An Optimized Approach for k-means Clustering

Sadhana Tiwari
Galgotias College of Engineering & Technology
Greater Noida, INDIA

Tanu Solanki
Galgotias College of Engineering & Technology
Greater Noida, INDIA

ABSTRACT

Cluster analysis method is one of the most analytical methods of data mining. The method will directly influence the result of clustering. This paper discusses the standard of k-mean clustering and analyzes the shortcomings of standard k-means such as k-means algorithm calculates distance of each data point from each cluster centre. Calculating this distance in each iteration makes the algorithm of low efficiency. This paper introduces an optimized algorithm which solves this problem. This is done by introducing a simple data structure to store some information in every iteration and used this information in next iteration. The introduced algorithm does not require calculating the distance of each data point from each cluster centre in each iteration due to which running time of algorithm is saved. Experimental results show that the improved algorithm can efficiently improve the speed of clustering and accuracy by reducing the computational complexity of standard k-means algorithm.

Keywords

Cluster Analysis, k-means clustering, kd-tree, Lloyd’s algorithm, Standard k-means algorithm, Constrained k-means algorithm.

1. INTRODUCTION

A cluster is generally thought of as a group of items(objects, points)in which each item is closed .A simple cluster representation is shown in Fig. 1. To a central item of a cluster and that members of different clusters are “far away” from each other. Clusters can be viewed as “high density regions” of some multidimensional space. A method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. Cluster analysis is another but different from those in other groups. In marketing there is keen interest among managers in developing products and strategies to target segments. The challenge with cluster analysis is that it involves both art and science, and it always produces an answer whether there really are clean and separable segments or whether consumers are positioned in a continuous cloud. Complicating matters further, there are numerous cluster analysis routines, which can lead to different results. Popular statistical tool for finding groups of respondents, objects, or cases that are similar to one There are given a set of n data points in dimensional space Rd and an integer k and the problem is to determine a set of k points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest center. A popular heuristic for k-means clustering is Lloyd's algorithm. In this paper, a simple and efficient implementation of Lloyd's k-means clustering algorithm is presented, which is called the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. We establish the practical efficiency of the filtering algorithm in two ways. First way is a data-sensitive analysis of the algorithm's running time, which shows as the separation between clusters increases the algorithm runs faster.

Second way is to present a number of empirical studies both on synthetically generated data and on real data sets from applications in color quantization[1].

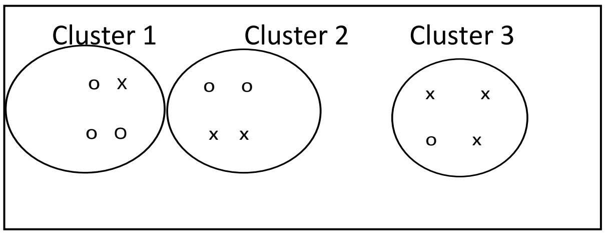

Fig. 1 Cluster Representation

On World Wide Web there is a computer program which allows such a statistical technique to be carried out in a very simple way. This paper also shows how this approach can be used with cross cultural data to extract similarities and differences between societies in a systematic fashion. Although the example used focuses on the economic systems of foragers, the methodology is also applicable to a wide variety of other cross-cultural research problems [2].In paper [3] there is introduced an improved k-mean algorithm which is based on background knowledge. Background knowledge is used as constraints to produce desired result and named as “Constrained k- means algorithm with background knowledge.” This paper includes four parts. Second part is analysis of standard k-means algorithm and shows the shortcomings of the standard k-means algorithm [4] [5]. The third part introduces the optimized k-means algorithm and fourth part shows experimental results and conclusions.

2. ANALYSIS OF STANDARD KMEANS ALGORITHM

A: *The process of standard k-means algorithm*:
The algorithm consists of two phases. The first phase selects k centers randomly and the next phase is to group the each data object to its nearest centre where k is decided in advance. Euclidian distance is generally is taken to determine the distance between cluster’s centre and data object [6].

The Euclidian distance d(xi ,yi) between two vector x=(x1,x2,…….,xn) and y=(y1,y2,……,yn) can be calculated as follows:

$$d(x_i, y_i) = [\sum_{i=1}^{n} (x_i - y_i)^2]^{1/2}$$

The algorithm is as follows:

Input:
Number of desired clusters k and a database D={d1,d2,…….,dn} with n data objects.

Output:
A set of k clusters
Steps:

```
1) Randomly select k data objects from dataset D as initial clusters.
2) Repeat;
3) Calculate the distance between each data object di(1<=i<=n) and all k cluster centres cj(1<=j<=k) and assign data object di to its nearest cluster.
4) For each cluster j (1<=j<=k), recalculate the cluster centre.
5) Until no changes in the cluster centres.
```

The k-means algorithm always converges to the local minimum. Before converging it calculates distance of each data object from each cluster centre in each loop execution. Suppose there are t iterations in the execution of algorithm then it requires kt comparisons for each data objects. It means time complexity of the standard k-means algorithm will be O(nkt) where n is the total number of data objects[7].

