ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Threshold Based K-Mean Adaptive Clustering for High Dimensional Data Analysis

N. SreeRam

Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India

sriramnimmagadda@gmail.com

Dr. M. H. M. Krishna Prasad

Prinicipal, University College of Engineering, JNTUK, Kakinada,

krishnaprasad.mhm@gmail.com

Abstract.

Due to rapid growth of data sources exponential have evolved the necessity in developing new techniques for gathering useful information. To gain and gather knowledgeable information clustering is an investigable technique which is used in finding hidden homogeneous patterns from the data sources. Partition based clustering techniques are employed to acquire clustering information from user specified parameters such as similarity threshold value and number of clusters. For effective clustering, we suggest a deterministic algorithm called ad Threshold based k-means Adaptive Clustering (TAC). Artificial and real data sets have been tested by it. The algorithm k-means has also been contrasted. A parameter, neighborhood distance is used to cluster data items in this suggested TAC algorithm. The user doesn't specify neighborhood distance, but it is calculated automatically, and it is also an adaptive parameter. The minimum support value of the tiny clusters is another parameter in TAC algorithm. TAC's performance is also evaluated by using real and artificial datasets, which have found outliers can be detected and overlapped and non-overlapped clusters are generated. The results show that TAC algorithm produces clusters of distinct dimensions while k-means generates clusters of almost identical dimensions.

Keywords: Adaptive clustering, Hidden and homogeneous patterns, Neighborhood distance, and Similarity index.

1. Introduction

Clustering is the method by which unlabelled data objects segments are found based only on information found in the data that describes the objects and their relationships. In all application fields k-means algorithm is commonly recognized. But it is nondeterministic

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

because it needs the initial centroids supplied by the user, it is suspectable to outliers and produces overlapped clusters. However, the use of this algorithm was not limited by these aspects of k-means rather data scientists encouraged to improve its functionality. The k-means algorithm depends on initial centroids or the data objects being randomly selected. This means that the findings of these algorithms are different in sequence of executions and they are deemed to be non-deterministic algorithm. So, in order to make k-means as deterministic algorithm an effective Threshold based Adaptive Clustering (TAC) has been suggested in this proposed work. TAC produces both overlapped and non-overlapped clusters. It uses an adaptive threshold value for similarity and detects outliers.

2. PROPOSED TAC ALGORITHM

Provide sufficient detail methods to allow the work to be reproduced. Methods already published should be indicated by a reference: only relevant modifications should be described.

A. Parameters used in TAC algorithm

Clustering parameters play an important role. The clustering algorithms functionality depends on them, so they should be carefully specified by the user. Two parameters were introduced in the TAC algorithm: neighborhood distance threshold and minimum support.

Neighborhood distance threshold: This is the first parameter used in the proposed TAC algorithm and is used to identify all the neighbours of a data object. Partitioning based algorithms are also employ this parameter but as non-adaptive and user specified. In this proposed TAC algorithm, it is an adaptive parameter, the value this parameter increases during the formation of a cluster, as well as user doesn't specify it. TAC algorithm needs neighborhood distance value (Nd) in order to form a cluster.

Mathematically Nd is defined as

$$Nd = \max(minp, minq) \qquad (1)$$

Where minp = min [|p-r| : r € set of un-clustered data objects and Minq = min [|q-r| : r € set of un-clustered data objects. Here P and Q are the farthest data objects in the un clustered data objects. Nd is not fixed value, rather it is increased by a small value (Δ) during the formation of a cluster. The small increment to be made in Nd is the distance between the closest member of a cluster. It helps in increasing the size of a cluster. It is defined as

$$\delta = \min \{[u - v] : u, v \in \text{cluster}\} \qquad (2)$$

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Moreover, value of Nd is initialized every time a new cluster is being formed.

Minimum Support Parameter: The proposed TAC algorithm uses this parameter to improve the result of the clustering. This parameter specifies the minimum number of data objects in a cluster. The usage of minimum support has been adopted by the Association Rule mining technique. Its value can be specified by the user after clusters have been generated. It improves the result of clustering by removing the small clusters

B. TAC algorithm

TAC algorithm consists of the following steps:

Step 1: Initially the farthest data objects are identified from the given dataset.

Step2: Now, the distance of the closest data object to one of the farthest objects and the distance of the closets data objects to other farthest data objects are calculated. The formation of a new cluster starts from one of the farthest data objects whose closest data object is closer than its counterpart.

Step 3: Initialize the value of Nd and assign the selected farthest data object and all other data objects within the neighborhood distance as a member of cluster being formed.

Step 4: Increase the value od Nd by a small value Δ and assign more data objects as the members of cluster being formed. This step is repeated till new data objects are added to cluster.

Step 5: If more than one data objects are left un clustered then again find the farthest data objects from un clustered data objects and repeat steps 2,3, and 4.

