by repetitively clustering the same data with random initializations to generate stable cluster regions such that each pattern fits exactly into one of those regions and no single pattern can be fitted in two clusters regions. More recently, [Fred and Jain, 2002a] used the idea of evidence clustering to combine the results of multiple clusterings (*N* times) into a single data partition, by viewing each clustering result as an independent evidence of data organization. They did so by running a *K*-means algorithm many times with different parameters or initializations. First, the data is split into a large number of compact and small clusters; different decompositions are obtained by random initializations of the *K*-means algorithm. The final data partition is obtained by clustering this new similarity matrix, corresponding to the merging of cluster [Fred, 2001]. Topchy et al [2003] presented an algorithm to combine multiple weak clusterings and formulated that combined clustering becomes equivalent to clustering a categorical data based on some chosen consensus function. They showed efficacy of combining partitions generated by weak clustering algorithms that uses random data splits. All of this research work is based on numerical data. In this paper, we extend the idea of Evidence Accumulation to categorical data sets by generating multiple partitions as different data organization by seeding *K*-modes algorithm, every time, with random initial modes. The resultant modes are then stored in a Mode Pool and the most diverse set of modes were computed, which were used as initial modes.

The rest of the paper is organized as follows. Section 2 briefly discusses the *K*-modes algorithm. Section 3 describes the proposed approach in computing the initial modes of the data sets using Evidence Accumulation. Section 4 presents the experimental results on applying the proposed approach to compute initial modes for different categorical data sets [UCI data repository] and demonstrates improved and consistent clustering results. Section 5 concludes the presentation.

2 The *K*-modes algorithm for clustering categorical data

The *K*-means clustering algorithm cannot cluster categorical data because of the dissimilarity measure it uses. The *K*-modes clustering algorithm is based on *K*-means paradigm but removes the numeric data limitation whilst preserving its efficiency. The *K*-modes algorithm extends *K*-means paradigm to cluster categorical data by removing the limitation imposed by *K*-means through following modifications:

- Using a simple matching dissimilarity measure or the hamming distance for categorical data objects
- Replacing *means* of clusters by their modes

The simple matching dissimilarity measure [Jain and Dubes, 1988] can be defined as following. Let X and Y be two categorical data objects described by F categorical attributes. The dissimilarity measure $d(X,Y)$ between X and Y can be defined by the total mismatches of the corresponding attribute categories of two objects. Smaller the number of mismatches, more similar the two objects are. Mathematically, we can say

$$d(X,Y)=\sum_{j=1}^{F}\delta(x_j,y_j) \qquad (2.1)$$

where $\delta(x_j,y_j)=\begin{cases}0 & (x_j=y_j)\\ 1 & (x_j\neq y_j)\end{cases}$

$d(X,Y)$ gives equal importance to each category of an attribute.

Let Z be a set of categorical data objects described by categorical attributes, $A_1, A_2, \ldots A_F$, a mode of $Z=\{Z_1,Z_2,\ldots Z_n\}$ is a vector $Q=[q_1,q_2,\ldots q_F]$ that minimizes

$$D(Z,Q)=\sum_{i=1}^{n}d(Z_i,Q) \qquad (2.2)$$

Here, Q is not necessarily an element of Z. When the above is used as the dissimilarity measure for categorical data objects, the cost function becomes

$$C(Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{F} \delta\left(z_{ij}, q_{lj}\right) \qquad (2.3)$$

where $Q_l = [q_{l1}, q_{l2}, \ldots q_{lm}] \in Q$

The *K*-modes algorithm minimizes the cost function defined in equation 2.3.

The *K*-modes algorithm consists of the following steps: -

a) Select *K* initial modes, one for each of the cluster.
b) Allocate data object to the cluster whose mode is nearest to it according to equation 2.1.
c) Compute new modes of all clusters.
d) Repeat step 2 to 3 until no data object has changed cluster membership.

3 Computing Initial Modes Using Evidence Accumulation

The idea of Evidence Accumulation clustering [Fred and Jain, 2005] is to combine the results of multiple clusterings into a single data partition, by viewing each clustering result as an independent evidence of data organization. Fred and Jain [2002] used the *K*-means algorithm as the basic algorithm for decomposing the data into a large number of compact clusters; evidence on pattern association is accumulated, by a voting mechanism, over *N* clusterings obtained by random initializations of the *K*-means algorithm. There are several possible ways to accumulate evidence in the context of unsupervised learning: (a) combine results of different clustering algorithms; (b) produce different results by re-sampling the data, such as in bootstrapping techniques (like bagging) and boosting; (c) running a given algorithm many times with different parameters or initializations. In this paper we take the last approach, using *K*-modes algorithm as the underlying clustering for creating multiple partitions of the categorical data.

