International Conference on Innovative Applications in Engineering and Information Technology(ICIAEIT-2017)

International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017

A study on monothetic Divisive Hierarchical Clustering Method

P.Praveen[1].*, B.Rama[2] ,Uma Dulhare[3]

1 Research Scholar && Assistant Professor in Department of Computer Science, SR Engineering College, Warangal, Telangana. prawin1731@gmail.com

2 Assistant Professor, Department of Computer Science, KakatiyaUniversity, Warangal,Telangana, rama.abbidi@gmail.com

3. Professor, Department of Computer Science, MuffakhamJah Engineering College, Hyderabad

Abstract*-- DIVCLUS-T is a divisive hierarchical leveled clustering Algorithm in view of a monothetic bipartition approach permitting the dendrogram of the progressive system to be perused as a choice tree. It is intended for either numerical or straight out information. Like the agglomerative progressive clustering Algorithms and the k-means dividing Algorithm, it depends on the minimization of the dormancy standard. Notwithstanding, dissimilar to Ward and k-means, it gives a straightforward and regular elucidation of the groups. In this paper we study what are the clustering algorithms and what are problems to split a cluster of Divisive Clustering using monothetic method.

***Index Terms*— Decision dendrogram, Divisive clustering, Inertia criterion, K-means, Monothetic clustering.**

I.INTRODUCTION

Classification and cluster are important techniques that partition the objects that have many attributes into meaningful disjoint subgroups [7] so that objects in each group are more similar to each other in the values of their attributes than they are to objects in other group [4].There is a serious distinction between cluster analysis and classification. In supervised classification, the categories are a unit outlined, the user already is aware of what categories there are a unit, and a few training data that's already tagged by their category membership is out there to training or build a model. In cluster analysis, one doesn't recognize what categories or clusters exist and also the downside to be resolved is to cluster the given data into purposeful cluster. Rather like application of supervised classification, cluster analysis has applications in many various areas corresponding to in promoting, medicine, business. Sensible applications of cluster analysis have additionally been found in character recognition, internet analysis and classification of documents, classification of astronomical information. The first objective of cluster is to partition a collection of objects into homogenized teams. a good cluster wants an appropriate live of similarity or unsimilarity. Thus a partition structure would be known within the sort of natural teams [9].

Clustering has been exuberantly applied in various fields as well as care systems, client relationships management, producing, biotechnology and geographical data systems. Several algorithms that type clusters in numeric domain are planned; however few algorithms are appropriate for mixed knowledge like collective [7], the most aim of this paper a way to unify distance illustration schemes for numeric knowledge. Numeric cluster adopts distance metrics whereas emblematic uses a tally theme to calculate conditional probability estimates for defining the relationship between groups [8].

II. LITERATURE SURVEY

Divisive hierarchical leveled clustering turns around the procedure of agglomerative hierarchical Clustering, by beginning with all articles in one group, and progressively partitioning every group into littler ones. A characteristic approach for isolating a group into two non-exhaust subsets is considering all the conceivable bi-segments. Plainly such a total identification methodology gives a worldwide ideal yet is computationally restrictive. An assortment of divisive clustering strategies which don't consider the sum total of what bipartitions have been proposed. MacNaughton -Smith et al. (1964) and Kaufman and Rousseeuw (1990) professional postured iterative divisive strategies utilizing a normal uniqueness between a protest and a gathering of items. Different techniques utilizing a disparity lattice as information depend on the enhancement of criteria, for example, the split or the width of the bipartition (Gu'enoche et al., 1991; Wand et al., 1996). For the inactivity rule, divisive partners to Ward's agglomerative calculation have been proposed: for instance, rather than part by aggregate specification it is conceivable to apply the k-implies calculation, with k=2 (Mirkin, 2005).

In divisive Clustering, a few techniques are polythetic, though some others are monothetic. A bunch is called monothetic if a conjunction of sensible properties, every one identifying with a solitary variable, is both essential and sufficient for participation in the group (Sneath and Sokal, 1973). A clustering strategy which works, by development, monothetic groups is then monothetic. In divi-sive Clustering, monothetic bunches are acquired by utilizing, for every division, a solitary variable and by isolating articles having particular variable qualities from the individuals who don't. Monothetic divisive clustering techniques are typically variations of the affiliation investigation strategy (Williams and Lambert, 1959) and are de-marked for parallel information. We can refer to among others Lance and Williams (1968), Kaufman and Rousseeuw (1990). Not at all like the principal strategies referred to over, these monothetic techniques are not in view of the improvement of a "polythetic" criterion like the inactivity or the measurement of the bipartitions. These techniques depend on the determination, at every stage, of the twofold factor which amplifies a measure of relationship to alternate factors. The articles are then separated utilizing the qualities (0 and 1) of the twofold factor.

