ISSN: 2321-7782 (Online)
Volume 3, Issue 1, January 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Performance Evaluation of Clustering Algorithm Using Different Datasets

Peerzada Hamid Ahmad[1]
Research Scholar,
MMICT&BM (MCA), M. M. University (Mullana),
Amballa, Haryana,
India.

Dr. Shilpa Dang[2]
Assistant Professor,
MMICT&BM (MCA), M. M. University (Mullana),
Amballa, Haryana,
India.

Abstract: ***With the advancement of technology, Cluster analysis plays an important role in analyzing text mining techniques. It divides the dataset into several meaningful clusters to reflect the dataset's natural structure. In this paper we analyze the four major clustering algorithms namely Simple K-mean, DBSCAN, HCA and MDBCA and compare the performance of these four clustering algorithms. Performance of these four techniques are presented and compared using a clustering tool WEKA. The results are tested on different datasets namely Abalone, Bankdata, Router, SMS and Webtk dataset using WEKA interface and compute instances, attributes and the time taken to build the model. I have also highlighted the advantages, disadvantages and applications of each clustering technique.***

Keywords: Density based clustering algorithm; Hierarchical clustering algorithm; Make density based clustering; Simple K-mean.

I. INTRODUCTION

Clustering is an unsupervised classification mechanism where a set of patterns (data), usually multidimensional is classified into groups (clusters) such that members of one group are similar according to a predefined criterion [1]. Clustering is a separation of data into groups of related objects. Each group, called a cluster, consists of data that are similar (homogenous) between them and dissimilar (heterogeneous) compared to data of other groups [2]. Clustering of a set forms a partition of its elements chosen to minimize some measure of dissimilarity between members of the same cluster. It is mainly helpful for organising documents to retrieval and support browsing.

Cluster analysis is a very important technology in text mining. It is an iterative process of information detection or interactive multi-objective optimization that involves test and failure. It divides the datasets into several meaningful clusters to reflect the dataset's natural structure. There are several commonly used clustering algorithms namely as Simple K-means, DBSCAN and Hierarchical and so on. A clustering algorithm partitions a dataset into several groups such that the similarity within a group is larger than among groups. Clustering algorithms are often useful in various fields like spatial data analysis, earthquake study, image processing, data mining, learning theory, pattern recognition, etc [3].

The rest of the paper is organized as follows. Section II introduces introduction of clustering techniques used, its advantage, and disadvantage and also highlights main application areas. Section III gives us description of dataset. Section IV gives the interpretation and results. Finally Section V gives us conclusions and future scope.

II. CLUSTERING TECHNIQUES

Clustering in text mining was brought to life by intense developments in information retrieval, extraction [4] and data mining. They resulted in a large amount of application-specific developments and also in some general techniques. These techniques and classic clustering algorithms that relate to them shown below:

