ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 7, July 2016

Clustering:

An art of grouping related objects

Sumit Kumar, Sunil Verma

Abstract- **In today's world, clustering has seen many applications due to its ability of binding related data together but there are many challenges while accomplishing the task which seems to be simple. Need for clustering arisen due to desire for useful data at one location out of huge amount data increasing day by day. Clustering has proved itself fruitful in many areas like search engines, e-governance and e-business etc. This paper intends to give overview of many clustering algorithms with their characteristics. It also attempts to spot out various clustering techniques and briefly reveals which technique is more suitable for particular requirement.**

Index Terms Clustering, data mining, dataset, mean, medoid.

I. INTRODUCTION

Data is increasing explosively at large scale but not all the collected data is valuable. We have to extract useful knowledge which is a cumbersome task in itself. Throughout the years, many techniques have been developed to extract interesting and hidden knowledge from large dataset. Several different methodologies like Clustering, Classification and Association rule[1]are there to solve the problem.

Data analysis divides data into meaningful clusters. If meaningful clusters are the ultimate goal then the resultant clusters should use the "natural" structure of the data. Cluster Analysis is based on the information found in the data describing the objects and their relationships.

Sumit Kumar*, M.Tech student, computer science, Guru Jambheshwar University of Science and Technology, Hisar. Hisar, India.*

Sunil Verma*, Computer Science Jambheshwar University of Science and Technology, Hisar, India.*

Clustering can be seen as data segmentation which is adaptable to diversity and aids to point out different features that can identify a particular cluster.

Clustering is often considered as a form of data compression. Clustering is basic technique of grouping related data values based upon some similarity measures and similar values are assigned a particular cluster label whereas similarity between values are defined with some distance formula. Various data mining applications are clustering based and clustering methods are available that made it possible to label related data separately.

II. DATA MINING

Data mining is also known as knowledge mining where intelligent methods are applied to extract data patterns that can be useful to discover useful knowledge. The data source might include databases, data warehouses, the web, other repositories, or the data that are dynamically streamed into the system.

In data mining, the electronically stored data is searched in automated and least augmented fashion by computer in a bid to solve problems by data analysis. It is popular with successful applications in telecommunication, marketing and web pages but in these days data mining has proven itself in other fields also.

Pattern discovery process in data mining must be meaningful that lead to knowledge mining automatically or semi-automatically without any human intervention. The pattern discovered must be

useful in some or the other way leading to support future decision concerning analysis of new database.

Data mining offers promising ways to find out hidden patterns from huge data that can potentially be used

to predict future behaviour. New data mining algorithms should be used with caution so as to produce desired result against small computational cost. Prior to the mining process, it is essential to have sufficient amount of data exploration that might be integrated from multiple heterogeneous sources and converting it into a form precise to a target decision support application. After that the data needed to be prepared by applying pre-processing for knowledge extraction, the next step is to choose the technique that is most appropriate to mining. However, there is always a trade-off to consider when to choose the appropriate data mining technique to be used in a certain application[1]. Trial and error is considered to be best method to proceed.

III. APPLICATIONS OF DATA MINING

Data mining research and development has importance in many areas including digital libraries, E-governance, industries, E-business, search engines, marketing, finance, health care, bioinformatics etc. where large amount of data is available to be stored and operated[1].

IV. CLUSTERING

Clustering is the process of organizing data into groups with members having similarity in one or the other way. Clustering can be considered as the most important unsupervised learning technique where rather than using model, clustering of actual data is performed directly.

The goal is that the objects in a group will be similar to other objects in the same group and different from the objects in different groups[1]. The greater the similarity within a particular group and the greater the difference between different groups, the better the clustering will be.

Purpose of clustering analysis is to divide available data into useful and meaningful clusters. Cluster analysis can be used as an initial step for further data analysis, e.g. data compression where the goal is to maximize the intra-cluster similarity and to minimize the inter-cluster similarity. Most cluster analysis techniques divide data separation into non overlapping groups.

V. CLUSTERING APPROACHES

Some working definitions of cluster are discussed below:

Well-Separated Cluster: A cluster with set of points such that any point in a cluster is more similar to every other point in the same cluster than to any point in some other cluster. Sometimes a threshold is specified that every point in a cluster must be suitably close to one another.

Centre-based Cluster: A cluster with set of objects such that an object in a cluster is closest to the central value of that particular cluster in which it is assigned.

Contiguous Cluster (Transitive Clustering or Nearest neighbour): A cluster is a set of points such that a point in a cluster is similar to one or more points in a cluster than to any point not in the cluster.

Density-based: A cluster with dense region of points is separated by less dense regions. This definition is more often used when the clusters are irregular or intertwined.

Similarity-based Cluster: A cluster with similar objects is categorized under this category. This defines a cluster as a set of points which altogether create a region with a uniform local property, e.g. density or shape.

