Volume 2, Issue 3, March 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets

M.Vijayalakshmi,MCA,M.Phil
Asst.Prof in PG(CS) Department,
Sree Saraswathi Thayagaraja College, Pollachi
Vijiyes4@gmail.com

M.Renuka Devi,MCA,M.Phil,(Phd)
Asst.Prof in MCA Department,
Sree Saraswathi Thayagaraja College, Pollachi

Abstract— **Data mining is the process of extracting patterns from data. Data mining is seen as an increasingly important tool by modern business to transform data into an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. This paper gives an overview of different clustering algorithms used in large data sets. It describes about the general working behavior, and the methodologies followed on these approaches and the parameters which used in these algorithms with large data sets.**

Keywords— **Clustering, Supervised Learning, Unsupervised Learning Hierarchical Clustering, K-Mean Clustering Algorithm, Density Based Clustering Algorithm**

I.INTRODUCTION

Clustering is a process of grouping objects with similar properties. Any cluster should exhibit two main properties; low inter-class similarity and high intra-class similarity. Clustering is an unsupervised learning i.e. it learns by observation rather than examples. There are no predefined class label exists for the data points. Cluster analysis is used in a number of applications such as data analysis, image processing, market analysis etc. Clustering helps in gaining, overall distribution of patterns and correlation among data objects [**1**]. In this paper describes about the general working behaviour, the methodologies to be followed and the parameters which used in these clustering algorithms.

In This paper is organized as follows gives an overview of different clustering algorithms. Then different clustering algorithms are used as follows Hierarchical clustering algorithms, K-means clustering algorithms, and Density Based Clustering Algorithm and thehow the is methodology applied on these algorithms and the parameter used in these algorithms are described. Finally the conclusions are provided.

II.OVERVIEW OF DIFFERENT CLUSTERING ALGORITHMS

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data[2]. Clustering is a division of data into groups of similar objects. [1] Clustering algorithm can be divided into the following categories:

1. Hierarchical clustering algorithm
2. K-means clustering algorithm
3. Density Based Clustering algorithm
2. Partition clustering algorithm
3. Spectral clustering algorithm
4. Grid based clustering algorithm

A. Hierarchical Clustering

Hierarchical clustering algorithm groups data objects to form a tree shaped structure. It can be broadly classified into agglomerative hierarchical clustering and divisive hierarchical clustering. In agglomerative approach which is also called as bottom up approach, each data points are considered to be a separate cluster and on each iteration clusters are merged based on a criteria. The merging can be done by using single link, complete link, centroid or wards method. In divisive approach all data points are considered as a single cluster and they are splited into number of clusters based on certain criteria, and this is called as top down approach[1]. Examples for this algorithms are LEGCLUST [3], BRICH [4] (Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using Representatives) [5], and Chemeleon [6]. Under the hierarchical clustering we have different clustering algorithms as follows,

1) Classifying Large Data Sets Using SVM with Hierarchical Clusters:

In this clustering "Classifying Large Data Sets Using SVM with Hierarchical Clusters" present a new method, Clustering-Based SVM (CB-SVM), which is specifically designed for handling very large data sets. This paper proposes a new method called CB-SVM (Clustering- Based SVM) that integrates a scalable clustering method with an SVM method and effectively runs SVMs for very large data sets. The existing SVMs are not feasible to run such data sets due to their high complexity on the data size.CB-SVM tries to generate the best SVM boundary for very large data sets given limited amount of resource based on the philosophy of hierarchical clustering where progressive deepening can be conducted when needed to find high quality boundaries for SVM. This experiments on synthetic and real data sets show that CB-SVM is very scalable for very large data sets while generating high classification accuracy [7.]

2) Efficient Hierarchical Clustering of Large Data Sets Using P-trees:

Hierarchical clustering methods have attracted much attention by giving the user a maximum amount of flexibility. Rather than requiring parameter choices to be predetermined, the result represents all possible levels of granularity. In this paper a hierarchical method is introduced that is fundamentally related to partitioning methods, such as k-medoids and k-means, as well as to a density based method, namely center-defined DENCLUE. It is superior to both k-means and k-medoids in its reduction of outlier influence. Nevertheless it avoids both the time complexity of some partition-based algorithms and the storage requirements of density-based ones. An Implementation is presented that is particularly suited to spatial-, stream-, and multimedia data, using P-trees1 for efficient data storage and access [8].

B. K-Mean Clustering Algorithm

K-means clustering is a partitioning method. ***K-means* clustering** is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.

The k –mean algorithm

The *k*-means algorithm has the following important properties:

1. It is efficient in processing large data sets.
2. It often terminates at a local optimum
3. It works only on numeric values.
4. The clusters have convex shapes

1) Parallel *k/h*-Means Clustering for Large Data Sets:

In this paper presented a parallel version of the *k/h*-means clustering algorithm. The algorithm is designed to be used on large distributed data sets. Even on a very simple distributed computing environment, namely a PC cluster on a 10 MBits Ethernet, we are able to achieve about 90% efficiency for a configuration up to 32 processors. These results show that parallel *k/h*-means is scalable and thus enlarges its field of application to clustering tasks where it would be the preferred algorithm, but the task's computational complexity previously made it impossible [9].

