CHAPTER 3

IDENTIFICATION OF INITIAL DATA ITEMS FOR CLUSTERING

After identifying limitations in the previous chapter, this chapter focuses on resolving the issue of sensitivity of clustering algorithms to random initial data items. While generating clusters using k-means or k-modes algorithms, the selection of random initial cluster center is an important factor having high impact on accuracy of clusters.

3.1 WHY TO IDENTIFY INITIAL DATA ITEMS?

In most popular and simple k-means algorithm, this initial center selection is made randomly. Different initial data items of cluster centers lead to different clustering results. The process of clustering in the existing classic algorithms is shown in Figure 3.1 as follows: given the dataset for clustering, first the user has to specify the number of clusters that the dataset has to be partitioned. For each cluster, initial cluster center is chosen randomly and then select the distance measure like Euclidean, Manhattan and etc is selected with respect to type of data. Allocate each data item to the appropriate cluster to which it is nearer. This process is repeated for all data items and the cost of function is minimized. The process of clustering is same for both k-means and k-modes algorithms, but the terminology varies. That is, k-means algorithm uses initial centers, whereas k-modes algorithm uses initial modes. As k-means algorithm is for numerical data, it uses distance measures, while k-modes is for categorical data and uses similarity or dissimilarity measures. If one data item of dataset is selected as initial cluster center, then the other data items having the similar features with respect to that cluster center will fall into the same cluster. So, data items in the cluster depend on cluster center selected initially.
Selection of random cluster center impacts more on accuracy and purity of clusters. The data that is newly added to the database may also have a chance to act as a random cluster center. For example, let the transactional database consist of 1,00,000 transactions that have to be clustered into some user-specified number of clusters for decision making, say 100 clusters. Then, it is unfair if cluster center is picked from new data like hourly or daily transactions, because these transient transactions change based on customer preferences which in turn vary based on discount rates, any temporary offers, influenced by friends, neighbors etc. If any item is picked as cluster center from these drifting transactions and a decision, is taken, it will have high impact on business development. If k-modes algorithm, which is the first and foremost popular clustering algorithm for categorical data, is applied, it suffers from the problem of convergence at local optimal solutions because of random initial modes. Distinct random initial modes result in distinct clustering results. So, a proper selection of these initial modes is necessary. Therefore this chapter mainly focuses on new approaches for finding initial points for clustering algorithms.
In the literature, clustering algorithm selects random initial data items and generates initial clusters, whereas clustering algorithm in the proposed system identifies initial data items and generates initial clusters. The step of existing system is shown in red circle and step of proposed system is shown in green circle in Figure 3.2. The contribution of this feature to the proposed system lies in reducing the number of randomizations and time required for clustering.

The proposed methods for finding initial data items are incorporated at the initialization step of any clustering algorithms that generate clusters based on random initial points. So, these methods act as initializer or preprocessor for the clustering applications. Before applying directly the clustering algorithm for a given dataset, certain preprocessing steps are required, such as, analyzing the properties of dataset, formulating objective function and finding initial cluster center. This process varies based on the number of attributes using which clustering algorithm has to be performed and presence of restrictions. Then the remaining process of clustering depends on the chosen clustering algorithm. The final clusters are generated on the basis of the corresponding clustering algorithm selected. Finally it reduces running an algorithm for various randomizations, which in turn reduces the time and number of iterations for clustering process.

Now, let this process be explained in detail taking k-means algorithm since it selects initial centers randomly. K-means algorithm with its objective function tries to reduce the sum of squares of point-to-center distances. Diminishing the sum of squares of Euclidean distances makes convergence of procedure to a local minimum in finite number of iterations [111].
Building blocks of k-means algorithm includes the following steps.

**Step 1: Selection of initial k centers:** Arbitrarily select k data items as initial centers

In this step, an initial choice of candidate centers is made. This choice is critical because both quality of clusters and number of iterations required to converge are associated strongly to this choice.

**Step 2: Main loop**

In this step, given a set of initial k centers, repeats assigning each input data item to the cluster to which the data item is most similar. Collection of number of data items associated to each center forms one cluster. Update the mean of each cluster and the next iteration starts. Repeat this process until there is no change in the mean value of objects.

**Step 3: Termination condition**

Possible termination conditions are more; e.g. loop is terminated after predefined number of iterations, or centers change is below user specified threshold.