Various Clustering Algorithms: There are many clustering algorithms being used today, some of them are discussed here.

Partitioning algorithm: Desired clusters are obtained by partitioning data into subsets and then by recursively iterating and relocating points between subsets until specified criterion met[2].

Hierarchical algorithm: Hierarchical clustering algorithm seeks to build clusters by hierarchical decomposition of values using some criteria. Two basic hierarchical approaches are:

- **Agglomerative:** Desired clusters are obtained by repeatedly merging available clusters at each step on the basis of similarity.

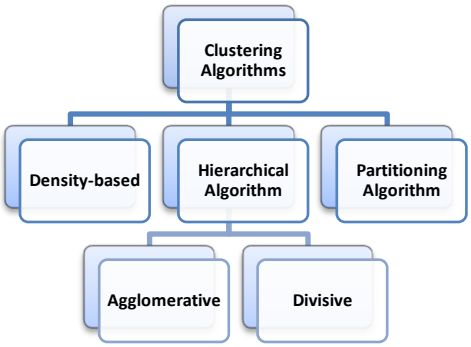

Figure Error! No text of specified style in document.**.1 Clustering Algorithms**

- **Divisive:** Desired clusters are obtained by repeatedly splitting available clusters until some similarity criterion met.

Density-based: Density-based clustering actually concerned with the space around various data points rather than around individual point. Dense region is specified with the help of core points, border points and separated from noise points. The region with many similar data values is known as dense region.

VI. LITERATURE REVIEW

Selim and Ismail[3] formulated clustering problem as mathematical problem. Restatement of k-mean algorithm is used to find local minimum solution. This algorithm gives partial optimal solution and if Minkowsky metrics is used there might be chances of not getting local minimum of the given problem under certain conditions.

Further modified clustering algorithm, CURE (Clustering Using Representatives) proposed by Guha*et al.*[4], identifies large shaped, non-spherical clusters that are generated by well scattered points and shrinking them toward the centre. Traditional clustering algorithms usually converge with spherical shaped and similar sized clusters and are sensitive to outliers. CURE uses Random sampling along with outlier handling techniques and effectively removes outliers contained in the data set. A random sample drawn from the data set is first partitioned and each partition is clustered partially. The partial clusters are then clustered further in a next pass to yield the desired clusters. Experimental result shows that proposed algorithm outperforms other existing algorithms.

Two new algorithms are introduced by Huang [5] are extension of k-mean algorithms. New algorithms work with categorical domain and mixed domain. Original k-mean algorithm works only on numerical values while proposed algorithms work with both numeric and categorical values. The *k*-modes algorithm can be applied on categorical data by using dissimilarity measure. To minimize the clustering cost, frequency-based method is used to update modes in clustering process. The *k*-prototypes algorithm works with mixed numeric and categorical attributes by combining algorithm k-mean and k-mode algorithms. It use combined dissimilarity measure for mixed domain. These algorithms works better with large datasets and are useful in data mining applications.

Jain *et al.*[2] made a survey on different clustering techniques and applications. Different clustering steps are described as following:

(1) Pattern representation

(2) Similarity computation

(3) Grouping process

(4) Cluster representation.

Finding all useful data values from high dimensional dataset is problematic and to tackle these problems data abstraction is used instead of using whole data set directly. Data abstraction is simple and useful representation of data and is achieved using clustering. Paper put forward various statistical measures using fuzzy, neural, evolutionary, and knowledge-based approaches to clustering. Clustering is a subjective process where subjectivity describes different partitions of same data sets for different applications. This subjectivity makes the clustering process difficult. The solution of subjectivity can be achieved if used in the form of knowledge either implicitly or explicitly in one or more phases. This knowledge based clustering algorithms used domain knowledge explicitly.

Likas *et al.*[6] proposed a global K-Means clustering algorithm which takes total number of executions equal to total data set. This algorithm executes with

one cluster centre at a time with incremental move for dynamic data exploration procedure to reduce computational load having no effect on solution quality. This technique is deterministic global optimization technique where the sequence of local search is employed. This technique is also applicable to high dimensional data.

Sun *et al*.[7] put forward global optimization based semi-supervised K-Means algorithm. This algorithm can produce suitable number of clusters by using small amount of labelled data and produces large amount of supervision data. Voting rule for clustered labelling is used by integration of distribution features of data sets. Although k-mean is one of most commonly used partitioning technique where several attempts are made to determine better proximity but some general features are still undiscovered.

Barakbah and Kiyoki[8] introduced an approach to optimize initial centroid motivated by a thought of maximizing the pillar distance while building a house in a view to support ceiling value. In traditional method the initial centroid are chosen so as to reach nearest local minima rather than concentrating global optimum value. While proposed algorithm uses interactive approach to select initial centre. Algorithm calculate initial centre as weighted average of farthest data points with outlier detecting mechanism.

