JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-15, No.-1, January (2020) pp 339-348 ISSN (Print) 0973-8975

An Optimized Clustering Method to Create Clusters Efficiently

P. Praveen[1], B. Rama[2]

[1]Department of Computer Science and Engineering, S R Engineering College, Warangal, Telangana State, India

[2]Department of Computer Science, Kakatiya University, Warangal, Telangana State.

Email: [1]prawin1731@gmail.com, [2]rama.abbidi@gmail.com

https://doi.org/10.26782/jmcms.2020.01.00027

Abstract

The problem of mining numerical data and to propose different approaches to efficiently apply clustering to such data According to an aspect of the method the Mean Base Divisive Clustering (MB-DivClues) method is developed to categories unstructured data into various groups. The a constructive mean-based divisive clustering method is developed to reduce comparison includes several steps which includes identification of mean value from a given dataset, to find the arithmetic mean value of base cluster-infrequent attribute and storing the found mean value in a tree which is represented as root. Further the steps include comparing the objects in the dataset with the said mean value and stored in the nearest cluster. A new cluster is created to place the sorted object in new cluster. In the process of proposed method includes steps of shifting the object value to the left cluster when it is less than the mean value, shifting the object value to right cluster when it is greater than the mean value and repeating the above procedure until singleton cluster is picked from the given dataset. Wherein before applying divisive Clustering method, initially all the data objects are available in a single cluster and a mean value is calculated on the dataset.

Keywords : Classification, Clustering, Data mining, Divisive Methods, Mean Based Divisive Method (MB-DivClues)

I. Introduction

This study describes the past studies carried out on analyzing the clustering, which includes the following concepts of study in previous years, such as K- means, BIRCH,etc how clustering concepts was implemented in finding an efficient cluster in a Divided cluster. Classification and clustering are two most essential methods that partition the entities that have numerous featuresas significant disjoint subclasses so

that entities in every group are highly similar to each other in the attribute values than they are to objects in other group(I).

In supervised classification, the classes are described, the user already knows what classes there are, and some schooling data this is already categorized with the aid of their elegance membership is available to train or build a model. In cluster analysis, one does no longer recognize what instructions or clusters exist and the trouble to be solved is to institution the given records into significant cluster. Just like an application of supervised classification, cluster evaluation has packages in many special regions including in advertising and marketing, medication, commercial enterprise. Practical packages of cluster analysis have also been determined in individual recognition, net analysis and classification of documents, category of astronomical statistics and category of objects located in an archaeological observe.

The traditional clustering algorithms established in statistics assumed small datasets but with the beginning of bar codes, computing, and the web, users request to smear cluster analysis to large datasets. Ancient methods have therefore been adapted to cob large datasets and innovative methods are being industrialized.

Various algorithms can be used for Cluster analysis that be different expressively in their perception of creates a group also how to proficiently discover them. Prevalent ideas of clusters contain groups with diminutive distances between the group objects, data space in dense areas, specific numerical distributions. The groups can thus express as a multi-objective optimization difficulty.

To get an anticipated results and individual dataset, a suitable clustering algorithm and parameter settings should be used. The distance function to use, no. of probabilistic clusters or a density threshold are the values used for clustering methods. This research is a stepping stone of optimization as it involves trial and failure concept. Cluster analysis as such is not an automatic task, but repetitious process of knowledge finding or interactive multi objective optimization.

To achieve the result with expected properties, the modification of data preprocessing and model parameters is necessary.

The clustering methods

- Partitioning method
- Hierarchical method
- Model Based method
- Grid Based method
- Density Based method

All the above mentioned approaches ensure its own specialization and traditional methods which are commonly used in the process of data mining real application.

II. Review of Literature

Clustering algorithms according to Patel and Mehta (2011), produce clusters consuming parallel among data objects based on features belong to the same cluster.

Clustering algorithms that have been most often used and spread in most of the fields of publication areas such as, machine learning, pattern recognition [II], artificial intelligence, information technology, medical, biology, image processing, marketing and psychology. Consequently, the main objective of clustering will isolate a finite unlabeled data set into a fixed and separate set of natural, hidden data arrangements, rather than to deliver a correct representation of hidden samples shaped the same possibility distribution [I].

Han et al., (2011) mentioned as "an outlier is an information object that differs extensively from the relaxation of the objects" as if it has been produced through a different mechanism. Also, Barnet and Lewis (1994) show that the outlying word utilized to depart significantly from other contributors of the pattern where it occurs[VI-VIII].

