Max-D clustering K-means algorithm for Auto-generation of Centroids and Distance of Data Points Cluster

Wan Maseri Binti Wan Mohd, A.H.Beg, Tutut Herawan, K.F.Rabbi

Faculty of Computer Systems & Software Engineering
University Malaysia Pahang
Gambang-26300, Pahang, Malaysia
ahbeg_diu@yahoo.com , {maseri,tutut}@ump.edu.my

Abstract. K-Means is one of the unsupervised learning and partitioning clustering algorithms. It is very popular and widely used for its simplicity and fastness. The main drawback of this algorithm is that user should specify the number of cluster in advance. As an iterative clustering strategy, K-Means algorithm is very sensitive to the initial starting conditions. In this paper has been proposed a clustering technique called MaxD K-Means clustering algorithm. MaxD K-Means algorithm auto generates initial k (the desired number of cluster) without asking for input from the user. MaxD k-means also used a novel strategy of setting the initial centroids. The experiment of the Max-D means has been conducted using synthetic data, which is taken from the Llyod's K-Means experiments. Another experiment has been done using real-life data focusing on student's results in higher-education institution in Malaysia. The results from the new algorithm show that the number of iteration improves tremendously, and the number of iterations is reduced. The improvement rate is around 78%.

Keywords: K-means algorithm, Partitioning algorithm, Clustering, MaxD k-means, Data mining.

1 Introduction

The K-means is one of the classical and well-researched algorithms for unsupervised learning to solve the essential clustering problem. It tries to find the possible classes of data objects, structured categories, whose associates are identical in some way. The cluster therefore corresponds to a collection of objects that are "equivalent" to each other and are "different" and objects belonging to other groups. The K-means can be considered as the most important unsupervised learning approach. K-means method has the following prospective benefits: (i) covering different types of attributes, (ii) to discover clusters of arbitrary shape, (iii) the minimum requirements for domain knowledge to determine input parameters (iv) can be uses with noise and outliers, and (5) to minimize the difference between the data. Therefore, it relates to many fields such as marketing, biology, and image recognition [1].

Clustering is an essential technique used unsupervised classification to recognize some of the structures involved in the use of objects. The purpose of cluster analysis is to recognize objects into subsets that have aspect in the viewpoint of a particular problem. In particular, the clustering, a set of patterns, usually vectors in a multidimensional place, are organized into clusters so that patterns in the same cluster are similar in some sense and patterns in different categories are different in the sense. In some clustering concerns, the number of clusters, K, this is known in advance. In such conditions, clustering can be developed as a distribution model n in N dimensions statistic locations between groups of K so that the goals of a group are more similar to each other than trends in different clusters. This contains the minimization of an optimization specification extrinsic. K-Means algorithm is very well-known and widely used clustering technique appropriate in such conditions [2]. Clustering is often the very first steps in data analysis. It can be used to recognize natural categories in data sets and to recognize very subjective elements that might reside there, without having any basic knowledge on characteristics of data. Therefore, many classification methods have been developed such as hierarchical clustering [3], the mixture densities [4], [5], graph partitioning [6], and spectral classification [7] and these methods have been used in a wide range of areas such as computer vision, data mining, bio - informatics and information retrieval, to name a few [8]. The Pseudo-code of the Lloyd's K-Means algorithm [9] shown in Algorithm 1shown in Fig.1.

Algorithm 1:

Input:
D= {t_1, t_2, …. T_n } // Set of elements
K // Number of desired clusters
Output:
K // Set of clusters
K-Means algorithm:
Assign initial values for m_1, m_2,.... m_k
repeat
assign each item t_i to the clusters which has the closest mean;
calculate new mean for each cluster;
until convergence criteria is met;

Fig.1. Pseudo-code of the Lloyd's K-Means algorithm

K-Means is a simple algorithm that has adapted to areas with many problems. Similar to other algorithm, K-Means clustering has some limitations [10], [11], [12].To solve the existing k-means's problem is the main vision of this research. Hence, a new approach has been proposed to overcome existing problem. The new clustering algorithm proposed a technique to define the initial parameter of k-means through the auto generation of the number of clusters using the maximum distance of data points and a novel approach of defining initial centroid for effective and efficient clustering process. The algorithm helps the user in estimating the number of clusters which is highly dependent on the domain knowledge, which is not so desirable.

2 Background

There are different method has been proposed to improve the efficiency of k-means algorithm [13], [14], [15]. Among them it can find the k-means algorithm is one of the more common ones. But we know that K-means algorithm is sensitive to the initial cluster centers and easy to get immovable in local optimal solutions [16]. Moreover, when the number of data points is large, it takes a tremendous amount of time to find a global optimal solution [17], [18].