Khan and Ahmad [2004] presented that in a data set there are some data objects that do not change class membership irrespective of the choice of initial point. In other words they belong to same clusters irrespective of choice of initialization. For example, let $D_i = \{D_{i1}, D_{i2} \ldots D_{iF}\}$ be a dataset consisting of *n* data objects with *F* attributes. Let us assume that data objects D_{k1}, D_{k2}, D_{k3}, where $1 \leq$ k1, k2, k3 $\leq n$ are very similar, then they have same cluster membership whenever *K*-modes algorithm is executed with different initial modes. This information is quite useful to compute initial modes. The two major steps of our algorithm are

(a) Generate *N* independent evidences of data organizations by performing *K*-modes clustering using random initialization of modes and store the resultant modes of each of the *N* iteration in a Mode-Pool, P_N.
(b) Find the most diverse modes for each cluster to be used as initial modes

3.1 Generating Independent Data Organization

We assume that the choice of numbers of clusters (*K*) is the same as the number of natural groupings present in the data set. The algorithmic steps are:

1. Set *K*→Number of clusters present in the data set, *N*→Number of clusterings
2. Choose a value of number of clusters (*K*), *i*=1.
3. While ($i \leq N$) do the following
   (a) Choose random initial modes and execute *K*-modes algorithm; till it converges and create *K* partitions.
   (b) Store the *K* modes thus obtained (from each of the *K* partitions) in a Mode-Pool, P_i
   (c) Increment *i*.

3.2 Extracting Initial Modes from Mode-Pool

After the execution of algorithm discussed in 3.1, we are left with a Mode-Pool, P_N with *N* $K \times F$ modes (*F* is the number of attributes of the data set). To extract the most diverse modes, employ this following consensus algorithm

1. Set *i*=1, *j*=1, *k*=1
2. While ($i \leq K$) do the following
3. While ($j \leq F$) do the following
4. While ($k \leq N$) do the following
5. Extract the most frequent mode and store it in the Initial Modes Matrix, $I_{i \times j}$
6. Increment *k*
7. Increment *j*
8. Increment *i*

The modes generated by each *N* clusterings are mostly representative of those data objects/patterns that are less vulnerable to change cluster membership irrespective of the choice of random initial mode selection. After extracting the frequent modes, $I_{K \times F}$, from the Mode-Pool, P_N, we shall have captured representations mostly from those patterns only. The modes thus obtained for each of the *K* partitions should be quite dissimilar from each other with more diversity embodied in them.

4 Experimental Results

To test our approach we use the following categorical data sets obtained from UCI Machine Learning Data Repository [UCI data repository]

(1) Michalski soybean disease data set [Michalski and Stepp, 1983]

The soybean disease data set consists of 47 cases of soybean disease each characterized by 35 multi-valued categorical variables. These cases are drawn from four populations, each one of them representing one of the following soybean diseases: D1-Diaporthe stem canker, D2-Charcoat rot, D3-Rhizoctonia root rot and D4-Phytophthorat rot. Ideally, a clustering algorithm should partition these given cases into four groups (clusters) corresponding to the diseases.

(2) Wisconsin Breast Cancer Data

This data has 699 instances with 9 attributes. Each data object is labeled as *benign* (458 or 65.5%) or *malignant* (241 or 34.5%). In our literature, all attributes are considered categorical with values 1, 2… 10. There are 16 instances in Groups 1 to 6 that contain a single missing (i.e. unavailable) attribute value, denoted by "?". For data symmetry we took 241 benign case and 241 malignant cases for out analysis.

(3) Zoo small data

It has 101 instances distributed into 7 categories. This data is same as Zoo data but with only the important eight attributes (feathers, milk, airborne, predator, backbone, fins, leg and tail). All of these characteristics attributes are Boolean except for the character attribute corresponds to the number of legs that lies in the set {0, 2, 4, 5, 6, 8}

(4) Congressional Vote Data

This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to YES), voted against, paired against, and announced against (these three simplified to NO), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition). All attributes are Boolean with Yes (denoted as y) and No (denoted as n) values. A classification label of Republican or Democrat is provided with each record. The dataset contains 435 records with 168 Republicans and 267 Democrats

In the presence of true labels, as in the case of the data sets we used, the clustering accuracy for measuring the clustering results was computed as follows. Given the final number of clusters, K, clustering accuracy r was defined as:

$$r = \frac{\sum_{i=1}^{K} a_i}{n}$$

where n is the number of patterns in the dataset, a_i is the number of data objects occurring in both cluster i and its corresponding class, which had the maximal value. In other words, a_i is the number of records with the class label that dominates cluster i. Consequently, the clustering error is defined as

$$e = 1 - r$$

Low value of e suggests better clustering.

