An Agent Oriented Approach for Implementation of the Range Method of Initial Centroids in K-Means Clustering Data Mining Algorithm

Dost Muhammad Khan, Nawaz Mohamudally

Abstract

Intelligent agents are today accepted as powerful tools for data mining in a distributed environment. Artificial Intelligence (AI) algorithms are utilized to render software agents more intelligent. Knowledge discovery and data mining methods have been applied to discover hidden patterns and relations in complex datasets where clustering is one of the most important approaches used for this purpose. There exists a number of clustering algorithms among which the K-means algorithm is commonly used to find clusters due to its simplicity of implementation and fast execution. It appears extensively in the machine learning literature and in most data mining suite of tools. The algorithm is significantly sensitive to the selection of initial centroids. In this paper, we will present an agent oriented approach for implementation of the Range Method of initial centroids in K-means data mining algorithm. This Range Method is based on the actual sample datapoints. We have tested this method with both Euclidean and City Block (Manhattan) distances formulae on a different number of real life datasets.

Keywords: Centroids, Datapoints, Range Method

1. Introduction

In this section we will discuss an agent and its functions, how it becomes intelligent and the association of agents with the artificial intelligence and data mining. The agent can be defined as: "An Agent is a software thing that knows how to do things that you could probably do yourself if you had the time" (Ted Seller of IBM Almaden Research Center). "A piece of software which performs a given task using information gleaned from its environment to act in a suitable manner so as to complete the task successfully. The software should be able to adapt itself based on changes occurring in its environment, so that a change in circumstances will still yield the intended results." (With thanks to G.W. Lecky – Thompson for the definition) [1][2][3][4]. An intelligent agent can be further divided into a weak notation and a strong notation. Table 1 shows the properties for both notations.

<table>
<thead>
<tr>
<th>Weak notation</th>
<th>Strong notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autonomy</td>
<td>Mobility</td>
</tr>
<tr>
<td>Social ability</td>
<td>Benevolence</td>
</tr>
<tr>
<td>Reactivity &amp; Proactivity</td>
<td>Rationality</td>
</tr>
<tr>
<td>Temporal continuity</td>
<td>Adaptivity</td>
</tr>
<tr>
<td>Goal oriented</td>
<td>Collaboration</td>
</tr>
</tbody>
</table>

The word intelligence refers to the ability of the agent to capture and apply domain specific knowledge and processing to solve problems. An intelligent agent uses knowledge, information and reasoning to take reasonable actions in pursuit of a goal. It must be able to recognize events, determine the meaning of those events and then take actions on behalf of a user. An intelligent agent, have the Knowledge Representation, Reasoning and Learning [5]. An intelligent and autonomous agent has the properties like Perception, Reasoning and Action which also form the life cycle of an agent as shown in figure 1[6][7].
Knowledge Discovery in Databases (KDD) and Information Retrieval (IR) are today determinant to access the right information in a global world, just to mention the importance in tracking of epidemic diseases and movement of suspected criminals. Knowledge discovery and data mining methods have been applied to discover hidden and new patterns and relations in complex datasets [8]. Data mining is a statistical term but in Information Technology (IT), it is: “Discovery of useful summaries of data” [9]. Data mining is known as “Machine Learning” in Artificial Intelligence (AI). Machine Learning is an area of Artificial Intelligence (AI), concerned with the development of techniques which allow the computer to ‘learn’. It is a method of creating computer programs by the analysis of datasets. The agents must be able to learn to do classification, clustering and prediction using learning algorithms. There is an explicit link between Artificial Intelligence (AI) and Intelligent Agents [10][11].

In section 2 we present the K-means clustering algorithm using the Range Method. Section 3 is the methodology; deploying the K-means clustering algorithm in an agent. In section 4 we discuss the results and discussion, finally section 5 presents the conclusion.

