Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 1 (2012) 367 – 371

Procedia Technology

INSODE 2011

An Improved Parameter less Data Clustering Technique based on Maximum Distance of Data and Lioyd k-means Algorithm

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

Faculty of Computer Systems and Software Engineering, University Malaysia Pahang, Gabnang-26300, Pahang, Malaysia

Abstract

K-means algorithm is very well-known in large data sets of clustering. This algorithm is popular and more widely used for its easy implementation and fast working. However, it is well known that in the k-means algorithm, the user should specify the number of clusters in advance. In order to improve the performance of the K-means algorithm, various methods have been proposed. In this paper, has been presented an improved parameter less data clustering technique based on maximum distance of data and Lioyd k-means algorithm. 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 60%.

Keywords: K-Means Algorithm, Clustering, Partitioning Clustering Algorithm, Data Mining.

1. Introduction

The K-means is one of the classical, well-researched algorithms for unsupervised learning to solve the fundamental clustering problem. It tries to find the possible classes of data objects, organized groups, whose members are similar 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 potential benefits: (1) covering different types of attributes, (2) to discover clusters of arbitrary shape, (3) the minimum requirements for domain knowledge to determine input parameters, (4) uses with noise and outliers, and (5) to minimize the difference between the data. Therefore, it applies to many fields such as marketing, biology, and image recognition [1].

Clusters are an important technique used unsupervised classification to identify some of the structures involved in the presence of objects. The purpose of cluster analysis is to classify objects into subsets that have a role in the context of a particular problem. In particular, the clustering, a set of patterns, usually vectors in a multidimensional space, are grouped into clusters so that patterns in the same cluster are similar in some sense and patterns in different clusters are different in the sense. In some clustering problems, the number of clusters, K, this is known in advance. In such situations, clustering can be formulated as a distribution model n in N dimensions metric spaces between groups of K so that the motives of a group are more similar to each other than trends in different groups. This involves the minimization of an optimization criterion extrinsic. K-Means algorithm is very popular and widely used clustering technique applicable in such situations [2]. Clustering is often the first step in data analysis. It can be used

* Wan Maseri Binti Wan Mohd, A.H.Beg. Tel.: +60-109022970
E-mail address: ahbeg_diu@yahoo.com

doi:10.1016/j.protcy.2012.02.076

to detect natural groups in data sets and to identify abstract structures that might reside there, without having any basic knowledge on the 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 techniques have been used in a wide range of areas including computer vision, data mining, bio - informatics and information retrieval, to name a few [8]. The Pseudo-code of the Lloyd's K-Means algorithm has shown in Algorithm 1 [9].

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;

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. Related Work

To improve the performance of k-means algorithm different methods have been proposed [13, 14, 15]. In one method, 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 stuck in local optimal solutions [16]. In addition, when the number of data points is large, it takes a tremendous amount of time to find a global optimal solution [17]. D. Chang et al. [18] combines K-means algorithm for genetic algorithms (GA) to obtain better results.

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 (2003) 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. Improved Parameter less K-Means Algorithm:

In our previous work [22], has been proposed a new parameter less k-means algorithm. Based on the previous algorithm, has been improved the k-means algorithm MaxD, auto-generation of the initial number of cluster and centroids. MaxD K-means consists of two parts, one of the pre-processing parameters and another is Lloyd of k-means algorithm. Pre-processing parameters consist of two parts that are auto-generate of the initial number of clusters and determine the centroids of the clusters. In the pre-processing algorithm, centroids are defined centers of the two points that starts from the maximum and minimum data points. The rational of generating centroids from maximum and minimum data points and subsequently sort them is that the probability of binding all items in the dataset to the nearest centriod is high since the generated centroids is near to all respective items. The algorithm begins with setting maximum and minimum data points as centroids, and then in the midst of these values is set to a new focus. Then, from the generated centroids, new centroids are further generated by reassigning the new maximum and minimum from the new centroids. The process produces centroids ends when the maximum number of iteration is reached.

4. Experiment and Result

In order to study the effectiveness of the proposed approach for setting the parameters of K-Means algorithm, the experiment has been conducted using synthetic data which has been taken from the Llyod's K-Means experiments [17].

