INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 05, May 2024, pp: 2077-2085

e-ISSN : 2583-1062

Impact Factor: 5.725

A STUDY ON VARIOUS DATA CLUSTERING APPROACH

Nitish Marathe[1], Dr. Harsh Lohiya[2], Mr. Ankit Navgeet Joshi[3]

[1]Research Scholar, Department of CSE , SSSUTMS, Sehore, Madhya Pradesh, India.

[2]Associate Professor, Department of CSE , SSSUTMS, Sehore, Madhya Pradesh, India.

[3]Assistant Professor, Department of CSE , SSSUTMS, Sehore, Madhya Pradesh, India.

ABSTRACT

A evaluate primarily based on different types of clustering algorithms with their corresponding data units has been proposed. in this paper, we've got given a entire comparative statistical evaluation of various clustering algorithms. Clustering algorithms generally appoint distance metric or similarity metric to cluster the facts set into special walls. well known clustering algorithms were broadly utilized in diverse disciplines. form of clustering algorithm used relies upon upon the software and facts set utilized in that field. Numerical facts set is comparatively easy to implement as data are always actual quantity and can be used for statistical applications. Others form of information set which include specific, time collection, boolean, and spatial, temporal have restricted packages. by using viewing the statistical analysis, it's far determined that there's no finest answer for coping with issues with massive data sets of mixed and specific attributes. some of the algorithms may be carried out however their performance degrades as the scale of statistics maintains on growing.

Keywords: Data Clustering, Statistical analysis, Special, temporal, hierarchical clustering

1. INTRODUCTION

Data clustering is the process of dividing or detecting outliers in order to find a pattern, points or objects. Objects within a valid cluster are more similar to each other than objects outside the cluster. Webster (Merriam-Webster Online Dictionary, 2008) defines cluster analysis as "a statistical classification technique for determining whether individuals in a population fall into different groups by quantitatively comparing multiple characteristics". Clustering is an unsupervised technique that does not have pre-defined labeled data. Currently, many fields use many different kinds of clustering algorithms to separate datasets into groups and achieve quality results. Clustering methods differ in the choice of the objective function as well as the used distance matrix and the approach to the construction of the dissimilarity matrix. Clustering algorithms can be broadly categorized as: Hierarchical and Partitional. Other categories based on different datasets also emerged.

2. REVIEW OF PARTITION CLUSTERING METHODS

Partition clustering algorithms [7] partition the n objects into a set of k non-overlapping groups. K is an input parameter which gives that in how many clusters we want to partition. Partitioned clustering algorithms use an iterative approach to group the data into a k number of clusters by minimizing the objective function. In partitioning algorithm if we want to create 2 clusters then we have to identify 2 points, we call them seed point. Find out nearest seed point to all the points, for that we need to find the distance between a point and both seed points and assign it to the nearest seed point. In this process, the choice of seed point is an important consideration. Incorrect choice of seed points may give us incorrect solution. Ways to choose good seed points is by Choosing two points out of these given points only instead of some other points. Its advantage is that we will not have null clusters. One thing to be kept in mind while choosing seed points is that they should also be sufficiently far away from each other so that correct clusters are formed. The Algorithms studied in this category include: k-mean, k-modes, PAM, CLARA, CLARANS, Fuzzy-C-means, DBSCAN etc.K-means is an algorithm to group your objects based on attributes/features into K number of group. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. The objective function of k-means is given as:

$$E-\sum \mid \mid X_i - m_i \mid \mid ^2$$

In the above expression, E is sum of square error for all objects in the data, Xi is the point in a cluster C, mi the mean of cluster ki. The goal of K-means is to minimize the sum of the squared error over all K clusters. The algorithm states that initially, place k points into space represented by objects that need to be clustered as initial group centroids. In the second step, Assign each object to its closest cluster center. Then calculate mean of each cluster so as to have a new centroids Repeat these steps until there is no change in centroids. K-means was first proposed over 50 years ago, it is still one of the most widely used algorithms for clustering [2]. The basic K-means algorithm has been expanded and modified in many different ways. The modified versions of k-means are FORGY, ISODATA, CLUSTER, and WISH.

Fuzzy-C means(FCM) algorithm is also a centroid based clustering as k means and it requires number of clusters 'k' to be specified in advance. In this an object is assigned a membership value between 0 and 1 to indicate its belongingness to each cluster rather than assigning each object to its unique cluster. Application of FCM is in data mining, pattern recognition, classification, image segmentation. The other partitioning algorithms are: PAM, CLARA and CLARANS [9] are the algorithm which use the concept of choosing the seed points out of the given points.

PAM (Partitioning Around Medoids) was developed by Kaufman and Housseeuw. PAM's find k- clusters by first determing a representative object point for each cluster. This representative object, is the most centrally located object within the cluster called a medoid. After selecting the medoids, Average dissimilarity to all the non-selected object to the selected medoids is calculated. Then it is grouped to the medoid to which it is most likely to belong. Algorithm then iteratively moves and calculates a better medoid.

