International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

A COMPARATIVE STUDY OF CLUSTERING ALGORITHMS IN MACHINE LEARNING

R Arunachalam[1], M Jayanth Kumar[2], Dr D Beulah David[3]

[1]UG Student, Department of Computer Science and Engineering, Jeppiaar Engineering College,
[2]UG Student, Department of Computer Scienceand Engineering, Jeppiaar Engineering College,
[3]Assistant Professor, Department of Computer Science and Engineering, Jeppiaar Engineering College,

---

Abstract - *Clustering or Cluster Analysis is a data mining technique or machine language algorithm which is used to divide the objects into groups according to the similarity between them. These groups are also called clusters which are basically used in statistical data analysis, data mining, and machine learning. It is an unsupervised algorithm which can be used to describe general behavioural patterns or working behaviour. It also faces many challenges such as computational complexity, noise, large dimensions of the data, Scalability etc. Various algorithms have been proposed till date to address the issue but no one algorithm can tackle all on its own, which poses a challenge of selecting an ideal algorithm for our problem. In this paper we are going to give a comparison between well known numerical clustering algorithms out there, which can be used as a guide for selecting the ideal algorithm to suit one's purpose.*

Keywords: Clustering algorithms, Partitioning Method, Hierarchical Method, Numerical Data, Number of Clusters.

1. INTRODUCTION

Clustering or Cluster Analysis is a data mining technique or machine language algorithm which is used to divide the objects into groups according to the similarity between them. The result of the cluster analysis is to get a different group of objects called clusters. Objects within a cluster are similar to each other and dissimilar to other objects in another cluster [1, 2] . This similarity and distinct measures between clusters is determined using Euclidean Distance, Cosine Similarity, Manhattan Distance and Jaccard similarity[3] based on the data objects the problem involves with. Clustering is used in many fields of study which includes statistics, data mining, Machine Learning, Artificial Intelligence etc.

Clustering is not an easy task as classification but a difficult one. The problems which are posed such as High dimension of the data, computational complexity, noise present in the data, scalability etc. Huge number of clustering algorithms have been proposed till date to meet some specific requirements, but no one algorithm can handle or tackle all the problems. Which makes it difficult to choose a single algorithm for a given problem or a specific task. In this paper we are going to give a comparison between well known numerical clustering algorithms out there, which can be used as a guide for selecting the ideal algorithm to suit one's purpose.[4][5]

2. TYPES OF CLUSTERING METHODS

Generally, all algorithms can be categorised into two broad categories : 1. Partitioning and 2. Hierarchical based on the properties of the clusters.Various algorithms which have been proposed may follow different methodology hence it is difficult to categorie them within these two categories.Detailed categorisation of the clustering algorithms is given in [1 , 2]. Following section provides a brief view of some very well known methods.

2.1. PARTITIONING METHOD

As the name suggests, this method divides the dataset into **K** partitions containing **N** objects, where K is always lesser than or equal to the number of objects, data is divided into partitions on some evaluation criteria.Partition can be made by various approaches but checking all of them is not feasible. Hence one of the greedy methods is used.

2.1.1 K-means

The partition algorithm in which each cluster is represented by the magnitude of their centres is known as K-means algorithm. It is one of the simplest unsupervised clustering algorithms. In this algorithm,

the number of clusters for which the dataset to be grouped is to be known in advance. K-means is an iterative algorithm where it tries to divide the objects into K partitions given that K is lesser than or equal to the N number of objects in the dataset. The basic algorithm is very simple :-

1. Assign all the objects associated with the closest centroid.
2. Compute the new centroid by calculating the mean of all the objects associated with that particular centroid.
3. Repeat step 1 and step 2 until there is no change.

Drawback of the K-means algorithms is with the smaller samples of the datasets the algorithm fails to cluster data accurately. It also fails for categorical data. Also if there exists datas which are highly overlapping then the algorithm fails to resolve them into clusters distinctly.

Some examples under the K-means method are :- K-means , bisecting K-means method.

2.1.2. K-medoids

The partitioning algorithm in which the cluster is represented by one of the objects located near its centre is called k-medoids. Medoids are defined by the location of predominant fraction of points inside the cluster which makes it less sensitive to the presence of outliers in the data. The clusters are defined as a subset of points close to their respective medoids and the objective function is defined as the average distance between the medoids and the points or any other similarity function is used based on the type of dataset.

K-medoids method is not suitable for clustering non-spherical object groups because this method relies on minimizing the distance between the non medoids objects and mediod of the cluster. It groups based on the closeness rather than the connectivity between objects.

