Novel Solutions for Initial Cluster Means Problem for K-Mean Algorithm

Faijalurahaman khan *1 , Virendra kumar*2 , Savita Shiwani *3
M.tech Scholar *1, Assistant Professor*2, Associate Professor*3
*1,2,*3 Suresh Gyan Vihar University, Jaipur India

Abstract

K-means algorithms are used in various applications there are some categories of the clustering algorithm.

K-means algorithm is part of the one of clustering method partitioning. K-means algorithm generally is used for the grouping of the data, and it is use in various fields like Gene data, banking data and some other fields due to its easiness and simplicity for the same purpose.

As K-means algorithm is simple, it has some disadvantages like, it takes initial cluster centre arbitrarily and also we need to predetermine number of cluster formed in the data set which is practically not possible it performance also decreases with the large data set.

In this paper we worked upon the initial cluster centre problem of the k-means algorithms, in our introduce approach initial cluster centre are taken by using the M-Way tree method by using the M-way tree approach it gives nearly the exact centre of the cluster.

This method is tested upon the synthetic data; it takes less time and iteration in clustering the data as compare to the K-means algorithms.

Keywords

Initial cluster centre Clustering, K-means, M-way tree, Knowledge-Discovery.

I. Introduction

As the size of the data repository system or databases increasing so rapid speed, these data are being used in future for decision making purpose in business or some other fields, patterns for which expert system analyst or automated data mining system is looking may be presented in these databases [1].

Knowledge discovery system are used to get the hidden information in the data, classification and clustering are the methods of the knowledge discovery, classification is a method in which group or class are predefined in which data will be classified, in classification there are two simple steps, one step is learning and other step is the decision making step.

One major difference between the Clustering and classification is that in classification we have information regarding the classes we are going to form vice-versa in the clustering process, clustering also known as the unsupervised method of the grouping of the data, in this method data are classified on the basis of the some distance measure, in this data are classified in those group to which cluster centre data points are nearer, clustering are applied in various fields like Banking, Gene expression and some other fields.

There are some clustering methods, K-means algorithms is the part of the one of the clustering methods which is known as the partitioning method, this algorithms are applied in various fields due to its simplicity and efficiency.

K-means algorithm is used in the fields of the medicine and gene data since long time, as the K-means algorithm is being used in the fields of medicine and gene data in both fields data size as well as the dimension is increasing due to which K-means performance decreases because traditional K-means algorithms is not capable of dealing with the large data set as the size of the data set and dimension is increasing traditional K-means performance decreases drastically, paper [2] and [3] deal with this problem of K-means algorithm.

Other problem of the K-means algorithms is that it performance goes down when it applied at the data set which contain missing value data, missing data value problem arises due to the equipment unavailability and not proper functioning of the equipment, due to this an accurate and efficient result does not come.

Missing values generally deal by the following three methods:

- Ignoring methods.
• Parameter estimation.
• Imputation.

Missing value problem is addressed in the paper [4].

Another problem with the K-means algorithm is that, this algorithms performed the clustering only on the basis of data set or instances but it does not consider the background information this problem is addressed in the paper [5].

Another problem with traditional K-means algorithms is that, it gives same weight to each feature, as every feature got the same importance due this some feature which are redundant in nature degrade the cluster quality, this problem is addressed in the paper [6].

In this paper, we worked upon the main problem of the K-means algorithms, initial cluster centre problem, K-means algorithm takes initial centre or mean from a set of data objects arbitrarily, and other data points are clustered according those centre or mean, means of each cluster calculated again-and-again until the K-means algorithm does not get the good quality cluster or mean stop deflected, in our approach we took the initial cluster centre user specified, we took the data in the three row and idea of picking initial cluster centre are taken from the M-way tree, we inserted data in each row odd numbers and shorted the data and then initial cluster centre are picked by applying the M-way tree approach this approach is better than the traditional K-means algorithm.

Rest of the paper is organised as follows, in section II we have given brief literature regarding the K-mean algorithm various problem and its solution, in section III we have briefly describe the K-means algorithms, in section IV we have briefly explain our approach, in section V we explain the experimentation and result and finally we have described in section VI conclusion and future work.

II. Background Study

One of the main concern of the K-means algorithm, is it takes initial cluster centre arbitrarily and number of cluster are taken from the user, paper [7-8] address the problem of taking number of cluster from user.

Paper [7] proposed a method, which provides two options one is to enter the number of cluster from the user and other is the fix the minimum number of the cluster, this algorithm provide better solution but this approach has a limitation computation complexity of this method increase.

Paper [8] due to simplicity of K-means algorithm employed in various field to address the problem of the K-means algorithm of number of cluster entered by the user one method which is based upon the KMBA, K-means using bat feather echolocation behaviour are used but this algorithms does not guarantee of the unique prediction of the K values.

In the paper [8], an approach is proposed on the basis of the hill climbing algorithms this algorithm is known as the modified hill climbing aided K-means algorithm.

This algorithm is better than the KMBA as well as ordinary K-means algorithm.

In this paper we are dealing with the initial cluster centre problem various research work has done to deal this problem of K-means algorithms.

Paper [9-12], address this problem, in the paper[9], they proposed a method which is based upon the observation that the re-assignment step which takes place again and again, those data point which change the cluster are at boundary of the cluster to which this data point are re-assigned.

