ISSN: 2321-7782 (Online)
Volume 2, Issue 7, July 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Survey on Clustering Algorithms

Parth P. Makadia[1]
Comp. Engg
R.K. University
Rajkot – India

Maulik V Dhamecha[2]
ASST. professor CE Dept
R.K. University
Rajkot – India

Abstract: ***The data mining is hugely used in the normal life .we find something from it. How to Select these thing from large group of data. They used clustering, classification, regression techniques, outlier analysis etc. the clustering is used some special application. In this paper we show about the hierarchical co-clustering.it is used in music related data analysis. The data is Meta data .how to generate group from it'll be describe in the below. We show about the ant based clustering .in It we show about ant habit and feeding etc. all this thing used in the data travelling, all function of network .the graph based network is not work the help of the initialization .but the method is applied and work without initialization. Actively self-training clustering. We compare all this parameter in our comparison.***

Keywords: ***clustering, hierarchical co-clustering, hierarchical divisive co-clustering, hierarchical agglomerative co-clustering, data mining.***

I. INTRODUCTION

The process of grouping set of physical or abstract object into classes of similar objects is called clustering. Cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. In the base of single characteristics we create new group.it is an important human activity. Early childhood we learn how to distinguish between cat and dogs or between animals and plant by continuously improving subconscious clustering schemes. By automated clustering, we can identify dense and sparse regions in object space and, therefore, discover overall distribution patterns and interesting correlations among data attributes. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. Clustering may also help in the identification of areas of similar land use in an earth observation database and in the identification of groups of houses in a city according to house type, value, and geographic location, as well as the identification of groups of automobile insurance policy holders with a high average claim cost. It can also be used to help classify documents on the Web for information discovery. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Clustering can also be used for outlier detection, where outliers (values that are "far away" from any cluster) may be more interesting than common cases. Applications of outlier detection include the detection of credit card fraud and the monitoring of criminal activities in electronic commerce. For example, exceptional cases in credit card transactions, such as very expensive and frequent purchases, may be of interest as possible fraudulent activity. As a data mining function, cluster analysis can be used as a stand-alone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Alternatively, it may serve as a preprocessing step for other algorithms, such as characterization, attribute subset selection, and classification, which would then operate on the detected clusters and the selected attributes or features. Data clustering is under vigorous development. Contributing areas of research include data mining, statistics, machine learning, spatial database technology, biology, and marketing. Owing to the huge amounts of data collected in databases, cluster analysis has recently

become a highly active topic in data mining research as a branch of statistics; cluster analysis has been extensively studied for many years, focusing mainly on distance-based cluster analysis. Cluster analysis tools based on k-means, k-medoids, and several other methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SAS. In machine learning, clustering is an example of unsupervised learning. Unlike classification, clustering and unsupervised learning do not rely on predefined classes and class-labeled training examples. For this reason, clustering is a form of learning by observation, rather than learning by examples. In data mining, efforts have focused on finding methods for efficient and effective cluster analysis in large databases. Active themes of research focus on the scalability of clustering methods, the effectiveness of methods for clustering complex shapes and types of data, high-dimensional clustering techniques, and methods for clustering mixed numerical and categorical data in large databases.

II. General Discussion

Hierarchical co clustering: It is combination of two methods "hierarchical and co-clustering". They use the advantage of both method and create new method "hierarchical co-clustering". Clustering means grouping of same property data. In it they create hierarchy for one or more different type of data. There is two sub method of this method: hierarchical agglomerative co-clustering & hierarchical divisive co-clustering. Above process is done in simultaneous manner. First one follow the top to bottom and second follow the bottom to top strategy. In this paper they use music data for our understanding. They are used the artist style mood, and tag. The feature is inter-related or not is checked by them with the help of tag/style. They compare all other cluster with hacc and hdcc. They concluded that hacc and hdcc is best for the large data. They are use social network tags and all things.in the social networking site we are share music video link and comment on it. This is very useful for the check for popularity of the song. The large size of this data is not classifying in general situation. So with the help of this method they did all process easy. They develop method called novel hierarchical cluster and develop query for two type of data organization. They incorporate instance level constrain in hcc. They case study in which they show that this method is used for the artist similarity quantification process. The all selection process is done with the help of cluster hetreoneity measure.it is find best layer among them. Closeness is best or not is decided by the help of it. Use instance level process in hcc method but issues: how to transfer constraint one data type from to the other data type. How to use them for grouping data points of the same data type. First issue solved by the help of Dunn index (for best layer) & constraint k-means. Second issue solved by the help of alternating exchange algorithm. Dunn index use based on idea "good clustering produces well separated & compact cluster". Constrained k-means used in the single data; in this they mpck-means .I think we done process with the help of dynamic programming. They perform the experiment in 403 artists, 8529 unique tags, and 350 styles. They collect all information from the last.fm, allmusic.com. Hacc & Hdcc compare with information theoretic co-clustering, Euclidean co-clustering & minimum residual co-clustering. Evaluation process done with the help of parameter: accuracy, adjusted rand index, purity, normalized mutual index (NMI).all experiment demonstration is only in the artist in the tag. But it is with the pair of artist-mood, artist-style. Similarity quantification with the help of resnik, jiang-conrath, lim & schliker. Case study provides one thing quantification is also with the help of it.

