Cluster Center Initialization for Categorical Data Using Multiple Attribute Clustering

Shehroz S. Khan* Amir Ahmad†

Abstract

The K-modes clustering algorithm is well known for its efficiency in clustering large categorical datasets. The K-modes algorithm requires random selection of initial cluster centers (modes) as seed, which leads to the problem that the clustering results are often dependent on the choice of initial cluster centers and non-repeatable cluster structures may be obtained. In this paper, we propose an algorithm to compute fixed initial cluster centers for the K-modes clustering algorithm that exploits a multiple clustering approach that determines cluster structures from the attribute values of given attributes in a data. The algorithm is based on the experimental observations that some of the data objects do not change cluster membership irrespective of the choice of initial cluster centers and individual attributes may provide some information about the cluster structures. Most of the time, attributes with few attribute values play significant role in deciding cluster membership of individual data object. The proposed algorithm gives fixed initial cluster center (ensuring repeatable clustering results), their computation is independent of the order of presentation of the data and has log-linear worst case time complexity with respect to the data objects. We tested the proposed algorithm on various categorical datasets and compared it against random initialization and two other available methods and show that it performs better than the existing methods.

1 Introduction

Clustering aims at grouping multi-attribute data into homogenous clusters (groups). Clustering is an active research topic in pattern recognition, data mining, statistics and machine learning with diverse application such as in image analysis [19], medical applications [21] and web documentation [2].

The K-means [1] based partitional clustering methods are used for processing large numeric datasets for its simplicity and efficiency. Data mining applications require handling and exploration of heterogeneous data that contains numerical, categorical or both types of attributes together. K-means clustering algorithm fails to handle datasets with categorical attributes because it minimizes the cost function by calculating *means*. The traditional way to treat categorical attributes as numeric does not always produce meaningful results because generally categorical domains are not ordered. Several approaches have been reported for clustering categorical datasets that are based on K-means paradigm. Ralambondrainy [22] present an approach by using K-means algorithm to cluster categorical data by converting 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 [7] use a similarity coefficient and other dissimilarity measures to process data with categorical attributes. CLARA (Clustering LARge Application) [15] is a combination of a sampling procedure and the clustering program Partitioning Around Medoids (PAM). Guha et al. [8] present a robust hierarchical clustering algorithm, ROCK, that uses links to measure the similarity/proximity between a pair of data points with categorical attributes that are used to merge clusters. However this algorithm has worst case quadratic time complexity.

Huang [12] presents the K-modes clustering algorithm by introducing a new dissimilarity measure to cluster categorical data. The algorithm replaces *means* of clusters with *modes*, and use a frequency based method to update *modes* in the clustering process to minimize the cost function. The algorithm is shown to achieve convergance with linear time complexity with respect to the number of data objects. Huang [13] also pointed out that in general, the K-modes algorithm is faster than the K-means algorithm because it needs less iterations to converge.

In principle, K-modes clustering algorithm functions similar to K-means clustering algorithm except for the cost function it minimizes, and hence suffers from the same drawbacks. Likewise K-means, the K-modes clustering algorithm assumes that the number of clusters, K, is known in advance. Fixed number of K clus-

*University of Waterloo, Ontario, Canada.
†King Abdulaziz University, Rabigh, Saudi Arabia.

ters can make it difficult to predict the actual number of clusters in the data that may mislead the interpretations of the results. It also fall into problems when clusters are of differing sizes, density and non-globular shapes. K-means does not guarantee unique clustering due to random choice of initial cluster centers that may yield different groupings for different runs [14]. Similarly, K-modes algorithm is also very sensitive to the choice of initial centers, an improper choice may yield highly undesirable cluster structures. Random initialization is widely used as a seed for K-modes algorithm due to its simplicity, however, this may lead to undesirable and/or non-repeatable clustering results. Machine learning practioners find it difficult to rely on the results thus obtained and several re-runs of K-modes algorithm may be required to arrive at a meaningful conclusion.