TAC uses a function new cluster(R,Nd,K,Adj[n][n]) to generate clusters. This function generates overlapping clusters and membership of overlapping data objects by using the above-mentioned algorithm. The size of each cluster generated is compared with given minimum support value and small size clusters to be considered insignificant. This algorithm always produces the same result for given dataset and minimum support value on successive runs [15].

3. Results and Discussion

The experiments were performed on artificial and real data sets respectively in the first and second subsections. The TAC algorithm is compared to the K-means algorithm in the third subsection

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

3.1. Experiment on Artificial Datasets

Five artificially generated data sets used to assess TAC algorithm efficiency. Table 1 Presents the result of TAC worked on Artificial Datasets

Table 1 Results of TAC algorithm

Datasets	# clusters identified	Value of minimum support (MIN_SUP)	# significant clusters	# Data objects in significant clusters	Total data objects covered
AD_2K_1	34	1.5%, i.e. Minimum 30 data objects/cluster	14	249, 36, 88, 47, 40, 202, 68, 48, 270, 61, 33, 50, 66 AND 286	77.2% of Dataset
AD_1.5K_2	29	1.5%, i.e., minimum 23 data objects/cluster	14	249, 88, 40, 48, 100,53, 171, 64, 28, 14679, 94, 22 AND 28	80.6 % of Dataset
AD_1.5K_3	31	1.5%, i.e. minimum 23 data objects/cluster	13	36, 47, 29, 202, 68, 33, 296, 31, 60, 41, 181, 29 AND 72	75 % of Dataset
AD_1.5K_4	33	1.5%, i.e. minimum 23 data objects/cluster	15	249, 88, 40, 48, 58, 28,162, 124, 29, 41, 4178, 24, 131 AND 37	78.5% of Dataset

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

AD_2K_1	34	1.5%, i.e., minimum 30 data objects/cluster	14	249, 36, 88, 47, 40, 202, 68, 48, 270, 61,33, 50, 66 AND 286	77.2% of Dataset

Wisconsi-Original breast cancer (Bcw-O), Ecoli, Glass Identification (Gi), Haberman's Survival, Iris, Seed, Wine and Yeast were used to evaluate TAC algorithm efficiency. These datasets are taken from the repository of UCI. For these datasets, the minimum support value is 5% of the dataset size.

4. COMPARISON OF TAC ALGORITHM WITH K-MEANS ALGORITHM

To Compare Clustering Results of TAC algorithm on Artificial and Real Datasets, the portion of dataset which lies within the ambit of significant clusters is taken as an input to the k-means algorithm. The value of k required in the k-means algorithm is taken from the dataset clustered using TAC algorithm. It is equal to the number of significant clusters generated. in this section, the results of this comparison are presented. clustering results of k-means algorithm on artificial datasets Ad_2k_1, Ad_1.5k_2, Ad_1.5k_3, Ad_1.5k_4, Ad_1.5k_5 And Ad_1.5k_6 are shown respectively. From these figures, it is evident that different partitions of a given dataset are generated. Clustering results, including comparison between k-means algorithm and tac algorithm are summarized in the following table . For these datasets K-means algorithm is executed for the same eight real datasets as used by the TAC algorithm. The clustering results obtained for these datasets are presented in table 3 and it has been observed that k-means algorithm partitions the dataset into K clusters such that the range of the number of data objects in a cluster is narrow whereas the proposed algorithm can handle a wider range. It indicates that k-means algorithm generates clusters of almost same sizes whereas tac algorithm generates clusters of different sizes.

Table 2. COMPARISON OF TAC WITH K-MEANS

	#	TOTAL	ALGORITHM	

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

DATASET	CLUSTERS (K)	DATA OBJECTS	# DATA OBJECTS IN CLUSTERS	# DATA OBJECTS IN CLUSTERS
BCW-O	4	617	179, 79, 94 AND 265	40, 449, 89 AND 39
ECOLI	2	276	204 AND 72	55 AND 221
GI	3	203	89, 23 AND 91	18, 174 AND 11
HABERMAN	4	295	31, 74, 81 AND 109	20, 200, 31 AND 44
IRIS	3	149	49, 39 AND 61	31, 96 AND 22
SEED	7	174	25, 18, 24, 26,27, 25 AND 29	26, 21, 38, 20, 25, 21, AND 23

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

WINE	4	159	70, 56, 25 AND 8	20, 78, 32 AND 29
YEAST	5	1253	341, 214, 124, 208 AND 366	150, 718, 81, 143 AND 161

5. Conclusions

From the experiments carried out in the job described in this article the following observations were observed. TAC algorithm automatically clusters the datasets without knowing the number of clusters. Many clusters generated using this algorithm, however the parameter minimum support plays an important role to obtain significant clusters. The proposed algorithm can generate both overlapped and non-overlapped clusters. Since the algorithm is deterministic in its nature it gives same result on successive runs for a given dataset and given minimum support value it can detect outliers. It doesn't generate no spatial clusters as it is based on centroid based clustering. It requires adjacency matrix as the prerequisite, which increases the time complexity of the algorithm. Overlapped and non-overlapped clusters. It has an advantage over the single pass and modified single pass clustering algorithms as it does not depend on the order of the selection of the data objects. Here, it has also been observed from the experiments that tac algorithm generates clusters of different sizes whereas k-means algorithm generates clusters of almost same sizes. Therefore, the proposed TAC algorithm can be used as an alternate clustering algorithm.