To conduct experimental comparison and to verify the efficacy of our proposed method, we supplied initial random modes to the K-modes algorithm as suggested by huang [1997]. Table 1 compiles the clustering results on the categorical data sets (described above) using random initial modes and the modes supplied by our proposed approach using Evidence Accumulation. Results presented for our approach are based on combination of N=100 K-modes clusterings, a considerable high value to ensure that convergence of the method is ensured. The reported clustering error and standard deviation is average of 50 executions of the whole process. It can be seen that the clustering results have improved with less standard deviation in error when the modes were chosen by our proposed method in comparison to the random selection of initial modes.

Figure 1 and 2 represents these results graphically. Figure 1 show the clustering error and its standard deviation when initial modes were randomly chosen. Figure 2 shows the same statistics when initial modes were picked up using our proposed method based on Evidence Accumulation and were fed to the K-mode clustering algorithm. A reduced clustering error with less variance can be seen from figure 2.

One important observation was that the initial modes computed by our proposed approach were quite similar to the actual/desired modes for these data sets and therefore better clustering and fast convergence was achieved. And since K-mode is executed large number of times (N=100), the weak clustering results were eliminated and most of the time we get representations from those patterns that are less susceptible to random selection of modes and therefore we get repetitive initial modes that leads to consistent clustering results with less variance in clustering error.

Data Set	Random Initialization of modes		Proposed method of Initialization using Evidence Accumulation	
	Avg. Clustering Error	Standard Deviation	Avg. Clustering Error	Standard Deviation
Soybean	0.055	1.89	0.021	1.02
Wisconsin Breast Cancer	0.155	1.77	0.132	0.44
Zoo small	0.162	0.966	0.166	0.54
Congressional Vote	0.141	4.52	0.132	0.707

Table 1. Clustering Error and Standard Deviation comparison using random initialization of modes and modes supplied using the proposed approach

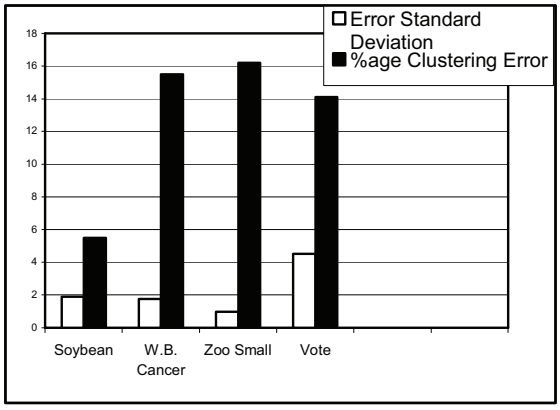

Figure 1. Graphical Representation of Clustering Error and Standard Deviation using Random Selection of Modes

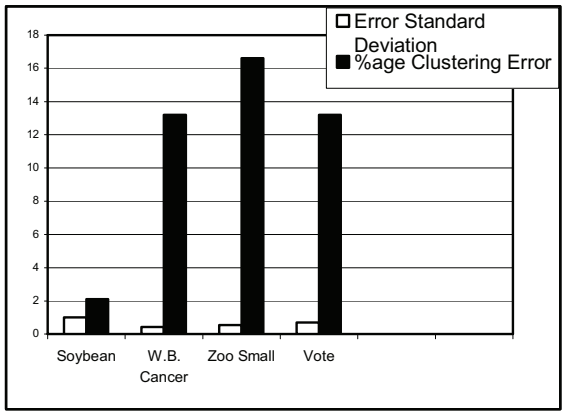

Figure 2. Graphical Representation of Clustering Error and Standard Deviation using Proposed Approach for Initial Mode Computation

5 Conclusions

K-modes algorithm suffers from the drawback of choosing random initial modes which may lead to formation of non-repetitive clustering structures that are undesirable for analysis. In this paper, we have presented an approach to compute the initial modes for *K*-modes clustering algorithm for clustering categorical data using Evidence Accumulation. The procedure is motivated by the observation that some data objects do not change their class membership even when subjected to different random initial conditions (modes). We utilized the idea of Evidence Accumulation for combining the results of multiple *K*-mode clusterings. The resultant modes of each of these runs were stored in a Mode-Pool. The most diverse set of modes were extracted from the Mode Pool as the initial modes for the *K*-mode algorithm. The computed modes were majorly being representative of those patterns that are less susceptible to random selection of initial modes. Also, the modes computed using this method were found to be quite similar to the actual/desired modes of the datasets. Therefore, consistent clustering with fast convergence was achieved with less variance in clustering error.