TABLE I

Algorithm	Advantage	Drawback
Hacc algorithm	Layer wise optimization	
	Accuracy	
Alternating exchange	Easily transfer constrain one type Data from another	
Ant base clustering	Deals with non-Gaussian function	Not done in simple processing
	Result is visible and. Does not Need of prior knowledge of clustering.	Take more time
Actively self-training clustering	Initialization independent	Not work in large data interval between data.

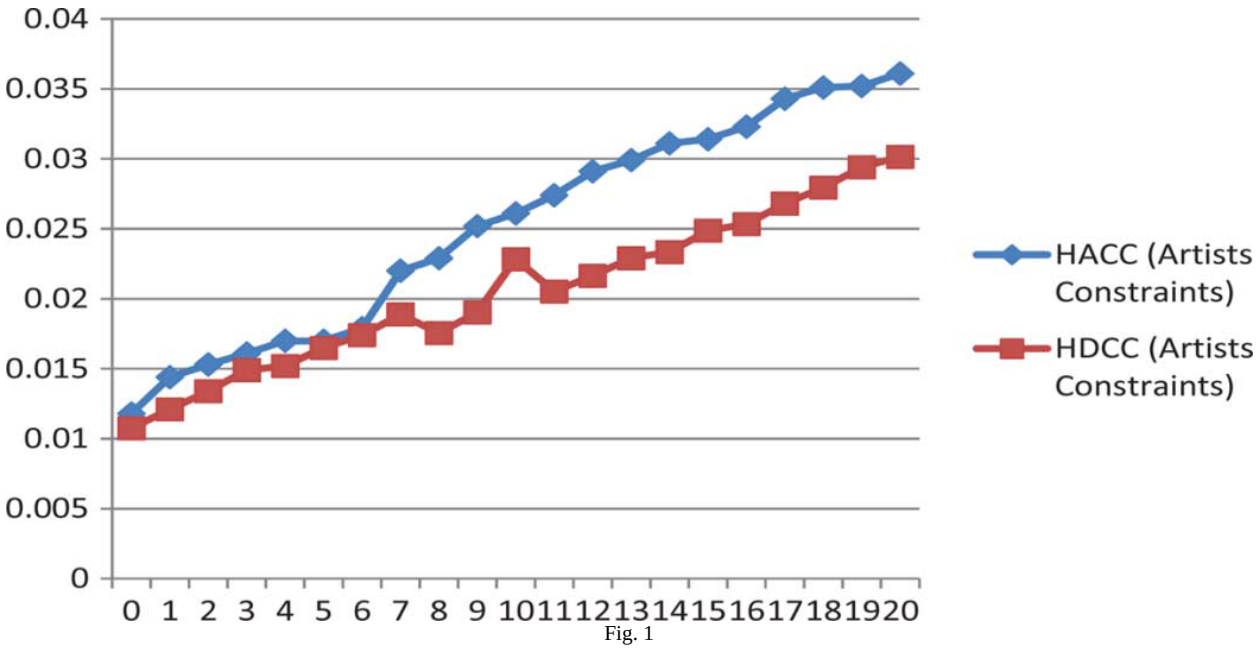


Fig. 1

Index	Artists
A1	Nine Inch Nails
A2	Radiohead
A3	Queen
A4	The Beatles

Index	Tags
B1	Alternative
B2	Classic Rock
B3	Industrial
B4	Rock

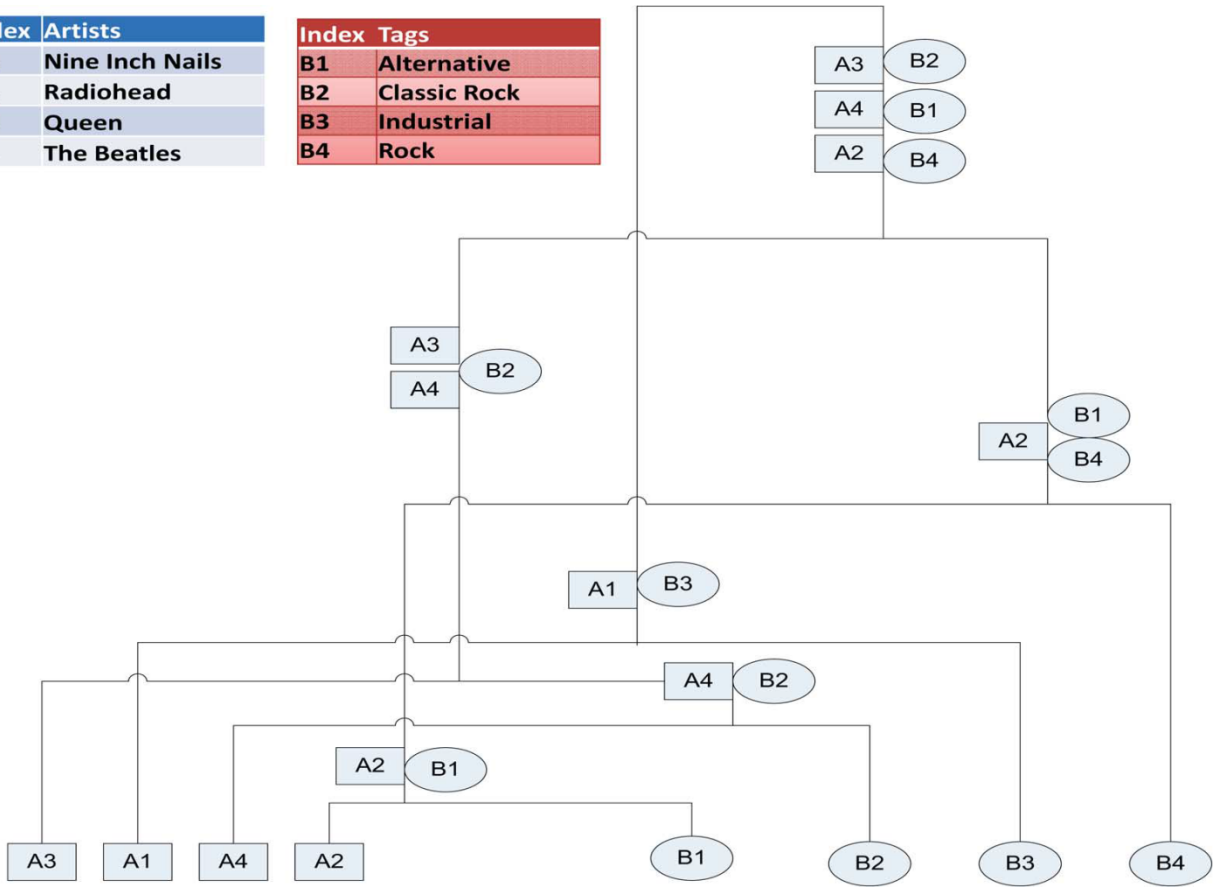
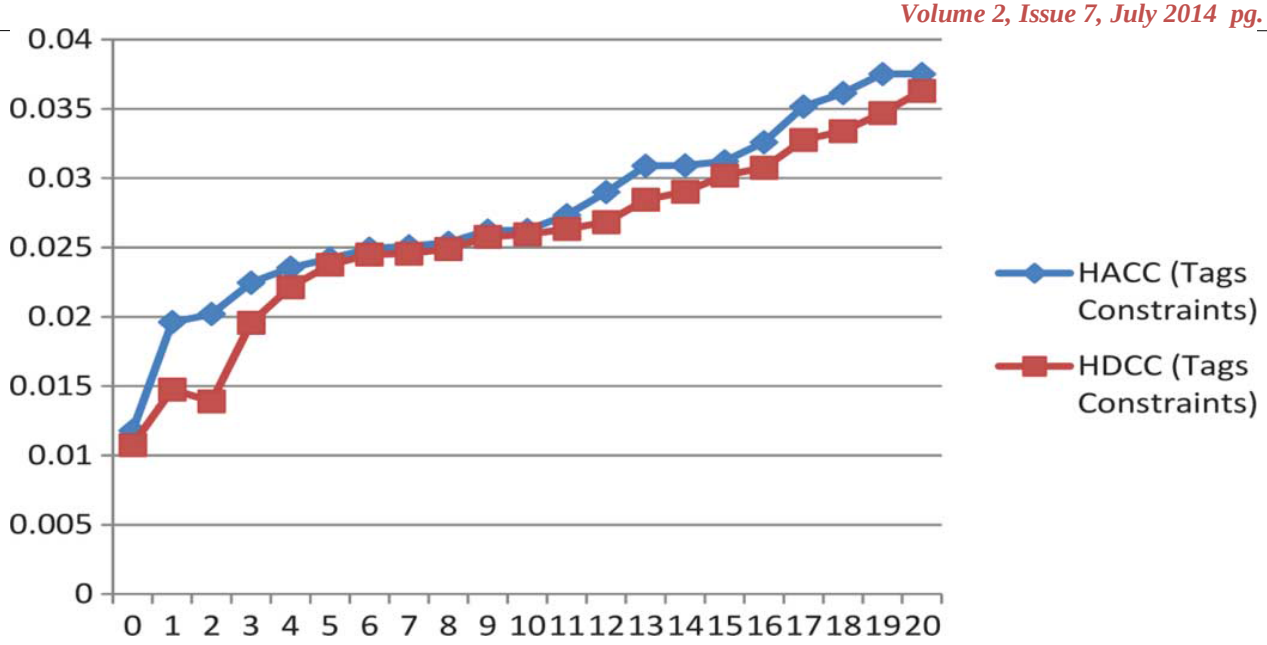