There are several attempts to initialize cluster centers for K-modes algorithm, however, most of these methods suffer from either one or more of the three drawbacks: a) the initial cluster center computation methods are non-linear in time complexity with respect to the number of data objects b) the initial modes are not fixed and possess some kind of randomness in the computation steps and c) the methods are dependent on the presentation of order of data objects (details are discussed in Section 2). In this paper, we present a multiple clustering approach that infers cluster structure information from the attributes using their attribute values present in the data for computing initial cluster centers. Our proposed algorithm performs mulitple partitional clustering on different attributes of the data to generate fixed initial centers (modes), is independent of the order of presentation of data and thus gives fixed clustering results. The proposed algorithm has worst case log-linear time complexity with respect to the number of data objects.

The rest of the paper is organized as follows. In Section 2 we review research work on cluster center initialization for K-modes algorithm. Section 3 briefly discusses the K-modes clustering algorithm. In Section 4 we present the proposed approach to compute initial modes using multiple clustering that takes contributions from different attribute values of individual attributes to determine distinguishable clusters in the data. In Section 5, we present the experimental analysis of the proposed method on various categorical datasets to compute initial cluster centers, compare it with other methods and show improved and consistent clustering results. Section 6 concludes the presentation with pointers to future work.

2 Related Work

The K-modes algorithm [12] extends the K-means paradigm to cluster categorical data and requires random selection of initial center or modes. The random initialization of cluster center may only work well when one or more chosen initial centers are close to actual centers present in the data. In the most trivial case, the K-modes algorithm keeps no control over the choice of initial centers and therefore repeatability of clustering results is difficult to achieve. Moreover, inappropriate choice of initial cluster centers can lead to undesirable clustering results. Hence, it is desirable to start K-modes clustering with fixed initial centers that resemble the true representatives centers of the clusters. Below we provide a short review of the research work done to compute initial cluster centers for K-modes clustering algorithm and discuss their associated problems.

Huang [13] propose two approaches for initializing the clusters for K-modes algorithm. In the first method, the first K distinct data objects are chosen as initial K-modes, whereas the second method calculates the frequencies of all categories for all attributes and assign the most frequent categories equally to the initial K-modes. The first method may only work if the top K data objects come from disjoint K clusters, therefore it is dependent on order of presentation of data. The second method is aimed at choosing diverse cluster center that may improve clustering results, however a uniform criteria for selecting K-initial centers is not provided. Sun Yin et al. [23] present an experimental study on applying Bradley et al.'s iterative initial-point refinement algorithm [3] to the K-modes clustering to improve the accuracy and repetitiveness of the clustering results. Their experiments show that the K-modes clustering algorithm using refined initial points leads to higher precision results much more reliably than the random selection method without refinement. This method is dependent on the number of cases with refinements and the accuracy value varies. Khan and Ahmad [16] use Density-based Multiscale Data Condensation [20] approach with Hamming distance to extract K initial points, however, their method has quadratic complexity with respect to the number of data objects. He [10] presents two farthest point heuristic for computing initial cluster centers for K-modes algorithm. The first heuristic is equivalent to random selection of initial cluster centers and the second uses a deterministic method based on a scoring function that sums the frequency count of attribute values of all data objects. This heuristic does not explain how to choose a point when several data objects have same scores, and if it randomly break ties, then fixed centers cannot be guaranteed. Wu et al. [24] develop a density based method to compute

the K initial modes which has quadratic complexity. To reduce its complexity to linear they randomly select square root of the total points as a sub-sample of the data, however, this step introduces randomness in the final results and repeatability of clustering results may not be achieved. Cao et al. [4] present an initialization method that consider distance between objects and the density of the objects. A major drawback of this method is that it has quadratic complexity. Khan and Kant [18] propose a method that is based on the idea of evidence accumulation for combining the results of multiple clusterings [6] and only focus on those data objects that are more less vulnerable to the choice of random selection of modes and to choose the most diverse set of modes among them. Their experiment suggest that the initial modes outperform the random choice, however the method does not guarantee fixed choice of initial modes.

In the next section, we briefly describe the K-modes clustering algorithm.

3 K-Modes Algorithm for Clustering Categorical Data

Due to the limitation of the dissimilarity measure used by traditional K-means algorithm, it cannot be used to cluster categorical dataset. 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 the K-means paradigm to cluster categorical data by removing the barrier imposed by K-means through following modifications:

1. Using a simple matching dissimilarity measure or the Hamming distance for categorical data objects
2. Replacing means of clusters by their modes (cluster centers)

The simple matching dissimilarity measure can be defined as following. Let X and Y be two categorical data objects described by m 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

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \tag{3.1}$$

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

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

Let Z be a set of categorical data objects described by categorical attributes, $A_1, A_2, \ldots A_m$. When the above is used as the dissimilarity measure for categorical data objects, the cost function becomes

$$C(Q) = \sum_{i=1}^{n} d(Z_i, Q_i) \tag{3.2}$$

where Z_i is the i^{th} element and Q_i is the nearest cluster center of Z_i. The K-modes algorithm minimizes the cost function defined in Equation 3.2.