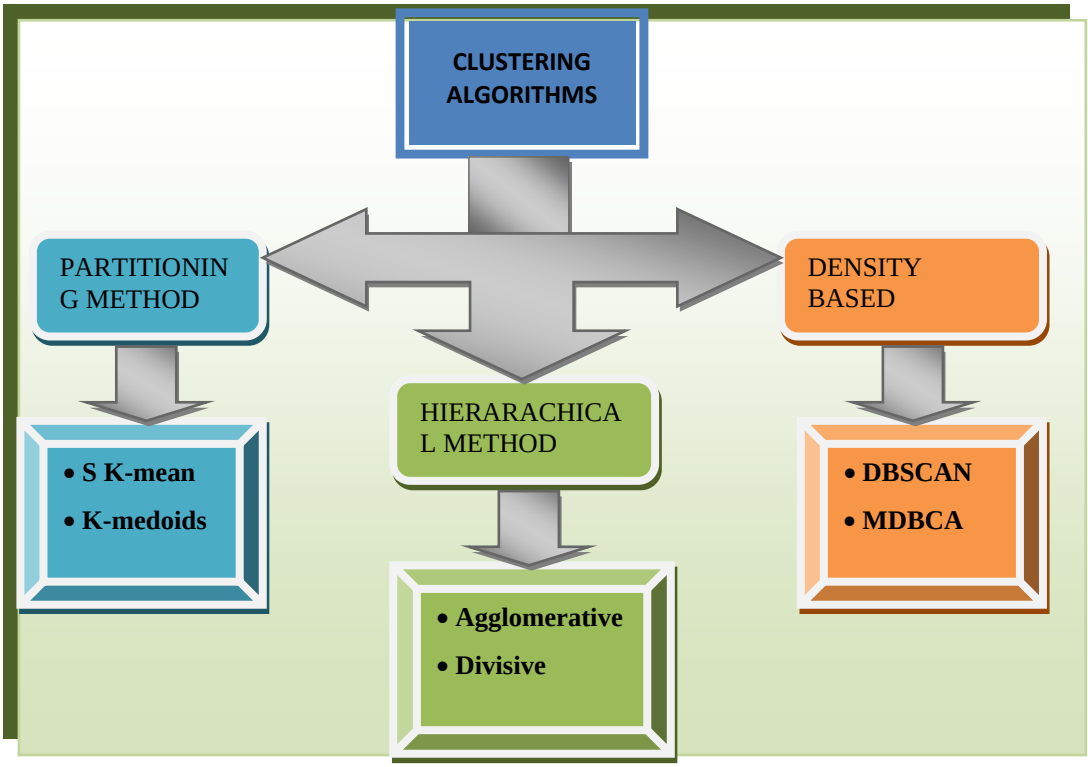

Figure 1 Clustering Algorithms

Clustering techniques are broadly divided into partitioning, hierarchical and density based [5].

a) **Partitioning algorithms:** Identify clusters as areas highly populated with data. They learn clusters directly.

b) **Hierarchical clustering:** Build clusters gradually and are less sensitive to noise.

c) **Density-Based clustering algorithm:** Discover dense connected components of data, which are flexible (shape). These algorithms are less sensitive to outliers and can discover clusters of irregular shapes.

1. Simple K-mean

The K mean algorithm was first projected by Stuart Lloyd, as a technique for pulse-code modulation in 1957 [6]. It is a classical and well known clustering algorithm. It is the most commonly used partitioned clustering algorithm because it can be easily implemented. It is efficient in terms of the execution time. Its time complexity is O (tKn) where n data point numbers, K is the cluster number and t is the iteration number. It is used to partition data points into discoverable K (non-overlapping) clusters by finding K centroids or centre points and then assigning each point to the cluster associated with its nearest centroid [7].

Table 1 Advantages, Disadvantages and Applications of SK- mean

ADVANTAGES	DISADVANTAGES	APPLICATIONS
• Commonly used and easily implemented • Computationally faster method • Scalable • Faster for low dimensional data • Produces tight clusters • Find more sub-cluster if data large cluster number is specified.	• Work only for well-shaped clusters • Fixed number of clusters can make it difficult to predict what K should be. • Not handle non-globular data of different size and densities. • Not identify outliers & noise • Restricted to data which has the notion of centre (centroid)	• Geostatic • Computation vision • Market segmentation • Earth quake study • Land use

2. DBSCAN

DBSCAN was proposed by Martin Ester et al in 1996. It is one of the most common clustering algorithms [8]. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of

corresponding nodes. This algorithm is based on connecting points within certain distance thresholds similar to linkage based clustering. However, it only connects points that satisfy a density criterion (minimum number of objects within radius). An arbitrary shape cluster is formed which consists of all density-connected objects. DBSCAN separates data points into three classes:

» **Hub points:** Points that are at the interior of a cluster (Centre).

» **Edge points:** Falls within the neighbourhood of a hub point which is not a hub point.

» **Noise points:** Any point that is not a hub point or an edge point.

To find a cluster, DBSCAN starts with an arbitrary instance (p) in data set (D) and retrieves all instances of D with respect to epsilon (Eps) and minimum points (minPts). minPoints, defined as the minimum number of points required to exist in a neighbourhood to be declared a cluster, and Eps defined as the radius of the neighbourhood of a point based on a distance (Euclidean, Manhattan or Minkowski) metric. The algorithm makes use of a spatial data structure to locate points within Eps distance from the core points of the clusters [9].

Table 2 Advantages, Disadvantages and Applications of DBSCAN

ADVANTAGES	DISADVANTAGES	APPLICATIONS
• Can discover arbitrarily shaped clusters • Find cluster completely surrounded by different clusters. • Robust towards outlier detection (noise) • Require just two points which are very insensitive to the ordering of the points in the database.	• Not partitionable for multiprocessor systems. • Datasets with altering densities are tricky. • Sensitive to clustering parameters minPoints and EPS. • Fails to identify cluster if density varies and if the dataset is too sparse. • Sampling affects density measures.	• Scientific literature • Images of satellite • Crystallography of x-ray • Anomaly detection in temperation data

3. Hierarchical Clustering Algorithm

The Hierarchical clustering algorithm (HCA) is also called as connectivity based clustering, which is mainly based on the core idea of objects that are being more relative to the nearby objects than to the objects far away. It is a method of cluster analysis which seeks to build a hierarchy of clusters. Its result is usually presented in a dendrogram. It is generally classified as Agglomerative and Divisive methods that depended upon how the hierarchies are formed [2].