K-means algorithm has more advantages because it is simple and the clustering quality is steady and sufficiently good for various settings and data. It has made k-means the most popularly used clustering algorithm. Because of its importance, vast literature is available with shortcomings and improvements for basic framework. The initialization step has profound impact on the accuracy of final clusters.

The proposed methods of identifying initial data items are incorporated in step 1 of k-means algorithm and they generate initial data item location instead of random location. The remaining steps of k-means algorithm are the same. The proposed methods are not only limited to k-means algorithm but also applied to clustering algorithms which depend on random initial data items.
3.2 IDENTIFICATION OF INITIAL DATA ITEMS FOR SINGLE-ATTRIBUTE DATA WITHOUT RESTRICTIONS

To apply this method, first analyze the properties of data items of a given dataset. Select suitable measure among these properties. Formulate into the objective function [112]. Then this method generates an interval in which an initial cluster center lies. For example, given a dataset with values of variable ‘age’ to classify into clusters like youth, middle aged and old. By observing the values, directly formulate the function that satisfies the given values of dataset and apply the proposed method on each cluster to generate interval. Otherwise, if multi-attribute dataset like student dataset is given, each student record can be taken as one data item. To understand student record, properties may be Stu_Roll_No, Stu_Name, Stu_Address Stu_Marks, etc. Suppose clustering needs to be performed with respect to only one attribute, then the key attribute must be identified by performing attribute subset selection or feature selection methods. Let that attribute be Stu_Marks. Let the percentage of all students in dataset be 30-95. Clusters are generated for describing the students with distinction, first class, second class and fail.

To address the dynamic issue of determining an initial center in an interval, this method successively removes subintervals of original interval so as to reduce the final interval having initial centers. When the length of the remaining subinterval is reduced to a small length, then the process gets terminated. Generation of interval works in two levels. In the first level, a rough interval is generated. This interval is taken as input to the next level that refines and results in a small interval which reflects in the quality of clusters.

Level 1: Rough Interval (RI): This step computes the rough interval in which an initial data item lies.

Level 2: ReFined Interval (RFI): This step takes the output of level 1 as input and reduces exactly half of the interval at each iteration and results in refined interval.

This method reduces the required time for generation of clusters if a data item is selected from rough interval. The quality of clusters is improved if data item is selected from refined interval. The generation process of interval and clustering process are discussed in the forthcoming chapter and the results prove its efficiency compared with k-means algorithm.
3.3 IDENTIFICATION OF INITIAL DATA ITEMS FOR TWO-ATTRIBUTE DATA WITH RESTRICTIONS

The procedure used for finding initial data items while clustering with respect to two attributes data and includes the following steps.

Step 1: Treat every inequality condition, either with ≤ or ≥, as an equality condition.

Step 2: Plot each equality condition as a straight line. Each and every point on that line satisfy the equation.

Step 3: For the inequality condition with ≤, darken the area below the line. Similarly, for the inequality condition of the line with ≥, darken the area above the line.

Step 4: The common region that satisfies all the conditions is to be identified.

Step 5: Take the boundaries of the common region. Take the coordinates of boundaries and find the minimum value of objective function.

This method can also be extended to three attributes by drawing planes instead of lines. But it becomes difficult to visualize the common area. Therefore this method is preferable for two attributes as it is easy to plot straight lines as the relationship between the attributes and restrictions, and it is linear.

3.3.1 ILLUSTRATIVE EXAMPLE

Consider the following example of two attributes to illustrate. A doctor suggests a patient to increase the intake of vitamin A and B. Two vitamin pills are suggested. Irozorb and Nimcee. Each Irozorb pill contains 40 mg of vitamin A and 30 mg of vitamin B. Each Nimcee pill contains 20 mg of vitamin A and 40 mg of vitamin B. Each Irozorb pill costs 7 rupees and each Nimcee pill costs 5 rupees. The patient is suggested to have at least 2,000 mg of vitamin A and 2,400 mg of vitamin B. To group different categories of patients in different areas, let I = Number of Irozorb pills and N = Number of Nimcee pills. Formulate the function as:

Minimize $U = 7I + 5N$ 

subject to constraints
$40I + 20N \geq 2000$ (For vitamin A)

$30I + 40N \geq 2400$  (For vitamin B), $I \geq 0$ and $N \geq 0$

Plot each inequality condition as a straight line assuming as an equality condition.