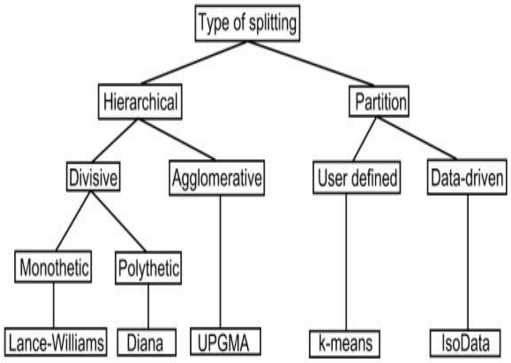
The divisive Clustering strategy proposed in this paper is monothetic however pro ceeds by streamlining of a polythetic paradigm. The bipartitional calculation and the decision of the group to be part depend on the minimization of the inside bunch dormancy. The total list of all conceivable bipartitions is stayed away from by utilizing the same monothetic approach as Breiman et al. (1984) who proposed, and utilized, double inquiries in a recursive partitioned handle, CART, with regards to segregation and relapse. With regards to clustering, there are no indicators and no reaction variable. Thus DIVCLUS-T is a Divisive Clustering strategy whose yield is not an order nor a relapse tree, but rather a Clustering-Tree. Since the dendrogram can be perused as a decision tree, it all the while gives segments into homogeneous groups and a basic understanding of those bunches.

In Chavent (1998) a disentangled rendition of DIVCLUS-T was introduced for the specific instance of quantitative information. Chavent et al. (1999) connected it, together with another monothetic divisive grouping technique in view of correspondence examination, to unmitigated information set on solid human skin. A first examination of DIVCLUS-T with Ward and k-means was given in this paper yet just for a solitary all out dataset and for the 6-bunch parcel. All the more as of late, it has been connected to bookkeeping revelation examination (Chavent et al., 2005) and a progressive divisive monothetic clustering strategy in light of the Poisson procedure has been proposed in Pircon (2004).

III. CLUSTERING ALGORITHMS

Cluster analysis was first projected in numeric domains, where distance is clearly defined. Later it extended to categorical data. However, much of data in real world contains a mixture of categorical and unbroken facts; as a result, the demand of cluster analysis on the diverse data is growing.

Cluster analysis has been an area of research for several decades and there are too many different methods for all to be covered even briefly. Many new methods are still being developed. In this section we discuss some popular and mostly used clustering algorithm and present their complexity.



K-means: K-means is that the best and established agglomeration strategy that is clear to actualize. The traditional will exclusively be utilized if the data in regards to every one of the items is found inside the fundamental memory. The strategy is named K-implies subsequent to everything about K groups is depict by the mean of the objects inside it [12].

Nearest Neighbor Algorithm: Associate in nursing formula kind of like the one link technique is termed the closest neighbor formula. With this serial formula, things are iteratively united into the present clusters that are closet. during this formula a threshold, it's accustomed confirm if things are further to existing clusters or if a replacement cluster is formed [2].

Divisive Clustering: With dissentious agglomeration, all things are at the start placed in one cluster and clusters are repeatedly split in two till all things are in their own cluster. The concept is to separate up clusters wherever some parts don't seem to be sufficiently near to alternative parts [2][6].

BIRCH Algorithm: BIRCH is meant for agglomeration an oversized quantity of numerical information by integration of stratified agglomeration (at the initial small agglomeration stage) and alternative agglomeration ways equivalent to reiterative partitioning (at the later macro agglomeration stage). It overcomes the 2 difficulties of clustered agglomeration methods: (1) measurability and (2) the lack to undo what was wiped out the previous step [3][6].

ROCK (Robust agglomeration mistreatment links) may be a stratified agglomeration formula that explores the thought of links (the variety of common neighbors between 2 objects) for information with categorical attributes [10].

CURE formula: One objective for the CURE agglomeration algorithm is to handle outliers well. It's each a stratified element and a partitioning element [3].

Chameleon: Chameleon may be a stratified agglomeration formula that uses dynamic modeling to work out the similarity between pairs of clusters [5]. it absolutely was derived supported the determined weakness of the Two stratified agglomeration algorithms: ROCK and CURE [4]. Distance based HC methods are widely used in unsupervised data analysis but few authors fake account of uncertainty in the distance data [8].

Distance between P1 to P2 = 5, d(P1, P2) =5.

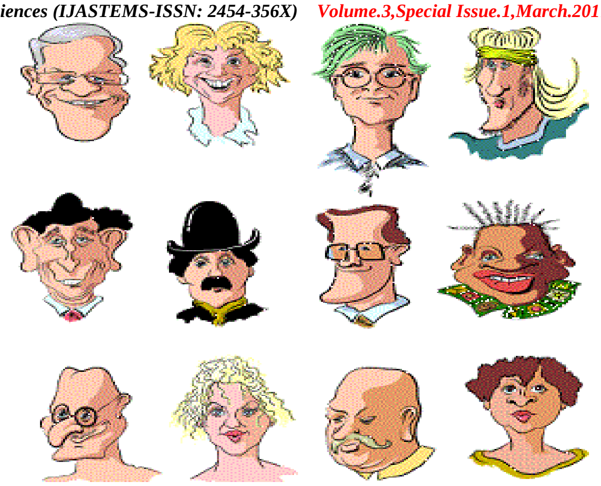
(p_1, q_1)
$l_2 = ((p_2 - p_1)^2 + (q_2 - q_1)^2)^{1/2}$

Where A and B are pair of elements considered as cluster, d (a,b) denotes the distance between the two elements . Distance function nature is defined by an integer q (q=2).for a data set of numeric values [1].

