INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

SURVEY ON PERFORMANCE ANALYSIS OF DIFFERENT CLUSTERING ALGORITHMS

Mrs Shobha D

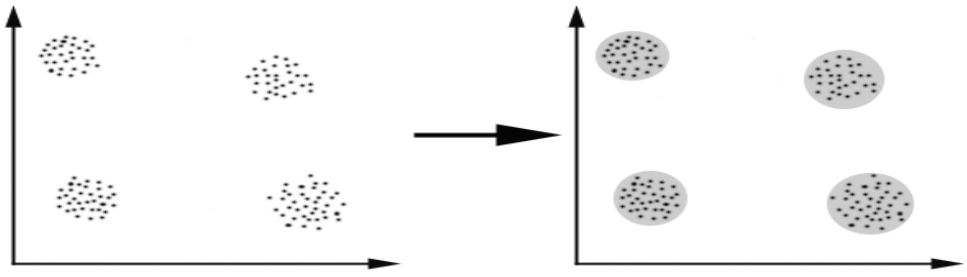
Assistant Professor, DoS in Computer Science,
PG wing of SBRR Mahajana First Grade College,
K R S Road, Metagalli Mysuru -16
E-mail: shobhadgps@gmail.com

Abstract: - Dividing the collected data into groups of similar objects is clustering. Every object in cluster exhibits the similar property between them. Clustering is a method to represent many data in few groups. Clustering model is based on supervised learning method or unsupervised learning but for some clustering we use both. Main goal of clustering is to obtain a new set of cluster which are descriptive and predictive. It provides a way to analyse the object without having the previous knowledge of data. In my paper, I made survey on research done on clustering in different area and its algorithms used for it.

Keywords: clustering, k-means, data mining, partitioning, knowledge discovery.

Introduction

Data Mining is an emerging technology; it made a revolutionary change in the information world. Data mining is nothing but knowledge discovery, to discover knowledge it analyzes the data from different perspectives and summarizes into useful format. Data mining involves the activity of the partitioning the set of data objects into subset, to perform this activity it uses the techniques called clustering analysis. They are used to structure the objects in a manner which exhibits similarity between themselves, the dissimilar objects into another group. A number of previous studies have discussed and evaluated the performance of different clustering algorithms in many research areas like image processing, pattern recognition, market research, information recovery and in machine learning concept. A *cluster* is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.



A good clustering algorithm should have the following properties:-

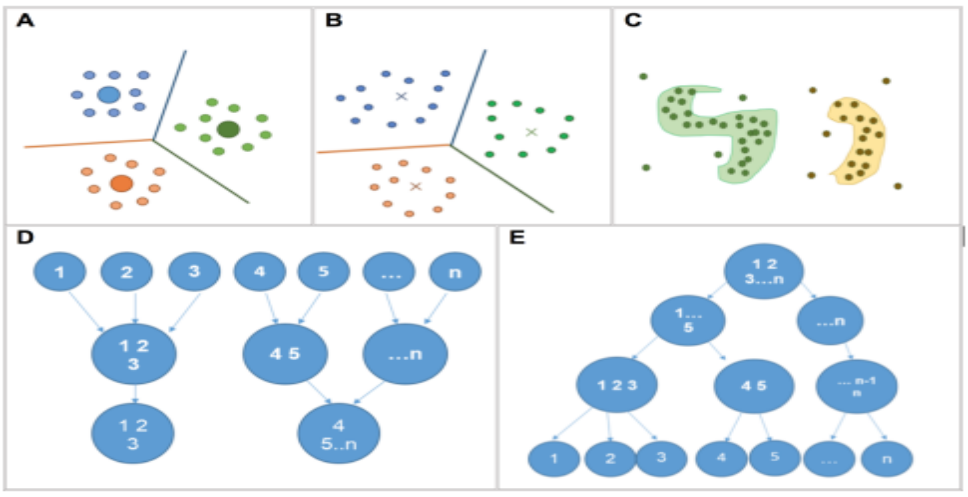
- Scalability: it is an ability to perform cluster well with large number of data elements.
- Analyze mixture of attribute types: It is the ability to analyze set of data with single as well as mixture property of the elements.
- Find arbitrary-shaped clusters: Different types of algorithms are should show inclination towards finding different types of cluster structures or shapes.
- Minimum requirements for input parameters: in order to analyse data many clustering algorithms require some parameters, such as the number of clusters. With large datasets and higher dimensionalities, it is desirable that a method require only limited guidance from the user, in order to avoid bias over the result.
- Handling of noise: in order to improve cluster quality Clustering algorithms should handle deviations; Deviations are defined as data objects as outliers.
- Sensitivity to the order of input records: when we provide the same data set in different order to certain algorithms may produce different results. The order of input may affects the algorithms that require a single scan over the data set, leading to locally optimal solutions at every step. Thus, it is crucial that algorithms be insensitive to the order of input.
- High dimensionality of data: As amount of resources increases then storing or represent them also grows. The distance of a given point from the nearest and furthest neighbor is almost the same it effects the efficiency of a clustering algorithm, since it would need more time to process the data, while at the same