CLARA(Clustering LARge Applications) based on sampling [9]. In CLARA a sample of the data set is choosen from the entire data and then PAM is used to select medoids out of the sample. The concept is that if the sample is chosen in a fairly random manner, then its medoid will correctly represent the whole data set. The most important advantage of CLARA is that runs on multiple samples and gives the best clusters out of the given set of samples.

CLARANS is more efficient and effective than PAM and CLARA[8]. It runs efficiently on databases of thousands of object. It tries to combine PAM and CLARA. CLARANS(Clustering Large Applications based on RANdomized Search) is a improved k-medoid method. Disadvantage of CLARANS is that it assumes that all objects to be clustered can reside in the main memory at the same time, which is not possible every time if the database is large enough. This leads to decrease in run time of CLARANS on large databases. PAM, CLARA and CLARANS are based on distance from medoid. CLARA and CLARANS are applied to make PAM more scalable and effective.

3. REVIEW OF HIERARCHICAL CLUSTERING METHODS

Partitional clustering algorithms find all the clusters simultaneously as a partition of the data and do not impose a hierarchical structure. Hierarchical clustering algorithms find nested clusters. Types of Hierarchical clustering algorithms are:

1)Agglomerative mode- It is a bottom up method of clustering, we start with single data point as its own cluster and merging the most similar pair of clusters successively till a final cluster is obtained that has all the data points.

2) Divisive mode- It is top down clustering method, we start with all the data points contained as one cluster and recursively dividing each cluster into smaller clusters.

Basis of hierarchical clustering is that the solution is in hierarchy starting from 'n' groups to 1 group or vice versa. Initially each point itself is a group, we have a distance matrix between each point which is actually the distance matrix among the groups. Then we picked the distance which is the smallest and brought those two points together or those two groups together and formed a new group. Now find out the next group and the distance matrix is changed. Now find the distance between the group formed and all other points. There are several ways to compute this distance between a group and point. This equivalent distance can be done in more than one ways i.e. either to take minimum distance, average distance or maximum distance. If we choose single linkage clustering, we tend to choose minimum distance of the point. Average linkage clustering chooses average distance within the cluster to some other point outside the cluster. Complete linkage chooses the longest distance from any member of one cluster to any member of other cluster. Set of solution is given if we are given a way to find distance and given a way to relate group distance versus individual distance. A large number of hierarchical clustering algorithms exist in literature but they only differ in two ways, first the way similarity coefficient or distance measure is calculated and secondly, it may be single linkage, complete linkage or average linkage. So, we can say that there are nine different versions of clustering algorithm. Which includes three ways of calculating similarity co-efficient or distance (using Jacards coefficient, calculating similarity by number of components that visit both the machines, dissimilarity or distance matrix) and three ways of defining group distance versus individual distance(minimum, average, maximum).

The major drawback of hierarchical clustering is that once the two points are linked, they do not go to other group in a hierarchy or tree. There are few algorithms use hierarchical clustering with some variations. They are: BIRCH, CURE, ROCK, and CHAMELEON. BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies. It is based on the concept of Cluster Features (CF).CF is the triple (n, LS ,SS) where n is the number of data objects in the cluster, LS _ is the linear sum of the attribute values of the objects in the cluster and SS is the sum of squares of the attribute values of the objects in the cluster. These are kept in a tree called CF-tree. As it is stored in a tree form we do not need to keep whole tuples or whole clusters in main memory, but only, their tuples[5].

CURE (Clustering Using Representatives) [19] is a data clustering algorithm for large databases that is more robust to outliers and captures clusters of different shapes and sizes. It performs good on 2- dimensional data set. Its time complexity is O(n2 log n).BIRCH and CURE both handle outliers well. BIRCH has a better time complexity but is lacking in cluster quality than CURE algorithm.

ROCK was based on agglomerative hierarchical clustering algorithm for categorical data set. It is based on the number of links between two records; links capture the number of other records that the two are both sufficiently similar. This algorithm is not based on any distance function.

CHAMELEON[21] is also a hierarchical clustering algorithm. Two clusters are merged only if the interconnectivity and closeness (proximity) between two clusters are high relative to the internal interconnectivity of the clusters and closeness of items within the clusters.

4. OTHER CLUSTERING ALGORITHMS

A. Density-Based Clustering Method

This was introduced to discover clusters of arbitrary shape. It is based on the fact that within each cluster there is a typical density of points and this density is higher than outside the cluster. Outside points with lower density are recognized as noise points. One of the most commonly known algorithm in this category is, DBSCAN: Density Based Spatial Clustering of Applications with Noise [10]. DBSCAN requires only one input parameter and supports the user in determining an appropriate value for it. User needs to specify the radius and minimum number of objects it should have in its neighborhood. The key idea of density based clustering is that the numbers of data objects in the “neighborhood” are considered to determine density. For each object of a cluster the neighborhood of a given radius (r) has to contain at least a minimum number of objects, i.e. the cardinality of the neighborhood has to exceed a threshold. Regions of high density are defined in a separate cluster from the regions of no or low density. To find the neighborhood of a data point a spatial index is employed, which has improved the complexity of finding clusters by other methods from O(n^2) to O(nlogn). It doesnot perform well if the data is high dimensional and Euclidean distance is used to find proximity of objects.

