K-MODE CLUSTERING ALGORITHM TO ANALYZE DATA IN DIFFERENT CATEGORIES

Dinju V Joseph[1], Jomol Jose[2], Ashwin C Tom[3], Ajith G.S[4]

[1, 2, 3] P.G Scholar of Amal Jyothi College of Engineering, Kanjirappally, Kerala
[4] Asst. Professor of Amal Jyothi College of Engineering Kanjirappally, Kerala

Abstract— Now a days data mining and knowledge acquisition has emerged as an important process. The data that is used can be a mixture of several categories and of very large size. So, to cluster such data sets, simple or numerical clustering algorithm cannot be used. For that, a clustering algorithm to cluster large data sets of categorical data : K-mode clustering algorithm is designed. It cluster very large categorical data and gives an optimal solution.

Keywords—clustering, k-modes, k-means, Categorical data, Data mining

I. INTRODUCTION

Data mining is the process of analyzing the patterns hidden in large datasets according to various perspectives for categorization into useful information. Clustering is one of the technique to categorize data in data mining.

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters).The k-modes clustering algorithm is an extension to the standard k-means clustering algorithm basically for clustering categorical data, introduce a different dissimilarity measure and update the modes with a frequency based method [2]. In the case of data mining, k-means is the mostly used algorithm for clustering data because of its efficiency in clustering very large data set. Clustering methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria.

The k-modes approach modifies the standard k-means process for clustering categorical data by replacing the Euclidean distance function with the simple matching dissimilarity measure, using modes to represent the centroid of clusters and updating modes with the most frequent categorical values in each of iterations of the process [3]. These modifications guarantee that the clustering process converges to a local minimal result and the efficiency of the clustering process is maintained [3].

In most cases the datasets available will be a mixture of numerical as well as other data types. Therefore, using K-means or related clustering algorithms to categorize such data types is not possible. In that case all required is a method to cluster categorical data and k-modes is best suited algorithm for that. K-modes algorithm cluster data from different categories and give an optimal result.

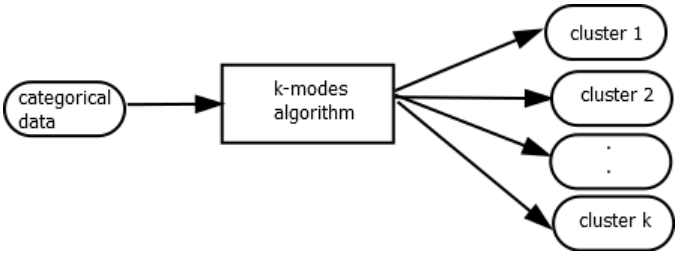

Fig1: Stages in k-modes clustering

II. LITERATURE REVIEW

Zhexue Huang [1] proposes an algorithm, called k-modes, to extend the standard k-means to categorical domains. K-means can be used to cluster only numerical data, The k-modes algorithm presented in this paper has removed this limitation preserving the efficiency of K-means.

The k-modes algorithm has made the following extensions to the k-means algorithm:

1. replacing means with modes of clusters,
2. using simple dissimilarity measures to deal with categorical objects,
3. a frequency based method is used to update modes of clusters.

These extensions allow to use the k-means paradigm directly to cluster categorical data without need of data conversion.

Goyal et al.[2] proposes that k-mode algorithm is an extension of k-means algorithm and is the partitioning based clustering algorithm.

Instead of using Euclidean distance it uses simple matching dissimilarity function. K-Mode works well for categorical datasets whereas K-Means Algorithm does not work well for Categorical datasets. It is famous for simplicity, speed and is linearly scalable with respect to the dataset.

Prakash et al.[3] proposes the k-modes clustering algorithm as an extension to the standard k-means clustering algorithm for clustering categorical data. The k-modes approach modifies the standard k-means process for clustering categorical data by replacing the Euclidean distance function with the simple matching dissimilarity measure, using modes to represent centroids of cluster and updating modes with the most frequent categorical values in each of iterations of the clustering process. These modifications guarantee that the clustering process converges to a local minimal result and the efficiency of the clustering process is maintained as it doe snot fully changes the standard k-means algorithm. Determination of grouping in a set of unlabeled information on the basis of its features is the main objective of clustering.
San et al. [4] proposes that the k-modes algorithm is unstable due to non-uniqueness of the modes. That is, the clustering results depend strongly on the selection of modes during the clustering process. This paper aims at eliminating the above drawback in the k-modes algorithm by introducing a new notion of "cluster centers" called representatives for categorical objects. Formulate the clustering problem of categorical objects as a partitioning problem in the fashion similar to k-means clustering. Arithmetic operations are completely absent in categorical objects, use the Cartesian product and union operations for the formation of "cluster centers" based on the notion of means in the numerical setting.

Joshua Zhexue Huang [5] , In data mining research, much effort has been put on development of new techniques for clustering categorical data. The k-modes clustering algorithm is one of the first algorithms for clustering very large categorical data. An extension of k-modes algorithm called fuzzy k-modes clustering algorithm and other variants were introduced. Here, instead of assigning each object to one cluster, the fuzzy k-modes clustering algorithm calculates a cluster membership degree value for each object to each cluster. This is achieved by introducing the fuzziness factor in the objective function similar to that of fuzzy k-means. The clustering result can be improved whenever the inherent clusters overlap in a data set. The k-prototypes clustering algorithm combines k-means and k-modes to cluster data with mixed numeric and categorical values.