Llyod's K-Means algorithm:

Input – k - number of desired clusters
Initial values for centroids – $c_1, c_2, \ldots c_k$

Given the following items to cluster:
{3, 5, 11, 13, 4, 21, 31, 12, 26}

Suppose, $k = 2$ *and* $c_1 = 2$, $c_2 = 4$

Using the Llyod's K-Means algorithm, has been obtain the following result:

Table 1. Result of Llyod's K-Means Algorithm

Iteration	Clusters (k)	Cluster's Members
1	1	{3,4}
	2	{5,11,13,21,31,12,26}
2	1	{3,4,5}
	2	{11,13,21,31,12,26}
3	1	{3,4,5,11}
	2	{13,21,31,12,26}
4	1	{3,4,5,11,12,13}
	2	{21,31,26}
5	1	{3,4,5,11,12,13}
	2	{21,31,26}

New K-Means algorithm:

Input – *k* - number of desired clusters

Initial values for controids- $c_1, c_2, \ldots c_k$ using new approach of defining centroids as in algorithm 1.

Given the following items to cluster:

{3, 5, 11, 13, 4, 21, 31, 12, 26}

And suppose $k = 2$ *and* $c_1 = 2$, $c_2 = 30$ (using new approach of defining centroids)

Using the new K-Means algorithm, we obtain the following result:

Table 2. Result of New K-Means Algorithm (New approach of defining centroids)

Iteration	Clusters (k)	Cluster's Members
1	1	{3,5,11,13,4,12}
	2	{21,26,31}
2	1	{3,5,11,13,4,12}
	2	{21,26,31}

The result in Table 1 and Table 2 shows that the use of new approach to defining the centroids, the number of iterations is reduced from 5 to 2, which is 60% improvement.

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 before hand, 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 60%.

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. H. Zhou and Y. Liu. Accurate integration of multi-viewrange images using k-means clustering. Pattern Recognition. 41 (2008) 152-175.
2. S. Bandyopadhyay and U. Maulik. An evolutionary technique based on K-Means algorithm for optimal clustering. Information Sciences 146 (2002) 221-237.
3. R. Duda, P. Hart and D. Stork. Pattern Classification, second ed. John Wiley and Sons. New York. 2001.
4. A.P. Dempster, N.M. Laird and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Spc. 39 (1977)1-38.
5. G.L. McLachlan and K.E. Basford. Mixture Models: Inference and Application to clustering. Marcel Dekker. 1987.
6. S. Jiambo and M. Jitendra. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Machine Intell. 22 (2000) 288-905.
7. Y. Stella and S. Jianbo. Multiclass spectral clustering. In: Proc. Internat. Conf. on Computer Vision. pp. 313-319. 2003.
8. L. Murino, C. Angelini, I. De Feis, G. Raiconi and R. Tagliaferri. Beyond classical consensus clustering: The least squares approach to multiple solutions. Pattern Recognition Letters. 32 (2011) 1604-1612.
9. M. Dunham. Data Mining: Introductory and Advance Topics. N.J. Prentice Hall. 2003.
10. M. Chiang, C. Tsai and C.Yang. A time-efficient pattern reduction algorithm for k-means clustering. Information Sciences 181 (2011) 716-731.
11. R. Xu and D. Wunsch. Survey of clustering algorithms. IEEE Transaction on Neural Netowrks. 16 (3) (2005) 645-678.
12. A.K. Jain, M.N. Murty and P.J. Flynn. Data clustering: a review. ACM Computing Surveys 31 (3) (1999).
13. T. Kanungo, D. Mount, N.S. Netanyahu, C. Piatko, R. Silverman and A. Wu. An efficient K-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 881-892.
14. A. Likas, N. Vlassis and J.J. Verbeek. The global K-means clustering algorithm. Pattern Recognition. 36 (2003) 452-461.
15. D. Charalampidis. A modified K-means algorithm for circular invariant clustering. IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (2005) 1856-1865.
16. S.Z. Selim and M.A. Ismail. K-means type algorithms: a generalized convergence theorem and characterization of local optimality, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) 81-87.
17. H. Spath. Cluster Analysis Algorithms. Ellis Horwood, Chichester. UK. 1989.
18. D. Chang, D Xian and W. Chang. A genetic algorithm with gene rearrangement for K-means clustering. Pattern Recognition. 42 (2009) 1210-1222.
19. M. Otsubo, K. Sato and A.Yamaji. 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. 28 (2006) 991-997.
20. S. Kalyani and K.S.Swarup. Particle swarm optimization based K-means clustering approach for security assessment in power systems, Expert Systems with Applications. 38 (2011) 10839-10846.
21. A.M. Bagirov, J.Ugon and D. Webb. Fast modified global k-means algorithm for incremental cluster construction, Pattern Recognition. 44 (2011) 866-876.
22. W. M. Wan Mohd, J. Mohd Zain, and A. Embong. Parameterless K-Means: Auto-generation of Centroids and Initial Number of Cluster based on Distance of Data Points. In the proceeding of Malaysia-Japan international Symposium on Advanced Technology. Kuala Lumpur Malaysia. 2007.