DBCLASD[23] is also a density based algorithm, works on locality of points as DBSCAN but the difference is that it assumes that the objects inside of the clusters are randomly distributed and it does not need any input parameter.

OPTICS (Ordering Points To Identify the Clustering Structure) is an extension to *DBSCAN* but in this the requirement of the input parameters is not so strict. It creates an ordering of a database, additionally storing the core-distance and a suitable reachability-distance for each object. A clustering structure is created which defines a broad range of possible values and it automatically and interactively cluster the data. OPTICS computes an augmented cluster-ordering which has the information on a vivid variety of parameters as in the density based clustering [20]. DENCLUE (DENsity-based CLUstEring)[22] is anaggregate of partitioning and hierarchical clustering approaches. It is more effective than other approaches for the same input set. This algorithm works on arbitrary noise levels and on high-dimensional multimedia data sets on which other algorithms are not able to work. It performs much better than DBSCAN.

B. Grid-Based Clustering Method

Grid based clustering techniques include:

STING (STatistical INformation Grid approach) by Wang, Yang and Muntz (1997), it is one of the highly scalable algorithm and has the ability to decompose the data set into various levels of detail. STING retrieves spacial data and decomposes it into number of cells using rectangular hierarchical structure. Then mean, variance, minimum, maximum of each cell is computed .A grid structure is formed and new objects are inserted in the grid. It gives information about the spatial data by visiting appropriate cells at each level of the hierarchy.

WaveCluster by Sheikholeslami, Chatterjee, and Zhang (VLDB’98). Like STING it also uses multi-resolution approach (decompose the data set into different levels of hierarchy).

It uses a signal processing technique that decomposes a signal into different frequency sub-band (can be applied to n-dimensional signals). Data are transformed to preserve relative distance between objects at differentlevels of resolution. It allows natural clusters to become more distinguishable. It is highly scalable and can handle outliers well. It is not suitable for high dimensional data set [5].It can be considered as both grid-based and density based.

CLIQUE: It is developed by Agrawal, et al. (SIGMOD’98). It can be used to cluster high-dimensional data. CLIQUE can be considered as both density-based and grid-based. It partitions each dimension into the same number of equal length interval. It partitions an m dimensional data space into non-overlapping rectangular units.

A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter. A cluster is a maximal set of connected dense units within a subspace.

C. Model Based Clustering

In model based clustering optimization attempts to the fit between the given data to some mathematical model. It is based on the assumption that data are generated by a mixture of underlying probability distribution. It includes: EM (Expectation maximization) algorithm:

A popular iterative refinement algorithm. It is an extension to k-means. It assigns each object to a cluster according to a weight(probabilistic distribution) and the new means are computed based on weighted measures.

COBWEB: It is developed by Fisher in 1987. It is a popular and a simple method of incremental conceptual learning. It creates a hierarchical clustering in the form of a classification tree. Each node refers to a concept and contains a probabilistic description of that concept.

It automatically adjusts the number of classes in a partition. It does not rely on the user to provide such an input parameter. CLASSIT: It is an extension of COBWEB for incremental clustering of continuous data. It also suffers from similar problems as COBWEB.

Auto Class: It is developed by Cheeseman and Stutz in 1996. It uses Bayesian statistical analysis to estimate the number of clusters. It is very popular in industry.

SOM (Soft-Organizing feature Map): SOMs, also called topological ordered maps, or Kohonen Self Organizing Feature Map (KSOMs).It maps all the points in a high- dimensional source space into a 2 to 3-d target space, such that, the distance and proximity relationship (i.e., topology) are preserved as much as possible. It is Similar to k-means where cluster centers tend to lie in a low-dimensional manifold in the feature space. Here, clustering is performed by having several units competing for the current object. The unit whose weight vector is closest to the current object wins. The winner and its neighbors learn by having their weights adjusted.SOMs are believed to resemble processing that can occur in the brain. It is useful for visualizing high-dimensional data in 2- or 3D space.

D. Categorical data clustering

Most of the Clustering algorithms discussed so far are oriented towards numerical data set and they use Euclidean distance measures. For, categorical data we need to develop a different strategy. Here these distance measures cannot be used. A categorical variable is one for which the measurement scale consists of a set of categories. Example: liberal, moderate, conservative, they belong to Political philosophy category. Hot, cold belong to Choice of breakfast cereal. Algorithms implemented in literature here include:

K-prototype: It similar to K-means algorithm. Only difference is that here a different dissimilarity measure is used, mean is replaced by modes and a frequency based method is used to update modes. The algorithm requires a linear number of in- memory operations and thus can be used for large inputs. K- prototype is an integration of k- means and k-modes.