2. K-means Clustering Algorithm using the Range Method

The algorithm has as input a predefined number of clusters that is the ‘K’ from its name. The ‘means’ stands for an average, an average location of all the members of a particular cluster. The K-means algorithm is an iterative procedure where by final results depend on the values selected for initial centroid. A Centroid is an artificial point in the space of records which represents an average location of the particular cluster. The coordinates of this point are averages of attribute values of all examples that belong to the cluster. Finding useful pattern in large datasets has attracted considerable interest recently and of the most widely studied problems in this area is the identification of clusters or densely populated regions, in a multi-dimensional dataset. Given the desired number of clusters ‘K’ and a dataset of ‘N’ points and a distance-based measurement function. The major problem with the K-mean algorithms is that of choice of initial centroids. The algorithm depends upon this selection. If the selection is correct or accurate then the result will be fine otherwise it may fail. It is most commonly used algorithm, using a square-error criterion. It has two primary steps first, the assignment step where the instances are placed in the closest class and the second one, re-assignment step where the class centroids are recalculated from the instances to the class [12][13][14].

Figure 2 shows the basic steps as a flow chart of K-means clustering algorithm using the Range Method for the generation of initial centroids.
The interpretation of the flow chart is as follow:

**Step 1:** Enter the number of clusters and number of iterations, which are the required and basic inputs of the K-means clustering algorithm.

**Step 2:** Compute the initial centroids by using the Range Method shown in equations 1 and 2 below.

\[ c_i = ((\text{max} X - \text{min} X) / k) * n \]  
\[ c_j = ((\text{max} Y - \text{min} Y) / k) * n \]  

The initial centroid is \( C(c_i, c_j) \). Where: max X, max Y, min X and min Y represent maximum and minimum values of X and Y attributes respectively. ‘k’ represents the number of clusters and i, j and n vary from 1 to k where k is an integer. In this way, we can calculate the initial centroids; this will be the starting point of the algorithm. The value (maxX – minX) will provide the range of ‘X’ attribute, similarly the value (maxY – 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 the 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.

**Step 3:** Calculate the distance either by Euclidean’s distance or City Block (Manhattan) distance formulae (equations 3 and 4). On the basis of these distances, generate the partition by assigning each sample to the closest cluster.

**Euclidean Distance Formula:**

\[ d(x_i, x_j) = \left( \sum_{k=1}^{m} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}} \]  

**Manhattan Distance Formula:**

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

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 and j vary from 1 to N where N is total number of attributes of a given object. i,j and N are integers.

**Step 4:** Compute new cluster centers as centroids of the clusters, again compute the distances and generate the partition. Repeat this until the cluster memberships stabilizes [14].

There are different techniques available to use the initial points (it is required as an input). The followings are the different methods to calculate the initial points:

- **Corner:** In this method all the values in the data sets are scaled to be in [-1, 1], the set of all the clusters close to the vertices (-1,…,-1) is considered. This is usually a ‘bad’ set of initial points since it lies on the boundary of the data, and can be considered an outlier.
- **Bins:** This method consists in divide the space in bins and then takes random points inside each bin, this assures that the set of initial points are distributed covering the entire dataset.
- **Centroid:** This method consists of choosing all the starting clusters close to the mass centroid of the dataset. Each cluster center is calculated adding a small random perturbation to the centroid of the dataset.
- **Spread:** The cluster centers are distributing randomly and trying to cover the entire space. This method is similar to bins.
- **PCA:** The data points are projected in the space of the principal component, a clustering procedure is applied to this one-dimensional set. The cluster centers are calculated depending on the obtained clusters in the one-dimensional space [15].

Among the above mentioned methods, in order to calculate the initial points, we opt for the ‘centroid’ method because it has been advocated throughout the literature that the centroid method produces the best results.
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 hyperspheres.

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

3. Methodology: Deploying K-means Clustering Algorithm in an Agent

The agent for the K-means clustering algorithm using the Range Method of initial centroids is developed under ‘kaariboga framework’, an experimental Java-based mobile agent framework [16][17]. Choosing a clustering algorithm, however, can be a difficult task, even finding just the most relevant approaches for a given dataset is not obvious. Finding a good similarity function depends strongly on the dataset and the determinant elements are the nature of the data and the desired number of clusters. Another element is the type of input and tools that the algorithm requires because some algorithms only handle the numeric inputs and others categorical. For simplicity the tests were performed on datasets which have numerical attributes values. The figure 3 depicts the architecture of this agent system using the K-means clustering algorithm.