The K-modes assumes that the knowledge of number of natural grouping of data (i.e. K) is available and consists of the following steps: -

1. Create K clusters by randomly choosing data objects and select K initial cluster centers, one for each of the cluster.
2. Allocate data objects to the cluster whose cluster center is nearest to it according to equation 3.2.
3. Update the K clusters based on allocation of data objects and compute K new modes of all clusters.
4. Repeat step 2 to 3 until no data object has changed cluster membership or any other predefined criterion is fulfilled.

4 Multiple Attribute Clustering Approach for Computing Initial Cluster Centers

Khan and Ahmad [17] show that for partitional clustering algorithms, such as K-Means, some of the data objects are very similar to each other and that is why they share same cluster membership irrespective to the choice of initial cluster centers. Also, an individual attribute may provide some information about initial cluster center. He et al. [11] present a unified view on categorical data clustering and cluster ensemble for the creation of new clustering algorithms for categorical data. Their intuition is that attributes present in a categorical data contributes to the final cluster structure. They consider the attribute values of an attribute as cluster labels giving "best clustering" without considering other attributes and created a cluster ensemble. We take motivation from these research works and propose a new cluster initialization algorithm for categorical datasets that perform multiple clustering on different attributes and uses distinct attribute values as cluster labels as a cue to find consistent cluster structure and an aid in computing better initial centers. The proposed approach is based on the following experimental observations (assuming that the desired number of clusters, K, are known):

1. Some of the data objects are very similar to each other and that is why they have same cluster membership irrespective of the choice of initial cluster centers [17].

2. There may be some attributes in the dataset whose number of attribute values are less than or equal to K. Due to fewer attribute values per cluster, these attributes shall have higher discriminatory power and will play a significant role in deciding the initial modes as well as the cluster structures. We call them as *Prominent Attributes (P)* .

3. There may be few attributes whose number of attribute values are greater than K. The many atrribute values in these attributes will be spread out per cluster, add little to determine proper cluster structure and contribute less in deciding the initial representative modes of the clusters.

The main idea of the proposed algorithm is to partition the data, for every prominent attribute based on its attribute values, and generate a *cluster string* that contains the respective cluster allotment labels of the full data. This process yields a number of cluster strings that represent different partition views of the data. As noted above, some data objects will not be affected by choosing different cluster centers and their cluster strings will remain same. The algorithm assumes that the knowledge of natural clusters in the data i.e. K is available and merges similar cluster strings into K partitions. This step will group similar cluster strings into K clusters. In the final step, the cluster strings within each K clusters are replaced by the corresponding data objects and modes of every K cluster is computed that serves as the initial centers for the K-modes algorithm. The algorithmic steps of the proposed approach are presented below.

Algorithm: Compute Initial Modes. Let Z be a categorical dataset with N data objects embedded in M dimensional feature space.

1. Calculate the number of Prominent Attributes ($\#P$)

2. If $\#P > 0$, then use these Prominent Attributes for computing initial modes by calling *getInitialModes(Attributes P)*

3. If $\#P = 0$ i.e. there are no Prominent Attributes in the data, or if $\#P = M$ i.e. all attributes are Prominent Attributes, then use all attributes and call *getInitialModes(Attributes M)*

Algorithm: *getInitialModes(Attributes A)*

1. For every $i \in A$, i=1,2...A, repeat step 2 to 4. Let j denotes the number of attribute values of i^{th} attribute. Note that if A is P then $j \leq K$, else if A is M then $j > K$.

2. Divide the dataset into j clusters on the basis of these j attribute values such that data objects with different values (of this attribute i) fall into different clusters.

3. Compute j M-dimensional modes, and partition the data by performing K-modes clustering that consumes them as initial modes.

4. Assign cluster label to every data object as S_{ti}, where t=1,2...N

5. Generate cluster string G_t corresponding to every data object by storing the cluster labels from S_{ti}. Every data object will have A class labels.

