A Multiagent System (MAS) for the Generation of Initial Centroids for K-means Clustering Data Mining Algorithm based on Actual Sample Datapoints

Dost Muhammad Khan  
PhD Student, School of Innovative Technologies and Engineering, University of Technology, Mauritius.  
dostmuhammad_khan@yahoo.com

Nawaz Mohamudally  
Associate Professor & Head School of Innovative Technologies and Engineering, University of Technology, Mauritius.  
alimohamudally@utm.intnet.mu

doi: 10.4156/jnit.vol1.issue2.8

Abstract

Clustering is a technique in data mining to find interesting patterns in a given dataset. A large dataset is grouped into clusters of smaller sets of similar data using k-means algorithm. Initial centroids are required as input parameters when using k-means clustering algorithm. There are different methods to choose initial centroids, from actual sample datapoints of a dataset. These methods are often implemented through intelligent agents, as the later are very commonly used in distributed networks given that they are not cumbersome for the network traffic. More over, they overcome network latency, operate in heterogeneous environment and possess fault-tolerant behavior. A multiagent system (MAS) is proposed in this research paper for the generation of initial centroids using actual sample datapoints. This multiagent system comprises four agents of k-means clustering algorithm using different methods namely Range, Random number, Outlier and Inlier for the generation of initial centroids.

Keywords: Range Method, Outlier Method, Inlier Method

1. Introduction

The ‘k’, in the k-means algorithm stands for number of clusters as an input and the ‘means’ stands for an average, location of all the members of a particular cluster. The algorithm is used for finding the similar patterns due to its simplicity and fast execution. This algorithm uses a square-error criterion for re-assignment of any sample from one cluster to another, which will cause a decrease in the total squared error. It is easy to implement, and its time and space complexity are relatively small.

There are different techniques available to generate the initial points which are mandatory at the outset of the process of the algorithm. However, we are hereby focusing on the centroid method. The centroid method consists of choosing all the starting clusters close to the mass centroid of the dataset. The center of each cluster is calculated by adding a small random perturbation to the centroid of the dataset. A centroid represents an average location of the particular cluster.

The ‘centroid’ method is further divided into two different approaches to choose initial centroids first one is Synthetic method starting point having in total ten different methods and the second one is based on Actual sample starting point having four different methods [1]. This paper envisages four different methods for the generation of initial centroids of k-means clustering algorithm based on Actual sample datapoints using intelligent agents.

Intelligent Mobile Agents are used in data mining algorithms because of robustness, intelligence and suitability in heterogeneous environment. The term mobile agent is derived from two different disciplines, the ‘intelligent agent’ is created from Artificial Intelligence and the ‘code mobility’ is defined from the distributed systems. The architecture of the multiagent system (MAS) proposed in this research paper for the generation of initial centroids using actual sample datapoints is shown in figure 1. This multiagent system comprises four agents of k-means clustering algorithm for the generation of initial centroids using different methods. The architecture is based on a client-server model of distributed computing system where four intelligent agents are deployed on different clients and they are controlled through another agent on the central server.
A Multiagent System (MAS) for the Generation of Initial Centroids for k-means Clustering Data Mining Algorithm based on Actual Sample Datapoints
Dost Muhammad Khan, Nawaz Mohamudally

In section 2 we present an overview of k-means clustering algorithm. Section 3 is about the four different methods namely Random number, Range, Outlier and Inlier for generation of initial centroids based on actual sample data, the last three methods are introduced for the first time. In section 4 we discuss the results and discussion, finally section 5 presents the conclusion.

2. An Overview of k-means Clustering Algorithm

The flow chart in figure 2 below illustrates the different steps of k-means clustering algorithm.

The steps of the above flow chart are as follows:

**Step 1:** Select the value of ‘k’, the number of clusters.
**Step 2:** Calculate the initial centroids from the actual sample of dataset. Divide datapoints into ‘k’ clusters.
**Step 3:** Move datapoints into clusters using Euclidean’s distance formula. Recalculate new centroids. These centroids are calculated on the basis of average or means.
**Step 4:** Repeat step 3 until no datapoint is to be moved [2][3].
We will now explain in details how the user should proceed for the execution of the process described in the flowchart above in figure 2. We start by entering the number of clusters and the number of iterations, which are the required and basic inputs of the k-means algorithm, then compute the initial centroids by using the methods discussed in section III. After that we calculate the distance through Euclidean’s distance formula in equation 1.

