[Volume 2 issue 5 May 2014]
Page No.422-427 ISSN :2320-7167

INTERNATIONAL JOURNAL OF MATHEMATICS AND COMPUTER RESEARCH

N-Mst Based Split and Merge Clara Clustering

Swati Joshi, Farhat Ullah Khan, Thakur

Department of CSE , ASET, Amity University, Noida(Uttar Pradesh),India
cse.swatijoshi@gmail.com
Department of CSE,ASET, Amity University, Noida(Uttar Pradesh),India fukhan@amity.eduNarina
Department of CSE , Bharati Vidyapeeth College of Engg.,New Delhi, India, narinat@gmail.com

ABSTRACT
Clustering is the act of assembling N data points into K clusters in order that, data points inside the same cluster are analogous, while data points in unlike clusters are dissimilar from each other. The majority of clustering algorithms befall ineffective when unsuitable parameters are provided, or implemented on datasets which are consist of clusters with varied form, dimension, and masses. To lessen these insufficiencies, we propose a new split-and-merge clustering methodology. In which an n-MST (Minimum Spanning Tree) is formed to lead the splitting and merging process. The proposed method doesn't need any prior domain knowledge of dataset. Experimental consequences exhibit its efficiency on real datasets.

Keywords
Clustering, K-means, CLARA, MST (Minimum Spanning Tree), SAM(Split and Merge method).

1. INTRODUCTION
Clustering is act of determining groups of data points such that the points in a group will be similar (or related) to one another and dissimilar from (or unrelated to) the data points in other groups. A cluster is a set of entities which are alike, and entities from different cluster are not alike[1]. Cluster may be described as connected areas of a multi- dimensional space comprised of a relatively high density of points, and are separated from other clusters (having high density of points)by regions of low density of points[3].

Due to the imprecise nature of clustering, one algorithm or one approach isn't sufficient to calculate the best cluster in all the possible situations. Hence over time a number of clustering algorithms [4] have been created. But the most of algorithms occur inefficient when inappropriate parameters are given or when datasets are provided which are of diverse shape, sizes and densities. In most of the existing algorithms domain knowledge of dataset is required.

To minimize these limitations we propose a new split and merge clustering methodology which is based on n-mst based CLARA clustering. In it we first make n-mst to lead the split and merging process. Later, in split stage, nodes having highest degree are selected as initial prototypes to split. Afterwards, during merging, neighboring pairs of trees are elected to be merged. Then K-value is determined from n-mst, which is the basic requirement for making Clusters. Next we apply CLARA algorithm using K for creating efficient clusters. So in our methodology domain knowledge of dataset is not required.

2. RELATED WORK
K-means Algorithm (Partitional Clustering) represents a prototype in terms of a centroid, which is normally the mean of a set of data points. It is relevant to the objects in a continuous and n dimensional space. It is simple and has low complexity, whereas number of clusters need to be known beforehand and easily influenced by outliers [4].CLARA also a partitional clustering, is simple and works well on large dataset. It is not as easily influenced by outliers. Whereas it requires number of clusters in advance and has high complexity [5]. Partitional clustering produces inaccurate results when the objective function used does not capture the intrinsic structure of the data. So this is the reason for partitional clustering to be incapable in handling clusters of arbitrary shapes, distinct sizes and densities [7].

F. Murtagh and W. Day describe, Hierarchical Clustering forms a ladder of clusters or tree of clusters, so discovering data on diverse level of granularity. This algorithm doesn't need to know the number of clusters beforehand, while it doesn't deal well with huge dataset. Hierarchical clustering deals with comparatively soaring technological cost. Single linkage and complete linkage are two well-known examples of hierarchical clustering algorithms, and they take O(N^2 logN) time [6][8]. Clustering algorithms that unite the pros of hierarchical and partitional clustering have been proposed in the Literature [8][9]. In these hybrid methods work is done in two stages. In the first stage dataset is divided by partitioning algorithm and in the second stage merging is done by similarity measures. In the first stage K-means may create unlike partitions in different runs, the last results may be unbalanced. A minimum spanning tree (MST) is a functional graph structure, which has been used to confine imaginary grouping [10][13]. Zahn defined numerous criteria of edge variation for searching clusters of different shapes [11]. But the method cannot give an adaptive selection of the criteria for the dataset consisting of diverse shaped clusters. Xu et al.[12] gave three MST-based algorithms: eliminating lengthy MST-edges, an iterative algorithm based on centroid concept, and a global optimal algorithm based on the concept of medoids. But for a particular dataset which algorithm should be applied it is not known to user.