![Figure 3. The Architecture of an Agent System](image-url)
The chosen datasets are:

- **Iris dataset**: This dataset contains information about the flowers. There are 3 different classes of flowers. It has 150 records with 5 attributes. After applying the process of normalization, only 100 are selected.
- **Vote dataset**: This dataset is about the polls in USA. There are 2 different classes of polls i.e. Republicans or Democrats. There are total 300 records and 17 attributes. We normalize this dataset and select the same number of records for clustering.
- **Breast dataset**: This is a medical dataset, containing information about the Breast’s diseases. There are 2 different classes of Breast’s diseases. It has total 700 records and 11 attributes. After applying the normalization process, only 233 records are selected for clustering.
- **Diabetes dataset**: This is also a medical dataset, containing information about Diabetes’ diseases. We have divided this dataset into 5 different classes. It has total 513 records and 8 attributes. Similarly, we normalize this dataset and select only 256 records for clustering [18].

We created the vertical partitions of each datasets by selecting the proper number of attributes which are suitable for the K-means clustering algorithm. The clusters are created on the basis of the number of classes in these four datasets i.e. for dataset ‘iris’ 3 clusters, for dataset ‘vote’ 2 clusters, for dataset ‘breast’ 2 clusters and for dataset ‘diabetes’ 5 clusters; which is also the value of ‘k’ in this algorithm. Another required input is the number of iterations, which is ‘n’, for these datasets we used ‘n = 100’. There is no predefined rule, how many clusters will be created and what will be the number of iterations. It always depends upon the user. If the result is not according to the requirements, the values of ‘k’ (number of clusters) and ‘n’ (number of iterations) can be changed every time until the required and better results are found. This is a weakness of this algorithm. Once the clusters are discovered, they have to be interpreted in order to have some value. There are different ways to utilize clustering results; clusters membership can be used as a label for the separate classification problems, some descriptive data mining techniques like ID3 [19] (decision rules) can be used to find descriptions of clusters and clusters can be visualized using 2D or 3D scattered graphs. We have used 2D scattered graphs and ID3 for visualizing and interpreting the results of the K-means clustering algorithm.

### 4. Results and Discussion

The K-means clustering algorithm is tested on both peer-to-peer and client-server network, using Euclidean’s and City Block distance formulae. There is no difference of the results obtained from both modes. Table 2 shows the results of all the four datasets using both the distance formulae.

**Table 2. The Results of all the four datasets**

<table>
<thead>
<tr>
<th>K-Means/Datasets</th>
<th>Iris</th>
<th>Vote</th>
<th>Breast</th>
<th>Diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean’s distance</td>
<td>C1=38, C2=50, C3=12</td>
<td>C1=181, C2=119</td>
<td>C1=142, C2=91</td>
<td>C1 = 52, C2 = 26, C3 = 11, C4 = 91, C5 = 76</td>
</tr>
<tr>
<td>City-Block (Manhattan) distance</td>
<td>C1=27, C2=38, C3=35</td>
<td>C1=115, C2=185</td>
<td>C1=178, C2=55</td>
<td>C1 = 83, C2 = 11, C3 = 58, C4 = 81, C5 = 23</td>
</tr>
</tbody>
</table>

We are demonstrating the results obtained from the dataset ‘Iris’ in table 3.
We have noticed in all the four datasets the results obtained from the Range Method are satisfactory and consistent. So, the further tests have been demonstrated, the Range Method is transcendental i.e. independent of size of the data.

For the visualization of these obtained clusters we will first apply scattered graph technique only on dataset ‘iris’, by taking ‘petal_length’, ‘petal_width’ and ‘class’ attributes of this dataset. Figure 4 shows a scattered graph of all data points in ‘iris’ dataset before applying the K-means algorithm.

![All data points 'Iris' dataset](image)

Figure 4. 2D Scattered Graph for the ‘Iris’ a flower dataset

We can see there is a high density of data situated between 0 and 100. We will now apply K-means algorithm by using the Range Method of initial centroids and create 3 clusters for this dataset (after examining ‘iris’ dataset there are only 3 classes) with 100 iterations. Scattered Graphs shown in figures 5, 6 and 7 of each clusters are obtained by using City-Block (Manhattan) distance formula.