6. Find distinct cluster strings from G_t, count their frequency, and sort them in descending order. Their count, K', is the number of distinguishable clusters.

   There arise three possibilities:

   (a) $K' > K$ – Merge similar distinct cluster string of G_t into K clusters and compute initial modes (details presented in Section 4.1)

   (b) $K' = K$ – Distinct cluster strings of G_t matches the desired number of clusters in the data. Glean the data objects corresponding to these K cluster strings, which will serve as initial modes for the K-modes clustering algorithm.

   (c) $K' < K$ – An obscure case may arise where the number of distinct cluster strings are less than the chosen K (assumed to represent the natural clusters in the data). This case will only happen when the partitions created based on the attribute values of A attributes groups the data in the same clusters every time. A possible scenario is when the attribute values of all attributes follow almost same distribution, which is normally not the case in real data. This case also suggests that probably the chosen K does not resemble with the natural grouping and it should be changed to a lesser value. The role of attributes with attribute values greater than K has to be investigated in this case. This particular case is out of the scope of the present paper.

4.1 Merging Clusters As discussed in step 6 of algorithm *getInitialModes(Attributes A)*, there may arise a case when $K' > K$, which means that the number of distinguishable clusters obtained by the algorithm are more than the desired number of clusters in the data. Therefore, K' clusters must be merged to arrive at K clusters. As these K' clusters represent distinguishable clusters, a trivial approach could be to sort them in order of cluster string frequency and pick the top K cluster strings. A problem with this method is that it cannot be ensured that the top K most frequent cluster strings are representative of K clusters. If more than one cluster string comes from same cluster then the K-modes algorithm will fall apart and will give undesirable clustering results. This observational fact is also verified experimentally and holds to be true.

Keeping this issue in mind, we propose to use the hierarchical clustering method [9] to merge K' distinct cluster strings into K clusters. Hierarchical clustering has the disadvantage of having quadratic time complexity with respect to the number of data objects. In general, K' cluster strings will be less than N. However, to avoid extreme case such as when $K' \approx N$, we only choose the most frequent $N^{0.5}$ distinct cluster strings of G_t. This will make the hierarchical algorithm log-linear with the number of data objects (K' or $N^{0.5}$ distinct cluster strings here). The infrequent cluster strings can be considered as outliers or boundary cases and their exclusion does not affect the computation of initial modes. In the best case, when $K' \ll N^{0.5}$, the time complexity effect of log-linear hierarchical clustering will be minimal. The hierarchical clusterer merges K' ($N^{0.5}$ in worst case) distinct cluster strings of G_t by labelling them in the range of $1 \ldots K$. For every cluster label $k = 1 \ldots K$, group the data objects corresponding to the cluster string with label k and compute the group modes. This process generates K M-dimensional modes that are to be used as initial modes for K-modes clustering algorithm.

4.2 Choice of Attributes. The proposed algorithm starts with the assumption that there exists prominent attributes in the data that can help in obtaining distinguishable cluster structures that can be merged to obtain initial cluster centers. In the absence of any prominent attributes (or if all attributes are prominent), *Vanilla Approach*, all the attributes are selected to find initial modes. Since attributes other can prominent attributes contain attribute values more than K, a possible repercussion is the increased number of distinct cluster strings G_t due to the availability of more cluster allotment labels. This implies an overall reduction in the individual count of distinct cluster strings and many small clusters may arise side-by-side. Since the hierarchical clusterer imposes a limit of $\sqrt{N}$ on the top cluster strings to be merged, few distinguishable cluster could lay outside the bound during merging. This may lead to some loss of information and affects the quality of the computed initial cluster centers. The best case occurs when the number of distinct cluster strings is less than or equal to $\sqrt{N}$.

4.3 Evaluating Time Complexity The above proposed algorithm to compute initial cluster centers has two parts, namely, *getInitialModes(Attributes A)* and merging of clusters. In the first part, the K-modes algorithm is run P times (in the worst case M times). As the K-modes algorithm is linear with respect to the size of the dataset [12], the worst case time complexity will be $M.O(rKMN)$, where r is the number of iterations needed for convergence and $\ll$ N. In the second part of the algorithm, the hierarchical clustering is used. The worst case complexity of the hierarchical clustering is $O(N^2 logN)$. As the proposed approach chooses distinct cluster strings that are less than or equal to $N^{0.5}$, the worst case complexity becomes $O(NlogN)$. Combining both the parts, the worst case time complexity of the proposed approach becomes $(M.O(rKMN) + O(NlogN))$, which is log-linear with respect to the size of the dataset.