ROCK (RObust Clustering using linKs) [17] is a hierarchical algorithm for categorical data.

STIRR(Sieving through Iterated Relational Reinforcement): It uses an iterative approach where the values, rather than the tuples, are the data objects to be clustered[24]. COOLCAT algorithm: It was given by Barbara, Couto and Li is an information-theoretic algorithm most similar to K- means algorithm. To determine the similarity of objects it uses its entropy. Given the number of clusters to be produced, the objective of the algorithm is to partition a data set such that the entropy of the resulting clustering is minimized, or equivalently, the values within clusters can be predicted with maximum certainty [24].

CACTUS(Clustering Categorical data using Summaries algorithm)[7]: It is an agglomerative algorithm for categorical data.

LIMBO: It is hierarchical clustering algorithm handling categorical data values. It can be used when the objects to be clustered are either the tuples of a data set or the values of one of its attributes. LIMBO can cluster data of various sizes. The size of the model builds to summarize the data can be controlled to match the space available for use. It is the most scalable categorical clustering algorithm available till date [24].

5. STATISTICAL ANALYSIS

In accordance with the literature survey done, the charts and diagram below can show a comparative study. Different algorithms which have been studied have been described according to their categories:

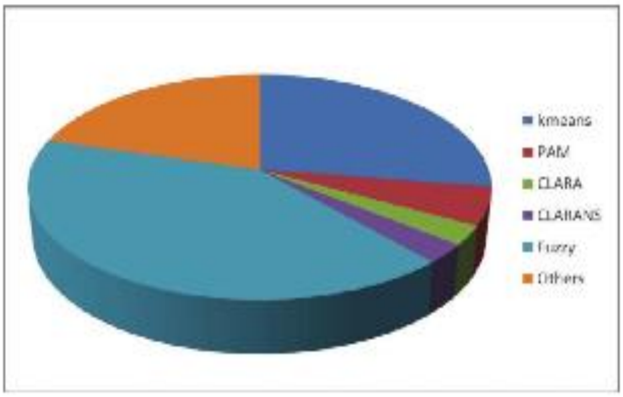
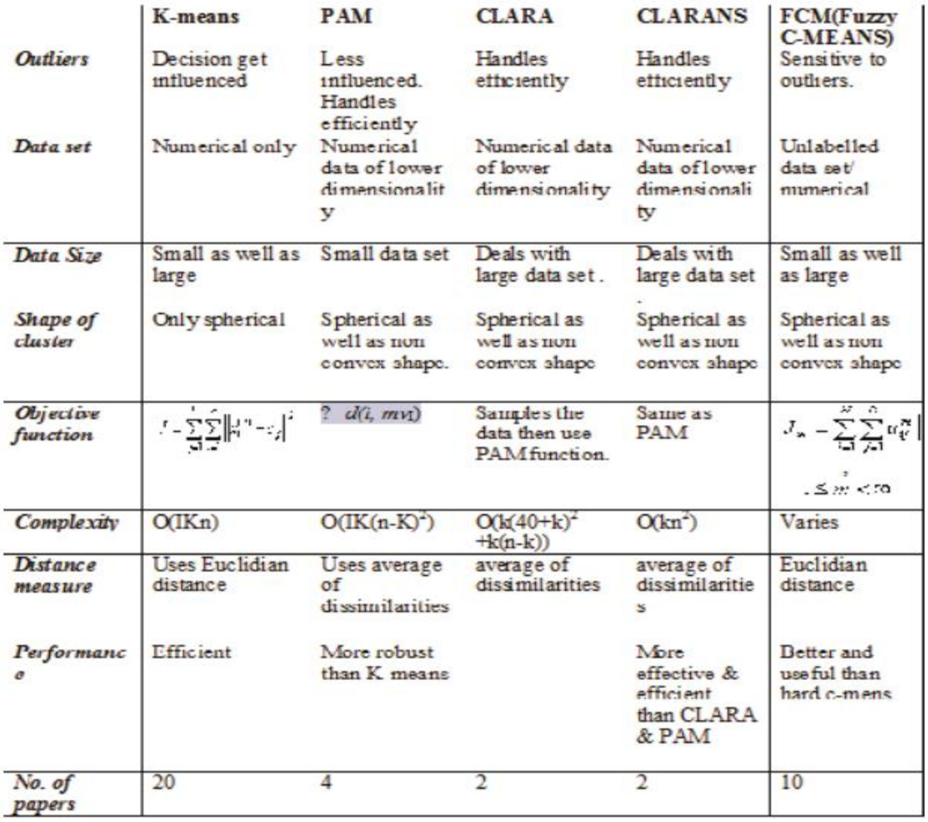