Bandyopadhyay and Maulik[9]introduced KGA clustering technique that coupled k-mean clustering algorithm with genetic algorithm. This technique overcomes the major limitation of k-mean algorithm which emphasis on local minimum solution. Centroid of large values available in the data set represents the centre of the cluster which is used to enhance the searching capability of the cluster. For data representation in this revolutionary approach floating point representation has been adopted. Experimental results shows that KGA algorithm leads to desired results in finite number of iterations with better speed and accuracy than that of traditional GA based algorithm. But problem is that this algorithm requires number of clusters required to be defined manually and provides clusters with crisp in nature.

Fahim *et al.*[10] proposed an algorithm that uses distance calculated in the previous iteration to check similarity with particular cluster in iterative approach of k-mean algorithm. This technique basically focuses on improving the time complexity without making any compromise with quality of cluster. This technique proposed that by calculating distance of all the elements with mean value in previous iteration can be used in the following step that lead to reduce computational cost and save time with same cluster quality.

An algorithm based on k-mean algorithm given by Ahmad and Dey[11] performs clustering on data with mixed numeric and categorical features that provide better cluster quality while traditional k-mean algorithm is limited to numeric data only. Numeric distance measure for numeric and categorical data is calculated by first converting into numerical values although exact numerical value for a particular categorical data is almost impossible yet efforts applied. Cost function and distance measure used in the algorithm ensure final cluster quality.

Arthur and Vassilvitskii[12] introduced a k-mean++ algorithm that focus on improvement in efficiency and accuracy of k-mean algorithm by augmentation of simple randomized seeding technique. Initial centre is chosen uniformly at random while next centre for other cluster is chosen with probability measure. The proposed algorithm performed consistently better than traditional k-mean in both speed as well as accuracy.

Vattani[13] proposed an algorithm that reduced number of iterations by improving well known lower bound from (n) to $2(\sqrt{n})$ and improved k-mean in terms of dimensionality also. Although k-mean is widely used algorithm for partitioning of large number of data points in d-dimensional space into desired number of clusters for a long time yet number of iterations used in traditional algorithm is almost unpredictable.

An algorithm that focuses on improvement in accuracy and efficiency of k-mean algorithm with reduced cost and complexity was given by Nazeer and Sebastian [14]. Traditional clustering algorithm does not ensure quality of clusters because different initial centroid leads to different clusters when k-mean algorithm on same data points is employed. Many improvements have been proposed in the

history but all with negotiations either for accuracy or efficiency. Proposed algorithm aiming to perform entire clustering with assured cluster quality.

Khan et al., [15] proposed Multi Agent System in which they used actual sample data points for the generation of initial centroid. The proposed system used four agents for the generation of initial centroid viz. Range, Random, Number Outlier and Inliers. K-Means clustering is extensively used due to ability of better local convergence but this requires the initial centroid should be taken carefully. Range method has been proposed to overcome the problem of sensitivity to initial condition in which outliers and inliers are avoided to be taken as initial centroid.

Yi *et al.*[16] put forward an approach to find the initial centre to reduce its effect on final clusters using density based approach. Traditional clustering techniques like k-mean, k-medoid suffers from many limitations like number of clusters and initial values of centroids. But by using this approach, k data objects from high density area are selected as the initial centre. These initial centres are used further in selection of cluster variables.

Gupta and Shrivastava[17] compared various clustering approaches. In this paper, Advantage of using k-mean algorithm is with small number of clusters even with large dataset and produces tight global clusters but problem arise for the prediction of values with different size and density. k-medoid

algorithm is better than k-mean for selecting central dataset. BIRCH performs hierarchical clustering at the first phase while refinement in the next phase in value but performance decreases in the case of large

Table 1 Comparision of clustering algorithms.

Sr. no.	Clustering	Advantage	Disadvantage	Complexity
1	**K-mean**	• Simple • Understandable • Items automatically assigned to clusters	• Number of clusters to be defined beforehand • Items are forced into clusters • Sensitive to outliers • Local minima	O(nkid)
2	**Pam**	• Less influenced by outliers • Work well for small data set	• Not good for large dataset	$O(k(n-k)^2$
3	**DBSCAN**	• Automatically find number of clusters • Arbitrary shape clusters • Robust to noise and outliers	• Sensitive to parameters • Difficult to wor with high dimensional data	$O(n \log n)$
4	**CURE**	• Adjust well to the geometry of non-spherical shapes • Employs a combination of random sampling and partitioning.	• Uses random samples that may effects final cluster quality.	$O(n^2 \log n)$
5	**BIRCH**	• Good clustering with single scan	• Sensitive to order of data records • Handle numeric data only	$O(n^2)$

which leaf clusters are clustered again. Better memory utilization is achieved but problem is that it requires different measures for different connectivity and closeness. DBSCAN is a density based clustering and does not require specifying number of clusters. But DBSCAN can't cluster dataset with high density differences and the quality depends upon distance measure used.