Some of the most common algorithms under K-medoids method are : CLARA (Clustering LARge Applications) , CLARANS (Clustering Large Applications using RANdomised Search) and PAM (partitioning Around Medoids)

2.2. HIERARCHICAL METHOD

In this method, it tries to group the object by decomposing the dataset into a group hierarchy. The decomposition is done by constructing a tree of data objects by their similarities. Hierarchical method iteratively merges clusters either into a single cluster or to different individual nodes. Dendrogram is used as a diagrammatic representation of the connection between the data objects and is used to determine the number of clusters which is done by cutting the dendrogram at a similarity measure [7]. Distances between the data objects are calculated using Euclidean distance or Manhattan distance. The distances can be calculated using single, average or complete linkage. When the distance between two points in a cluster is defined by the shortest distance is single linkage. When the distance is greatest between the two points in a cluster is called complete linkage. Similarly when the distance between two clusters is defined as the average distance between each point in one cluster to every point in another cluster is average linkage [8]. Single linkage suffers from chaining while the complete linkage suffers from crowding. As to what kind of linkage to be chosen should be entirely based on the type and complexity of data objects.

2.2.1. Agglomerative Hierarchical method

The Bottom up method is called agglomerative hierarchical clustering method. Each data point is considered as a single cluster and it merges the clusters based on the similarity between individual clusters until all the data objects are constructed under a single cluster. For N data objects , N * N distance matrix is created, then the basic algorithm for the method can be described as follows :-

1. Initialise the method with N clusters, where each individual object is considered as a cluster. Then calculate the initial similarities between the cluster and store it in the matrix.
2. Determine a pair of clusters based on the distance between them, unify two clusters which have minimum distance. Now the unified pair of clusters is considered a single cluster.
3. Calculate the similarities between the cluster which was newly formed to the priory available clusters and store it into the matrix.
4. Repeat steps 2 and 3 until all data objects are merged into a single cluster consisting of N data objects. [9]

Complexity of the Agglomerative method is $O(n^3)$ in general, but we can reduce the complexity to $O(n^2 \log n)$ if we use priority Queue.[10]

2.2.2 Divisive hierarchical method

In this method hierarchical trees are constructed using a top down approach for the data

points. All the data objects are considered as a single cluster initially and then the clusters are divided until all the data objects are considered as a single individual cluster. This method is much more complex than the Agglomerative method because a second method is required for dividing the clusters. Since there are $2^{N-1} - 1$ ways to divide a group of N clusters into individual clusters, it is computationally not feasible for finding the optimal solution for dividing the clusters, hence one of the heuristic approaches is used. In most of the cases K-means or bisecting K-means algorithms are used for dividing the clusters. Another one of the strategies which can be used is by constructing a dissimilarity graph using a minimum spanning tree and then by making a new cluster by breaking down the connections between the largest dissimilarity.

Algorithm of Divisive hierarchical method is as follows :-

1. Intialise the process by making a single cluster containing all the objects.
2. Choose a cluster with a largest diameter.
3. Detect the objects in the cluster chosen in step 2 with the minimum average similarity to all the other objects in the same cluster.
4. The object which was detected in step 3 is the element which will be added to the fragment group.
5. Detect the object in the original group which shows the highest average similarity with the objects in the fragment group.
6. If the average similarity for the detected object in step 5 is greater for the fragment group than the original group then assign the object to the fragment group and go to step 5 ; otherwise do nothing.
7. Repeat step 2 - step 6 until each data point is considered as a single cluster. [12]

Complexity of this method with exhaustive search is O(2^N) but various heuristics used in the second method result in varying complexity.

Hierarchical methods can suffer from noise and outliers present in the objects, it also struggles in handling different sized clusters and convex shapes of the objects.

Some of the examples of Hierarchical methods are :- BIRCH, ROCK, CURE, Chameleon, AGNES, DIANA etc.

2.3. GRID BASED METHOD

In this method, the object space is divided into a finite number of cells forming a grid structure where clustering operations for the objects is performed. One of the main advantages in using this method is rapid processing time due to the reason that this method does not depend on the number of objects but the number of cells in the grid structure. Grid based method requires a large number of parameters but this method has fast processing time which depends on the number of cells in each dimension in the quantised space.[1]

Some of the algorithms based on this method are :- STING, CLIQUE and WaveCluster.