Caiming, Zhong, proposed a minimum spanning tree based split-and-merge method (SAM)[2]. It works on numerical data and assumes that the graph can be calculated in a vector space. In split stage three iterations of MSTs are used to construct a neighborhood graph called 3-MST graph. In the merge stage, the adjacent groups with respect to the MST are selected out and considered for merge. The computational complexity of the proposed method (SAM) is O(N^2), which is conquered by the building of the 3-MST graph. If the factor of dimensionality d is also taken in account, the accurate complexity would be O(D*N^2). Major drawback for this algorithm is that though it prevents outlier detection it also causes a number of important data points to get missed. For a huge dataset too many smaller subsets would be obtained which could lead it to being ineffective.

3. PROPOSED METHODOLOGY

We have proposed the new methodology based on Minimum Spanning Trees i.e. Split and Merge Clustering Methodology. The Methodology comprised of three main stages: 1)Construct an n-MST graph 2)Split 3)Merge 4)Determining k value for n-mst

The descriptions of the stages are given below:

1) Construct an n-MST graph: Make an n-mst Graph where n is determined by the number of columns in the dataset where N=columns/2. Let Gmst(X,n) denote the n-MST graph, which is defined as a union of the n MSTs: Gmst(X,n) = T1 U T2… U Tn. In this paper, we use Gmst(X0,n) to determine the initial prototypes in the split stage and to calculate the merge index of a neighboring partition pair in the merge stage.

2) Split:
In the split stage, nodes having highest degree in the graph Gmst(X0,n),are elected as initial prototypes. Initially pruning is done by removing nodes having degree one from the n-mst graph. Later, on pruned dataset K-means is applied using built prototypes. The formed partitions are accustomed to maintain the clusters be connected with respect to the MST.

3) Merge:The merge stage is executed to acquire the resulting clusters; after X0 has been divided into subgroups. On the basis of MST-based clustering unconnected subtrees cannot be merged. As a result, just the neighboring pairs relating to Tree are the candidates.

4) Determining k value for n-mst: Using the cluster function, determine the value of k for the created n-mst graph. Using this value of k, any clustering algorithm like K-means, CLARA can be applied to dataset to find desired clusters, as we got the required k-value. In this paper we apply CLARA[14][15] on the dataset.

4. DATASET INFORMATION

- The IRIS [19] data set consist of 3 classes, and each class has 50 instances.

- One class is linearly independent of the other two; the latter two are not linearly independent of each other.

Attribute Information:

1. sepal length in cm(SL)
2. sepal width in cm(SW)
3. petal length in cm(PL)
4. petal width in cm(PW)
5. class:
   -- Iris Setosa
   -- Iris Versicolour
   -- Iris Virginica

Number of Attributes (Columns): 4

5. IMPLEMENTATION

5.1 Technology adopted

In this research we implemented proposed algorithm in R-studio[16][17], which works in conjunction with R-tool. The R statistical programming language is a free open source package based on the S language developed by Bell Labs. Many statistical functions are already built in. Contributed packages expand the functionality to cutting edge research[17][18]. As it is a programming language, so to complete tasks, computer code need to be generated.

Advantages of R-

- R has over 800 built in packages providing numerous inbuilt functions.
- Fast and free.
- Interfaces with database storage software (SQL).
- Excellent especially for data analysis.

5.2 Applying Methodology:

5.2.1 Creation of n-mst graph :

First we have to divide the graph depending on the number of columns. The IRIS dataset has 4 columns hence n=2 here. We need to determine the n-mst graph with n=2 here. So we have to create 2-mst, which is derived as :
union of (mst-1, mst-2)

a) Creation of mst-1 (PL, PW):
We create mst-1, by taking dataset -1 as input in the code for making minimum spanning tree in R. Dataset-1 contains instances for attributes PL and PW.

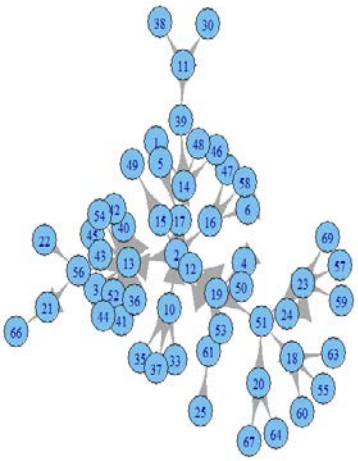

Fig 1: Plot of MST-1

b) Creation of mst-2(SL, SW):
We create mst-2, by taking dataset -2 as input in the code for making minimum spanning tree in R. Dataset-2 contains instances for attributes SL and SW.