\[ d(x_i, x_j) = \sqrt{\sum_{k=1}^{N} (x_{ik} - x_{jk})^2} \]  

(1)

Where \( d(x_i, x_j) \) is the distance between \( x_i \) and \( x_j \). \( x_i \) and \( x_j \) are the attributes of a given object, where \( i, j \) and \( k \) vary from 1 to \( N \) where \( N \) is total number of attributes of that given object, indexes \( i, j, k \) and \( N \) are all integers. On the basis of these distances, we need to generate the partition by assigning each sample to its closest cluster. Then we compute new cluster centers as centroids of the clusters, again compute the distances and generate the partition. We must repeat this until the cluster membership stabilizes. This algorithm uses a square-error criterion for re-assignment of any sample from one cluster to another, which will cause a decrease in the total squared error. The sum of square-error value ‘E’ is shown in equation 2.

\[ E = \sum_{i=1}^{L} \sum_{j=1}^{M} \sum_{k=1}^{N} (F - C)^2 \]  

(2)

Where \( (F - C)^2 \) is the distance between the datapoints.

The following are major issues in k-means clustering algorithm:

i. The algorithm is only applicable to datasets where the notion of the mean is defined. Thus, it is difficult to apply to categorical datasets. There is however, a variation of the k-means algorithm called k-modes, which clusters categorical data. The algorithm uses the mode instead of the mean as the centroid.

ii. The user needs to specify the number of clusters \( k \) in advance. In practice, several \( k \) values are tried and the one that gives the most desirable result is selected.

iii. The algorithm is sensitive to outliers. Outliers are datapoints that are very far away from other datapoints. Outliers could be errors in the data recording or some special datapoints with very irrelevant values.

iv. The algorithm is sensitive to initial seeds, which are the initially selected centroids. Different initial seeds may result in different clusters. Thus, if the sum of squared error is used as the stopping criterion, the algorithm only achieves local optimal value. The global optimal value is computationally infeasible for large datasets.

v. The k-means algorithm is not suitable for discovering clusters that are not hyper-ellipsoids or hyper-spheres [4][8].

vi. The value of \( k \) and the number of iterations should be small otherwise the time and space complexity of the algorithm will become high. This is a major issue of k-means clustering algorithm.

The following are areas but not limited to where k-means clustering algorithm can be applied:

- **Marketing**: Finding groups of customers with similar behaviour given large database of customer containing their profiles and past records.
- **Biology**: Classification of plants and animals given their features.
- **Libraries**: Book ordering.
- **Insurance**: Identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds.
- **City-planning**: Identifying groups of houses according to their house type, value and geographically location.
- **Earthquake studies**: Clustering observed earthquake epicenters to identify dangerous zones.
- **WWW**: Document classification; clustering web log data to discover groups of similar access patterns.
A Multiagent System (MAS) for the Generation of Initial Centroids for k-means Clustering Data Mining Algorithm based on Actual Sample Datapoints
Dost Muhammad Khan, Nawaz Mohamudally

- **Medical Sciences**: Classification of medicines; patient records according to their doses etc. [5].

In this research paper will apply this k-means clustering algorithm on a medical dataset ‘Diabetes’ [6][7]. The data is collected only from the female gender with a dataset/testbed of 1500. Before applying k-means clustering algorithms on this dataset, the data is pre-processed, called data standardization. The interval scaled data is properly cleansed by applying the range method [4]. The attributes of the dataset/testbed ‘Diabetes’ are:

- a. Number of times pregnant (min. age = 21, max. age = 81). The table 1 below shows the relationship between age and number of times pregnant:

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Number of times pregnant</th>
</tr>
</thead>
<tbody>
<tr>
<td>21 – 29</td>
<td>1</td>
</tr>
<tr>
<td>30 – 39</td>
<td>2</td>
</tr>
<tr>
<td>40 – 49</td>
<td>3</td>
</tr>
<tr>
<td>50 – 59</td>
<td>4</td>
</tr>
<tr>
<td>60 – 69</td>
<td>5</td>
</tr>
<tr>
<td>70 – 79</td>
<td>6</td>
</tr>
<tr>
<td>80 – 89</td>
<td>7</td>
</tr>
</tbody>
</table>

b. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
c. Diastolic blood pressure (mm Hg)
d. Triceps skin fold thickness (mm)
e. 2-Hour serum insulin (mu U/ml)
f. Body mass index (weight in kg/(height in m)^2)
g. Diabetes pedigree function
h. Age (years)
i. Class (1: +ive for diabetes, 2: -ive for diabetes) [6][7].