B: ***The shortcomings of k-means algorithm***

The K-means algorithm has drawback in its computation that can be seen easily by above analysis that in standard algorithm distance is calculated in each iteration while it is sure that in some iteration data object will belong to the same cluster. If data object x remains z times in same cluster then algorithm takes x(k-1) time more to execute the algorithm.

3. OPTIMIZED K-MEANS CLUSTERING ALGORITHM

The standard k-means algorithm takes extra time in calculating distance from each cluster's centre in each iteration. This extra time can be saved by adapting this method. If the distance of a data object from new cluster centre is smaller than the distance of the data object from previous cluster centre then it will belong to same cluster means there is no need to calculate the distance of the data object from other cluster centre. It can be done by maintaining two simple data structure which will be used to store the label of cluster and distance of data object to corresponding cluster centre and it can be used in next iteration.

The process of optimized k-means algorithm is as follows:

Input:
The number of desired clusters, k, and a dataset D=(d1, d2 ,………….,dn) containing n data objects.

Output:
A set of k clusters.
Steps:

```
1) Randomly select k data objects from dataset D as initial clusters.
2) Calculate the distance between each data object di (1<=i<=n) and each cluster centre cj (1<=j<=k) as Euclidian distance d(di, cj).
3) For each data object di, find the closest centre cj and assign di to the cluster centre cj.
4) Store the label and distance as:
   Set cluster[i]=j, and dist[i]= d(di, cj).
   Where j is the cluster label in which data di reside and d (di, cj) is the distance of data object di to the cluster centre labeled by j.
5) For each cluster j (1<=j<=k) recalculate the cluster centre.
6) Repeat.
7) For each data object di .Compute its distance to the centre of the present cluster;
   a) If this distance is less than or equal to dist[i], the data object stays in the initial cluster;
   b) Else
      For every cluster centre cj(1<=j<=k),it compute the distance d(di, cj) and assign the data object di to the nearest cluster.
      Set cluster[i] = j;
      Set dist[i] = d(di, cj).
8) For each cluster centre j(1<=j<=k),recalculate the centres;
9) Until the centre is same.
10) Output the clustering result.
```

The optimized algorithm requires two data structure cluster and dist to store label of cluster and distance respectively for each iteration which is used in next iteration.

This paper does not require calculating distance in each iteration. The time complexity of this algorithm is O(nk). If a data point remains in its initial cluster then the time complexity will be O(1) otherwise O(k). If half of the data points move from its initial cluster then the time complexity will be (nk/2). So the proposed algorithm effectively increases the speed of standard k-means algorithm. But this algorithm also requires the value of k in advance. If one wants the optimal solution then he must test for different values of k.

4. EXPERIMENTAL RESULT

This paper selects three different datasets from the UCI repository of machine learning database to test the efficiency of the standard k-means algorithm and this proposed algorithm. In the following experiment time taken by two algorithms is computed. Characteristics of the datasets are given in the table I. The same dataset is given as input for both algorithms. Experimental operating system is window 7.Visual Studio 10.0 is used for performing the clustering.

Table I.

Dataset	**Number of attributes**	**Number of records**
Iris	5	200
Glass	4	250
Letter	10	1500

This paper uses iris and glass data sets because they are suitable datasets for clustering. The number of cluster k sets to 3.

Standard k-mean algorithm and optimized k-mean algorithm is shown in the following table II.

Table II

Data set	**K means running time(s)**	**Optimized k-means running time(s)**	**k-means accuracy %**	**Optimized k-means accuracy %**
Iris	0.0568	0.0526	82.6	91.5
Glass	0.0879	0.0824	76.8	85.5

5. CONCLUSION

K-means is a typical and widely used algorithm for clustering. This paper analyses the standard algorithm and gives an approach to reduce the time complexity of the standard k-means algorithm. This paper proposes an algorithm which is of O(nk) time complexity. This time complexity is less than that of standard k-means algorithm. Experimental result shows that the running time is less than that of standard k-means. Thus it can conclude that the algorithm explained is feasible.

7. REFERENCES

[1] T. Kanungo, D. M. Mount, N. Netanyahu, C.Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: Analysis and implementation” IEEE Transaction Pattern Analysis and Machine Intelligence, 2002.

[2] Bruce A. Maxwell, Frederic L. Pryor, Casey Smith, “Cluster analysis in cross-cultural research” World Cultures 13(1): 22-38, 2002.

[3] Kiri Wagstaff and Claire Cardie Department of computer science, Cornell University, USA “Constrained k- means algorithm with background knowledge”.

[4] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest, Introduction to Algorithms, Prentice Hall, 1990.

[5] Anil K. Jain, M. N. Murty, P. J. Flynn, “Data Clustering: A Review,” ACM Computing Surveys, 31(3): 264-323 (1999).

[6] Anil K. Jain and Richard C. Dubes, Algorithms for Clustering Data, Prentice Hall (1988).

[7] Ahmet Alken, Department of Electrical and Electronics Engineering, KSU, Turkey, “Use of K-means clustering in migraine detection by using EEG records under flash stimulation” International Journal of the Physical Sciences Vol. 6(4), pp. 641-650, 18 February, 2011.