INTERNATIONAL JOURNAL OF ENHANCED RESEARCH IN MANAGEMENT AND COMPUTER APPLICATIONS
VOL. 2, ISSUE 3, MARCH 2013 ISSN NO: 2319-7471

# Clustering Algorithm Technique

M. R. Sindhu[1], Rachna Sable[2]
[1,2]M.E., Computer Science & Engineering, G. H. Raisoni College of Engineering and Mgmt., India
[1]sindhu.mrs@gmail.com, [2]rachna.sable@gmail.com

**Abstract**: **Data mining and retrieval has always been a cumbersome task for the computer professionals for which the need of scientific method of data mining was evolved. Use of Partitioned and hierarchical clustering techniques like K-mean and agglomerative clustering technique provides the efficient way to cluster the unlabelled data which is explained in the below study paper. The advantages, applications and the weakness are also discussed during the study.**

**Keywords**: **Data Mining, Clustering, Partitioned and Hierarchical clustering, K-Mean, Agglomerative.**

## I. INTRODUCTION

Clustering can be considered the most important unsupervised learning technique. It is the process of organizing objects into groups whose members are similar in some way‖. Clustering is also called as data segmentation in some applications because clustering partitions large datasets into groups according to their similarity. A cluster is therefore a collection of objects which are ―similar‖ between them and are ―dissimilar‖ to the objects belonging to other clusters. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. Clustering can be used in Data mining, Information retrieval, text mining, Marketing, Medical diagnostic etc. Representing the data by fewer clusters necessarily loses fine details, but achieves simplification.

There are two main categories of the document clustering algorithm that is Partitioned Clustering and Hierarchical Clustering.

**Partitioned clustering**: Partitioned clustering algorithm partitioned the documents in to k number of clusters. Example of partitioned clustering is k-means clustering.

**Hierarchical clustering**: In hierarchical clustering, the clusters of documents are arranged in tree like structure. Hierarchical clustering can be divided into Agglomerative hierarchical clustering (AHC) and divisive clustering.

Data mining refers to extracting or mining knowledge from large amounts of data. Data and Knowledge Mining is learning from data. In this context, data are allowed to speak for themselves and no prior assumptions are made. This learning from data comes in two flavours: supervised learning and unsupervised learning. In supervised learning (often also called directed data mining) the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent variables. The target of the analysis is to specify a relationship between the explanatory variables and the dependent variable. In unsupervised learning situations all variables are treated in the same way, there is no distinction between explanatory and dependent variables.

The following are typical requirements of clustering algorithm in data mining:

- **Types of attributes algorithm can handle**: Many algorithm are design to cluster interval based (numerical) data. However, applications may require clustering other types of data which are binary categorical or mixtures of these data types.

- **Scalability:** Clustering algorithms works well on small datasets. Clustering on a sample of a given large data set may lead to biased results. Highly scalable clustering algorithms are needed.

- **Discovery of clusters with arbitrary shape:** Many clustering algorithms based on Euclidean distance measures tend to find spherical clusters with similar size and density. However the cluster could be of any shape so it is important to develop algorithms that can detect clusters of arbitrary shapes.

- **Handling Outliers:** Outliers are objects that do not belong to any of the clusters. So these outliers should be taken care of, which can be detected using statistical tests.

- **Incremental clustering:** It is difficult to incorporate new data into existing clustering structures and instead must determine a new clustering from scratch.

- **Time complexity:** Time complexity of algorithm should be taken into account for proper and fast execution.

- **Data order dependency:** Given a set of data objects, an algorithm may written dramatically different clustering depending on the order of presentation of the input objects, so an incremental clustering should be developed that are insensitive to the order of input.

**Clustering algorithms may be classified as listed below:**

a) Hierarchical Methods

- Agglomerative Algorithms
- Divisive Algorithms

b) Partitioning Methods

- Relocation Algorithms
- Probabilistic Clustering.
- K-medoids Methods.
- K-means Methods.
- Density-Based Algorithms.

  i. Density-Based Connectivity Clustering
  ii. Density Functions Clustering.

c) Grid-Based Methods.

d) Methods Based on Co-Occurrence of Categorical Data.

e) Constraint-Based Clustering.

f) Clustering algorithm used in machine learning

- Gradient descent and artificial neural network
- Evolutionary methods

g) Scalable clustering algorithms

h) Algorithms for high dimensional data

- Subspace clustering
- Projection techniques
- Co-clustering Techniques

## II. Overview of K-Mean and Hierarchical Clustering

The most popular partitional algorithm among various clustering algorithms is K-mean clustering.

K-means is an exclusive clustering algorithm, which is simple, easy to use and very efficient in dealing with large amount of data with linear time complexity. The objective function of K-mean algorithm is to minimize the intra-cluster distance and maximize the inter-cluster distance based on Euclidean distance.

The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

Steps for K-means clustering algorithm: Input:

D = {d1, d2, d3,…….., dn} // set of n data items. k // desired number of clusters.

**Output:** A set of k clusters.

Steps:

1. For initial centroid, select the k data items from documents D ;

2. Repeat

Allocate each item d1 to the cluster which has the closest centroids;

New mean values are calculated for each Cluster; Until specified condition criteria is met.
The mean value is calculated on the basis of the formula such as Euclidean distance and Manhattan distance that is defined as:

In Euclidean distance the distance is measured between two points such as X (x1, x2) and Y (y1, y2).

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

In Manhattan distance the distance is measured between two pair of objects are:

d (i, j) =|xi1-xj1| + |xi2-xj2|+………..+|xin-xjn| (2)

When the Euclidean distance and Manhattan distance are used for measuring the distance the following properties are satisfied:

d (i, j)>=0: distance is a non-negative number.

d (i, i) =0: The distance of an object to itself is 0. d (i, j) = d (j, i): distance is a symmetric function.