Yu-Chen Song (2008) as a result, proximity-centered outlier detection process has two varieties: distance-established and density-founded approaches. A distance founded outlier detection approach accesses the region of an object, which is demarcated through a given radius. An object is taken as an outlier if its near cluster does not have sufficient elements. A density [III] founded outlier detection method examines the density of an object and that of its neighbors. Right here, an object is recognized as an outlier if its density is moderately much decrease than that of its neighbors [III][IV][V-IX].

Consequently, with the intention to discover outliers by using k-mean [X] algorithm acknowledged that, with k-means, data are separated into k partitions via allocating them to the regional cluster centers. Subsequently able to calculate the distance (or dissimilarity) between the items and its cluster center, and select those with far distances as outliers. On this part, they were going to speak about a couple of means proposed with the aid of plenty of researchers to discover outliers in K-mean and K-Medians clustering algorithms[XI-XII].

Vijayarani and Nithya (2011), awarded some algorithms as PAM [X][XI], CLARA , and CLARANS with a new proposed clustering algorithm known as ECLARANS for outliers detection[XII]. The experimental effect indicates that, the proposed algorithm ECLARANS [XIII] increases the accurateness of detection and CLARANS decreases the time complexity when associated with other algorithms.

Babenko and V. Lempitsky (2015) further classified the clustering algorithm into a hierarchical and partitional algorithm [XV]. The hierarchical method applied to produce a tree-like structure using two approaches, divisive and agglomerative. In divisive approach, the algorithm considered all data items as a single cluster and by using the top-down approach; it continued to divide the data into the smaller cluster.

III. Proposed Method MB-DivClues

The MB-DivClues is a novel linkage method as it aims to identify all the singleton clusters for performing k observations over a cluster C_r with k observations are performed over other cluster denoted by C_v. In the subject of MB-DivClues, this

thesis proposes two novel aspects: MB-DivClues for creating clusters and MB-DivClues for searching an object in clusters

1. MB-DivClues: here we are creating singleton clusters to compute the mean of all objects between k distinct closest interpretations in clusters for identifying the average of nearest mean. Definition:1 defines MB-DivClues concept for the first time.

   Definition . Let X = { $I_1, I_2,\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,I_n$} is initially a single cluster, a set of m observations in left cluster C_l contains {$I_1, I_2,\ldots\ldots\ldots ,I_m$} and a set of m+1 to n observations in right cluster C_r contains {$I_{m+1}, I_{m+2},\ldots\ldots\ldots ,I_n$}, where $x\in C_l$ and $x\in C_r$, as shown in Equations(4.5.1).

   Objective function for MB-DivClues:

   $$C(m_i) = \frac{1}{|k|}\sum_{i=1}^{k} xi \ \{C(I_1, I_n)\} \quad (1)$$

   On the basis of MB-DivClues method, one cluster which has the mean member on average is split into equal stages of the process where more than one pair of observations are identified and implied on the same process.

Constructing MB-DivClues Algorithm for Clustering Data

Algorithm MB-DivClues(n, Dct)

Input: X = { $I_1, I_2,\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,I_n$}

Output: Y = { { I_1 }, {I_2}, ……..{I_n} }

Step 1 . Begin

Step 2. I ← 0

Step 3 . Dct ← { $I_1, I_2,\ldots\ldots\ldots,I_n$}

Step 4 . Nro_c ← n

Step 5 . C_i ← 0, {$\forall\, i = 1,2\ldots.n$}

Step 6. C_j ← 0, {$\forall\, j = 1,2\ldots.n$}

Step 7 . repeat

Step 8. Flag=1

Step 9. for each 1 to n

Step 10. m=divclues_mean{ $I_1, I_2,\ldots\ldots\ldots ,I_n$ }

Step 11. If flag is even

Step 12. Create_clust(C_{ij}top, C_{ij}bottom)

Step 13. C_{ij}top= {$I_1, I_2,\ldots\ldots\ldots ,I_m$}

Step 14.C_{ij}bottom = {I_{m+1}, I_{m+2},……. ,I_n}

Step 15. If flag is odd

Step 16 .Create_clust(C_{ij}left, C_{ij}right)

Step 17. C_{ij}left= {I_1, I_2,……. ,I_m}

Step 18. C_{ij}right= {I_{m+1}, I_{m+2},……. ,I_n}

Step 19. until Nro_c <= 1.