2)A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets:

The algorithm presents a method to use both advantages of HC and K-Means by introducing equivalency and compatible relation concepts. By these two concepts we defined similarity and our space and could divide our space by a specific criterion. Many directions exist to improve and extend the proposed method. Different applications can be used and examined the framework. Text mining is an interesting arena. Based on this method data stream processing can be improved. Data type is another direction to examine this method. In this study K-Means has been used for second phase whereas we can use other clustering algorithms e.g. genetic algorithm, HC algorithm, Ant clustering [11], Self Organizing Maps [12], etc. Determining number of sub spaces can be studied as important direction for the proposed method.

C. Density Based Clustering Algorithm

Density based algorithm continue to grow the given cluster as long as the density in the neighbourhood exceeds certain threshold [6]. This algorithm is suitable for handling noise in the dataset. The following points are enumerated as the features of this algorithm.

1. Handles clusters of arbitrary shape
2. Handle noise
3. Needs only one scan of the input dataset.
4. Needs density parameters to be initialized.

DBSCAN, DENCLUE and OPTICS [6] are examples for this algorithm.

1) DESCRY: a Density Based Clustering Algorithm for Very Large Data Sets

This paper described a new method, named DESCRY, to identify clusters in large high dimensional data set having different size and shape. The algorithms parametric the agglomerative method used in the pre-clustering step and the similarity metrics of interest. DESCRY has a very low computational complexity, indeed it requires O(Nmd) time, for high-dimensional data sets, and O(N logm) time, for low dimensional data sets, where m can be considered a constant characteristic of the data set. Thus DESCRY scales linearly both the size and the dimensionality of the data set[13]. Despite its

low complexity, qualitative results are very good and comparable with those obtained by state of the art clustering algorithms. Future work includes, among other topics, the investigation of similarity metrics particularly meaningful in high-dimensional spaces, exploiting summaries extracted from the regions associated to midpoints.

2) Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications

In this paper, presented the clustering algorithm GDBSCAN generalizing the density-based algorithm DBSCAN (Ester et al., 1996) in two important ways. GDBSCAN can cluster point objects as well as spatially extended objects according to both, their spatial and their non-spatial attributes. After a review of related work, the general concept of density-connected sets and an algorithm to discover them were introduced [14]. A performance evaluation, analytical as well as experimental, showed the effectiveness and efficiency of GDBSCAN on large spatial databases.

III CONCLUSION

The paper describes different methodologies and parameters associated with different clustering algorithms used in larger data sets. And it gives an overview of different clustering algorithms used in large data sets. Then describes about the general working behaviour, and the methodologies followed on these approaches and the parameters which used in these algorithms with large data sets..

REFERENCES

[1] S.Anitha Elavarasi and Dr. J. Akilandeswari and Dr. B. Sathiyabhama, January 2011, A Survey On Partition Clustering Algorithms.

[2] Survey of Clustering Data Mining Techniques, Pavel Berkhin, Accrue Software, Inc.

[3] Santos, J.M, de Sa, J.M, Alexandre, L.A , 2008. LEGClust- A Clustering Algorithm based on Layered Entropic subgraph. Pattern Analysis and Machine Intelligence, IEEE Transactions : 62-75.

[4] M. Livny, R.Ramakrishnan, T. Zhang, 1996. BIRCH: An Efficient Clustering Method for Very Large Databases. Proceeding ACMSIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery :103-114.

[5] S. Guha, R. Rastogi, and K. Shim, 1998. CURE: An Efficient Clustering Algorithm for Large Databases. Proc. ACM Int'l Conf. Management of Data : 73-84.

[6] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques" Elsevier Publication.

[7] Hwanjo Yu AND Jiong Yang AND Jiawei Han, "Classifying Large Data Sets Using SVM with Hierarchical Clusters"

[8] Efficient Hierarchical Clustering of Large Data Sets Using P-trees, Anne Denton, Qiang Ding, William Perrizo And Qin Ding

[9] Parallel *k/h*-Means Clustering for Large Data Sets, Kilian Stoffel and Abdelkader Belkoniene

[10] Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets, Madjid Khalilian, Norwati Mustapha, MD Nasir Suliman, MD Ali Mamat

[11] U. Boryczka, "Finding groups in data: Cluster analysis with ants," *Applied Soft Computing Journal*, vol. 9, pp. 61-70,2009.

[12] D. Isa, V. P. Kallimani, and L. H. Lee, "Using the self organizing map for clustering of text documents," *Expert Systems With Applications*, vol. 36, pp. 9584-9591, 2009.

[13] DESCRY: a Density Based Clustering Algorithm for Very Large Data Sets, Fabrizio Angiulli, Clara Pizzuti, Massimo Ruffolo

[14] Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications, Jörg Sander, Martin Ester, Hans-Peter Kriegel, Xiaowei Xu