Figure 1 : Above shows the number of papers available and studied

Table1. Partition Clustering Algorithms



	K-means	PAM	CLARA	CLARANS	FCM(Fuzzy C-MEANS)
Outliers	Decision get influenced	Less influenced. Handles efficiently	Handles efficiently	Handles efficiently	Sensitive to outliers.
Data set	Numerical only	Numerical data of lower dimensionality	Numerical data of lower dimensionality	Numerical data of lower dimensionality	Unlabelled data set/ numerical
Data Size	Small as well as large	Small data set	Deals with large data set .	Deals with large data set .	Small as well as large
Shape of cluster	Only spherical	Spherical as well as non convex shape.	Spherical as well as non convex shape	Spherical as well as non convex shape	Spherical as well as non convex shape
Objective function	$J=\sum_{j=1}^{k}\sum_{i=1}^{n}\|x_i^{(j)}-c_j\|^2$	? d(i, mvi)	Samples the data then use PAM function.	Same as PAM	$J_m=\sum_{i=1}^{N}\sum_{j=1}^{C}u_{ij}^m\|$ $1\leq m<\infty$
Complexity	O(IKn)	$O(IK(n-K)^2)$	$O(k(40+k)^2 +k(n-k))$	$O(kn^2)$	Varies
Distance measure	Uses Euclidian distance	Uses average of dissimilarities	average of dissimilarities	average of dissimilarities	Euclidian distance
Performance	Efficient	More robust than K means		More effective & efficient than CLARA & PAM	Better and useful than hard c-mens
No. of papers	20	4	2	2	10

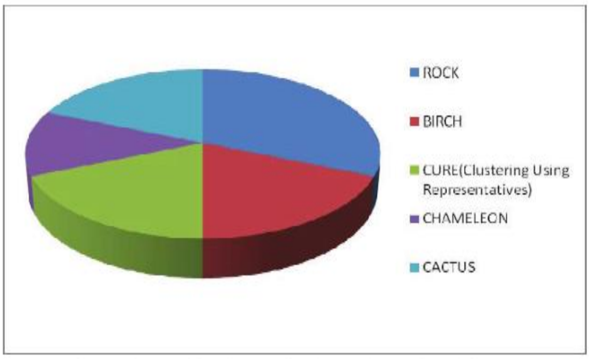
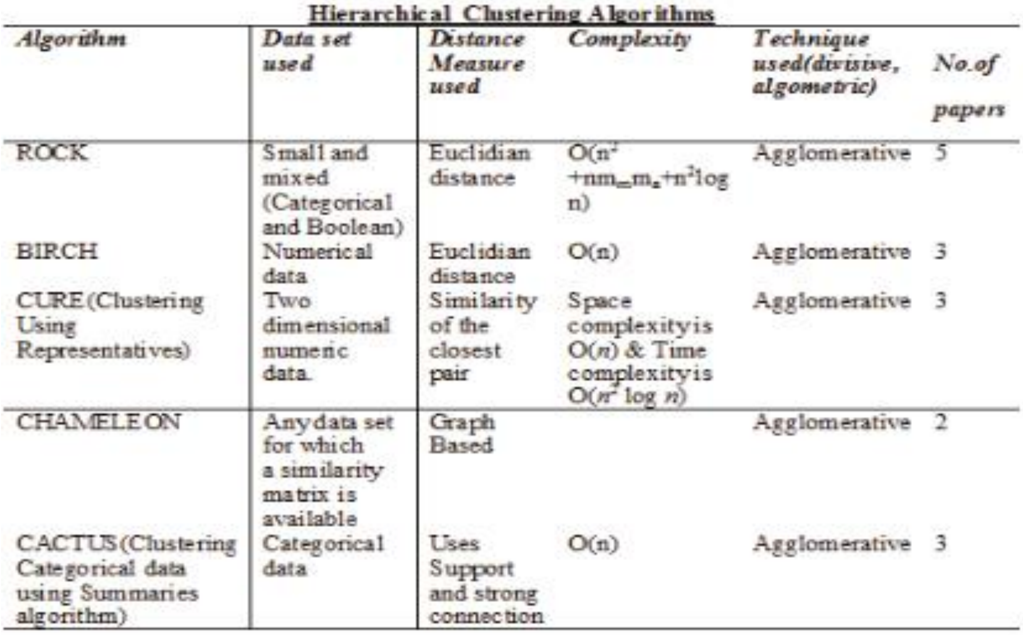