Kaushik and Mathur[18] given comparative study of k-mean with other techniques. Paper states that quality of traditional k-mean increases with increase

in number of clusters but useful in case of numerical data only. In order to deal with categorical data, hierarchical algorithm was developed in which rank is assigned by converting categorical data into numerical values. Noisy data is problematic in every clustering techniques but its effect decreases with increase in huge dataset. Performance of k-mean algorithm is better in case of large dataset.

VII. CONCLUSION

Clustering can be seen as data segmentation which is adaptable to diversity and aids to point out different features that identify similarity. Traditional methods used in data clustering generally focused on local minima e.g. k-mean but without making any concern about global data while others stick to global data rather than local similarity. Many algorithms are available that require number of clusters to be defined e.g. k-mean, k-medoid, KGA etc. Manually stating number of clusters leads to merging of heterogeneous data if number of clusters defined is less and in case, if number of clusters is more than actual distribution, it leads to fragmentation of homogenous data. In many algorithms, output cluster generated depends upon the initial splitting value that leads to different result even if same data values are used. Thus output quality of algorithm is very much affected by the initial splitting value. While some of the algorithms suffer from problem of high computational cost, others work better with continuous range of data leading to empty clusters when data with irregular values are used.

Many clustering algorithms have been proposed but the main concern is efficiency and accuracy. Pre-requirement for the specification of number of clusters is to be removed making final cluster dependent on actual data itself rather than initial central value and number of clusters specified.

REFRENCES

[1] "Han and Kamber: Data Mining---Concepts and Techniques, 3nd ed., Morgan Kaufmann, 2011."

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," in *ACM computing surveys*, vol. 31, no. 3, 1999, pp. 264–323.

[3] S. Z. Selim and M. A. Ismail, "K-means-type algorithms: a generalized convergence theorem and characterization of local optimality," in *Pattern Analysis and Machine Intelligence on IEEE Transactions*, no. 1, 1984, pp. 81–87.

[4] S. Guha, R. Rastogi, and K. Shim, "CURE: an efficient clustering algorithm for large databases," in *ACM SIGMOD Record*, vol. 27, 1998, pp. 73–84.

[5] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," in *Data mining and knowledge discovery*, vol. 2, no. 3, 1998, pp. 283–304.

[6] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," in*Pattern recognition*, vol. 36, no. 2, 2003, pp. 451–461.

[7] X. Sun, K. Li, R. Zhao, and X. Hu, "Global Optimization for Semi-supervised K-means," in *Asia-Pacific Conference on Information Processing*, vol. 2, *2009*, pp. 410–413 .

[8] A. R. Barakbah and Y. Kiyoki, "A pillar algorithm for k-means optimization by distance maximization for initial centroid designation,"

in *Computational Intelligence and Data Mining on IEEE Symposium*, 2009, pp. 61–68,.

[9] S. Bandyopadhyay and U. Maulik, "An evolutionary technique based on K-means algorithm for optimal clustering in RN," in *Information Science*, vol. 146, no. 1, 2002, pp. 221–237.

[10] A. M. Fahim, A. M. Salem, F. A. Torkey, and M. A. Ramadan, "An efficient enhanced k-means clustering algorithm," *Journal of Zhejiang University SCIENCE A* , vol. 7, no. 10,2006, pp. 1626–1633.

[11] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data & Knowledge Engineering* ,vol. 63, no. 2, 2007, pp. 503–527.

[12] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007,pp. 1027–1035.

[13] A. Vattani, "K-means requires exponentially many iterations even in the plane," *Discrete & Computational Geometry*, vol. 45, no. 4, 2011, pp. 596–616.

[14] K. A. Nazeer and M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm," in *Proceedings of the World Congress on Engineering*, vol. 1, 2009, pp. 1–3.

15] D. M. Khan and N. Mohamudally, "A multiagent system (MAS) for the generation of initial centroids for k-means clustering data mining algorithm based on actual sample datapoints," in *2nd International Conference on Software Engineering and Data Mining*, 2010, pp. 495–500.

[16] B. Yi, H. Qiao, F. Yang, and C. Xu, "An improved initialization centre algorithm for k-Means clustering," in *International Conference on Computational Intelligence and Software Engineering*, 2010, pp. 1–4.

[17] M. Gupta and V. Shrivastava, "Review of various Techniques in Clustering," in *International Journal of Advanced Computer Research*, vol. 3, no. 2, 2013, pp. 134-137.

[18] M. Kaushik and M. B. Mathur, "Comparative Study of K-Means and Hierarchical Clustering Techniques," in*International Journal of Software & Hardware Research in Engineering*, vol. 2, no. 6, 2014, pp. 93-98.