- Interpretability and usability: Most of the times, it is expected that clustering algorithms produce usable and interpretable results. But when it comes to comparing the results with preconceived ideas or constraints, some techniques fail to be satisfactory. Therefore, easy to understand results are highly desirable.

Clustering methods:

Depending on the analysis of data type, similarity exhibited by themselves that is similarity measure and the theory used to define clustering on the basis of technique used to create the partition. In this paper I made a survey on different clustering algorithm used by different researcher in different area. Some of them are partitioning clustering; Hierarchical clustering and Density based clustering.

- **Partitioning method**: This algorithm make K partitions for n objects, where K<=n similar to one another and in other cluster. It consists dissimilar to other objects of other clusters. In this method objects within the data set are represented by a centroid in K- Medoids as shown in fig A or value within the domain space that is K-means as shown in fig B



K-means algorithm is depends on centre and always converges to the nearest local optimum from the starting position of the search. Mulik and Bandyopodhyay[12] have suggested a technique using genetic algorithm to decide the clustering issue which was experimental on synthetic and real life dataset to calculate the performance. Krisna Murthy [11] suggested this is a Model which expresses a vital mutation operatior controlled clustering. It is a simple fast relatively efficient and centre based algorithm. But it applicable only when mean is defined and for non convex shapes partition around Medoids works effectively for small data set does not scale well for large data sets. The complexity is $o(k(n-k)^2$ of n objects for K clusters.

Hierarchical clustering method: It clusters' the objects in hierarchy or tree representation as shown in fig D. It initializes a cluster as a set of singleton or single cluster of all points and proceeds the process by slitting the objects into group until it achieve the last criterion. Agglomerative or singleton cluster and divisive hierarchical clustering are as shown in fig D and fig E.
Hierarchical clustering uses distance metrics as clustering criteria. Similarity of the closest pair of data points belongs to different cluster is used to measure the similarity between clusters. In this method Hierarchical structure are constructed with a clustering feature tree for multiple clustering. First it scans the database to build the in memory clustering feature tree than leaf nodes of the clustering feature tree are clustered using some arbitrary clustering algorithm. A good clustering is obtained by one scan but to improve the quality few additional scans are used. It handles only numerical data and very sensitive to the order of the data records.

Density – based method: In this method object are grouped to obtain cluster by considering the local densities that is nothing but neighbourhoods as shown in fig C for the identification of clusters of arbitrary shapes. It uses distance function to measure the quality of cluster. It provides a natural protection against outliers. Given cluster continues growing as long as the number of objects in the neighborhood exceeds some parameter.
In this we have two types

- Density-Based Connectivity Clustering: In this method using local distribution of nearest neighbours, clustering technique density and connectivity are measured. Depending upon the density all the objects are reachable from one another. Connectivity is a symmetric relation from core objects which can be factorized into maximal connected components serving as clusters. The points that are not connected to any core point are not covered by any cluster.
- Density Functions Clustering: In this density function clustering method, local maxima of the overall density functions are used to compute density. Overall density is modeled as the sum of the density functions of all objects. Clusters are obtained by *density attractors*. The influence function can be an arbitrary one.

RELATED WORK:

The purpose of this study is to make Survey on Performance Analysis of different Clustering algorithms. I studied various journals and articles regarding performance evaluation of Data clustering algorithms in different application with different tools, some of them are described here, Ying Liu et all worked on Classification algorithms while Osama abuabbas worked on clustering algorithm, and Abdullah compared various classifiers with different types of data set on WEKA,

Ying Liu, wei-keng Liao et al[15] in his article they mentioned mine bench and benchmarking suite containing datamining applications. It includes categories & applications that are commonly used in industry there by achieving a realistic representation of the existing applications. JiangjiaoDuan et al., in 2005 discussed the Model-based clustering is very important ways to represent time series data mining. The process of clustering may meet several problems. Here a novel clustering algorithm of time-series which include recursive Hidden Markov Model (HMM) training was proposed.