Step 20. end

Referring the above algorithm there is a vital scope to execute some of the steps in a iterative or parallel way in a distributed environment. Some of the possible steps that can be executed parallelly are briefly defined below in following steps:

In Step 2 we initialize the item value to zero, stating that the item set is empty. In next step that is step 3 we take the dataset objects that are numerous in number from database into Dct. In next step we have defined Nro_c a parameter represent the number of objects taken into consideration for example objects such as {A, B, C, D, E, F }. In step 10 we call the DivClues-mean algorithm that generates mean value of the given data set. In step 11 we create the cluster in a horizontal manner where we create a top cluster and bottom cluster and in step 15, we create the cluster in a vertical manner where we create a left cluster and right cluster.

Finding Mean Values for DivClues-mean Algorithm

Algorithm: DivClues-mean

Step 1. Start

Step 2. Read total number of items or objects from database, say n number
Step 3. Initialize counter variable say, i ← 1 and initialize variable SUM ← 0
Step 4. Repeat the below steps till i < n, otherwise goto step 9
Step 5. Read the element or object from data set say x.
Step 6. Sum ← Sum+ x
Step 7. i ← i + 1
Step 8. Repeat the above steps from step 3
Step 9. mean ← Sum/n
Step 10. Stop

Graphical Representation of Proposed Model

In this research, the formal definition of MB-DivClues on dataset, where H represents set of objects such that H = {O_1, O_2, . . . , O_i, . . . , O_N}, where N represents the total number of objects in H and O_i is the i^{th} object or element in H. Such that each object O_i is defined by a unique Object Identifier $O_i.ID$. The dot notation is used to represent the access the identifier and other various component parts of a given object. set of N observations in left cluster C_l contains {I_1, I_2,……. ,I_m} and a set of m+1 to n

observations in right cluster C_r contains $\{I_{m+1}, I_{m+2}, \ldots\ldots\ldots, I_n\}$, these all are shown in below Fig 1.

The cluster plotting diagram which is generated is:

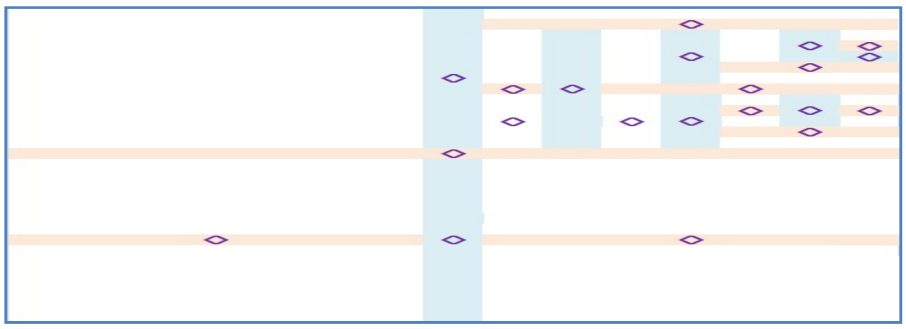

Fig 1 Cluster plotting using Vector quantization on Dataset

The values plotted for the example are:

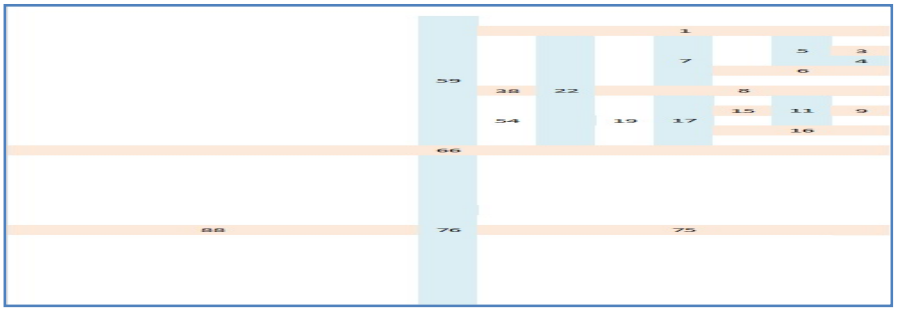

Fig 2 MB-DivClues Algorithm for significant value plotting for clusters

In the above scaling implementation the random sample of values are: (66, 59, 1, 22, 8, 7, 6, 17, 16, 11, 9, 76, 5, 3, 19, 38, 54, 15, 88, 4, 75) whose values have qualified minimum threshold of confidence and support. The process of implementation of algorithm starts by considering the first value as the mean value that is 66 and plotted or stored in the middle of the diagram. Based on the next value horizontal scaling is performed. If the value is greater than the mean value then the value will be plotted in the above partition otherwise scaling is performed in the bottom partition. As the value is 59 and is less than the mean it is plotted in the top partition. Then the next value 1 is compared with the mean as the value is lesser than mean then the search moves to the top partition and if value is available then the comparison is done with 22 as it is lesser, it is compared with the right partition value which is 1, as the value is greater than 1 and no bottom partition is created, horizontal scaling is implemented and 22 is stored in the bottom partition.