5 Experimental Analysis

5.1 Datasets. To evaluate the performance of the proposed initialization method, we use several pure categorical datasets from the UCI Machine Learning Repository [5]. All the datasets have multiple attributes and varied number of classes, and some of the dataset contain missing values. A short description for each dataset is given below.

Soybean Small. The soybean disease dataset 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. The clustering results on soybean data are shown in Table 2.

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%). There are 9 instances in Attribute 6 and 9 that contain a missing

(i.e. unavailable) attribute value. The clustering results of breast cancer data are shown in Table 3.

Zoo Data. It has 101 instances described by 17 attributes and distributed into 7 categories. All of the 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. The clustering results of zoo data are shown in Table 4.

Lung Cancer Data. This dataset contains 32 instances described by 56 attributes distributed over 3 classes with missing values in attributes 5 and 39. The clustering results for lung cancer data are shown in Table 5.

Mushroom Data. Mushroom dataset consists of 8124 data objects described by 22 categorical attributes distributed over 2 classes. The two classes are edible (4208 objects) and poisonous (3916 objects). It has missing values in attribute 11. The clustering results for mushroom data are shown in Table 6.

5.2 Comparison and Performance Evaluation Metric. We compared the proposed cluster initial center against the random initialization method and the methods described by Cao et al. [4] and Wu et al. [24]. For random initialization, we randomly group data objects into K clusters and compute their modes to be used as initial centers. The reported results are an average of 50 such runs.

To evaluate the performance of clustering algorithms and for fair comparison of results, we used the performance metrics used by Wu et al [24] that are derived from information retrieval. If a dataset contains K classes for any given clustering method, let a_i be the number of data objects that are correctly assigned to class C_i, let b_i be the number of data objects that are incorrectly assigned to class C_i, and let c_i be the data objects that are incorrectly rejected from class C_i, then precision, recall and accuracy are defined as follows:

$$PR = \frac{\sum_{i=1}^{K} \left(\frac{a_i}{a_i + b_i} \right)}{K} \tag{5.3}$$

$$RE = \frac{\sum_{i=1}^{K} \left(\frac{a_i}{a_i + c_i} \right)}{K} \tag{5.4}$$

$$AC = \frac{\sum_{i=1}^{K} a_i}{N} \tag{5.5}$$

Table 1: Effect of choosing different number of attributes.

~~Dataset~~	Proposed		Vanilla		$\sqrt{N}$
	#P	#CS	#A	#CS	
Soybean	20	21	35	25	7
Zoo	16	7	17	100	11
Breast-Cancer	9	355	9	355	27
Lung-Cancer	54	32	56	32	6
Mushroom	5	16	22	683	91

5.3 Effect of Number of Attributes. To test the intuition discussed in Section 4.2, we performed a comparative analysis on the effect of number of selected attributes on the number of distinct cluster strings. In Table 1, #P is the number of prominent attributes, #A is the total number of attributes in the data, #CS is the number of distinct cluster string and $\sqrt{N}$ is the limit on the number of top cluster strings to be merged using hierarchical clustering. The table shows that choosing a *Vanilla* approach (all attributes) leads to higher #CS, whereas with the proposed approach the number of distinct cluster strings are much lesser. For breast cancer data, all the attributes were prominent therefore #P and #A are same and hence same #CS. For lung cancer data #P $\approx$ #A therefore #CS are same. Due to the limit of merging top $\sqrt{N}$ cluster string (and reasons described in Section 4.2), clustering results using a *Vanilla* approach is worse than the proposed approach and are not reported in the paper. It is to be noted that the #CS using proposed approach for Zoo and Mushroom data are within the bounds of $\sqrt{N}$ limit.