III. COMPARISON OF K-MODES WITH K-MEANS CLUSTERING ALGORITHMS

The k-means algorithm is built upon four basic operations to cluster numerical datasets: (1) selection of the initial k means for k clusters, (2) the dissimilarity measure (distance) is calculated between an object and the mean of a cluster, (3) an object is allocated to the cluster whose mean is nearest to the object, (4) Re-calculation of the mean of a cluster from the objects allocated to each cluster so that the intra cluster dissimilarity is minimized [1]. (2), (3) and (4) steps are repeatedly performed in the algorithm until all the elements are categorized.

Some of the important properties of k means algorithm is that:

1. It is efficient in processing large data sets.
2. It often terminates at a local optimum.
3. It works only on numeric values because it minimizes a cost function by calculating the means of clusters.
4. The clusters are formed in different shapes.

When it comes to k modes algorithm, following modifications are made to k-means algorithm [1],

1. Dissimilarity measures is used instead of Euclidean distance function.
2. Means are replaced with modes.
3. Use of frequency based method to update modes of a cluster in each iterations.

A local minimal result is guaranteed with these modifications [5]. The efficiency of the clustering process is maintained since the k-means clustering process is not entirely changed [5].

IV. THE K-MODES ALGORITHM METHODOLOGY

The k-modes algorithm is an extension to the standard k-means algorithm to cluster categorical data by using different methods as follows [5]:

- A simple matching dissimilarity measure for categorical objects.
- Means are replaced with modes to form clusters.
- Clustering cost function is minimized using a frequency-based method to update modes in the k-means fashion.

Steps in k-modes algorithm to cluster categorical data:

- k initial modes are selected from which one will be assigned to each cluster.
- Then group the objects into various clusters and find the dissimilarity measure of each object with mode.
- Update the mode after each iterations and reassign objects to clusters which are more close to modal values.

• Repeat the process till no objects are close to any other modes in different clusters and all objects are grouped in any of the clusters. Hence, we will obtain different clusters which are unique from one another.

The k-modes also produce local optimal solutions that are dependent on the initial modes and the order of objects in the data set as in k-means.

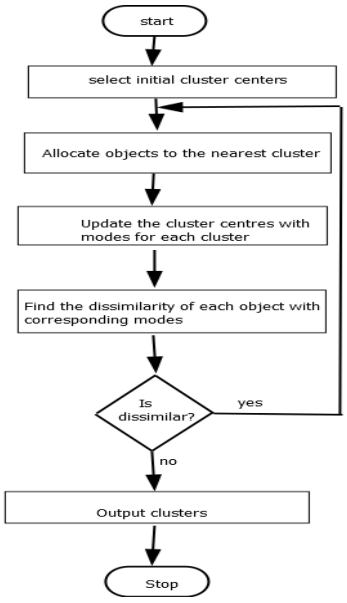

Fig2: Clustering using k-modes

Let there be two categorical objects X and Y described by m categorical attributes [1]. The dissimilarity measure between X and Y can be defined by the total mismatches of the corresponding attribute categories of the two objects [1]. The smaller the number of mismatches is, the more similar the two objects [1].

V. METHOD OF IMPLEMENTATION

The methodology can be implemented using any programming language. First a prototype should be designed for the implementation of research element. Then connect the dataset with the corresponding database and use the above mentioned methodology to group the elements in the database. Finally, it will result in grouping the data into corresponding clusters.

VI.EXPERIMENTAL RESULT

K-modes technique is used to cluster categorical data in a data set. Illustrating an example, the customers visiting the restaurant are categorized into several groups such as frequently visited customers, VIP customers etc. based on certain values.

	Category	Customer1	Customer2	…	CustomerN
1	Number of visits	210	547	..	80
2	Payment made	56759	70750	…	10100
3	Status of customer	VIP	NVIP	…	STAR

Table 1: Showing the data set of values of different customers visiting the restaurant.

For example, if in a restaurant only 10 tables are available for reservation and 100 customers reserved for the same time slot. Using k-modes clustering, we will group these 100 customers into K clusters based on certain values chosen. By doing so the system can automatically approve or disapprove the reservation made by customers who are randomly chosen based on the tables available in the restaurants and from the clusters that are formed using k-modes clustering.

Complexity of K-modes algorithm is O (tkn), where n is instances, c is clusters, and t is iterations and relatively efficient.

VII.CONCLUSION

K-modes algorithm is best suited for clustering categorical data. It preserves the efficiencies of k-means clustering to find the local minimum solution. There is several possible research directions may be worked on in the future to further extend and enhance the work presented in this paper.

VIII.FUTURE WORK

The method of implementation can variate based on the advancement in technologies. A newer version of K-modes could be proposed which preserves the efficiency of the previous one. To

cluster categorical data certain other algorithms can be used.

IX. ACKNOWLEDGEMENT

The authors would like to acknowledge the entire Master of Computer Applications Dept, Amal Jyothi College of Engineering, Koovappally for their valuable guidance and support that were used to improve the quality of this paper.

REFERENCES

[1] A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining- Zhexue Huang

[2] A Review on K-Mode Clustering- Manisha Goyal , Shruti Aggarwal

[3] A Review on K-Mode Clustering-Antara Prakash, Simran Kalera, Archisha Tomar , Aarushi Rai, Pooja Reddy, Prof. Ramesh Babu

[4] An alternative extension of the K-Means Algorithm for clustering categorical Data- Ohn Mar San , Van-Nam Huynh , Yoshiteru Nakamori

[5] Clustering Categorical Data with k-Modes- Joshua Zhexue Huang