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

Prathmesh Gaonkar
gaonkar.1@iitj.ac.in

Mechanical Engineering Department,
Indian Institute Of Technology Jodhpur

1 Critical Analysis

The distinct characteristic of clustering operation in data mining is that the dataset often contains numerical and categorical attribute values. This challenges the algorithm to handle such complexity of data while preserving the inter-and intra-relation of the data sets, no matter numeric or categorical. Ralambondrainy [1995] proposed an approach to cluster categorical data by using the K means algorithm. The idea is to convert multiple category attributes into binary attributes (using 0 and 1 to represent a category absent or present) and treat the binary attributes as numeric in the K-means algorithm.

The K-modes algorithm extends the K-means paradigm to cluster categorical data. The K-modes algorithm removes the numeric limitation while preserving the efficiency of the model. The following modifications have been done to use K modes for categorical data.

1. To make use of hamming distance as a dissimilarity measure for categorical data objects.
2. The means of the cluster were replaced by modes.

This simple matching dissimilarity measure can be explained as follows. Let us consider two data objects, X and Y, each having F categorical attributes. The dissimilarity measure between them is denoted by d(X, Y). It can be defined as total mismatches between corresponding attribute categories of two objects. The smaller the value of d(X, Y) (mismatches), the more similar the two objects are. This can be represented mathematically as

$$d(X,Y) = \sum_{i=1}^{F} \delta(x_j, y_j) \tag{1}$$

where,

$$\delta(x_j, y_j) = \begin{matrix} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{matrix} \tag{2}$$

Let us assume Z to be a set of categorical data objects. The categorical attributes defined be A_1, A_2, A_3... A_f. The mode of this data objects, Z=[Z_1, Z_2, Z_3,...Z_n] is a vector Q=[Q_1,Q_2...Q_3] that minimizes

$$D(Z,Q) = \sum_{i=1}^{n} d(Z_i, Q) \tag{3}$$

Thus the above equation denotes the distance between each data object and mode initialized. Also, the final mode of the data objects does not need to be an element of Z.

The K modes algorithm can be summarized as follows:

a) Selection of K initial modes for each cluster.
b) Based on the value of equation (1), we allocate the data points to each of the clusters whose mode is nearest to it.
c) After all the data objects are allocated to a cluster, we will compute the new modes of the group formed.
d) Repeat steps b and c until convergence i.e. the modes of the cluster do not change on further iterations.

In this algorithm, we chose our initial center randomly. Moreover, inappropriate choice of initial modes gives clustering, which is undesirable. The above-discussed algorithm gives the best results when the initial cluster center chosen is very close to the final cluster center. In this paper, the idea proposed is based on Evidence accumulation to categorical data sets. In evidence accumulation, we acquire evidence in the decision-making process until sufficient evidence has been collected to favor one decision over another. Here, we cluster the results of multiple clustering into a single model, in which we consider each clustering result as independent evidence of data organization.

There are many ways in which we can gather evidence in the context of unsupervised learning :
a) To combine the results of various clustering models.
b) To produce different results by resampling the data and then combining them.
c) To run an algorithm several times with other parameters and initialization.
This paper uses the last approach to collect evidence to create multiple partitions of the categorical data.

The two major steps of the proposed algorithm are

- By applying the K modes algorithm, we generate N independent evidence of data organization by initializing random modes and the resultant modes thus obtained are stored in a Mode-Pool, P_n.
- From the Mode-Pool obtained, find the most diverse set of modes which will be our final initial mode for clustering our dataset.

In mathematical terms,

- Let K be the number of clusters, and N represents the number of iterations of the algorithm with different initial modes.
- i = 1 While (i<=N)
  - Consider random initial modes and execute the K modes algorithm till it converges.
  - Store the final mode obtained in mode pool, P_i
  - i=i+1
- After we are done with this execution, we are left with a mode pool, P_n, containing N number of modes, each having a dimension of $K \times F$. (F is the number of attributes).
- Now, to extract the most diverse mode employs the following algorithm
  - Set i=1, j=1, k=1
  - While $i <= K$ do the following
  - While $j <= F$ do the following

- While $k <= N$ do the following
- Extract the most frequent feature and store it in a Initial Mode Matrix $I_{i \times j}$
- k = k + 1
- j = j + 1
- i = i + 1

Let us understand what this algorithm is implementing,

- Firstly we have N set of modes to start with.
- Next let assume each mode matrix to be in a stack one behind the other.
- The code in step 4 is telling us to take the mode of each grid along the z axis and that mode will be stored in a new Initial Mode Matrix (I)
- Finally we obtain a mode matrix $I_{K \times F}$.

2 Pros

- Each of the N clusterings generates its modes. These modes are representative of those data points/patterns that are less susceptible to change of cluster membership irrespective of the choice of the random initial mode selection.
- The initial modes calculated by this method are very similar to the actual/desired modes of the data set.
- It can be observed that when the proposed methods are used to choose the modes, the algorithm achieves better clustering and faster convergence compared to traditional methods.
- The final modes obtained for each K cluster should be pretty dissimilar, with more diversity embodied within them.
- Consistent results in terms of clustering can be obtained by this method.
- The clustering results have been improved with minor variance in clustering error.

3 Cons

- The proposed idea is computationally quite expensive as we apply the same algorithm repeatedly to obtain the initial modes from the mode pool.
- The proposed idea has not provided any clue as to how to decide the number of iteration (N) in selecting random initial modes and applying the K modes algorithm over it to eventually the mode pool.
- This lack of knowledge can result in the non-convergence of this method, and we will not be able to obtain optimum initial modes.
- The simple distance metric gives equal importance to each category of an attribute.

4 Future Research

As discussed in the cons of the approach presented by the author, there is a need for a technique that can help understand how many times we have to iterate and obtain initial modes to fill in the mode pool.