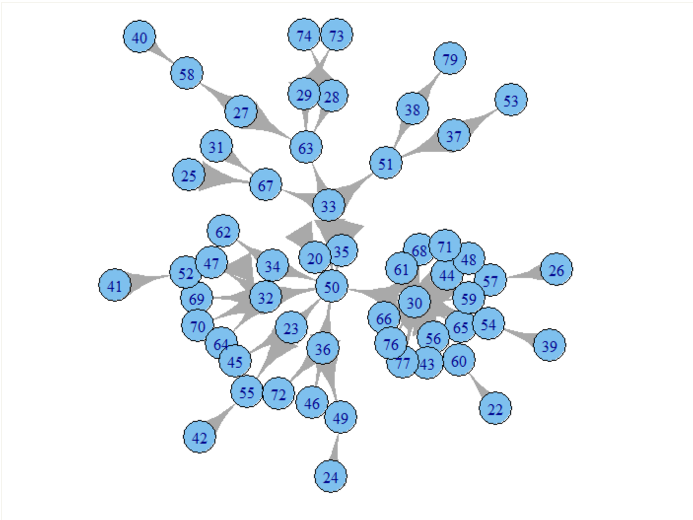

Fig 2: Plot of MST-2

c) Creation of 2-mst graph:
2-mst graph is created as n=2; from IRIS dataset- 2-mst = union of (mst-1, mst-2)

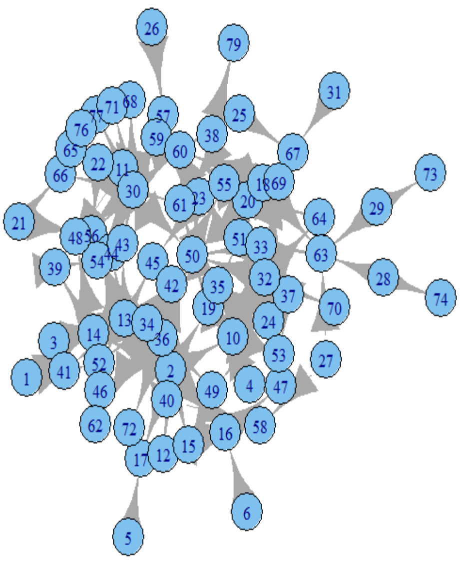

Fig 3: Plot of 2-MST graph

5.2.2 Determining K value from n-mst:

a) K value is determined from **:** K<-clusters(value in 2-mst) where 2-mst value taken from above(c) .

b) $membership [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 [42] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1[83] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

5.2.3 Applying CLARA using the Value of K:

After getting the K value we apply CLARA algorithm to the cluster function in R to find the desired clusters. In the function, iris dataset, containing all the attributes: PL, PW, SL, and SW, is applied as input.

6. RESULT AND ANALYSIS:

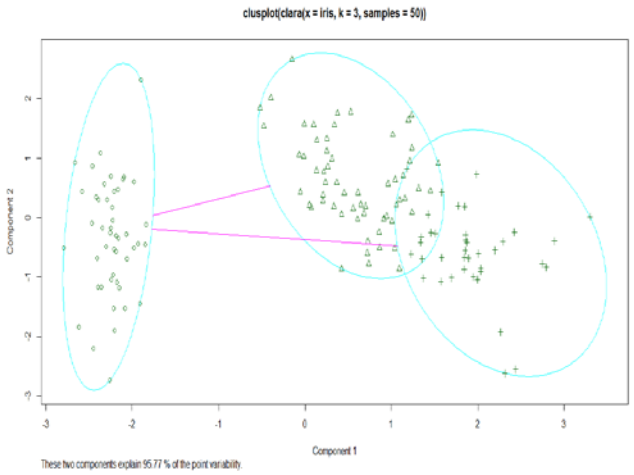

Fig 4: Performance-plot1 after putting K value

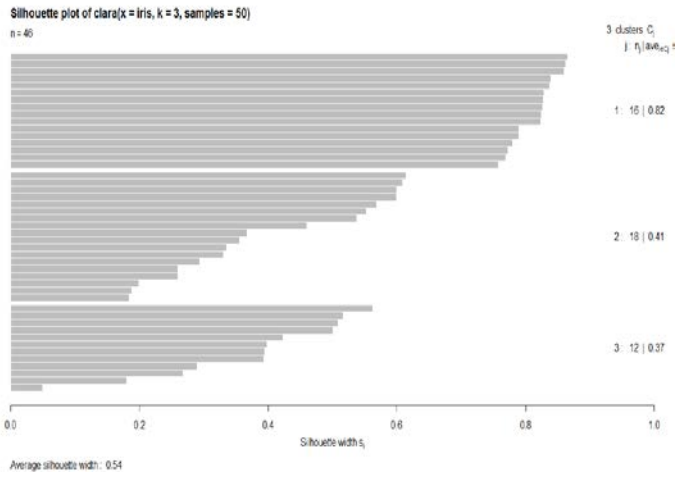

Fig 5: Performance-plot2 after putting K value

1. From Fig:4 , it seems that our methodology is better, as it identifies three clusters as corroponding to three classes: Setosa, Versicolour and Virginica. As K value has been derived not taken as default, the above dataset has given 95.77% accuracy. In CLARA the K value is initially needed to pass in the cluster function.