2.4. MODEL BASED METHOD

In this method, rather than forming clusters based on data objects but are formed by providing a probability of each data object belonging to a particular cluster. Moreover there is an additional advantage of using this method is that the number of clusters need not be determined beforehand because this algorithm automatically identifies it itself. The typical method found in the literature can be classified in three groups; statistical approach, machine learning approach and neural network approach. The main drawback for using this method is that it requires large data sets or it becomes computationally not feasible when the data sets contain very few observed data objects. Also it can converge to local optimal points which can be overcome with running the algorithm multiple times with random initialization.[1]

EM or (Expectation - Maximization) is one of the algorithms proposed under this method.

2.5. DENSITY BASED METHOD

This method based on the density of the objects in the data, the algorithm tries to group the data objects into clusters based on the idea that a group of similar objects or objects belonging to the same cluster will have a contiguous regions of high point density which are separated from other groups of objects or cluster by contiguous regions of low density. A value has to be defined beforehand known as the threshold value or 'ϵ' which helps in determining the neighbourhood containing the objects.

The algorithm also requires a value which represents the minimum number of data objects we want in the neighbourhood of radius 'ϵ' known as minPts.

Using these values we provided, the objects are divided into three categories :-

Core Points : - A object is considered to be a Core Point if the ϵ - neighbourhood of the this particular object contains minimum number of points or minPts.

Border Points : - A object is considered to be a Border Point if the ϵ - neighbourhood of this object doesn't contain minimum number of points or minPts but within ϵ - neighbourhood of some other Core Point.

Outlier : - If an object doesn't come under Core or Border points, then this object is considered as an outlier.

Density Based method struggles in handling high dimensional data objects or when the density between the clusters are varying. Also this method is highly sensitive to the parameters provided and fails to perform when the data is too sparse.[13]
Some algorithms under this method are : DBSCAN, OPTICS, DENCLUE

3. METHODS TO DETECT OPTIMAL NUMBER OF CLUSTERS

3.1. ELBOW METHOD

The basic idea of this method is to minimise the variation within the clusters which is done by using the square of distance between the sample points in each cluster and the centroid of the cluster to give various K values. The steps is as follows:-

1. Compute clustering algorithm for different values of k, for instance starting from 1 to 15.
2. For each K calculate the total variation within a cluster using Eq (1).

$$\sum_{K=1} KW(C_K) \quad (1)$$

Where C_K is the Kth cluster and $W(C_K)$ is the variation within the cluster.
3. Then the variation is plotted against each k values.
4. The location of a bend which resembles an elbow
bend is generally considered as an indicator of the ideal number of clusters for the data objects as shown in Fig.1.

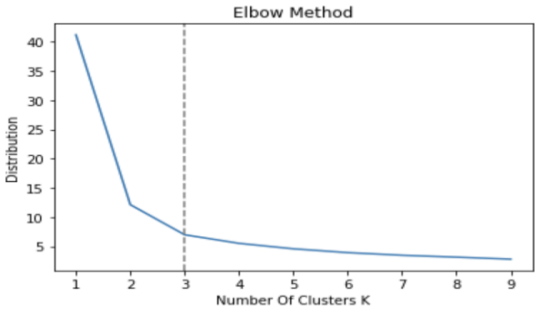

Fig.1. Graphical representation of the results of the Elbow method.

3.2. THE GAP STATISTIC ALGORITHM

This algorithm was proposed by Tibshirani [15] to determine the number of clusters of datasets with unknown classification numbers. It uses the Monte Carlo method to introduce reference measurements and to calculate the distance between two measurements in each class. The clustering results of the constructed reference measurements are compared to determine the number of clusters in the data set.
The formula is as in Eq (2):-

$$Gap_n(k) = E_n^*(\log(W_k)) - \log W_k E_n^*(\log(W_k))$$

$$= (1/P)\sum_{b=1}^{P} \log(W^*_{kb}) \approx (1/P)\sum_{b=1}^{P} \log(W^*_{kb})s(k)$$

$$= \text{sqrt}((1+P)/P)s(k) \quad (2)$$

where $E_n^*(\log(W_k))$ refers to $\log(W_k)$ expectations. In order to get the average, first you will get an approximate $E_n^*(\log(W_k))$ value. P is the number of samplings, s(k) is the standard of joining, and finally Gap_k can be calculated. The K value corresponding to the maximum value of Gap_k is the best k; that is, it satisfies the minimum k of $Gap_k \geq Gap_{k+1} - S_{k+1}$ as the optimal number of clusters which can be found in Fig.2.