5.4 Clustering Results. It can be seen from Table 2 to 6 that the proposed initialization method outperforms random cluster initialization for categorical data in accuracy, precision and recall. Another advantage of the proposed method is that it generates fixed cluster centers, whereas the random initialization method does not. Therefore, repeatable and better cluster structures can be obtained using the proposed method. In comparison to the initialization methods of Cao et al. and and Wu et al., the findings can be summarized as:

- in terms of accuracy, the proposed method outperforms or equals other methods in 4 cases and perform worse in one case.
- in terms of precision, the proposed method performs well or equals other methods in 2 cases while performs worse in 3 cases.
- in terms of recall, the proposed method outperforms or equals other methods in 4 cases whereas

it perform worse in 1 case.

Table 2: Clustering results for Soybean data

	Random	**Wu**	**Cao**	**Proposed**
AC	0.8644	**1**	**1**	0.9574
PR	0.8999	**1**	**1**	0.9583
RE	0.8342	**1**	**1**	0.9705

Table 3: Clustering results for Breast Cancer data

	Random	**Wu**	**Cao**	**Proposed**
AC	0.8364	0.9113	0.9113	**0.9127**
PR	0.8699	**0.9292**	**0.9292**	**0.9292**
RE	0.7743	0.8773	0.8773	**0.8783**

Table 4: Clustering results for Zoo data

	Random	**Wu**	**Cao**	**Proposed**
AC	0.8356	0.8812	0.8812	**0.891**
PR	0.8072	**0.8702**	**0.8702**	0.7302
RE	0.6012	0.6714	0.6714	**0.8001**

Table 5: Clustering results for Lung Cancer data

	Random	**Wu**	**Cao**	**Proposed**
AC	**0.5210**	0.5	0.5	0.5
PR	0.5766	0.5584	0.5584	**0.6444**
RE	0.5123	0.5014	0.5014	**0.5168**

Table 6: Clustering results for Mushroom data

	Random	**Wu**	**Cao**	**Proposed**
AC	0.7231	0.8754	0.8754	**0.8815**
PR	0.7614	**0.9019**	**0.9019**	0.8975
RE	0.7174	0.8709	0.8709	**0.8780**

The above results are very encouraging due to the fact that the worst case time complexity of the proposed method is log-linear, whereas the method of Cao et al. has quadratic complexity and the method of Wu et al. induces random selection of data points. The accuracy values of proposed method are mostly better than or equal to other methods, which implies that the proposed approach is able to find fixed initial centers that are close to the actual centers of the data. The only case where the proposed method perform worse in all three performance metric is the soybean dataset. We observe that on some datasets the proposed method gives worse values for precision, which implies that in those cases some data objects from non-classes are getting clustered in given classes. The recall values of proposed method are mostly better than the other methods, which suggests that the proposed approach tightly controls the data objects from given classes to be not clustered to non-classes. Breast cancer data has no prominent attribute in the data and uses all the attributes and produces comparable results to other methods. Lung cancer data, though smaller in size has high dimension and the proposed method is able to produce better precision and recall rates than other methods. It is also observed that the proposed method perform well on large dataset such as mushroom data with more than 8000 data objects. In our experiment, we did not get a scenario where the distinct cluster strings are less than the desired number of clusters. The proposed algorithm is also independent of the order of presentation of data due to he way mode is computed for different attributes.

6 Conclusions

The results attained by the K-modes algorithm for clustering categorical data depends intrinsically on the choice of random initial cluster center, that can cause non-repeatable clustering results and produce improper cluster structures. In this paper, we propose an algorithm to compute initial cluster center for categorical data by performing multiple clustering on attribute values of attributes present in the data. The present algorithm is developed based on the experimental fact that similar data objects form the core of the clusters and are not affected by the selection of initial cluster centers, and that individual attribute also provide useful information in generating cluster structures, that eventually leads to computing initial centers. In the first pass, the algorithm produces distinct distinguishable clusters, that may be greater than, equal to or less than the desired number of clusters (K). If it is greater than K then hierarchical clustering is used to merge similar cluster strings into K clusters, if it is equal to K then data objects corresponding to cluster strings can be directly used as initial cluster centers. An obscure possibility arises when cluster strings are less than K, in which case either the value of K is to be reduced, or assumed that the current value of K is not true representative of the desired number of clusters. However, in our experiment we did not get such situation, largely because it can happen in a rare occurence when all the attribute values of different attributes cluster the data in the same way. These initial cluster centers when used as seed to K-modes clustering algorithm, improves the accuracy of the traditional K-modes clustering algorithm that uses random modes as starting point. Since there is a def-

inite choice of initial modes (zero standard deviation), consistent and repetitive clustering results can be generated. The proposed method also does not depend on the way data is ordered. The performance of the proposed method is better than or equal to the other two methods on all datasets except one case. The biggest advantage of the proposed method is the worst case log-linear time complexity of computation and fixed choice of initial cluster centers, whereas both the other two methods lack either one of them.