![City Block's distance cluster 1](image)

Figure 5. 2D Scattered Graph of cluster1 of dataset ‘Iris’ (City Block)

The distance is constant between ‘petal_length’ and ‘petal_width’ for all data points.
An Agent Oriented Approach for Implementation of the Range Method of Initial Centroids in K-Means Clustering Data Mining Algorithm
Dost Muhammad Khan, Nawaz Mohamudally

Figure 6. 2D Scattered of cluster 2 of dataset ‘Iris’ (City Block)

As compared to the first cluster in figure 5, the distance between ‘petal_length’ and ‘petal_width’ is variable. This cluster has a structure close to figure 4 before applying K-means clustering algorithm.

Figure 7. 2D Scattered of cluster 3 of dataset ‘Iris’ (City Block)

The distance between ‘petal_length’ and ‘petal_width’ is larger as compared to the other two clusters in figures 5 and 6. It is more or less constant up to ‘petal_length’ 20 and then it varies. There is a linear combination among these three clusters in figure 5, 6 and 7 to form the original dataset ‘iris’ before clustering shown in figure 4. Scattered Graphs shown in figures 8, 9 and 10 of each cluster are obtained by using Euclidean’s distance formula.

Figure 8. 2D Scattered of cluster 1 of dataset ‘Iris’ (Euclidean’s distance)

There is almost a constant distance between ‘petal_length’ and ‘petal_width’ up to 20 and then it varies.
As compared to the first cluster in figure 8, the distance between ‘petal_length’ and ‘petal_width’ is variable. There is a constant trend between the distances after ‘petal_length’ 40. This cluster has a structure more or less close to figure 4 before applying K-means clustering algorithm.

There is a variable distance between ‘petal_length’ and ‘petal_width’ in this cluster. We applied ID3 on each cluster in order to interpret the results obtained from the clusters. The decision rules for cluster 1 of ‘iris’ dataset using Euclidean’s distance are shown in figure 11.

```
if petal_width = 0   then  Class = Irissetosa
else
  if petal_width = 2 then Class = Irissetosa , Irisversicolor
  else
    if petal_length = 3   then Class = Irisversicolor
    else
      if petal_length=1 then Class = Irisvirginica, Iriscol
      else
        Class = Irisvirginica
```

**Figure 11.** Decision rules for cluster 1 of dataset ‘iris’ (Euclidean’s distance)

There are four decision rules for cluster 1 of dataset ‘iris’ after applying Euclidean’s distance formula. The decision rules for cluster 1 of ‘iris’ dataset using City Block’s distance are shown in figure 12.

```
if petal_length = 1 and  petal_width = 0  then 
  Class= Irissetosa
```

**Figure 12.** Decision rules for cluster 1 of dataset ‘iris’ (City Block’s distance)

There is only one decision rule for cluster 1 of dataset ‘iris’ after applying City Block’s distance formula because the distance between ‘petal_length’ and ‘petal_width’ is constant as shown in figure 5.

Similarly the decision rules can be found for the other clusters of this dataset and the number of rules will vary from one cluster to another cluster depending on the values of data points in that cluster. As distribution of data points within a cluster depends upon the choice of initial centroids, therefore it is
An Agent Oriented Approach for Implementation of the Range Method of Initial Centroids in K-Means Clustering Data Mining Algorithm
Dost Muhammad Khan, Nawaz Mohamudally

paramount importance to apply the most appropriate values of initial centroids.

5. Conclusion

In this paper, we discussed the K-means clustering algorithm using centroids method with the Range Method of initial centroids. The Range Method of initial centroids is based on the actual sample datapoints. It is transcendental i.e. independent of size. We tested this method on four different datasets with both distance formulae (Euclidean and Manhattan). 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 nor should be very low. The value must be between high and low, which can be achieved with the use of the Range Method. The selection of distance formula has no impact on the K-means clustering algorithm, the number of records per cluster obtained for all the four datasets by using both distance (Euclidean and City Block) formulae are almost the same. The results obtained from the Range Method are satisfactory and consistent. Our methodology is a prototype tested on peer-to-peer and local distributed networks. The study could be extended to large scale distributed databases so as to validate the effectiveness of the proposed methodology. For further investigation in this direction, one will undoubtedly has to take into account the parameters such as data caching and the validity of the agent framework.

6. References

