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

The technique proposed in this paper results in better clustering compared to the traditional method of randomly initializing the modes. However, there are certain aspects that can be added to this technique that will result in even better results.

1 Weighted matching distance metric

In calculating the distance between two objects, we need to calculate the distance between two objects described by categorical attributes. To find this distance the method used by us is 'simple matching distance metric'. As defined earlier the simple matching distance metric is

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

where,

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

There are limitations to this metric as all attributes have the same level of importance as we are simply adding 0 or 1 in the case of equality and inequality, respectively. However, in a real dataset, the effect of one attribute may be different from other attributes. To find the significance of each attribute we use the concept of entropy-based significance. The update in the proposed method is to change the metric by another metric called the 'weight matching distance metric'. Let 'a' be defined by any attribute from the set of attributes 'A'. Let weight(a) denote the weight of the attribute. This metric is defined as follows:

$$wd(x,y) = \sum_{a \in A} weight(a) \times \delta_a(x,y), \tag{3}$$

Here δ will be calculated by a simple matching distance metric, but we are multiplying it with the weight of that particular attribute. These weights for every attribute can be calculated by

$$weight(a) = \begin{matrix} 1/2 \times (1 + \frac{count_{zero}}{|A| + \sqrt{|A| - count_{zero}}}) & if Sig(a) = 0; \\ 1 + Sig(a) & if Sig(a) > 0 \end{matrix} \tag{4}$$

Where,
Sig(a) denotes the partition entropy-based significance of attribute ‘a’.
mod(A) denotes the cardinality of set A
$count_{zero}$ denotes the number of attributes in A whose significance equals 0.

$$count_{zero} = |a \in A : Sig(a) = 0|$$

From the above equation, we can observe that when Sig(a) is greater than 0 then weight(a) is equal to 1 + Sig(a) i.e. the weight is directly proportional to Sig(a) and when it is less than 0 then weight(a) is proportional to $count_{zero}$. By doing so we are assigning a small weight to those attributes whose significance equals 0 and assigning a bigger value of weights to those attributes whose significance is greater than 0.

Thus, by applying weights to various attributes, we consider each attribute’s importance, which contributes to deciding the initial modes for our dataset.