Figure 2: Hierarchical Based Clustering Algorithms

Table 2. Hierarchical Clustering Algorithms



Hierarchical Clustering Algorithms

Algorithm	*Data set used*	*Distance Measure used*	*Complexity*	*Technique used(divisive, algometric)*	*No.of papers*
ROCK	Small and mixed (Categorical and Boolean)	Euclidian distance	$O(n^2 + nm_m m_a + n^2 \log n)$	Agglomerative	5
BIRCH	Numerical data	Euclidian distance	O(n)	Agglomerative	3
CURE (Clustering Using Representatives)	Two dimensional numeric data.	Similarity of the closest pair	Space complexityis $O(n)$ & Time complexityis $O(n^2 \log n)$	Agglomerative	3
CHAMELEON	Anydata set for which a similarity matrix is available	Graph Based		Agglomerative	2
CACTUS (Clustering Categorical data using Summaries algorithm)	Categorical data	Uses Support and strong connection	O(n)	Agglomerative	3

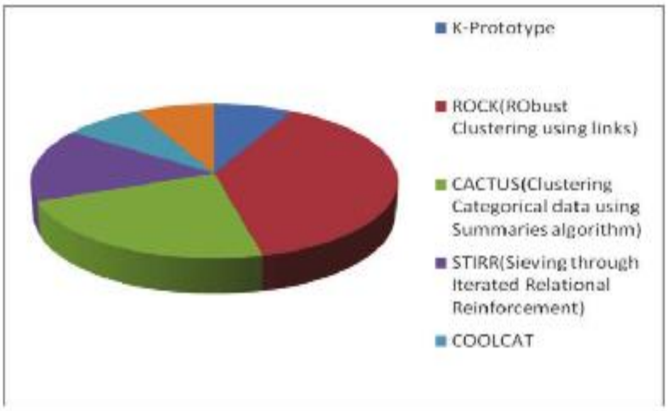
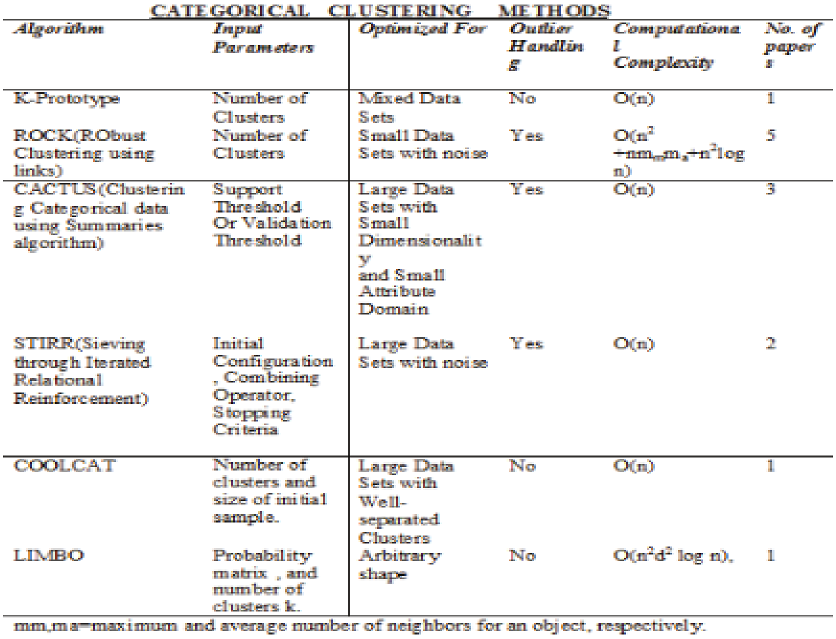


Figure 3: Categorical Algorithms

Table 3. Categorical Clustering Algorithms



CATEGORICAL CLUSTERING METHODS

Algorithm	*Input Parameters*	*Optimized For*	*Outlier Handling*	*Computational Complexity*	*No. of papers*
K-Prototype	Number of Clusters	Mixed Data Sets	No	O(n)	1
ROCK(RObust Clustering using links)	Number of Clusters	Small Data Sets with noise	Yes	$O(n^2 + nm_m m_a + n^2 \log n)$	5
CACTUS(Clustering Categorical data using Summaries algorithm)	Support Threshold Or Validation Threshold	Large Data Sets with Small Dimensionality and Small Attribute Domain	Yes	O(n)	3
STIRR(Sieving through Iterated Relational Reinforcement)	Initial Configuration, Combining Operator, Stopping Criteria	Large Data Sets with noise	Yes	O(n)	2
COOLCAT	Number of clusters and size of initial sample.	Large Data Sets with Well-separated Clusters	No	O(n)	1
LIMBO	Probability matrix , and number of clusters k.	Arbitrary shape	No	$O(n^2d^2 \log n)$,	1

mm,ma=maximum and average number of neighbors for an object, respectively.