The next value to be created is 8, it is compared with the mean as the value is lesser than mean then the search moves to the top partition then it is compared with 8 as it is lesser, it is compared with the right partition value which is 1, as the value is greater than 8 is compared with bottom partition whose value is 22, as 8 is lesser and no right

partition is created, vertical scaling is implemented and 8 is stored in the right partition.

As the unique values are taken hence there is no scope of data complications and loss of data. All the values are created in vectors which are in other words the dynamic arrays whose size increases or decreases and allows us to efficiently scale the data or allows us to effectively partition the data.

IV. Results by using synthetic data set

In this section, the application of proposed MB-DivClues method in Divisive hierarchical clustering is implemented using two datasets shown in Table 4.1 represents probable datasets consisting of X and Y. For coordinating with all the possible instances that comprises of 18 unique tuples are identified by their significance ID. The 18 objects that are considered in the algorithm are: [A , B , C , D , E , F , G , H , I , J , K , L , M , N , O, P , Q , R] where every item is kept into one group or chose to be a group.

Table 1 Synthetic Dataset for Generated Clusters using MB-DivClues

Dataset COBB			
Instance	**Significance of**	**Significance of**	**Significance of**
A	2	7	1
B	2	6	1
C	3	6	1
D	3	8	2
E	4	6	2
F	5	8	2
G	7	9	3
H	12	16	4
I	13	16	2
J	14	16	4
K	14	17	4
L	15	16	4
M	16	16	3
N	15	18	3
O	12	9	3
P	13	9	1
Q	15	9	4
R	16	8	3

Table 1 shows the clustering steps which are complete for *MB-DivClues* that consists of one dataset with two user defined parameters are selected using Binary search approach.

In the initial step, every perception in the dataset is expected as one particular cluster which as contain objects like as {A, B , C, D , E ,F G , H , I , J , K , {L , M, , N , O ,

P , Q, R}}. From that point onward, we locate the mean estimation of the above informational collection of clusters as indicated by the chose mean criteria is part into two groups in each progression. For exemplar; the groups in the early stair are group one LC1 is {A, B , C, D , E ,F G} and the second group RC1 is { H , I , J , K , L , M, , N , O , P , Q, R} so groups LC1 is part into LC2 and RC2 .

The process is repeated in an iterative fashion till all clusters are divided into clusters that are of singleton. Table 3.1 represents the generation of new cluster creation process by dividing into two existing clusters which are identical in nature.

Visualizes the clusters and Figure 3.1 shows that the relevance of MB-DivClues algorithm based on dissimilar mean method can construct dissimilar clustering outcome at the subsequent steps. Here the MB-DivClues procedure (Table 3.2) comprises of means among the LC and RC that are denoted by **LC#2** and denoted by **RC#2** will be compared with the constant values and then clusters **C1** and **C3** are divided if and only if **m2** is greater than **m1**.

Table 2 Clustering Processes using Proposed Method

Steps	Cluster_Name	Clustering_Objects	Cluster_Significance Mean value
1	C#1	{A, B , C, D , E ,F G , H , I , J , K , L , M, , N , O , P , Q, R}.	10
2	LC#2	C#1→L {A, B , C, D , E ,F G}	3.7
3	RC#2	C#1→R { H , I , J , K , L , M, , N , O , P , Q, R}	14
4	L-LC#3	LC#2→L {A, B , C, D }	2.5
5	L-RC#3	LC#2→R { E ,F G}	5.3
6	L-LC#4	L-LC#3→L {A, B }	2
7	L-RC#4	L-LC#3→R{ C, D }	3
8	L-LC#5	L-LC#4→L{A }	Final singleton cluster
9	L-RC#5	L-LC#4→R{ B }	Final singleton cluster
10	L-LC#5	L-RC#4→L { C }	Final singleton cluster
11	L-RC#5	L-RC#4→R { D }	Final singleton cluster
12	L-LC#4	L-RC#3→L { E ,F}	4.5
13	L-RC#4	L-RC#3→R{ G}	Final singleton cluster
14	L-LC#5	L-LC#4→L { E }	Final singleton cluster
15	L-RC#4	L-RC#4→R { F}	Final singleton cluster
16	R-LC#3	RC#2→L{ H , I , J , K , O , P }	13
17	R-RC#3	RC#2→R{ L , M, , N , Q, R}	15.5
18	R-LC#4	R-LC#3→L { H , I , O , P }	12.5