P.T. Kavitha and Dr. T.Sasipraba [13] in their research paper using java platform they analysed the performance of distributed data mining framework. The aim of framework was to developed and efficient association rule mining tool are used to support effective decision making. To find the patterns from huge amount of data available in the data warehouses they used Association rule. Pramod S. and O.P.vyas[14] in their research paper to analyze the sorted and unsorted data sets using association rule mining algorithm. They

worked on Continuous Association Rule Mining Algorithm (CARMA) and DataStream Combinatorial approximation Algorithm (DSCA). They also implemented the algorithms in JAVA and the results were plotted, all algorithms were tested with 5 datasets and are available in Frequent Item set Mining data set (FIM) repository. The transactions of each data set were looked up one by one in sequence to simulate the environment of an online data stream. The DSCA algorithm used sorted transaction items while other 2 algorithms used unsorted transaction items. For performance evaluation of the popular clustering techniques, Performance Measure for reach characteristic, we analyzed how the results differ whenever test mode is changed.

CONCLUSION:

In data analysis clustering is very important. Clustering lies at the heart of data analysis and data mining applications. It finds highly mutual relationship between the regions of objects. When their number becomes incredibly large it is highly useful, previous research try to discover the underlying behavior without having any prior knowledge about the data. A good clustering method will produce high quality clusters with similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The selection of the clustering algorithm is based on the knowledge of the structure of data. The quality of a clustering is also depends on its ability to discover all the hidden patterns. Many research papers used K-means algorithm for evaluation and investigation.

REFERENCES

[1] Ahmed M., Yamany S., Mohamed N., Farag A., and Moriarty T., "A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data," *IEEE Transactions on Medical Imaging*, vol. 21, no. 3, pp. 193-199, 2002.
[2] Sun Shibao, Qin Keyun,"Research on Modified kmeans Data Cluster Algorithm"I. S. Jacobs and C. P. Bean, *"Fine particles, thin films and exchange anisotropy,"* Computer Engineering, vol.33, No.13, pp.200–201,July 2007.
[3] Merz C and Murphy P, UCI Repository of Machine Learning databases, Available :ftp://ftp.ics.uci.edu/pub/machinelearning-databases
[4] Liu, Y. et al., 2013. Understanding of Internal Clustering Validation Measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. Sydney, New South Wales, pp. 1–6
[5] VOORHEES, E.M. 1986. Implementing agglomerative hierarchical clustering algorithms for use in document retrieval. Information Processing and Management, 22, 6, 465-476.
[6] PELLEG, D. and MOORE, A. 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In Proceedings 17th ICML, Stanford University.
[7] RAMASWAMY, S., RASTOGI, R., and SHIM, K. 2000. Efficient algorithms for mining outliers from large data sets, Sigmoid Record, 29, 2, 427-438.
[8] BRADLEY, P. S., BENNETT, K. P., and DEMIRIZ, A. 2000. Constrained k-means clustering. Technical Report MSR-TR-2000-65. Microsoft Research, Redmond, WA.
[9] Mutanen. T et all, (2010), Data Mining for Business Applications , Customer churn prediction – a case study in retail banking , Frontiers in Artificial Intelligence and Applications, Vol 218
[10] Pramod S., O. Vyas(2010), Performance evaluation of some online association rule mining algorithms for sorted & unsorted datasets, International Journal of Computer Applications, vol 2,no. 6
[11] Krishna K. and Narasimha M., "Genetic K-means Algorithm," *IEEE Transactions on Systems Man and Cybernetics B Cybernetics*, vol. 29, no. 3, pp. 433-439, 1999.
[12] Mualik U. and Bandyopadhyay S., "Genetic Algorithm Based Clustering Technique," *Pattern Recognition*, vol. 33, no. 9, pp. 1455-1465, 2002.
[13] Kavitha P.,T. Sasipraba (2011), Performance evaluation of algorithms using a distributed data mining frame work based on association rule mining, International Journal on Computer Science & Engineering (IJCSE)
[14] Pramod S., O. Vyas(2010), Performance evaluation of some online association rule mining algorithms for sorted & unsorted datasets, International Journal of Computer Applications, vol 2,no. 6
[15] *www.ics.**uci**.edu/~mlearn/*
[16] Xiaozhe Wang, Kate Smith and Rob Hyndman: "Characteristic-Based Clustering for Time Series Data", *Data Mining and Knowledge Discovery, Springer Science + Business Media, LLC Manufactured in the United States,* 335–364, 2006.