International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 5 | ISSN : 2456-3307

Novel way of finding initial means in k-means clustering and validation using WEKA

Amit Mithal[1], Rohit Mittal[2]
[1]Department of Computer Science & Engineering, Jaipur Engineering College & Research Centre, Jaipur, Rajasthan, India
[2]Department of Computer Science & Engineering, Arya College of Engineering & Information Technology, Jaipur, Rajasthan, India

ABSTRACT

The work proposes a novel choice for the randomly chosen initial means in the k-means clustering. The dataset used for the implementations and validations is the Iris flowers dataset, which contains 150 labeled instances on 5 attributes of the three Iris species. In the k-means clustering, to find the proposed initial means, certain objects are found and eliminated in the clustering, which are very far away from the rest of the objects in their respective clusters. The centroid values of these reduced k clusters are then taken as the initial means in the k-means clustering. The results have shown that the number of iterations required by the algorithm is significantly lesser using the proposed initial chosen means.

Keywords: Clustering, k-means, Weka, Iris

I. INTRODUCTION

Clustering has its applications in many areas like data mining, statistical data analysis, compression etc. Clustering has been introduced in many aspects of machine learning, pattern recognition, optimization and statistics [1]. The basic problem in clustering is to group data instances which are similar in nature. The number of iterations required by the k-means clustering algorithm is greatly influenced by the choice of the initially chosen random means. The current work proposes a novel choice for the randomly chosen initial means in the k-means clustering. The dataset used in the work is the Iris flowers dataset which is taken from the UCI (University of California, Irvine at California, United States) machine learning repository. The dataset contains 150 labeled instances of the three Iris species.

The following are the objectives of the current work, which uses the Iris flowers dataset:

- Proposal of novel way of finding initial means in the k-means clustering
- Implementation of k-means clustering
- Implementation of k-means clustering with proposed initial chosen means
- Validation of k-means clustering using WEKA

Much of the work has been done for improving the efficiency, accuracy and stability of the k-means clustering algorithm. In [1], a method is proposed for computing a refined starting condition from a given initial one that is based on an efficient technique for estimating the modes of a distribution. In [2], the initial clustering number is the square root of the number of instances and the user is given in advance the number of clusters. In [3], the authors propose a method that helps in increasing the searching probability around the best centroid and is less sensitive to the initial centroid. In [4], a method is proposed to choose suitable initial points, which are well separated and have the potential to form high-quality clusters.

In [5], the authors use a kd-tree as the only major data structure and show that the algorithm runs faster as the separation between clusters increases. In [6], the authors state that new clustering algorithms and their results are often externally evaluated with respect to an existing class labeling which may not be adequate for the structure of the data or the evaluated cluster model.

The related research areas that have observed this problem are surveyed.

II. METHODOLOGY

A. K-means clustering

In the k-means clustering, k of the objects are selected randomly which initially represents the means of k number of clusters. The distance of all the objects is calculated from each of the means and then the objects are assigned to that cluster whose distance from the initial chosen mean is minimum. Finally, all the points or objects in the dataset will be assigned to k clusters. Then the centroid values of each of the k clusters are calculated. The whole process is repeated for more number of iterations. In the iterations, some of the objects can move from one cluster to another. The algorithm terminates as soon as the points stop moving from one cluster to another. The final centroid values for each of the new k clusters are then calculated.

Algorithm: k-means clustering
Assumptions:
n = total number of objects or points in the dataset
k = total number of means or partitions in the dataset, so the dataset will be divided into k clusters. Initially, let k = 2.
t = total number of iterations required by the algorithm
distance1 = distance of first mean to all the points
distance2 = distance of second mean to all the points

Input:
Normalized dataset of Iris flowers containing 150 labeled instances over 5 attributes.

Output:

1) Total number of iterations as required by the algorithm i.e. t.
2) Centroid values of cluster 1 and cluster 2.
3) Number of clustered instances for cluster 1 and cluster 2.