S. Bandyopadhyay and U. Maulik [2] described a GA based clustering algorithm. In their strategy the chromosome encodes the centers of the clusters instead of a possible partition of the data points. The algorithm attempts to develop appropriate cluster centers, while optimizing a given clustering metric. In addition, the usefulness of KGA-clustering algorithm for classification of pixels of a satellite image to distinguish between the different areas of land has been designated. Note that even if the GAs is usually done with binary strings, they have implemented the encoding of floating point chromosome. M. Otsubo et al. [19] presented a computerized the identification of the clusters by using the k-means clustering technique. In their research they present a computerized technique to recognize clusters separately to determine the tensor representing a reduction of stress and the spread of tensors. To this end, uses a technique called k-means for the purpose of the division to reduce the stress tensor obtained by inversion methods into multiple clusters. Currently, the number of clusters, k, must be specified by the user. The k-means requires a well-defined distance between the objects to classify. The stress difference defined by Orif and Lisle [14] is a useful distance between the tensors of stress reduction. The parameter space is adequate, since the Euclidean distance between points in the parameter space is equal to the stress difference between the stresses that are represented by points. They tested the technique by artificial data sets. It has been shown that the resolution of visual identification of the clusters was often insufficient, and that the present technique correctly detected highlights from artificial data were generated with known stress.

S. Kalyani and K.S. Swarup [20] presented a modified K-means algorithm (PSOKM) using particle swarm optimization technique for the evaluation of static security, transient. Training set of vectors generated from offline simulations are presented as input to the PSO algorithm based K-means classification using supervised active

learning to adjust its weight vectors (cluster centers). The proposed algorithm was implemented in IEEE 30 bus, 57 bus, 118 bus and 300 bus standard of test cases, and its performance was compared with other K-means algorithm. Their results showed that the high-accuracy classifiers with lower rate of misclassification can be exchanged with the classification PSOKM.

A.M. Bagirov et al. [21] have developed a new version of the modified global k-means algorithm. This algorithm computes step by step through the clusters k_1 cluster centers from the previous iteration to solve the problem of k-partitions. An important step in the calculation of this algorithm is a starting point for the center of the cluster k-th. This starting point was calculated by minimizing the additional function known as clusters. The results of their numerical experiments show that in most cases, the proposed algorithm is faster and more accurate than the global k-means algorithm. At the same time, similar results the proposed algorithm requires much less evaluations and CPU time than changing the global k-means algorithm. Therefore, the proposed algorithm is a significant improvement in changing the global k-means algorithm. Moreover, this improvement is even more important that all size of the data set increases.

3. MaxD-Kmeans Algorithm

The MaxD-Kmeans algorithms are shown as bellow:

Algorithm: Calculate_centroid

Input: X is the set { x_1, x_2, x_3…….x_n}, where n represents the number of input values
Output: Y is the set {} of clusters

1. Let x_s= {} represent the sorted values
2. Let k= number of total centroids
3. Let C= {} represents set of the number of total centroids
4. Let X_{max} represents the maximum value in X
5. Let X_{min} represents the minimum value in X
6. Let C_t represents total centroids
7. X_s = sort (x);
8. X_{min}= read first value of (x_s);
9. X_{max}= read last value of (X_s);
10. C_t []= Xmin;
11. For each X_f in X_s
12. If (X_f > X_{min})
13. {
14. Ct = (X_f-X_{min})/2 + 2
15. }
16. Else
17. {
18. (C_f=X_{min})/ 2+ X_{min}

```
19.  }
20.  K= count (t);
21.  }
```

Fig.2. Max-D k-means algorithm (Calculate_centroid)

Algorithm Build_Cluster

Input: Y = { } represents the cluster
X = { } represents the input values
Output: Cm= {} {} represents the cluster members

```
1.  For each (Yf in Y)
2.  {
3.  P = [Yf];
4.  If([Yf]> =[xf] and [Yf]<= [Xf+1])
5.  P= [Yf];
6.  Cm= P[];
7.  }
```

Fig.3. Max-D k-means algorithm (Build_Cluster)

4 Result and Discussion

In order to study the effectiveness of MaxD-Kmeans Algorithm experiment has been conducted using synthetic data which is taken from the Llyod's K-Means experiments [10]. To show the significant improvement of the new algorithm, the experiment was divided into several cycles. Table 1, Shows the comparative result between Max-D K-means and Llyod's K-Means algorithm. The comparative results show that, with the Max-D k-means algorithm, the number of iteration improves tremendously when N is larger. It shows that the number of iterations is reduced from 18 to 4, which is an improvement of 78%.