19	R-RC#4	R-LC#3→R{ J , K }	13
20	R-LC#5	R-LC#4→L { H ,O }	12
21	R-RC#5	R-LC#4→R { I , P }	13
22	R-LC#6	R-LC#5→L { H }	Final singleton cluster
23	R-RC#6	R-LC#5→R{ O }	Final singleton cluster
24	R-LC#6	R-RC#5→L { I }	Final singleton cluster
25	R-RC#6	R-RC#5→R { P }	Final singleton cluster
26	R-LC#5	R-RC#4→L { J }	Final singleton cluster
27	R-RC#6	R-RC#4→R{ K }	Final singleton cluster
28	R-RC#3	R-RC#3→R { L , M, , N , Q, R}	15.5
29	R-LC#4	R-RC#3→L{ L, Q }	15
30	R-RC#4	R-RC#3→R { M , N , R }	16.3
31	R-LC#5	R-LC#4→L { L }	Final singleton cluster
32	R-RC#5	R-LC#4→R { Q }	Final singleton cluster
33	R-LC#5	R-RC#4→L { M , R }	16
34	R-RC#5	R-RC#4→R { N }	Final singleton cluster
35	R-LC#6	R-LC#5→L { M }	Final singleton cluster
36	R-RC#6	R-LC#5→R { R}	Final singleton cluster

One of the promising advantage of MB-DivClues is the results attained are considerably similar with almost all the existing well known clustering techniques that are divisive. Because the input to the divisive algorithm consists of several objects then the split operation is performed over all the possible instances into a singleton clusters.

V. Conclusion

This Method describes about the research methodology in detail where the concept is illustrated in the form of a proposed tree structured algorithm where the corresponding applications available in the process of data clustering for handling larger data sets (in millions). Where the algorithm includes or considers various fields or attributes of objects that tends to process data in an efficient clustering method which includes the implementation of an algorithm in various sequence of steps pictorially or graphical implementation in detail which is based on the mean based divisive clustering.

References

I. A. Babenko and V. Lempitsky, "Tree Quantization for Large-Scale Similarity Search and Classification," in CVPR, 2015.

II. A. K. Agogino and K. Tumer. Ensemble clustering with voting active labels. Pattern Recognition Letters, 29(14):1947–1953, 2008.

III. Brandt, "Transform coding for fast approximate nearest neighbor search in high dimensions," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 1815–1822.

IV. J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A Survey on Learning to Hash," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 9, 2017.

V. M. Laszlo and S. Mukherjee, "Minimum Spanning Tree Partitioning Algorithm for Micro aggregation", IEEE Trans. on Knowledge and Data Engg., 17(7), 2005, 902- 911.

VI. Mohamed A. Mahfouz, d M. A. Ismail. Fuzzy Relatives of the CLARANS Algorithm With Application to Text Clustering. International Journal of Electrical and Computer Engineering. 2009 370-377.

VII. N. Paivinen, "Clustering with a Minimum Spanning Tree of Scale- free-like Structure", Pattern Recognition Letters, Elsevier, 26(7), 2005, 921-930

VIII. P.Praveen, B.Rama, "An Efficient Smart Search Using R Tree on Spatial Data", Journal of Advanced Research in Dynamical and Control Systems, Issue 4,ISSN:1943-023x

IX. P. Praveen, B. Rama and T. Sampath Kumar, "An efficient clustering algorithm of minimum Spanning Tree," 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, 2017, pp. 131135. doi: 10.1109/AEEICB.2017.7972398

X. Pushpa.R. Suri, Mahak. Image Segmentation With Modified K-Means Clustering Method. International Journal of Recent Technology and Engineering 2012 176-179.

XI. S.Vijayarani, S.Nithya. An Efficient Clustering Algorithm for Outlier Detection. International Journal of Computer Applications. 2011 22-27.

XII. Xiaochun Wang, Xiali Wang and D. Mitchell Wilkes, "A Divide-and conquer Approach for Minimum Spanning Tree-based Clustering", IEEE Transactions on Knowledge and Data Engg., 21, 2009.

XIII. Y. Chen, T. Guan, and C. Wang, "Approximate nearest neighbor search by residual vector quantization," Sensors, vol. 10, no. 12, pp. 11259– 11273, 2010.

XIV. Yu-Chen Song, J.O'Grady, G.M.P.O'Hare, Wei Wang, A Clustering Algorithm incorporating Density and Direction, IEEE Computer Society, CIMCA 2008

XV. Zhang, D. Chao, and J. Wang, "Composite Quantization for Approximate Nearest Neighbor Search," in ICML, 2014.