Procedure:
Step 1: Randomly choose initial k-means as any k points present in the dataset.
Step 2: Find the distance of all the points to every k-mean using Euclidean distance.
Step 3: Find the least distance corresponding to every point from the k-means to get distance1 and distance2.
Step 4: If distance1 is less than or equal to distance2, then associate point 1 with cluster 1, otherwise with cluster 2. Repeat this step for all the n points of the dataset.
Step 5: Find the new mean or centroid of each of the k-clusters.
Step 6: Repeat steps 2 to 5 until convergence has been reached i.e. until points do not move from one cluster to another cluster. (During step 4, some points can move from one cluster to another).

The above algorithm can be modified accordingly for k = 3, 4 and 5 to partition the dataset into the respective k clusters.

B. K-means clustering with proposed initial means

The number of iterations required by the k-means algorithm and hence the performance of the algorithm depends upon the initial random choice of the chosen k-centers or means. The performance of the algorithm is influenced by outlier data objects. Outliers are those points which are very far away from the rest of the points in the dataset. To find the improved initial means, the k-means algorithm is executed and the k-clusters and their centroids are found. Then, approximately one percent of the outlier points are found from the k-clusters as obtained, to get the k-outlier clusters. The improved k-clusters will be the k-clusters minus the k-outlier clusters. The centroid values of the improved k-clusters are found. These centroid values are then taken as the initial centers and the k-means algorithm is again executed. A significant improvement in the number of iterations required can be observed as in Table 3.1.

Algorithm: k-means clustering with proposed initial means

Assumptions:

n = total number of objects or points in the dataset
k = total number of means or partitions in the dataset, so the dataset will be divided into k clusters. Initially, let k = 2.
timproved = total improved number of iterations required by the algorithm.

Input:

1) Normalized dataset of Iris flowers containing 150 labeled instances over 5 attributes.
2) Clusters found in the k-means clustering.

Output:

1) Total improved number of iterations as required by the algorithm i.e. timproved.
2) Partitioned data objects into improved cluster 1 and improved cluster 2.
3) Improved initial mean 1 from improved cluster 1 and improved initial mean 2 from improved cluster 2.

Procedure:

Step 1: Use k-means clustering to partition the dataset into k clusters. Find one percent outlier points i.e. points which are very far away from each of these k-clusters.
Step 2: Now, the points in improved cluster 1 = points in cluster 1 - outliers in cluster 1 and the points in improved cluster 2 = points in cluster 2 - outliers in cluster 2.
Step 3: Find the improved initial mean 1 from improved cluster 1 and improved initial mean 2 from improved cluster 2. Choose these improved initial k-means and again execute the k-means algorithm, which can now be executed in lesser number of iterations.

C. K-means clustering using WEKA

Algorithm: k-means clustering using WEKA

Assumptions:
n = total number of objects or points in the dataset
k = total number of means or partitions in the dataset, so the dataset will be divided into k clusters. Initially, let k = 2.

Input:
Normalized dataset of Iris flowers containing 150 labeled instances over 5 attributes.

Output:

1) Total number of iterations as required by the algorithm i.e. t.
2) Centroid values of cluster 1 and cluster 2.
3) Number of clustered instances for cluster 1 and cluster 2.

Procedure:

Step 1: Load the normalized data file of Iris dataset in arff format in the Preprocessing panel of WEKA. This step is the training phase where a model is built or trained depending on the training instances.

Step 2: Under the cluster panel of WEKA, select SimpleKMeans. The value of k and the distance metric to be used can be specified in the numClusters and distanceFunction respectively.

Step 3: Select Test option or Cluster mode as "Use Training Set". Click on start and the results will be obtained.

III. RESULTS

The results of k-means clustering, k-means clustering using proposed initial chosen means and validation using WEKA for k = 2 are compared in the Table 3.1.