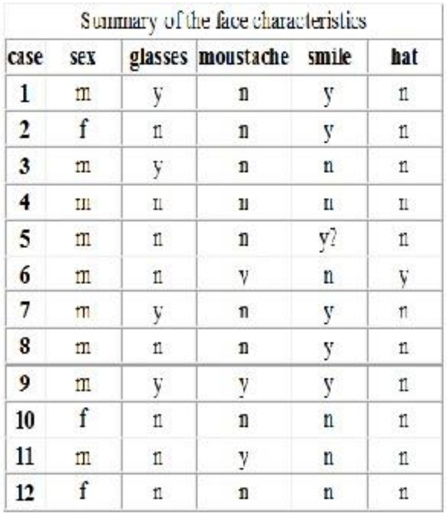
V .EXAMPLE

In a monothetic scheme cluster membership is based on the presence or absence of a single characteristic. Polythetic schemes use more than one characteristic (variables). For example, classifying people solely on the basis of their gender is a monthetic classification, but if both gender and handedness (left, right handed) are used the classification is polythetic.

Cluster analysis (CA) is a rather loose collection of statistical methods that is used to assign cases to groups (clusters). Group members share certain properties in common and it is hoped that the resultant classification will provide some insight into a research topic. The classification has the effect of reducing the dimensionality of a data table by reducing the number of rows (cases).



The above faces are the cases in their 'raw' state, i.e. before measurements and attributes have been recorded. More generally the starting point will be data that are already in a coded or numeric format. For example



Summary of the face characteristics

case	sex	glasses	moustache	smile	hat
1	m	y	n	y	n
2	f	n	n	y	n
3	m	y	n	n	n
4	m	n	n	n	n
5	m	n	n	y?	n
6	m	n	y	n	y
7	m	y	n	y	n
8	m	n	n	y	n
9	m	y	y	y	n
10	f	n	n	n	n
11	m	n	y	n	n
12	f	n	n	n	n

l_2

In these circumstances it may be less easy to see an obvious way of clustering the cases. Cluster analysis provides the tools that can cluster these cases in an objective manner. In above Example we divide the Clusters by using nominal attribute Y/N and it is grouped as four clusters.

V. CONCLUSION

In the above study the Clustering techniques can be divisive or agglomerative. A divisive method begins with all cases in one cluster. This cluster is gradually broken down into smaller and smaller clusters. Agglomerative techniques start with (usually) single member clusters. These are gradually fused until one large cluster is formed. This paper proposes a divisive monothetic hierarchical clustering method designed for either numerical or categorical data. In above example studied Group members share certain properties in common and it is hoped that the resultant classification will provide some insight into a research. Similar Face characteristics are assigned to a Cluster. In future have extended these approaches and find the computational complexity.

REFERENCES

[1]. Chavent, M., Guinot, C., Lechevallier Y., Tenenhaus, M., 1999. M´ethodes divisives de classification et segmentation non supervis´e: recherche d'une typologie de la peau humaine saine. Revue Statistique Appliqu´ee, XLVII(4), 87-99.

[2]. Chavent, M., Ding, Y., Fu, L., Stolowy, H., Wang, H., 2005. Disclosure and Determinants Studies: An extension Using the Divisive Clustering Method (DIV). European Accounting Review, 15(2), 181-218.

[3]. Duda R.O., Hart P.E., Strok D.G., 2001. Pattern Classification.Wiley-Interscience, second edition.

[4]. Mirkin, B., 2005. Clustering for Data Mining. A Data Recovery Approach. Chapman & Hall/CRC.

[5]. Eisen, M. B., Spellman, P. T., Browndagger, P. O. and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings National Academy of Science, USA, 95(25): 14863-14868. Available electronically from http://www.pnas.org/cgi/reprint/95/25/14863.

[6]. Chavent, M., De Carvalho, F. A. T., Lechevallier, Y., and Verde, R. (2006). New clustering methods for interval data. Computational Statistics, 21(2):211–229.

[7]. De Carvalho, F. A. T. and De Souza, R. M. C. R. (2010). Unsupervised pattern recognition models for mixed feature-type symbolic data. Pattern Recognition Let-ters, 31(5):430–443.

[8]. Hardy, A. and Kasaro, N. (2009). A new clustering method for interval data.MSH/MSS, 187:79–91.

[9]. Fang, H. and Saad, Y. (2008). Farthest centroids divisive clustering. InProc. ICMLA, pages 232–238.

[10]. Hardy, A. and Baune, J. (2007). Clustering and validation of interval data. In Brito, P. et al.,