2. The overall average silhouette width of the silhouette plot is also counted as quality index for analyzing the quality of the cluster structure. If silhouette width is greater than 0.5, then structure is of reasonable quality [20]. So in Fig:5 , silhouette width is 0.54.

3. The above method also gave better than the 92.5% result of SAM.

7. CONCLUSION

Experiments have confirmed the significance of each step of the methodology. The proposed method utilizes minimum spanning trees in different phases. N-MSTs on an iteratively developed graph are computed and united to find the initial prototypes for K-means, because arbitrarily chosen initial prototypes would direct to unbalanced partitions. The proposed method does not have any straight limitations for being

applied to datasets with high dimensions. No parameters are to be decided by the user and method doesn't need any prior domain knowledge of dataset. Moreover, this approach gives higher accuracy as compared to SAM

REFERENCES

[1]Swati Joshi, Farhat Ullah Khan, Narina Thakur,"Contrasting and Evaluating different Clustering Algorithm:A Literature Review",
[2] Caiming Zhong ,Duoqian Miao," Minimum spanning tree based split-and-merge: A hierarchical clustering method"
Journal of Information Sciences, Volume 181 Issue 16,August 2011,
Elsevier ScienceInc.New York,USA,pages:3397-3410.
[3]E. Mooi and M. Sarstedt, "A Concise Guide to Market Research",DOI 10.1007/978-3-642-12541-6_9,Springer-Verlag Berlin Heidelberg 2011.
[4] Xindong Wu · Vipin Kumar · J. Ross Quinlan," Top 10 algorithms in data mining", International Conference on Data Mining (ICDM) in December 2006.
[5].A. K. Jain and R. C. Dubes. "Algorithms for Clustering Data." Prentice-Hall, Englewood Cliffs, NJ, 1988.
[6] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms.Computer Journal, 26(4):354–359, 1983.
[7] S.Anitha Elavarasi and Dr. J. Akilandeswari and Dr. B.
Sathiyabhama, January 2011, A Survey On Partition Clustering Algorithms.
[8] D. Cheng, R. Kannan, S. Vempala, G. Wang, A divide-and-merge methodology for clustering, ACM Trans. Database Syst. 31 (2006) 1499–1525.
[9]G. Karypis, E.H. Han, V. Kumar, CHAMELEON: a hierarchical clustering algorithm using dynamic modeling, IEEE Trans. Comput. 32 (1999) 68–75.
[10] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[11] C.T. Zahn, Graph-theoretical methods for detecting and describing gestalt clusters, IEEE Trans. Comput. C-20 (1971) 68–86.
[12] Y. Xu, V. Olman, D. Xu, Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning tree, Bioinformatics 18(2002) 536–545.
[13]W. Day and H. Edelsbrunner., "Efficient algorithms for agglomerative hierarchical clustering methods". Journal of Classification, 1(7):7–24, 1984.
[14] Raymond T. Ng and Jiawei Han.," CLARANS: A Method for Clustering Objects for Spatial Data Mining. "IEEE Transactions on Knowledge and Data Engineering, 14(5):1003{1016, 2002.
[15] R. T. Ng and J. Han. ,"Efficient and Effective clustering methods for spatial Data Mining", Proc. of the 20th Int'l Conf.on Very Large Databases, Santiago, Chile, pages 144–155,1994.
[16] Maechler, M. Package 'cluster'. http://cran.r-project.org/web/packages/cluster/cluster.pdf.
[17] R tool .< http://cran.r-project.org/>
[18]R packages . <http://cran.r-project.org/web/packages/>
[19]Iris dataset . http://archive.ics.uci.edu/ml/datasets/Iris
[20] Anja Struyf," Clustering in an Object-Oriented Environment",Journal of statistical software, Vol. 1, Issue 4, Feb 1997.
[21]J. C. Bezdek et al., Fuzzy models and algorithms for pattern recognition and image processing, KluwerAcademic, 1999.
[22]L. Zadeh, "Furzy sets," Information and Control, vol. 8, pp. 338-353, 1965.
[23] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algori-thms, PlenumPress, New York, 1981.
[24] R. Krishnapuram, and J. Keller, "A Possibilistic Approach to Clustering," IEEE Trans. Fuzzy Systems, vol. 1(2), pp. 98-110, 1993.
[25]Data Mining WikiBook, http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/