The value of ‘k’, number of clusters is set to maximum number 7 and the number of iterations ‘n’ in each case is 50 i.e. value of k=7 and value of n=50. The sum of square error value ‘E’ is calculated using equation 2 for every cluster i.e. from k=1 to 7. These values of ‘E’ are then used to plot a graph against the number of clusters on 2D graph. It is observed that as the number of clusters increases, the value of sum of square value ‘E’ decreases.

3. **Initial Centroids Methods based on Actual Sample Datapoints**

In this section we will discuss the generation of initial centroids based on actual sample datapoints using Range Method, Random number Method, Outlier Method and Inlier Method.

3.1. **Range Method**

According to this method, the steps of k-means clustering algorithm are:

**Step 1**: Set value of ‘k’ to maximum number. k=7 for this dataset.

**Step 2**: Generate the ‘k’ initial centroids from the actual sample of dataset using equations 3 and 4. The seven initial centroids will be generated. Divide datapoints into ‘k’ clusters on the basis of these generated initial centroids.

**Step 3**: Move datapoints into clusters using Euclidean’s distance formula in equation 1. Recalculate new centroids on the basis of average or means.

**Step 4**: Repeat step 3 until no datapoint is to be moved.

The Range Method relates to equations 3 and 4.

\[ c_i = \frac{(\max X - \min X)}{k} \times n \text{ where } 1 \leq i \leq k \]  

(3)
The initial centroid is \( C(c_i, c_j) \). Where \( \text{maxX} \) and \( \text{minX} \) are maximum and minimum values of attribute \( X \), \( \text{maxY} \) and \( \text{minY} \) are maximum and minimum values of attribute \( Y \) respectively and ‘k’ represents the number of clusters. ‘k’ is always a positive and non zero integer. The values of ‘i’, ‘j’ and ‘n’ vary from 1 to ‘k’.

The value \( (\text{maxX} - \text{minX}) \) will provide the range of ‘X’ attribute, similarly the value \( (\text{maxY} - \text{minY}) \) will give the range of ‘Y’ attribute. If both the attributes have zero value then this formula will not work. The value of ‘k’ must be at least 1 if ‘k’ is zero then again it will give an error, the division by zero. The value of ‘n’ varies from 1 to ‘k’. The number of iterations should be small otherwise the time and space complexity will be very high and the value of initial centroids will also become very high and may be out of the range in the given dataset. This is a major drawback of k-means clustering algorithm.

Data standardization is the most important step for clustering. The Euclidean distance method requires the standardization of attributes so that all attributes can have equal impact on the distance computation. This is to avoid obtaining clusters that are dominated by attributes with the largest amounts of variation. The Range Method provides itself data standardization. The initial centroids are calculated by the division of ranges of the attributes and the number of clusters. This is an average or means of an attribute. The graph is plotted between the number of clusters ‘k’ and sum of square error value ‘E’ using this method shown in figure 3.

![Figure 3](image.png)

**Figure 3.** Graph between number of clusters ‘k’ and sum of square error value ‘E’ using Range Method

The graph shows that sum of square error value ‘E’ decreases as the number of clusters increases. This is due to the selection of initial centroids using Range Method.

### 3.2. Random number Method

This is the most popular method used to generate the initial centroids from the given actual sample datapoints. According to this method, the step 2 of k-means clustering algorithm will change, the other three steps, step 1, step 3 and step 4 remain the same: 

**Step 2:** Generate the ‘k’ initial centroids from the actual sample of dataset using equations 5 and 6. The seven initial centroids will be generated. Divide datapoints into ‘k’ clusters on the basis of these generated initial centroids.

\[
c_i = \text{random}(X) \quad \text{where} \quad 1 \leq i \leq k
\]  

(5)
A Multiagent System (MAS) for the Generation of Initial Centroids for k-means Clustering Data Mining Algorithm based on Actual Sample Datapoints
Dost Muhammad Khan, Nawaz Mohamudally

\[ c_j = \text{random}(Y) \quad \text{where} \quad 1 \leq j \leq k \]  

(6)

The initial centroid is \( C(c_i, c_j) \). The values of ‘i’, and ‘j’ vary from 1 to ‘k’. The graph is plotted between the number of clusters ‘k’ and sum of square error value ‘E’ using this method shown in figure 4.