Table 3.1: k-means clustering results for k = 2

Parameter	Our results	Weka's results	Our results with proposed initial means
Number of iterations	9	7	2
Cluster 1's centroid	0.455000, 0.636667 0.337966, 0.343333	0.455, 0.6367 0.338, 0.3433	0.455000, 0.636667 0.337966, 0.343333
Cluster 2's centroid	0.803889, 0.409167 0.921356, 0.94	0.8039, 0.4092 0.9214, 0.94	0.803889, 0.409167 0.921356, 0.94
Number of clustered instances for cluster 1	100	100	100
Number of clustered instances for cluster2	50	50	50

It is observed that the number of iterations required by the algorithm is significantly lesser with the improved initial chosen means and the results are being validated using WEKA.

Table 3.2 shows the results of k-means clustering for k = 3, 4 and 5.

IV. CONCLUSION

In the current work, which uses the Iris flowers dataset, the following objectives are achieved:

- Proposal of novel way of finding initial means in the k-means clustering
- Implementation of k-means clustering
- Implementation of k-means clustering with proposed initial chosen means
- Validation of k-means clustering using WEKA

A method is proposed to find a better choice of the randomly chosen initial means in the k-means clustering. The results have shown that the number of iterations required by the algorithm is significantly lesser (only 2, as compared to 9 using random initial means, for k = 2) using these proposed initial chosen means in the implementation.

Table 3.2: k-means clustering results for k = 3, 4, 5

Value of k	3	4	5
Number of iterations	4	6	5
Cluster 1's centroid	0.292735, 0.549145 0.202955, 0.175214	0.292735, 0.549145 0.202955, 0.175214	0.328431, 0.559641 0.239615, 0.214869
Cluster 2's centroid	0.558743, 0.692623 0.424285, 0.450820	0.558743, 0.692623 0.424285, 0.450820	0.586735, 0.716837 0.440332, 0.477041
Cluster 3's centroid	0.803889, 0.409167 0.921356, 0.940000	0.741072, 0.305060 0.917676, 0.922619	0.793055, 0.404167 0.921187, 0.929167
Cluster 4's centroid	-	0.883838, 0.541667 0.926040, 0.962121	0.893791, 0.571078 0.929212, 0.965686
Cluster 5's centroid	-	-	0.702991, 0.205128 0.911343, 0.923077
Number of clustered instances for cluster 1	39	39	51
Number of clustered instances for cluster 2	61	61	49
Number of clustered instances for cluster 3	50	28	20
Number of clustered instances for cluster 4	-	22	17
Number of clustered instances for cluster 5	-	-	13

The values of k used are from k = 2 to k = 5. It is observed that the minimum number of iterations are obtained for k = 3.

The work has used the 3 species of the Iris flowers dataset which contains 150 labeled instances over 5 attributes. In WEKA, "Use training set" is selected as a test option and cluster mode in the k-means clustering.

V. REFERENCES

[1] Paul S. Bradley, Usama M. Fayyad "Refining Initial Points for K-Means Clustering" Microsoft Research, May 1998, Technical Report, MSR-TR-98-36

[2] Zhang Chen, Xia Shixiong "K-means Clustering Algorithm with improved Initial Center" Second International Workshop on Knowledge Discovery and Data Mining, 2009 IEEE

[3] Zhe Zhang, Junxi Zhang, HuifengXue "Improved K-means Clustering Algorithm" 2008 Congress on Image and Signal Processing, IEEE

[4] Wei Zhong, Gulsah Altun, Robert Harrison, Phang C. Tai, and Yi Pan "Improved K-Means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property" IEEE Transactions on Nanobio-Science, vol. 4, no. 3, September 2005

[5] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu "An Efficient K-Means Clustering Algorithm: Analysis and Implementation" IEEE Transactions on Pattern

Analysis and Machine Intelligence, vol. 24, no. 7, July 2002

[6] Ines Faerber, Stephan Guennemann, Hans-Peter Kriegel, Peer Kroeger, Emmanuel Mueller, Erich Schubert, Thomas Seidl, Arthur Zimek "On Using Class-Labels in Evaluation of Clusterings" 2010 ACM