If the number of data objects is less than the number of cluster than assign the data as a centroid of the cluster. Each centroid will have a cluster number. If the number of data objects is bigger than the number of Cluster, calculate the distance to all centroid and minimum distance. The location of the centroid is based on the current updated data; assign all data to the new centroid. This process is repeated till that no data is moving to another cluster any more.

K-mean clustering used to minimize the squared error function. This method is relatively scalable and efficient in the processing of large data sets. The complexity of the algorithm is O (nkt). Where n is the total number of objects, k is the number of clusters and t is the number of iterations.

**Hierarchical Clustering Algorithm:**

Hierarchical Clustering method works by grouping data objects into a tree of clusters. This method uses hierarchical decomposition of a given set of data objects.

Hierarchical method can be categorized into:

- Agglomerative Hierarchical Clustering and,
- Divisive Hierarchical Clustering

**Agglomerative Hierarchical Clustering:**

Agglomerative hierarchical clustering is bottom-up strategy; it starts with each object a separate cluster itself and merges the clusters according to distance measures. Clusters are merge until all the objects in to a single cluster till the termination condition is satisfied. It merges the clusters iteratively. Most of the hierarchical clustering method belongs to this clustering category. Hierarchical agglomerative clustering is represented by the Genograms.

It is tree like structure show the relationship between objects. In Dendogram each merge is represented by the horizontal line. The y-coordinates of horizontal line are the similarity of two clusters that were merged and documents can be viewed as single cluster. The similarity measures can be calculated by-

- Single-linkage clustering
- Complete-linkage clustering
- Group-average clustering

## III. Related Work

Survey of Clustering Data Mining Techniques [1]. Applications of K-mean Clustering algorithms for prediction of Student's Academic Performance [2].

In Educational field monitoring the overall academic performance of the student is a cumbersome task. A system that analyzes students result and uses the statistical algorithm to arrange their scores according to their performance is used. K-mean algorithm with the deterministic model is combined.

Application Research of K-mean clustering algorithm in Image Retrieval System [3]. Ignoring the similarities among images in the database and retrieving the features similar to the required query is done using the K-Mean.

K-mean Clustering algorithm for Mixed Numeric and Categorical Data Sets [4]. In this paper the limitation of using the K-mean for only Numeric data set is overcome and a modified algorithm that works with mixed and categorical data set is proposed.-An Analytical Assessment on Document Clustering [5].

## Comparative Study

**Advantages of K-Mean & Agglomerative:**

- It has been seen that with large number of variables, K-Mean may be computationally faster than hierarchical.
- K-mean may produce tighter clusters than hierarchical clustering.
- Hierarchical document clustering is better than the partitioned clustering; its main work is to build the hierarchical structure in tree of clusters whose leaf node represents the subset of document collection.
- Embedded flexibility regarding the level of granularity
- Ease of handling of any forms of similarity or distance
- Consequently, applicability to any attribute types

**Disadvantages of K-Mean & Agglomerative:**

There is necessity for specifying K number of clusters. Fixed number of clusters can make it difficult to predict what K should be.

- Difficulty in comparing the quality of clusters.
- The advantage of hierarchical clustering comes at the cost of lower efficiency.
- Vagueness of termination criteria
- The fact that most hierarchical algorithms do not revisit once constructed (intermediate) clusters with the purpose of their improvement.

## IV. Conclusions

This paper presents the approach towards analyzing different clustering algorithms, Partitioned and hierarchical. K-Mean represents the partitioned clustering algorithm which is simple, easy and fast. Agglomerative is a hierarchical clustering algorithm which allows exploring data on different levels of granularity.

## V. Acknowledgements

I would like to express my sincere gratitude and thanks to Mrs. M. R Sindhu, Professor, department of Computer Engineering and Management and HOD, Mrs. Poonam Gupta for their help, encouragement and intellectual influence which made this paper possible. Their invaluable guidance, reviews and suggestions have been instrumental in successful completion of this paper.

## References

[1]. Pavel Berkhin, “Survey of Clustering Data Mining Techniques”.

[2]. Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C, “Application of k-Means Clustering algorithm for Prediction of Students Academic Performance”, (IJCSIS) International Journal of Computer Science and Information Security.

[3]. Hong Liu, Xiaohong Yu, “Application Research of k-means Clustering Algorithm in Image Retrieval System”, Proceedings of the Second Symposium International Computer Science and Computational Technology (ISCSCT ’09) Huangshan, P. R. China, 26-28, Dec. 2009, pp. 274-277.

[4]. Dharmendra K Roy and Lokesh K Sharma, “Genetic K-mean Clustering algorithm for mixed numeric and categorical data sets”, presented at International journal of artificial intelligence and application in April 2010.

[5]. Pushplata, Ram Chatterjee, “An Analytical Assessment on Document Clustering”, I. J. Computer Information Security, 2012, Published Online June 2012 in MECS.

[6]. Paulo Lopez-Meyer, Stephanie Schuckers, Oleksandr Makeyev, Juan M. Fontana, Edward Sazonov, “Automatic identification of the number of food items in a meal using clustering techniques based on the monitoring of swallowing and chewing”, Biomedical Signal Processing and Control, Volume 7, Issue 5, September 2012.

[7]. Jian Feng Cui, Heung Seok Chae, “Applying agglomerative hierarchical clustering algorithms to component identification for legacy systems”.

[8]. Constantinos S. Hilas, Paris As. Mastorocostas, “An application of supervised and unsupervised learning approaches to telecommunications fraud detection”.