In this algorithm they use a multiple random spatial tree, practical analysis shows that this method is more efficient methods as compare to the K-means algorithm.
In the paper [10], they proposed a method in which they use an auxiliary function which applied to get best initial cluster points after that they apply other step of the K-means algorithm, it is efficient as compare to the traditional K-means algorithm.

In the paper [11], to eliminate the initial cluster centre problem, they proposed a global clustering method, in this method to get the best initial cluster centre they use a greedy elimination method which is known as GEM.

In the paper [12], they proposed a method which is the combination of the K-means and hierarchical clustering; in this approach they use K-means algorithm for generating the initial cluster centres after that they applied hierarchical clustering, efficiency of K-means algorithm increases by applying this method.

Various modifications have been done regarding K-mean algorithm initial cluster centre problem, but still there is not any single globally recognised method which is used for initialization of the initial cluster centre in the K-means algorithm.

III.K-means Algorithms

K-means algorithm comes under the partitioning division of the one of the clustering method, K-means algorithm works as follows, initially it takes number of cluster from the user and pick initial cluster centre any arbitrary data points from the data set, and cluster the data in this arbitrary initialize cluster on the basic of distance similarity measure, data which are nearer any arbitrarily initialize cluster centre are become the part of that cluster and after that calculation of mean of each cluster is done, new mean of each cluster become the new cluster centre and reassignment of these data points again will take place with same distance measure this step will continue till the mean of cluster stop deflecting, as the k-means algorithms takes initial cluster centre arbitrarily it performance decrease because it may pick any arbitrary data due to which it iterate again and again to get the exact or optimal data cluster.

K-means algorithms step are as follows

1. In the first step initial cluster or means in K-means algorithms are taken arbitrarily.
2. In the second steps distance are calculated from each data points to the each initial cluster centre data points are assign to those initial cluster centre to which calculated distance is minimum.
3. Means of each cluster are calculated, now that calculated means become the new initial cluster centre.
4. Again the distance from each data point and each cluster centre will be calculated and new means will be generated again and again.
5. Step 3rd and 4th will be repeated again and again and it will stop when equilibrium will obtain. Process will be terminated when equilibrium will be obtained.

IV. Modified Approach

in our approach we have taken synthesised data for experimental purpose, in our data set we have 153 data points and we manually shorted these data points row wise and each row have 51 data points for obtaining the initial cluster centre we applied the M-way tree approach.

Suppose our data-set are as follows

<table>
<thead>
<tr>
<th>Data row 1</th>
<th>2</th>
<th>5</th>
<th>7</th>
<th>8</th>
<th>11</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data row 2</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>25</td>
<td>29</td>
<td>30</td>
<td>32</td>
<td>33</td>
</tr>
</tbody>
</table>

For obtaining the initial cluster centre we initially suppose each rows as a cluster, to obtain the initial cluster centre, we follow M-way tree approach which is use to get the route of the tree.
First initial cluster centre = Number of element in a row /2

We consider this division as an integer division so the output of this method will be 4, so we will take first initial cluster centre as the element whose position is at the fifth place. So by this approach, first initial cluster centre will be 11, next initial cluster centre will be obtained by the same approach, when initial cluster centre will be obtained, after obtaining the initial cluster centre from this approach, K-means algorithms other step will be performed as the same, this approach is mathematically and experimentally shows better result as compare to the traditional K-means algorithms.

Modified algorithms steps are as follows

1. Initially data are sorted manually.
2. Initial cluster centre will be taken by using M-way tree approach.
3. In the second steps distance are calculated from each data points to the each initial cluster centre, data points are assign to those cluster centre to which cluster centre calculated distance is minimum.
4. Means of each cluster are calculated, that calculated means become the new initial cluster centre.
5. Again the distance from each data point to each cluster centre will be calculated and new means will be generated again and again.
6. Step 3rd and 4th will repeat again and again and it will stop when equilibrium will obtain.
7. Process will be terminated when equilibrium will be obtained.

Modified approach is only different from traditional K-means algorithm in traditional K-means approach initial cluster centre are taken arbitrarily while in Modified approach initial cluster centre are taken by using the concept of M-way tree.

V. Experimental Result

For validating the efficiency of the our modified approach, we have taken the synthetic data set, we have implemented and compare our modified approach with the K-means algorithm in the NETBEANS tool in java programming language our approach is taking less time as well as the less iteration as compare to the traditional K-means algorithm maximum number of times.

Compare of traditional K-means and Modified Approach.

<table>
<thead>
<tr>
<th></th>
<th>Number of iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional K-means</td>
<td>13</td>
</tr>
<tr>
<td>Modified K-means</td>
<td>10</td>
</tr>
</tbody>
</table>

![Fig 1: Chart Comparison of Both Method](chart.png)

VI. Conclusion and Future Work

In this paper we propose a modified K-means clustering approach which is better solution as compare to the K-means algorithm, K-means algorithm one of the well organise and widely use algorithms in the various application like Genetic, Clinical, banking sector, main problem with the K-means algorithms is that it takes initial cluster centre as arbitrarily due to which the quality of the formed cluster is not good and also it takes more time so by our method it reduces the number of iteration due to which this approach is taking less time as compare to traditional K-means algorithm, some other method are inevitable which may also improve the efficiency of K-means algorithm regarding initial cluster centre.
References:


