Survey on Clustering Techniques

Mrs.K.RAMYA[1], M.ANUSUYA[2], A.PRIYANKA3, M.MONISHA[4]

Assistant Professor, Department of MCA, Gnanamani College of Technology, Tamilnadu, INDIA
PG Scholar, Department of MCA, Gnanamani College of Technology, Tamilnadu, INDIA

ABSTRACT

Clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. In this journal the clustering is applied to the image data. The feature values are taken, and the final solution depends upon, these values on which the categorization is done. The complexities for the different methods are also defined here. Clustering algorithms can be classified into partition-based algorithms, hierarchical based algorithms, density-based algorithms and grid-based algorithms. This journal paper focuses on a keen study of different clustering algorithms in data mining. A brief overview of various clustering algorithms is discussed.

KEYWORDS: Clustering, Cloud computing, Data analytics, Hierarchical Clustering.

INTRODUCTION

A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Besides the term data clustering as synonyms like cluster analysis, automatic classification, numerical taxonomy, botrology and typological analysis. The purpose of this journal is to provide a comprehensive description of the influential and important clustering algorithms rooted in statistics, computer science, and machine learning, with emphasis on new advances in recent years.

Supervised Learning

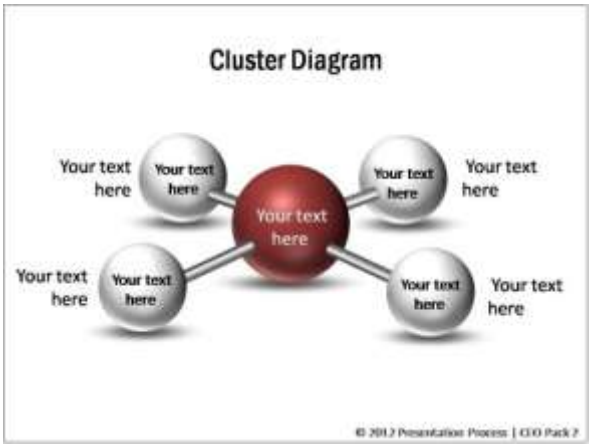
In supervised training, data includes together the input and the preferred results. It is the rapid and perfect technique. The accurate results are recognized and are given in inputs to the model through the learning procedure. Supervised models are neural network, Multilayer Perceptron and Decision trees.

Unsupervised Learning

The unsupervised model is not provided with the accurate results during the training. This can be used to cluster the input information in classes on the basis of their statistical properties only. Unsupervised models are for dissimilar types of clustering, distances and normalization, k-means, self organizing maps.

CLUSTERING

Clustering is a major task in data analysis and data mining applications. It is the assignment of combination a set of objects so that objects in the identical group are more related to each other than to those in other groups. Cluster is an ordered list of data which have the familiar characteristics.
Cluster analysis can be done by finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters.



CLUSTERING ALGORITHMS

A rough but widely agreed frame is to classify clustering techniques as hierarchical clustering and partitional clustering, based on the properties of clusters generated. Hierarchical clustering groups data objects with a sequence of partitions, either from singleton clusters to a cluster including all individuals or vice versa, while partitional clustering directly divides data objects into some prespecified number of clusters without the hierarchical structure. We follow this frame in surveying the clustering algorithms in the literature.

Distance and Similarity Measure

An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are all in the same physical units then it is possible that the simple Euclidean distance metric is sufficient to successfully group similar data instances. Distance between the two clusters can be measured by[1].

1.Euclidian Distance

2.City Block Distance

In addition to this some of the similarity and dissimilarity measures.

HIERARCHICAL CLUSTERING

A more common measure is Euclidean distance, computed by finding the square of the distance between each variable, summing the squares, and finding the square root of that sum. In the two-variable case, the distance is analogous to finding the length of the hypotenuse in a triangle; that is, it is the distance "as the crow flies".

CLASSIFICATION

Clustering algorithms may be broadly classified as listed below:

A. Hierarchical
---Agglomerative

Single linkage,

Complete linkage,

Group average linkage,

Median linkage,

Centroid linkage,

Ward's method,

Balanced iterative reducing and clustering using hierarchies (BIRCH),

Clustering using representatives (CURE),

Robust clustering using links (ROCK)

Divisive analysis (DIANA), monothetic analysis (MONA

Squared Error-Based (Vector Quantization)

K-means

Fuzzy
a.Fuzzy -means (FCM),
b. Mountain method (MM),

Possibilistic means clustering algorithm (PCM),
c. Fuzzy shells (FCS)

Neural Networks-Based

Learning vector quantization (LVQ),
Self-organizing feature map (SOFM), ART,
Simplified ART (SART),

Hyperellipsoidal clustering network
Self-splitting competitive learning network

Kernel-Based

Kernel -means,

Support vector clustering (SVC)

Data visualization/High-dimensional data

Iterative self-organizing data analysis technique (ISODATA),

Genetic -means algorithm (GKA),

Partitioning around medoids (PAM)

Similarly various clustering algorithms and their complexities are mentioned

Partitioning Algorithms

Partitioning clustering algorithm split the data points into k division, where each division represent a cluster and k<=n, where n is the number of data points. Partitioning methods are based on the idea that a cluster can be represented by a centre point. The cluster must exhibit two properties, they are (a) each collection should have at least one object (b) every object should belong to accurately one collection. The main drawback of this algorithm is whenever a point is close to the center of another cluster; it gives poor outcome due to overlapping of data points [4]. It uses a number of greedy heuristics schemes of iterative optimization.

The basic algorithm is very simple
1. Select K points as initial centroids.
2. Repeat.
3. Form K clusters by assigning each point to its closest centroid.
4. Re-compute the centroid of each cluster until centroid does not change.

K-Medoids Algorithm:

In this algorithm we utilize the actual entity to represent the cluster, using one representative entity per cluster. Clusters are generated by points which are close to respective methods. The partitioning is made based on minimizing the sum of the dissimilarities among every object and its cluster representative .

a) Randomly choose k objects in D as the first representative objects or seeds.

b) Repeat
i) Allocate each lasting object to the cluster with the nearby representative object
ii) Arbitrarily choose a non-represe
ntative object, orandom
iii) Calculate the total cost, S, of swapping representative objects oj with orandom iv) If S<0 then swap oj with orandom to form the new set of k representative objects.
c) Until no change.

HIERARCHICAL CLUSTERING ALGORITHM

Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical is this:

1.Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.

Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.

CONCLUSION

The objective of the data mining technique is to mine information from a large data set and make it into a reasonable form for the supplementary purpose. Clustering is a significant task in data analysis and data mining applications. It is the task of arranging a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Clustering algorithms can be classified into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. Under partitioning method, a brief description of k-means and k-medoids algorithms have been studied. In hierarchical clustering, the BIRCH and CHAMELEON algorithms have been described. The DBSCAN and DENCLUE algorithms under the density based methods have been studied. Finally, under grid-based clustering method, the STING and CLIQUE algorithms have been described. The challenge with clustering analysis is mainly that different clustering techniques give substantially different results on the same data. Moreover, there is no algorithm present which gives all the desired outputs. Because of this, there is extensive research being carried out in „ensembles‟ of clustering algorithms, i.e. multiple clustering techniques done on a single dataset

REFERENCES

Textbook on “Pattern Recognition and Image Analysis”,Earl Gose, Richard Johnsonbaugh ,Steve Jost.

1.A.K. Jain, ,M.N. Murty, P.J. Flynn.” Data Clustering: A Review”, ACM Computing Surveys, Vol. 31, No. 3, September 1999.

2.Rui Xu, Donald Wunsch “Survey of Clustering Algorithms”,IEEE Transactions on Neural Networks Vol 16, No. 3, May 2005.

3. S. Anitha Elavarasi and Dr. J. Akilandeswari (2011) *A Survey On Partition Clustering Algorithms*, International Journal of Enterprise Computing and Business Systems.

4. S.Vijayalaksmi and M Punithavalli (2012) *A Fast Approach to Clustering Datasets using DBSCAN and Applications* (0975 – 8887) Vol 60– No.14.

5. K.Pavithradevi, K.Ramya. S.Nandhini, G.Punitha, " History and Applications in Body Arear Network",International Journal for Research in Applied Science & Engineering Technology Vol 5, Issue II, February2017

6. Data Clustering. A Review: A.K. Jain Michigan State University and M.N. Murty Indian Institute of Science and P.J. Flynn The Ohio State University.

7. Pavel Berkhin, *Survey of Clustering Data Mining Techniques*, Accrue Software, Inc.

8. NavneetKaur, *Survey Paper on Clustering Techniques*, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013. [12]. Nisha and Puneet Jai Kaur, "*A Survey of Clustering Techniques and Algorithms*", IEEE (978-9-3805-4415- 1), 2015

9. Han, J. and Kamber, M. *Data Mining- Concepts and Techniques*, 3rd Edition, 2012, Morgan Kauffman Publishers.

10. Megha Mandloi, *A Survey on Clustering Algorithms and K-Means*, July-2014.

11.K.Ramya and K.Pavithradevi "Effective Wireless Communication," International Journal of Advanced Research, volume4(12),pp. 1559-1562 Dec 2016.

12.G.Arunachalam, K.Ramya, M.Vimala, M.Shanmugapriya,C.Krishnaveni,"Future Principle of TCP High-Speed Network "International Journal for Research and development & Technology" Volume-7,Issue-2 (Feb-17) ISSN (O) :- 2349-3585

13. Karthikeyan.R, Dr.Geetha.T ,Ramya.K ,Pavithradevi.K," A Survey on Sensor Networks", International journal for Research and Development in Technology, Volume 7,Issue 1 Jan 17.

14.K.Ramya,G.Arunachalam,M.Shanmugapriya,P.Sathya "Mobile Computing Broadband Networks "International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 5 Issue II, February 2017 ISSN: 2321-9653

15.L.Gomathi, K.Ramya "Data Mining Analysis using query Formulation In Aggregation Recommendation", Volume 2 Issue 1- October 2013.