![Random number Method](image)

**Figure 4.** Graph between number of clusters ‘k’ and sum of square error value ‘E’ using Random number Method

The graph shows that starting value of sum of square error value ‘E’ is higher as compared to graph in figure 3. This is due to the selection of initial centroids using this method. As the number of clusters increases, the value of sum of square error value ‘E’ decreases.

### 3.3. Outlier Method

This method is based on the outer or maximum values of given dataset as initial centroids. The outlier is the maximum value of an attribute in actual sample datapoints. According to this method, the step 2 of k-means clustering algorithm will change, the other three steps, step 1, step 3 and step 4 remain the same:

**Step 2:** Generate the ‘k’ initial centroids from the actual sample of dataset using equations 7 and 8. The seven initial centroids will be generated. Divide datapoints into ‘k’ clusters on the basis of these generated initial centroids.

\[ c_i = \max(X) - i \quad \text{where} \quad 0 \leq i \leq k \]  

(7)

\[ c_j = \max(Y) - j \quad \text{where} \quad 0 \leq j \leq k \]  

(8)

The initial centroid is \( C(c_i, c_j) \). The values of ‘i’ and ‘j’ vary from 0 to ‘k’. The graph is plotted between the number of clusters ‘k’ and sum of square error value ‘E’ using this method shown in figure 5.
The graph shows that starting value of sum of square error value ‘E’ is higher as compared to graph in figures 3 and 4. This is due to the selection of initial centroids using this method. This method uses the maximum values of attributes every time. As the number of clusters increases, the value of sum of square error value ‘E’ decreases.

3.4. Inlier Method

This method is based on the inner or minimum values of given dataset as initial centroids. The inlier is the minimum value of an attribute in actual sample datapoints. According to this method, the step 2 of k-means clustering algorithm will change, the other three steps, step 1, step 3 and step 4 remain the same:

Step 2: Generate the ‘k’ initial centroids from the actual sample of dataset using equations 9 and 10. The seven initial centroids will be generated. Divide datapoints into ‘k’ clusters on the basis of these generated initial centroids.

\[ c_i = \min(X) - i \text{ where } 0 \leq i \leq k \]  \hspace{1cm} (09)

\[ c_j = \min(Y) - j \text{ where } 0 \leq j \leq k \]  \hspace{1cm} (10)

The initial centroid is \( C(c_i, c_j) \). The values of ‘i’ and ‘j’ vary from 0 to ‘k’. The graph is plotted between the number of clusters ‘k’ and sum of square error value ‘E’ using this method shown in figure 6.

---

**Figure 5.** Graph between number of clusters ‘k’ and sum of square error value ‘E’ using Outlier Method

**Figure 6.** Graph between number of clusters ‘k’ and sum of square error value ‘E’ using Inlier Method
The graph shows that starting value of sum of square error value ‘E’ is smaller as compared to graph in figures 3, 4, and 5. This is due to the selection of initial centroids using this method. This method uses the minimum values of attributes every time. As the number of clusters increases, the value of sum of square error value ‘E’ decreases.

The graph between number of clusters ‘k’ and sum of square error value ‘E’ is shown in figure 7 using all the four methods. This summarizes all the four methods discussed in this section.

![Figure 7. Graph between number of clusters ‘k’ and sum of square error value ‘E’ for all the four methods](image)

The graph shows that starting value of sum of square error value ‘E’ for the value of k=1, using the Outlier method is higher than the other three methods, Random number, Range and Inlier. As the value of ‘k’ increases, the value of sum of square error value ‘E’ decreases by using all these four methods. The graph also shows that the convergence is achieved after k=2 in the Range method and k=3 in the Random number method. The other two methods, Outlier and Inlier are converged after k=7. This shows that the starting value of initial centroids in k-means clustering algorithm is very important.

4. Results and Discussion

The problem with Inlier Method is that as number of clusters increases the starting value of sum of square error value ‘E’ increases as shown in figure 8. The starting value of ‘E’ is low at k = 1 and increases as the value of k increases. Another issue in this method is that the value of starting initial centroids can be zero and negative if the attribute of given dataset have zero value. This method can perform better if the attributes have smaller sets of values.