Table 4. Comparison Of Density Based Methods

DENSITY BASED CLUSTERING METHODS

	Data set	*Complexity*	*Can handle outliers*	*Shape of cluster*	*No. of papers*
OPTICS(Ordering points To Identify Clustering Structure)	Numerical data set	O(nlogn)	Yes	Arbitrary	3
DBSCAN	Numerical	O(nlogn)	Yes	Arbitrary	3
DBCLASD	Numerical		Yes	Arbitrary	1
DenClue (DENsity-based CLUstEring)	High dimensional multimedia data set.	O(nlogn)	Yes	Arbitrary	2
CLIQUE(Clustering in QUEst)	High dimensional data	O(n)	Yes	Arbitrary	2

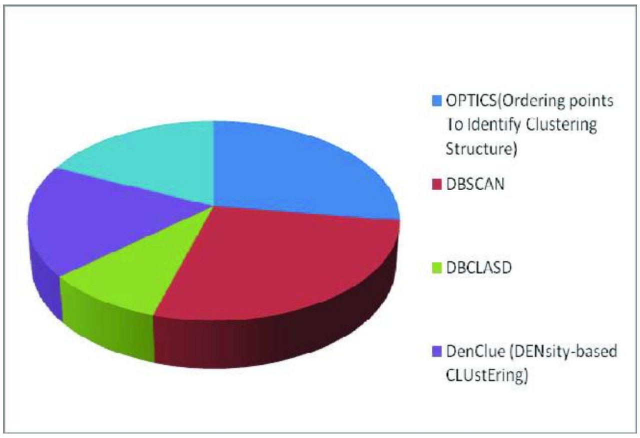


Figure 4: Density Based Algorithms

Table 5. Grid Based Clustering Methods

GRID BASED CLUSTERING METHODS

	Data set	*Complexity*	*Shape of cluster*	*Outlier Handling*	*No. of papers*
STING (STatistical INformation Grid)	Large spatial dataset	O(n)	Vertical and Horizontal boundaries	Yes	2
CLIQUE(CLusterin in QUEst)	High dimensional data	O(n)	Arbitrary	Yes	2
Wave Cluster	Only applicable to low dimensional data.	O(n)	Arbitrary shape	Yes	1

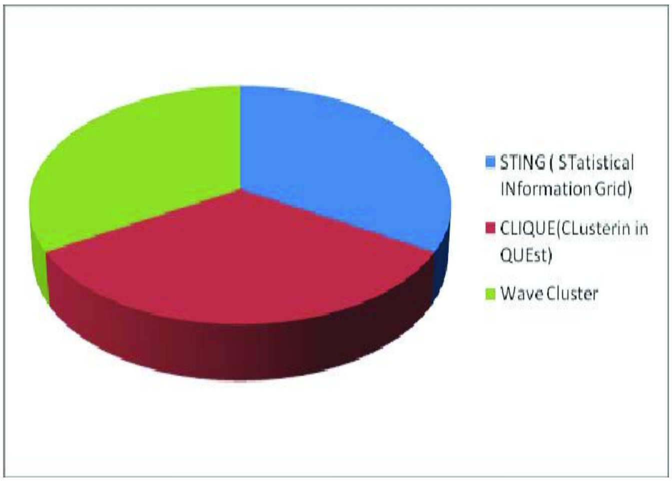
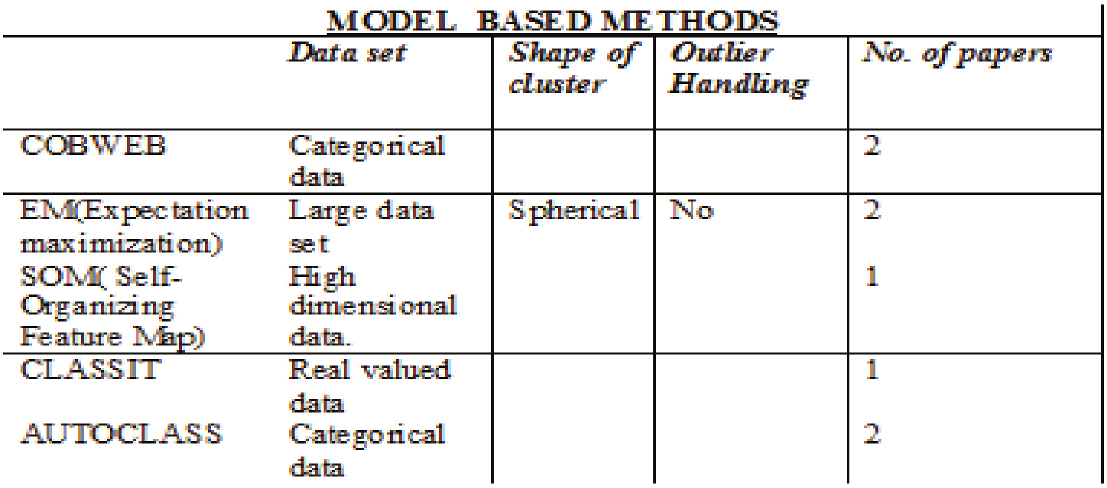