In scenarios when the desired number of clusters are not available at hand, we would like to extend the proposed multi-clustering approach for categorical data for finding out the natural number of clusters present in the data in addition to computing the initial cluster centers for such case.

References

[1] Michael R. Anderberg. *Cluster analysis for applications*. Academic Press, New York, 1973.

[2] Daniel Boley, Maria Gini, Robert Gross, Eui-Hong Han, George Karypis, Vipin Kumar, Bamshad Mobasher, Jerome Moore, and Kyle Hastings. Partitioning-based clustering for web document categorization. *Decision Support Systems.*, 27:329–341, December 1999.

[3] Paul S. Bradley and Usama M. Fayyad. Refining initial points for k-means clustering. In Jude W. Shavlik, editor, *ICML*, pages 91–99. Morgan Kaufmann, 1998.

[4] Fuyuan Cao, Jiye Liang, and Liang Bai. A new initialization method for categorical data clustering. *Expert Systems and Applications*, 36:10223–10228, 2009.

[5] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[6] Ana L. N. Fred and Anil K. Jain. Data clustering using evidence accumulation. In *ICPR (4)*, pages 276–280, 2002.

[7] K. Chidananda Gowda and E. Diday. Symbolic clustering using a new dissimilarity measure. *Pattern Recogn.*, 24:567–578, April 1991.

[8] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Rock: A robust clustering algorithm for categorical attributes. In *Proceedings of the 15th International Conference on Data Engineering, 23-26 March 1999, Sydney, Austrialia*, pages 512–521. IEEE Computer Society, 1999.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. In *in SIGKDD Explorations*, volume 11 of *1*, 2009.

[10] Zengyou He. Farthest-point heuristic based initialization methods for k-modes clustering. *CoRR*, abs/cs/0610043, 2006.

[11] Zengyou He, Xiaofei Xu, and Shengchun Deng. A cluster ensemble method for clustering categorical data. *Information Fusion*, 6(2):143–151, 2005.

[12] Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *Research Issues on Data Mining and Knowledge Discovery*, 1997.

[13] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.*, 2(3):283–304, 1998.

[14] Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[15] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990.

[16] Shehroz S. Khan and Amir Ahmad. Computing initial points using density based multiscale data condensation for clustering categorical data. In *Proc. of 2nd Int'l Conf. on Applied Artificial Intelligence*, 2003.

[17] Shehroz S. Khan and Amir Ahmad. Cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letters*, 25:1293–1302, 2004.

[18] Shehroz S. Khan and Shri Kant. Computation of initial modes for k-modes clustering algorithm using evidence accumulation. In *Proceedings of the 20th international joint conference on Artificial intelligence (IJCAI)*, pages 2784–2789, 2007.

[19] Jiri Matas and Josef Kittler. Spatial and feature space clustering: Applications in image analysis. In *CAIP*, pages 162–173, 1995.

[20] Pabitra Mitra, C. A. Murthy, and Sankar K. Pal. Density-based multiscale data condensation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):734–747, 2002.

[21] Euripides G. M. Petrakis and Christos Faloutsos. Similarity searching in medical image databases. *IEEE Transactions on Knowledge Data Engineering.*, 9(3):435–447, 1997.

[22] H. Ralambondrainy. A conceptual version of the k-means algorithm. *Pattern Recognition Letters*, 16(11):1147–1157, 1995.

[23] Ying Sun, Qiuming Zhu, and Zhengxin Chen. An iterative initial-points refinement algorithm for categorical data clustering. *Pattern Recognition Letters*, 23(7):875–884, 2002.

[24] Shu Wu, Qingshan Jiang, and Joshua Zhexue Huang. A new initialization method for clustering categorical data. In *Proceedings of the 11th Pacific-Asia conference on Advances in knowledge discovery and data mining*, PAKDD'07, pages 972–980, Berlin, Heidelberg, 2007. Springer-Verlag.