Figure: hcc diagram



III. Issues and Characteristics

Issues:

1. How to select the representative point for each class the seed the points.
2. The key problem in this clustering algorithm is how to find the centroid points to seed the clusters.
3. ASTC is not done work well when large interval of supervised data.

Characteristics:

1. If the labels of partial data are given, the class labeled data can be predicated more reliably by the semi supervised learning method.
2. Astc is independent of initialization thus, it is more convenient than the state of art clustering algorithm.
3. Astc algorithm can be readily applied in the case the labels of some data samples are provided by user, but the number of labeled classes is smaller than the number of clusters to be grouped.
4. It can be applied in the case that the only one sample is labeled in the data and the output will probability values.

IV. Conclusion

in this paper we are discuss about the comparison of graph based clustering and the normal clustering we are more discuss about the hierarchical co-clustering. Ant based clustering which is more generally used to work with parallel processing.

Acknowledgement

I really thankful to Mr. Maulik V Dhamecha my faculty. It also thank librarian of my college.

References

1. ”Hierarchical co-clustering: A New way to organize music data”; JingXuan li, Bo Shao, Tao li and mitsunori ogihara; IEEE transactions on multimedia, vol-14, no-2, april-2012
2. “A novel ant based clustering using kernel method”; Lei Zhang, Qixin cao science direct- 2010
3. “Initialization Independent clustering with actively self-training method”; Feiping Nie, dong Xu and Xuelong Li; IEEE transaction on system on man and cybernetics part b, vol-42, no-1 February-2012

AUTHOR(S) PROFILE

Parth P. Makadia received the B.E. degree in Computer engineering in 2011 from ATMIYA INSTITIUTE OF TECHNOLOGY & SCIENCE College, Rajkot, Gujarat, India. Currently he is pursuing M.Tech. In Computer Engineering from School of Engineering, R.K. University, Rajkot, Gujarat, India. His area of interest is clustering, hierarchical co-clustering, graph based mining.

Maulik V. Dhamecha has received his B.E degree in Computer Engineering from VVP Engg. College, Saurastra University, Gujarat, India and master Degree from Dharmsinh Desai University, Gujarat, India. He has been working at Department of Computer Engineering, School of Engineering, RK. University, Rajkot, Gujarat, where he is currently working as an Assistant Professor. His current research interest includes Multiple Classifier System, Sequence Pattern Mining, clustering.