References

1. katsavounidiset , Roger P., Muller-Gorman I. And Zimek A., (2017), "Detection And Visualization Of Subspace Cluster Hierarchies," In Advances In Databases: Concepts, Systems And Applications, Springer, Berlin, Germany, Pp. 152-163.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

2. Redmond and heneghan (2017), "Fast Algorithm For Projected Clustering," In Proc. AcmSigmod International Conference On Management Of Data, Acm Press, Pp. 61-72.
3. Celebiet ,Aggarwal C. C. And Yu P. S., (2017), "Finding Generalized Projected Clusters In High Dimensional Spaces," In Proc. AcmSigmod International Conference On Management Of Data, Acm Press, Pp. 70-81.
4. Celebiet al. (2014), , "Automatic Subspace Clustering Of High Dimensional Data For Data Mining Applications," In Proc. AcmSigmod International Conference On Management Of Data, Seattle, Wa, Pp. 94-105.
5. Bradley and fayyad (2016) "Fast Algorithms For Mining Association Rules," In Proc. Of 20th Vldb Conference, Pp. 487-499.
6. Reddy and jana (2015), "A K-Mean Clustering Algorithm For Mixed Numeric And Categorical Data," Data & Knowledge Engineering, Vol. 63, Pp. 503-527.
7. Aimet,.Ankerst M., Breuing M. M., Kriegel H-P And Sander J., (2016), "Optics: Ordering Points To Identify The Clustering Structure," In Proc. AcmSigmod International Conference On Management Of Data, Pp. 49-60.
8. Zhang and cheng (2016), "K-Means++: The Advantage Of Careful Seeding," In Proc. Of Symposium Of Discrete Analysis, Pp. 1027-1035.
9. Hatamlouet ,Arya S., Mount D. M., Netanyahu N. S., Silverman R. And Wu A. Y., (2014), "An Optimal Algorithm For Approximate Nearest Neighbor Searching In Fixed Dimensions," Journal Acm, Vol. 45, No. 6, Pp. 891-923.
10. Ashraf F., Ozyer T. And Kim et al. (2014), "Employing Clustering Techniques For Automatic Information Extraction From Html Documents," Ieee Transactions On Systems Man Cybernetics-Part C: Applications Reviews, Vol. 38, No. 5, Pp. 660-673.
11. Aslam J. A., Pelekhov E. And Yang and zhu (2014), "Star Clustering Algorithm For Static And Dynamic Information Organization," Graph Algorithms And Application, Vol. 8, No.1, Pp.95-129.
12. Assent I., Krieger R., Muller E. And Seidl T., (2017), "Visa: Visual Subspace Clustering Analysis," AcmSigkdd Explorations, Vol. 9, No. 2, Pp. 5-12.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

13. Bagirov A. M., Ugon J. And Bagirovet al. (2015, "Fast Modified Global K-Means Algorithm For Incremental Cluster Construction," Pattern Recognition, Vol. 44, No. 4, Pp.866-876.
14. Balcan M. F., Blum A. And Gupta A., (2013), "Clustering Under Approximation Stability," Journal Of The Acm, Vol. 60, No 2, Pp. 1-34.
15. Huang, Boley D., Gini M (2015), "Scalable Clustering Algorithms With Balancing Constraints," Journal Of Data Mining And Knowledge Discovery, Vol. 13, Pp. 365-395.
16. Nguyen ,Beg M. M. S. And Ahmad N., (2016), "Web Search Enhancement By Mining User Actions," International Journal Of Information Sciences, Elsevier, Vol. 177, No. 23, Pp. 5203-5218.
17. Nguyen et al. (2015), "Multidimensional Binary Search Trees Used For Associative Searching," Communications Acm, Vol. 18, No. 9, Pp. 509-517.
18. Berkhin P., (2016), Survey Of Clustering Data Mining Techniques, Berlin: Springer, Pp. 25-71.
19. Beyer K., Goldstein J., Ramakrishnan R. And Shaft U., (2014), "When Is 'Nearest Neighbor' Meaningful?," In Proc. Of 7th International Conference On Database Theory (Icdt-1999), Lncs 1540, Jerusalem, Israel, Pp. 217-235.
20. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)