Efficiency of K-Means Clustering Algorithm in Mining Outliers from Large Data Sets

Efficiency of outlier detection using several centroid initialization methods

Sridhar. A
B.Tech, Information Technology
SASTRA University, Tirumalaisamudram
Tanjore,, TamilNadu, India

Sowndarya. S
B.Tech, Information & Comm. Technology
SASTRA University, Tirumalaisamudram
Tanjore, TamilNadu, India

***ABSTRACT* - This paper presents the performance of k-means clustering algorithm, depending upon various mean values input methods. Clustering plays a vital role in data mining. Its main job is to group the similar data together based on the characteristic they possess. The mean values are the centroids of the specified number of cluster groups. The centroids, though gets changed during the process of clustering, are calculated using several methods. Clustering algorithms can be applied for image analysis, pattern recognition, bio-informatics and in several other fields. The clustering algorithm consists to two stages with first stage forming the clusters-calculating centroid and the second stage determining the outliers. There are three methods for assigning the mean values in k-means clustering algorithm. The three mean value assignment methods are implemented, performance is analysed and comparison of every method is done. Outliers, the disadvantage of the process are used in the analyzation to determine the performance with various mean inputs and methods.**

Keywords*: *k-means clustering algorithm, Mean values, Centroids, Outliers.

I. INTRODUCTION

We are living in a world filled with data. Every day, people collect large amount of data and store or represent it as information, for further and future analysis. There comes the field Data Mining, which is used to analyze large quantities of data; to derive with meaningful patterns and useful information from the raw data. One of the vital means in dealing with these data is to classify or group them into sets of categories or clusters. Generally human beings categorize data in some manner to yield useful information from them. Basically, systems come under supervised or unsupervised learning categories, depending upon various characteristics and attributes such as pre-defined output patterns leading to supervised processing. Unsupervised processing is done on input sets to yield efficient outputs. Clustering comes under the category of unsupervised learning. There are several clustering algorithms available. A clustering algorithm process on the given data, leading to clustered outputs. The clusters obtained as the output are generally grouping of elements done on some criteria.

Clustering algorithms are mainly classified into three types. They are:

- Partitional clustering
- Hierarchical clustering
- Spectral clustering

The clustering is done by using shortest distance method to cluster the given input set, depending upon the centroid values initialized.

The centroids initially are specified by the user with one of the described methods below. Further processing of centroids are calculations by finding the mean value of every cluster, leading to name the algorithm as k-mean clustering. K denotes the total number of clusters. The three methods for initializing the centroids are:

- Taking the first ‘k’ values as centroids.
- Random centroids generation
- User specified centroids

Outliers are normal elements when specified as input, but will lead in inefficient outputs when processed with them. They are elements which behave very differently from the norm, and are the major disadvantage of k-means clustering algorithm. Several actions can be performed when outlier appears.

- Do nothing
- Ignore the column
- Replace the outlying values
- Filter the rows containing them

There are two methods for outlier detection. They are:

- Univariate method
- Multivariate method.

Outliers can have anomalous causes. Several methods are present for the detection of the outlier.

A. *Algorithm (1st method):*

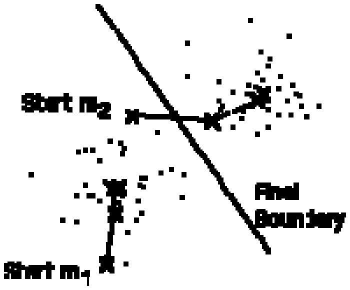
- Assign first k - values as initial centroids m1, m2….mn from the given set of inputs.
- Assign each item to the cluster which has nearest mean.
- Calculate new mean for each cluster until the convergence criteria is met.

B. *Algorithm (2nd method):*

- Assign initial centroids m1, m2….mn randomly.
- Assign each item to the cluster which has nearest mean.
- Calculate new mean for each cluster until the convergence criteria is met.

C. *Algorithm (3rd method):*

- Assign initial centroids by getting user inputs manually m1, m2….mn
- Assign each item to the cluster which has nearest mean.
- Calculate new mean for each cluster until the convergence criteria is met.