Figure 5. Grid Based Algorithms

Table 6. Model Based Clustering



MODEL BASED METHODS

	Data set	*Shape of cluster*	*Outlier Handling*	*No. of papers*
COBWEB	Categorical data			2
EM(Expectation maximization)	Large data set	Spherical	No	2
SOM(Self-Organizing Feature Map)	High dimensional data.			1
CLASSIT	Real valued data			1
AUTOCLASS	Categorical data			2

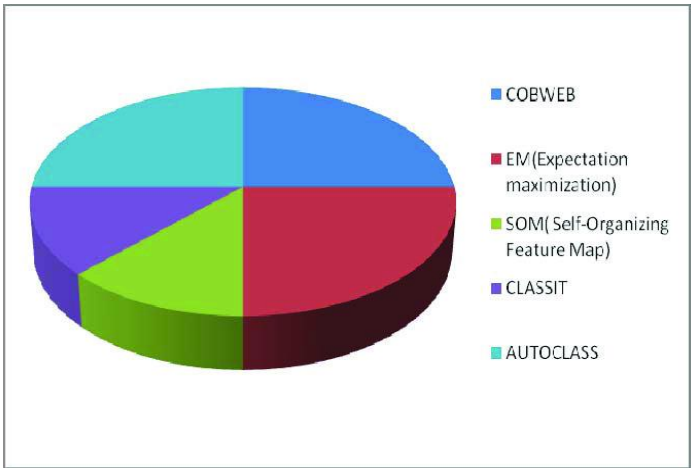


Figure 6: Model Based Algorithms

6. CONCLUSION

Clustering is the identification of similarities between data sets. Applications include data mining, pattern recognition, web document characterization, clustering of genes and proteins with similar functions, clustering of geographic data such as earthquakes, etc. This article discusses different clustering methods along with comparative studies. We can make an assumption that the most widely used and studied algorithm is the k-means algorithm. There are many variants of the algorithm. Most of the work done in this area of clustering focuses on data sets, and only a few methods exist for categorical and other databases.

7. REFERENCES

[1] Cominetti, Matzavinos, Samarasinghe, “DifFUZZY: A fuzzy clustering algorithm for complex data sets”, International Journal of computational intelligence in bioinformatics and system biology, 2010.

[2] Anil K. Jain, “Data clustering: 50 years beyond Kmeans”, 19th International conference in Pattern Recognition, 2009.

[3] Thair Nu Phyu, “Survey of Classification Techniques in Data Mining”, IMECS 2009, Hong Kong.

[4] I.K..Ravichandra Rao,“Data Mining and Clustering Techniques”, DRTC Workshop on Semantic Web December, 2003

[5] Periklis Andritsos, “Data Clustering Techniques”, University of Toronto, 2002

[6] Hung, Yang, “An efficient fuzzy C-means clustering algorithm”, IEEE 2001.

[7] Amir Ahmad, Lipika Dey, “A k-mean clustering algorithm for mixed numeric and categorical data”, Science Direct Data & Knowledge Engineering 63 (2007) 503–527.

[8] L. Kaufman, P.J. Rousseew, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, 1990.

[9] R. Ng, J. Han, Efficient and effective clustering method for spatial data mining, in: Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, 1994, pp. 144–155.

[10] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of KDD’96, 1996.

[11] Tangsen Zhan,Yuanguo Zhou, Clustering Algorithm on High-dimension Data Partitional Mended Attribute, 9th International Conference on Fuzzy Systems and Knowledge Discovery, IEEE 2012

[12] [12] S. J. Nanda, G. Panda, Accurate Partitional Clustering Algorithm Based on Immunized PSO,IEEE March 30, 31, 2012

[13] Shi-xia Ma, Jian-hua Liu,Dan Liu, Research of Case Retrieval Strategy Based on Partitional Clustering Algorithm, 2010 IEEE.

[14] JUN-HAO ZHANG, MING-HU HA*, JING WU, Implementation Of Rough Fuzzy K-Means Clustering Algorithm In Matlab, Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010.

[15] UCI repository of machine learning databases. [online]. Available http://archive.ics.uci.edu/ml/datasets.html, 2010.

[16] G. S. Thakur, R. S. Thakur and Ravi Singh Thakur, “2-Level Clustering Framework for Time Series Data Sets”, (SOCPROS 2011) DECEMBER 20-22, 2011.

[17] S. Guha, R. Rastogi, S. Kyuseok, “ROCK: A robust clustering algorithm for categorical attributes”, in: Proceedings of 15th International Conference on Data Engineering, Sydney, Australia, 23–26 March 1999, pp.512–521.

[18] Frank hoppner, what is fuzzy about fuzzy clustering Understanding and improving the concept of the fuzzifier.

[19] S. Guha, R. Rastogi, S. Kyuseok, “CURE An efficient clustering algorithm for large databases”,1998 ACM.

[20] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander, “OPTICS: Ordering Points to Identify the Clustering Structure”, In Proceedings of the International Conference on Management of Data, (SIGMOD), June 1996. ACM Press