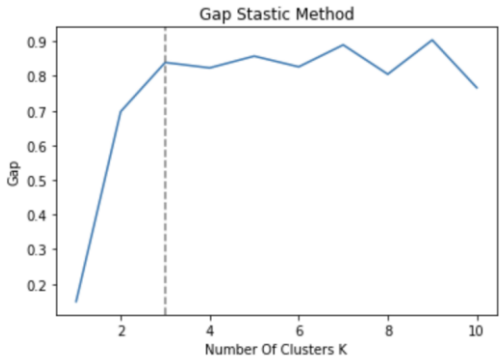

Fig.2. Graphical representation of variation of Gap value with different values of K. As shown when K=3 is when the optimal number of clusters is obtained.

3.3. Silhouette Coefficient Method

This method measures the quality of clustering. It defines how well each object lies within the clusters. A higher average silhouette width indicates an ideal clustering approach. Average silhouette width for different clusters are observed and ideal number of clusters is identified by choosing the cluster which has higher silhouette width. The graphical representation is given in Fig.3.

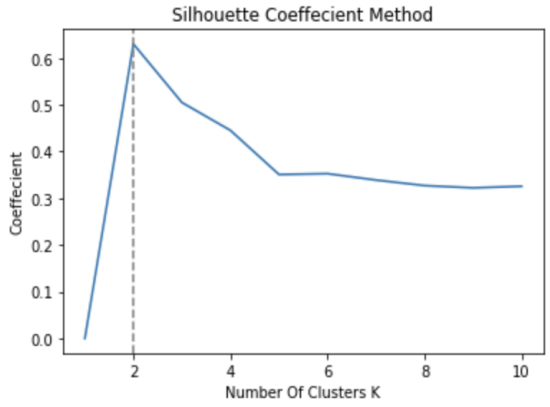

Fig.3. Graphical representation of Silhouette coefficient method.

As you can see the silhouette is greater for n = 2 clusters, hence we can take that as the ideal number of clusters for the given problem.

4. COMPARATIVE ANALYSIS

As mentioned before Clustering is a more difficult task than one of classification. The problems which are posed such as High dimension of the data, computational complexity, noise present in the data, scalability etc. Numerous algorithms have been proposed to overcome the problems but no one algorithm can tackle all on its own. First, detailed properties of some of the popular algorithms have been given in Table 1. Then we have taken 5 datasets with numerical values, where classification needed is taken: Iris plant Classification, Glass identification images, Image segmentation, Parkinson's disease detection and Breast cancer detection datasets.

All the datasets were taken from **UCI Machine Learning Repository**.

We have referenced various research papers and made a table consisting of the algorithm name, Time Complexity of the algorithm, whether the algorithm is capable of handling outliers, and what are the input parameters required for the algorithm when the algorithm was first proposed, which can help the users in selecting the Clustering algorithm. [1],[2],[6],[9],[10]

The information can be seen in Table.1. named description of the clustering algorithms below.

Table.1. Description of the clustering algorithms.

Algorithm Name	Time Compl - exity	Outlier	Input Parameter	Proposed year
K-Means	O(n)	No	Number of clusters	1955
CLARANS	$O(n^2)$	Yes	Number of clusters, Number of Local Minima, Max number of neighbour	1994
EM	O(n)	No	Number of components	1977
PAM	$O(n^2)$	Yes	Number of clusters	1990
Agglomerat-	O(n2	No	Number of	1963

ive	logn)		clusters, Linkage	
BIRCH	O(n)	Yes	Branching factor, Threshold	1996
DBSCAN	$O(n^2)$	Yes	ϵ - Radius, Minimum points	1996

Now we are going to see how these algorithms fare against the structured datasets.

Description of the datasets :-

Dataset # 1 - Iris Plant Classification Dataset containing 150 instances and 4 attributes.

Dataset # 2 - Glass Identification Dataset containing 213 instances and 10 attributes .

Dataset # 3 - Wisconsin Breast Cancer Dataset containing 569 instances and 32 attributes.

Dataset # 4 - Image Segmentation Dataset containing 2310 instances and 19 attributes.

Dataset # 5 - Parkinson's Disease Classification Dataset containing 756 instances and 754 attributes.

We calculated the accuracy for each algorithm on each particular dataset. We have tabulated the data so that this might give ideas for users to choose appropriate algorithms for their problem.

Table.2. Performance of the algorithm against the 5 datasets.