$40I + 20N = 2000$

Figure 3.3: Plot of inequality condition $40I + 20N \geq 2000$

First substitute $I=0$ then $N=100$. Therefore point = $(0,100)$. Next substitute $N=0$ then $I=50$ and point = $(50,0)$. Now represent these coordinates on a graph and join as a straight line. Darken the region above the straight line. It is shown in Figure 3.3.

$30I + 40N = 2400$

Figure 3.4: Plot of inequality condition $30I + 40N \geq 2400$
Now identify the common region satisfying both the equations. Collect the coordinates of extreme points and find minimum point.

![Graph showing common region with extreme points X, Y and Z and minimum point at Y = (33,35) and U_Y = 406](image)

Similarly plot the other inequality condition, 30I + 40N ≥ 2400. First substitute I=0, then N=60. Point= (0,60). Next substitute N=0, then I=80. Point= (80,0). Now represent these coordinates on a graph as a straight line and darken the area above the straight line. It is shown in Figure 3.4. Extreme points of the common region are: X= (0,100), Y= (33,35) and Z= (80,0) as shown in Figure 3.5. Substitute these coordinates in the function U and obtain the values. Substituting the point X= (0,100), in function U=7I+5N= 7*0 + 5*100= 500. Similarly, obtain the values of function at Y and Z as shown in figure 3.5. Compare the function values U_X, U_Y and U_Z, and get the minimum value. Therefore minimum value is at Y. Using this procedure an approximate range of

I = 32-34,
N = 34-36 and
U = 404-406

3.4 IDENTIFICATION OF INITIAL DATA ITEMS FOR MULTI-ATTRIBUTE DATA
This method is used in generating connected clusters in a tree structure for a large domain. The large domain is divided into a number of smaller domains. Preferably, larger dimensions are divided into two smaller dimensions and finally the points of all smaller dimensions are combined. For N dimensions of search space, $2^N + 1$ function evaluations are required. For two dimensions, square design is considered shown in Figure 3.6, that consists of five points. The best of these five points is chosen as the next base point using

![Diagram of square design for two dimensions](image)

Figure 3.6: Square design for two dimensions

which the next square design is located. If no trial points are better than base point, then reduce the scale of grid and the search process is continued. In higher dimensions, this refers to evaluation of the performance of each vertex and the center.

For N-dimensions, the pattern composed of N+1 points is located at equidistance to one another. These points form vertices of the pattern, i.e., for two dimensions, pattern forms an equilateral triangle and for three dimensions, it is a tetrahedron, and so on.

![Diagram of triangle pattern for two dimensions](image)

Figure 3.7: Triangle pattern for two dimensions

The process starts with the given starting or base point $y^{(1)}$ and computes the next two points $y^{(2)}$ and $y^{(3)}$ using the given increments, and forms the initial pattern as shown in Figure 3.7. A new pattern can be formed on any vertex of previous pattern by projecting the chosen vertex at
an appropriate distance through the midpoint of the remaining vertices of previous pattern as shown in Figure 3.8. The vertex with the highest function value is selected and reflected through the midpoint of the remaining vertices to get a new point which is used to complete the next pattern. Here, $y^{(1)}$, $y^{(2)}$ and $y^{(3)}$ form the initial pattern. If $y^{(1)}$ is selected as the point with the highest function, the value is reflected through the middle of $y^{(2)}$ and $y^{(3)}$ to get $y^{(4)}$. Then the points $y^{(2)}$, $y^{(3)}$ and $y^{(4)}$ will form a new pattern as shown in Figure 3.8.

![Figure 3.8: New pattern construction](image)

This method requires an array of (N+1, N+2) dimension for storage which is relatively low. As this method functions with respect to the highest function value, it provides effective results when outliers are important.

### 3.5 SUMMARY

Various methods for finding initial data items are provided in this chapter. These methods are based on the number of attributes using which clustering is to be performed in the presence or absence of restrictions. That means, these methods are based on properties of data and preprocessing is done to analyze these properties of dataset. Preprocessing of dataset is not performed in existing methods. Selection of initial data items is done in random manner while generating clusters in traditional clustering methods whereas the proposed system generate an initial data items thereby reducing the impact of randomness on clusters.