Table1. Comparison between Max-D K-Means and Llyod's K-Means using N=80

Comparative Algorithm	Total Clusters	Cluster Member
Max-D K-Means N= 80 K = 17 Number of iteration= 4 Total Cluster=10	1	2
	2	9
	3	9
	4	8
	5	6
	6	8
	7	16
	8	18
	11	3
	17	2
Llyod's K-Means N=80 K=10 Number of Iteration=18 Total Cluster =10	1	2
	2	12
	3	0
	4	12
	5	11
	6	8
	7	10
	8	13
	9	3
	10	9

5 Conclusion

In this paper, has been proposed a parameter less data clustering technique based on maximum distance of data and lioyd k-means algorithm, which requires a number of clusters, k, must be determined beforehand, which is not desirable, since the number of cluster configuration needs domain knowledge. In order to study the effectiveness of the proposed approach for setting the parameters of K-Means algorithm, the experiment has been done using synthetic data, which has been taken from the Llyod's K-Means experiments. The experimental results show that the use of new approach to defining the centroids, the number of iterations has been reduced where the improvement was 78%.

6 Acknowledgements

This work was supported by Fundamental Research Grant Scheme (FRGS-RDU110104), University Malaysia Pahang under the project "A new Design of

Multiple Dimensions Parameter less Data Clustering Technique (Max D-K means) based on Maximum Distance of Data point and LIoyd k-means Algorithm".

References

1. Zhou, H., Liu, Y.: Accurate integration of multi-viewrange images using k-means clustering. Pattern Recognition. Vol. 41, 152--175 (2008)
2. Bandyopadhyay, S., Maulik, U.: An evolutionary technique based on K-Means algorithm for optimal clustering. Information Sciences vol. 146, 221-237 (2002)
3. Duda, R., Hart, P., Stork., D.: Pattern Classification, second ed. John Wiley and Sons. New York (2001)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Spc. Vol. 39, pp.1-- 38 (1977)
5. McLachlan, G.L., Basford, K.E.: Mixture Models: Inference and Application to clustering. Marcel Dekker (1987)
6. Jiambo, S., Jitendra, M.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Machine Intell. Vol. 22, 288-905 (2000)
7. Stella, Y., Jianbo, S.: Multiclass spectral clustering. In: Proc. Internat. Conf. on Computer Vision. pp. 313--319 (2003)
8. Murino, L., Angelini, C., Feis, I.D., Raiconi, G., Tagliaferri, R.: Beyond classical consensus clustering: The least squares approach to multiple solutions. Pattern Recognition Letters. Vol. 32, 1604--1612 (2011)
9. Dunham, M.: Data Mining: Introductory and Advance Topics. N.J. Prentice Hall (2003)
10. Chiang, M., Tsai, C., Yang, C.: A time-efficient pattern reduction algorithm for k-means clustering. Information Sciences. Vol.181, 716--731 (2011)
11. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Transaction on Neural Netowrks. Vol.16 (3), 645--678 (2005)
12. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Computing Surveys. Vol. 31 (3) (1999).
13. Kanungo, T., Mount, D., Netanyahu, N.S., Piatko, C., Silverman, R., Wu, A.: An efficient K-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. Vol. 24 (7), 881--892 (2002)
14. Likas, A., Vlassis, N., Verbeek, J.J.: The global K-means clustering algorithm. Pattern Recognition. Vol. 36, 452-- 461(2003)
15. Charalampidis, D.: A modified K-means algorithm for circular invariant clustering. IEEE Trans. Pattern Anal. Mach. Intell. Vol. 27 (12),1856-1865(2005)
16. Selim, S.Z., Ismail, M. A.: K-means type algorithms: a generalized convergence theorem and characterization of local optimality, IEEE Trans. Pattern Anal. Mach. Intell. Vol. 6, 81--87(1984)
17. Spath, H.: Cluster Analysis Algorithms. Ellis Horwood, Chichester. UK (1989)
18. Chang, D., Xian, D., Chang, W.: A genetic algorithm with gene rearrangement for K-means clustering. Pattern Recognition. Vol. 42, 1210--1222 (2009)
19. Otsubo, M., Sato, K., Yamaji, A.: Computerized identification of stress tensors determined from heterogeneous fault-slip data by combining the multiple inverse method and k-means clustering. Journal of Structural Geology. Vol. 28, 991--997(2006)
20. Kalyani, S., Swarup, K.S.: Particle swarm optimization based K-means clustering approach for security assessment in power systems, Expert Systems with Applications. Vol. 38, 10839--10846(2011)
21. Bagirov, A.M., Ugon, J., Webb, D.: Fast modified global k-means algorithm for incremental cluster construction, Pattern Recognition. Vol. 44, 866--876 (2011)