Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 5 (2015), pp. 3371-3375
© Research India Publications
http://www.ripublication.com

# Various Clustering Techniques: A Survey

***M. Rajasekhar Reddy[1], Anishin Raj M M[2], B. Karthikeyan[1], Diana Baby[2], Dr. V. Vaithiyanathan[1], Ramgopalan V[1]**

[1] *School of Computing, SASTRA University, Thanjavur-613401, India*
[2]*Viswajyothi College of Engineering, Vazhakulam, Kerala*
**rajasekharmanyam04@gmail. com*

## ABSTRACT

The purpose of the paper is to explore versatile kinds of clustering techniques and their uses in different scenarios. The various clustering techniques have different and complex computational complexities due to the hyper dimensional points which are taken as the input parameters which will be a vector. Clustering has algorithms of different categories such as Agglomerative clustering, Divisive clustering have been discussed. Comparison of Hierarchical clustering types along with their advantages and limitations are discussed. A study of latest clustering techniques by various researchersalong with their merits and demerits are also mentioned.

**Keyword:** clustering, K Means, Agglomerative clustering, Divisive clustering.

## INTRODUCTION

Clustering refers to the dividing of data into its relevant groups or classes such that data which are present in same class are similar and data present in the different class are dissimilar. In other terms the mechanism by which a group of patterns are identified and separated into its characteristic group is called clustering. Clustering has a wide variety of applications in various fields such as image segmentation[11], market research, software evolutions, mathematical chemistry, climatology, transcriptomics, human genetic clustering, evolutionary algorithms, and so-on[4]. There are two types of clustering namely Hierarchical clustering[9], [10] and Partitionalclustering[10][12].

## I. PARTITIONING METHOD

In partitioning method, clustering datais based on a similarity such as distance and make sure that data within the cluster issimilar. It has to follow these two conditions in order to satisfy this method i. e. each cluster should contain at-least one data/object. A data/object must precisely reside only in one cluster. K means is the most commonly used partitioning technique. The primary objective is to partition the given N Observations to K groups where each observation enters to one group which is of its closest mean. Hence it is iterative in nature[12][14]. Forgy's clustering is also a type of partitional clustering where like K-means clustering centroid is made in use for iteration but with a small change in condition. Isodata algorithm is also an efficient clustering technique[14].

K-means cluster prove to be advantageous in various ways, the technique is easy to implement. It has a high intra-cluster similarity which is useful in detecting and clubbing the data's together in a single cluster. Furthermore to add up it automatically reduces the sum of squared deviations of pattern among clusters. It helps also relatively in expanding the efficiency as compared to others so its relatively efficient and has a complexity of O(Total iterations * Clusters * Objects). The limitations of K-means clustering include many aspects, only if mean is available then we could proceed with this, but we cannot apply with categorical data. Another disadvantage is that it terminates often at local optimum. It's been found that it is unable to handle noise data, limitation also include the need to specify K's value in advance. Initialization of K centre is also important.

Various advantages of K-means clustering has been exposed in various domains such as geo-statistics[1], market segmentation, vector quantization, agriculture, astronomy and various others. In unsupervised learning it can help to learn the features of input data. It's also assistive when working with processing of signals[2].

## II. HIERARCHICAL CLUSTERING

The idea behind hierarchical clustering is that the data are grouped as a tree of clusters. The hierarchical clustering can be split up into agglomerative (bottom-up) method[10], [12] and divisive (top-down) method[10], [12].

### A. Agglomerative Clustering

1) First each object is placed in its corresponding clusters, these are known as atomic clusters.
2) These atomic clusters are integrated together into large clusters.
3) The large clusters are combined until one single cluster is achieved [10], [12].

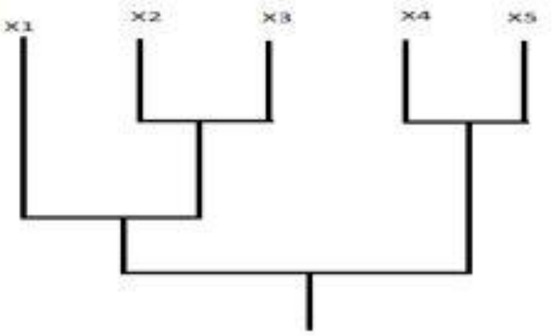

**Fig: Agglomerative clustering**

**B. Divisive Clustering**

**1)** Initially all the objects are regarded to be under one single cluster.
**2)** Then this large cluster dissevers into smaller clusters
**3)** These clusters are again split up until each object contains its own cluster [10][12].

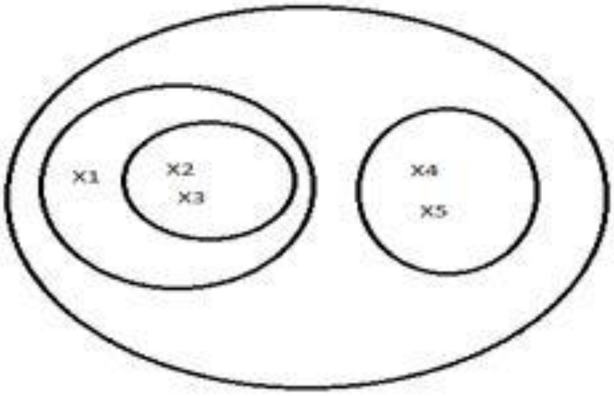

**Fig: Divisive clustering**

The limitations of the hierarchical clustering are selection the merge/split point is quite difficult. Once a cluster has been merged or separated it's hard to undo the process. The method may not scale well as it depends on decision of merge/splitting the cluster. There are numerous applications in which this type of clustering is used, medicine and genetics fields have been the major source which uses hierarchical clustering [12][13][14]. It provides the foundation for human genetic clustering and human genetic variation. Visualization technique, which is based on hierarchical clustering, has been built to allow users to take part in discovering nested clusters[3].