D. Convergence Criteria:

Theoretically the equation stated below, known as the convergence criterion is used. The algorithm is processed until the convergence criterion is met. But practically, when the old centroid value and new centroid value becomes equal, then it is said that the convergence criteria is met and the process is stopped.

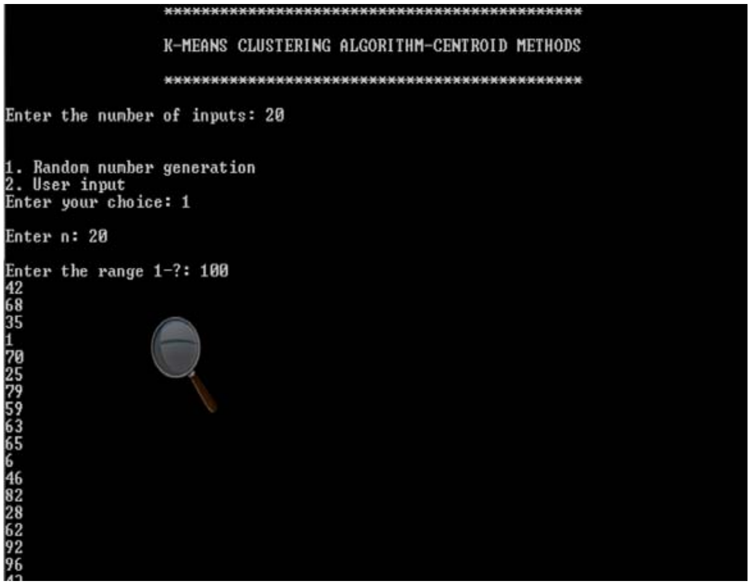
$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

II. Simulation and Result

No. of inputs	No. of clusters	k-means* 1st method (in sec)	k-means* 2nd method (in sec)	k-means* 3rd method (in sec)
1000	2	1.592000	2.196000	1.604000
1000	3	1.221000	2.981000	1.218000
1000	4	6.158000	2.331000	7.627000
10000	10	175.74300	143.61500	173.40300

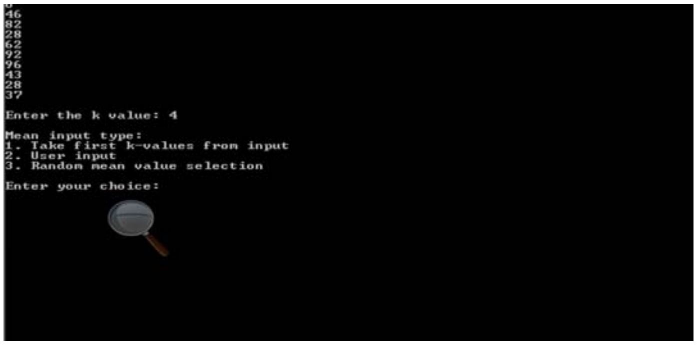
*- run time depends on the processor time. It varies every time as the processor time varies.

A. Initialization of values.



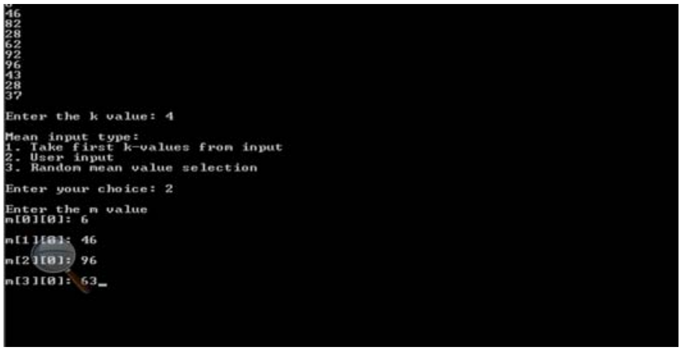
```
        *********************************************

        K-MEANS CLUSTERING ALGORITHM-CENTROID METHODS

        *********************************************

Enter the number of inputs: 20


1. Random number generation
2. User input
Enter your choice: 1

Enter n: 20

Enter the range 1-?: 100
42
68
35
1
70
25
79
59
63
65
6
46
82
28
62
92
96
```