![Figure 8. Graph between number of clusters ‘k’ and starting value of ‘E’ for each cluster using Inlier Method](image)
The graph shows that the convergence is not yet occurred after the value of \( k=7 \) and \( n=50 \). This shows that the selection of starting initial centroids is highly determinant in k-means clustering algorithm.

The problem with Outlier Method is that the starting value of ‘E’ is always very high due to the maximum values of attributes of the given dataset as starting initial centroids. Due to the high values of initial centroids, the convergence in this method takes longer time as compared to the other methods. This method cannot perform better on any attributes due to its higher values of initial centroids as starting point. This is illustrated in figure 9.

![Outlier Method](image)

**Figure 9.** Graph between number of clusters ‘k’ and starting value of ‘E’ for each cluster using Outlier Method

The graph shows that the convergence is not yet occurred after the value of \( k=7 \) and \( n=50 \). This shows that the selection of starting initial centroids plays an important role in k-means clustering algorithm.

The problem with the Random number Method is that the value of starting initial centroids can be same for all clusters. The calculated distances between the attributes will be the same. The convergence will take extra clock cycles which increase the number of iterations as well. This is shown in figure 10.

![Random Number Method](image)

**Figure 10.** Graph between number of clusters ‘k’ and starting value of ‘E’ for each cluster using Random number Method

The graph shows that the convergence is occurred after the value of \( k=5 \) and \( n=50 \). It is faster than then Outlier and Inlier methods. This shows the how starting initial centroids are important in k-means clustering algorithm.
A Multiagent System (MAS) for the Generation of Initial Centroids for k-means Clustering Data Mining Algorithm based on Actual Sample Datapoints
Dost Muhammad Khan, Nawaz Mohamudally

The Range Method is the best from the other methods discussed in this paper. The starting value of sum of square error ‘E’ is low as compared to the Outlier, Inlier and Random number methods. The value further decreases as the value of number of clusters ‘k’ increase. Another advantage of this method is it provides data standardization. The starting initial centroids using the Range Method is the best selection and save the convergence time. This is illustrated in figure 11.

![Figure 11](image1.png)

**Figure 11.** Graph between number of clusters ‘k’ and starting value of ‘E’ for each cluster using Range Method

The graph shows that the convergence is occurred after the value of k=2 and n=50, much faster than Outlier, Inlier and Random number methods. This also proves that the starting initial centroids play an important and pivotal role in k-means clustering data mining algorithm.

![Figure 12](image2.png)

**Figure 12.** Graph of number of clusters ‘k’ and starting value of ‘E’ for all the four methods

The figure 12 shows the graph between starting value of sum of square error value ‘E’ and number of clusters ‘k’ for all the four methods. Graph shows that the starting value of sum of square error ‘E’ is very high for Outlier Method. In case of Random number Method, the value is low for k=1 and it becomes high for k=2. As the value of k increases, the starting value of ‘E’ decreases in Random number Method after k=2. It again becomes high after k=5, this is a convergence point. The starting value of ‘E’ increases as the value of ‘k’ increases in Inlier Method. The starting value of ‘E’ is realistic in Range Method. It decreases with the increase in the number of clusters ‘k’. The starting sum of square error value ‘E’ again increases at k=4, this shows the convergence at this point.

The starting value of sum of square error value ‘E’ should decrease with the increase of the value of number of clusters ‘k’. This depends on the selection of initial centroids, if the selection is wrong...
then the value of sum of square error ‘E’ does not decrease quickly. This proves that the selection of initial centroids is important in k-means clustering algorithm.

5. Conclusion

In this paper we present four different methods of generation of initial centroids based on actual sample datapoints. Random number Method is a mature and the popular method. The other three methods Range, Outlier and Inlier are for the first time introduced in this research paper. The k-means clustering is a common algorithm to find clusters within given dataset. The selection of initial centroids has an effect on the results of this algorithm. We draw a conclusion that the starting value of initial centroids neither should be very high as in the case of Outlier Method nor should be very low as in the case of Inlier Method. The value must be between high and low, which can be achieved with the use of Range Method. The conclusion is that the Range Method performs better than the other methods based on actual sample datapoints. The preferred method for choosing initial centroids based on actual sample datapoints is therefore, the Range Method for the k-means clustering algorithm.

6. References