## III. LATEST CLUSTERING METHODS

### Distribution Based Clustering

This is an effective upcoming model which has been developed based on the statistics. Here the clusters have objects based on their distribution. One advantage of this model is that it closely matches to generation of artificial data sets. [5]

### Expectation-Maximum Clustering

In this method the data sets are postured with a defined number of Gaussian distributions that are initialized in a random manner. It will converge at a local optimum, so multiple iterations on it will result in different solutions. For hard clustering, the objects or data sets belong to its most likely Gaussian distribution which might not be required in soft clustering[6][8]. The advantage of using this Expectation-Maximum Clustering is that it is semantically strong and also produces complex models along with clusters.

## IV. K-MEANS DIVIDE and CONQUER CLUSTERING

Most clustering techniques doesn’t take into account about the various sizes or levels of the data, they only intend to group them with their similarities. By doing this they bring together different sizes under single cluster on a similarity neglecting the fact that they might be of different sizes. MadjidKhalilian et. al. proposed a method which could boost up the performance of k-means by using this idea. This method is proved to be accurate and efficient than a single one pass cluster. Few assumptions are required to follow in order to use this technique, they presume that the space is orthogonal, dimensions are same for all the objects. The efficiency and accuracy of original k-means is improved by considering the size of objects. When this method is followed they use subspaces for clustering, which accomplishes more accurate results[7].

## CONCLUSION

When compared to Hierarchical clustering[10], [12], K-means clustering [10], [12] is far better in terms of its performance. The advantage of K-means is that it produces better results for large scale quantities of data though the process is quite dragging[10]. Hierarchical clustering is quite efficient in providing accurate results for small sets of data. Distributive clustering technique helps in capturing co-relation and dependence of attributes with these models[13]. These algorithms which we have discussed is used in many applications such as web services, market research, recommender systems, analysis of images, bio-informatics, image processing. Clustering algorithms are used in real time applications like search engines, identifying cancerous data, drug activity prediction, wireless sensor networks etc.

## REFERENCES

[1] Kanungo, T., Mount, D. M.., Netanyahu, N. S., Piatko, C. D., Silverman, R., Wu, A. Y., 2002"*An Efficient K-Means Clustering Algorithm: Analysis and Implementation*", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7.

[2] Shang, R., Qi, L., Jiao, L., Stolkin, R., Li, Y., 2014 "*Change detection in SAR images by artificial immune multi-objective clustering*", Engineering applications of Artificail Intelligence 31(2014) 53-67.

[3] Rubin, D. L., 2012 "*Finding the meaning in images: Annotation and image markup*", Philos., Psychiatry, Psychol, vol. 18, no. 4, pp. 311-318.

[4] Bovolo, F.,. Bruzzone, L., Marconcini, M., 2008 "*A novel approach to unsuper-vised change detection based on a semi-supervised SVM and a similarity measure*", IEEE Trans. GeosciRemore Sens., vol. 46, no. 7, pp. 2010-2081.

[5] Anoop Kumar Jain., Satyam Maheswari., 2012 "*Survey of Recent Clustering Techniques in Data Mining*", International Archive of Applied Sciences and Technology, vol. 3, no, 2, pp. 68-75.

[6] Ravi, V. T., Agarwal, G, 2009 "*Performance Issues in Parallelizing Data-Intensive Applications on a Multi-Core Cluster*", 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, pp. 308-315

[7] Khalilian, M., Boroujeni, F. Z., Mustapha, N., Sulaiman, Md. N. 2009 "*K-Means Divide and Conquer Clustering*", IEEE, International conference on Computer and Automation Engineering, pp. 306-309.

[8] Yang, F., Sun, T., Zhang, C, 2009 "*An efficient hybrid data clustering method based on K-harmonic means, and Particle Swarm Optimization, Expert Systems with Applications*", pp. 9847-9852.

[9] Sharma, S., Gupta, V, 2012 "*Recent developments in text clustering techniques*", International journal of computer applications 37(6):14-19.

[10] Han, J., Kamber, M., 2003 "*Data Mining concepts and Techniques*", Morgan kaufmannpublishers, NewDelhi, India.

[11] Gonzalez, R. C., Woods, R. E, 2007 "*Digital Image Processing*", Pearson Education Third Edition.

[12] Narasimha Murthy, M., Susheela Devi, V., 2011 "*Pattern Recognition*", Springer Publication.

[13] kaur, M., Kaur, U, 2013 "*Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection*", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7.

[14] Gose, E., Johnsonbaugh, R., Jost, S, 2009 "Pattern Recognition and Image Analysis" PHI publications.