Method	**Dataset #1**	**Dataset #2**	**Dataset #3**	**Dataset #4**	**Dataset #5**
K-means	88.67	48.07	92.79	53.28	**73.94**
CLARANS	92.67	44.85	88.41	46.19	69.21
EM	**96.67**	48.60	**94.20**	**55.71**	62.31
PAM	90.02	36.78	93.49	48.06	55.42
Agglomera-tive	88.67	42.06	86.81	43.33	44.57
BIRCH	86.67	**50.32**	84.18	51.90	56.61
DBSCAN	66.72	29.47	71.35	42.62	40.21

As you can see K-means, PAM and CLARANS have fared good in almost all the dataset's. EM algorithm was most accurate in Dataset 1,3 and 4 as the EM algorithm is based on the model of the dataset. We can further improve the accuracy score by cleaning the data.

5. CONCLUSIONS

Clustering is not an easy task as classification but a difficult one. Choosing the right algorithm for the given problem plays a huge role in solving the problem. We have found that there is no one algorithm which can satisfy all the problems single handedly, hence we have given a comparison of performance between some of the popular algorithms out there on 5 different types of datasets to highlight each of the algorithm's performance in hope this help makes the selection process easier. The points we gathered from this experiment is as follows : -

- K-means, BIRCH algorithm performs consistently for all the 5 different datasets we have taken.
- The DBSCAN and Agglomerative algorithm fares poorly in Datasets 2,3 and 4 where the datasets were sparse and had higher variance than other two datasets.
- The EM algorithm performed well but struggled when the number of attributes became large.
- The PAM and CLARANS algorithm also performed consistently on almost all the datasets except Dataset 2 which contained a non-spherical group of objects.

The efficiency for the problems can be further improved by using multiple clustering combination approaches based on an iterative voting process.

6. REFERENCES

[1] Anindya Bhattacharaya, Nirmalya Chowdhury and Rajat K De," comparative analysis of clustering and biclustering algorithms for grouping of genes " Member at IEEE DOI : 10.2174/15748931279930440

[2] Han J, Kamber M. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann 2001.

[3] A. Alzu'bi, A. Amira, and N. Ramzan, "Semantic content-based image retrieval: A comprehensive study," Journal of Visual Communication and Image Representation, vol. 32, pp. 20-54, 2015.

[4] S. cha, "Comprehensive survey on distance/similarity measures between probability density functions" International Journal of Mathematical Models and Methods in Applied Sciences , 2007 .

[5] Leonardo N. Ferreira , A.R.Pinto and Liang Zhao , " QK-Means : A Clustering Technique based on Community Detection and K- Means for Deployment of Cluster Head Nodes ", Senior Member , IEE , July 2012

[6] S.R .Pande , Ms. S.S.Sambare and V.M.Tharke , "Data Clustering Using Data Mining Techniques"of international Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 8, October 2012January 2007

[7] Jayanthi Ranjan, "Applications of Data Mining Techniques in Pharmaceutical Industry",

[8] L. V. Bijuraj, "Clustering and its Applications," in Proc. National Conference on New Horizons in IT – NCNHIT, 2013.

[9] A. Jain and R. Dubes, "Algorithms for Clustering Data ", Englewood Cliffs , NJ: Prentice-Hall , 1988

[10] Dr. E. Chandra, V.P. Anuradha , "A survey on Clustering Algorithms for Data in spatial Database Management Systems, International Journal of Computer Application ", Vol. 24, pp, 19-26 , June 2011

[11] J.C. Bezdek, "Pattern recognition with fuzzy objective function algorithms ", Plenum Press , New York and London , 1987

[12] S.R Nanda, B. Mahanty , M.K. Tiwari " Clustering Indian stock market data for portfolio management of Expert System with Applications , 2010

[13] R. Tibshirani. (January 29, 2013). Clustering 2: Hierarchical clustering.

[14] Techniques in Data Mining "of IJCAT International Journal Computing and technology, Volume 1, Issue 4, ISSN : 2348 - 6090 , May 2014

[15] S.B. Kotsiantis, P.E . Pintelas , "Recent Advances in Clustering : A Brief Survey ll WSEAS Transactions on Information Science and Applications , Vol. 1, No. 1, pp, 73-81, Citeseer , 2004

[16] L. Kaufman and P. J. Rousseeuw "Partitioning Around Medoids (Program PAM), in finding groups in data : an introduction to cluster analysis " John Wiley & Sons, Inc., Hoboken, Nj ,USA , March 1990