B. K-Value (centroid) selection



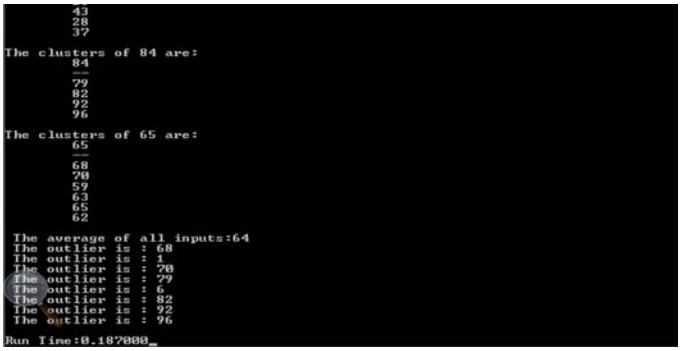
```
46
82
28
62
92
96
43
28
37

Enter the k value: 4

Mean input type:
1. Take first k-values from input
2. User input
3. Random mean value selection

Enter your choice:
```

C. Using user input method for giving the initial centroids.



```
46
82
28
62
92
96
43
28
37

Enter the k value: 4

Mean input type:
1. Take first k-values from input
2. User input
3. Random mean value selection

Enter your choice: 2

Enter the m value
m[0][0]: 6

m[1][0]: 46

m[2][0]: 96

m[3][0]: 63_
```

D. Final output with runtime, outlier, final clusters & centroids



```
          43
          28
          37

The clusters of 84 are:
          84
          --
          79
          82
          92
          96

The clusters of 65 are:
          65
          --
          68
          70
          59
          63
          65
          62

 The average of all inputs:64
 The outlier is : 68
 The outlier is : 1
 The outlier is : 70
 The outlier is : 79
 The outlier is : 6
 The outlier is : 82
 The outlier is : 92
 The outlier is : 96

Run Time:0.187000_
```

III. Conclusion

Generally, the runtime of the algorithm as well as the outliers depends upon the number of clusters specified and the initially selected centroids. In the above analyzation, the 3rd method of centroid (user input of centroid) takes a longer runtime but is much more efficient in detecting the outliers. The 2nd method is almost similar to the 1st method and is very less in efficiency.

Considering the example: the input numbers are 42, 68, 35, 1, 70, 25, 79, 59, 63, 65, 6, 46, 82, 28, 62, 92, 96, 43, 28, and 37. Taking the cluster ‘k’ value as 3 (total of 3 clusters), and taking the 1st method - first ‘k’ values are the initial centroid – 42, 68, 35 the clusters are obtained with the runtime: 0.106000 secs. The outliers obtained consist of 11 elements – 68, 1, 70, 79, 63, 65, 6, 82, 62, 92, 96. Whereas taking the 3rd method – user manual input of centroid values – 96, 70, and 37 the clusters are obtained with the runtime: 0.112000 secs. The outlier obtained consists of 4 elements – 1, 6, 92, and 96.

The conclusion is clearly explained with the above example.1st method takes less runtime but the outliers obtained is not efficient as it shows almost half the input values while the 3rd method takes longer runtime, but the outliers detected are efficient, detecting only the elements out of norm as outliers.

IV. References

[1] Yashwanth K Kanethker, *Let Us C*, 5th ed., BPB publications, New Delhi.

[2] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[3] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[4] Wikipedia search [Online]. Available: http://www.wikipedia.org//

[5] Rui Xu, Donald Wunsch, “Survey of clustering algorithms,” IEEE Transactions on Neural Networks, vol. 16, no. 3, May 2005, pp. 645-678

[6] Mu-Chun Su and Chien-Hsing Chou, “A modified version of the K-means algorithm with a distance based on cluster symmetry,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23,no. 6, June 2001, pp. 674-680.