Acknowledgements

The authors are thankful to Dr. P.K. Saxena, Director of SAG for his encouragement to pursue this research under his kind patronage. The authors are also highly indebted to Dr. Michael Madden, Lecturer, Department of Information Technology, NUI Galway for his support to publish this research work.

References

[Anderberg, M, 1973] Anderberg, M. *Cluster Analysis for Applications*, Academic Press, New York, 1973

[Bradley and Fayyad, 1998] Bradley, P.S, Fayyad, U.M. Refining Initial Points for K-Means Clustering, *Proceedings of the 15th International Conference on Machine Learning (ICML'98), San Francisco, Morgan Kaufmann*, 1998

[Bradley et al, 1997] Bradley, P.S., Mangasarian, O.L., Street, W.N. Clustering via Concave Minimization, in Advances in Neural Information Processing systems 9, MIT Press, 368-374, 1997.

[Duda and Hart, 1973]Duda, R.O., Hart, P.E., *Pattern classification and Scene Analysis*, John Wiley and Sons, N.Y. 1973.

[Fred and Jain, 2002]]Fred, A., Jain, A.K. Evidence Accumulation Clustering based on the K-means algorithm, in *Proceedings of the International Workshops on Structural and Syntactic Pattern Recognition (SSPR)*, Windsor, Canada, August 2002.

[Fred and Jain, 2002a] Fred, A., Jain, A.K.: Data Clustering Using Evidence Accumulation, in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Quebec City, August 2002.

[Fred and Jain, 2005] Fred, A., Jain, A.K. Combining Multiple Clustering Using Evidence Accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, number 6,835-850, 2005.

[Fred, 2001] Fred, A.L. Finding Consistent Clusters in Data Partitions. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, volume LNCS 2096, Springer, 309–318, 2001.

[Gower and Diday, 1991] Gower, J., Diday, E.: Symbolic Clustering Using a New Dissimilarity Measure, *Pattern Recognition Letters*,24(6), 567-578, 1991.

[Guha et al, 1998] Guha, S., Rastogi, R., Shim K.: CURE: An Efficient Clustering Algorithm for Large Databases, Published in the *Proceedings of the ACM SIGMOD Conference*, 1998.

[Huang, 1997] Huang, Z.: A Fast Clustering Algorithm to Cluster very Large Categorical Data Sets in Data Mining, *DMKD*, 1997.

[Jain and Dubes, 1988] Jain, A., Dubes, R.,: *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ , 1988.

[Kant et al, 1994] Kant, S., Rao, T.L., Sundaram, P.N., An Automatic and Stable Clustering Algorithm, *Pattern Recognition Letters*, vol. 15, Issue 6, 543-549, 1994.

[Khan and Ahmad, 2003] Khan, S.S., Ahmad, A.: Computing Initial points using Density Based Multiscale Data Condensation for Clustering Categorical data , *2nd International Conference on Applied Artificial Intelligence, ICAAI'03*, Kolhapur, India, 2003.

[Khan and Ahmad, 2004] Khan, S.S., Ahmad, A.: Cluster center initialization algorithm for K-means clustering. *Pattern Recognition Letters,* 25 (11), 1293-1302, 2004.

[Michalski and Stepp, 1983] Michalski, R., Stepp, R.: An automated construction of Classification: Conceptual clustering versus numerical taxonomy, *IEEE Trans. Pattern Anal. Machine Intelligence*. 5 (4), 396-410, 1983.

[Michalski et al, 1998] Michalski, R., Bratko, I., Kubat, M.: Machine *Learning and Data mining: Methods and Applications*. Wiley, New York, 1998.

[Mitra et al, 2002] Mitra, P., Murthy, C.A, Pal, S.K., Density Based MultiScale Data Condensation, *IEEE*

Trnasaction on Pattern Analysis and Machine Intelligence, Vol 24, no. 6, 734-747, 2002.

[Ralambondrainy, 1995] Ralambondrainy, H.: A conceptual version of the K-Means algorithm, *Pattern Recognition Letters*, 16., 1147-1157, 1995.

[Sing-Tze Bow, 2002]]Sing-Tze Bow.*: Pattern Recognition and Image Preprocessing*, Marcel Dekker, Inc , 2002.

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

[Topchy et al, 2003] Topchy, A., Jain, A.K., Punch, W., "Combining Multiple Weak Clusterings", in *Proceedings of the IEEE International Conf. Data Mining*, USA, 331-338, 2003.

[UCI data repository] *UCI data repository* http://www.sgi.com/tech/mlc/db/

[Zhexue Huang, 1998] Zhexue Huang : Extensions to the *K*-means Algorithm for Clustering Large Data Sets with Categorical Values, *Data Mining and Knowledge Discovery 2*, Netherlands , 283–304, 1998.