» **Agglomerative:** It is a "bottom up" approach. It starts by placing each object in its own cluster. Then merges these minute clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied. Its complexity is O (n^3) which makes then too slow for large data sets.

» **Divisive:** It is a "top down" approach. It starting with all objects in one cluster. Then splits are performed recursively as one move down the hierarchy. Its complexity is O (2^n) which is worse.

These algorithms join the objects and form clusters by measuring their distance. These algorithms cannot provide a particular partitioning in the dataset, but they provide a widespread hierarchy of clusters that are merged with each other at accurate distance [10].

Table 3 Advantages, Disadvantages and Applications of HCA

ADVANTAGES	DISAVANTAGES	APPLICATIONS
• Conceptually simple • Good for small data sets • Not require the number of clusters k in	• Cluster merging/splitting is permanent and the error occurring later is impossible to count	• Pattern recognition • Image segmentation • Wireless sensors networks

advance • Merging/splitting of cluster is permanent, alternative way is reduced • Less sensitive to noise in the data set. • Needs a termination/ readout condition	• Sensitivity to noise and outliers • Difficulty handling different sized clusters and convex shapes • Divisive methods can be computational hard • Methods are not scalable for large database • No objective function is directly minimized	• City planning • Spatial data analysis

4. Make Density Based Clustering Algorithm

The make density based clustering algorithm uses (wrapping) a clusterer algorithm internally. It returns both distribution and density. This clustering algorithm is very helpful when clusters are uneven. In this algorithm we try to find the cluster according to the density of data point in a region. The main idea of this clustering is for each of cluster the neighbourhood of given radius (Eps) has contain at least minimum number of instances (min Pts). It can also be used if the data has noise and when there are outliers in the data. The points of same density and present within the respective same areas will be connected while forming clusters. In this way, we get separate cluster of having low density regions (a set of points separated by low density) and high density regions (a set of points separated by high density). The high density region has are tight as compared to low dense regions [11].

Table 4 Advantages, Disadvantages and Applications of MDBCA

ADVANTAGES	DISADVANTAGES	APPLICATIONS
• Useful when clusters are not normal • Return both distribution and density • Used when data has noise • Used when outliers in the data • Gives result close to K-mean algorithms.	• Datasets with altering densities are tricky. • Sensitive to clustering parameters minPoints and EPS. • Sampling affects density measures.	• Scientific literature • Images of satellite • Crystallography of x-ray • Geostatic • Earthquake study

III. DESCRIPTION OF DATASET USED

For performing the comparison analysis we need input dataset which is an integral part of text mining applications. In this research data used in my experiment is either real world data obtained from UCI machine learning repository and widely accepted set available. We have taken five datasets containing continuous attributes (nominal type) that is all these datasets have.

» **Abalone:** Sea-nail based corpus. It consists of 2924 instances and 8 attributes.

» **Bankdata:** General information of a customer and consists of 513 instances and 12 attributes.

» **Reuter:** Collection of news paper articles on various topic. It contains 1554 instances and 1003 attributes.

» **SMS:** Spam messages extracted manually from grumble text website and contains 100 instances and 861 attributes

» **Webtk:** Web pages collected by World Wide Knowledge Base (various universities) and contains 2010 instances and 1013 attributes

IV. INTERPRETATION AND RESULTS

To verify improved performance of our research, we made experiments using datasets from UCI machine learning repository [15]. We used these five datasets ‘Abalone’, ‘Bankdata’, ‘Reuter’, ‘SMS’ and ‘Webtk’ in our experiment. The above discussed four clustering algorithms have been carried out in order to measure the comparative performance parameters of the algorithms over the datasets. Table 5 shows the numbers of instances and attributes of the used datasets.

Table 5 Number of instances and attributes used in the datasets

Data set	Clustering algorithm used	Instances	Attributes
Abalone	Simple K mean	2924	8
	DBSCAN		
	HCA		
	MDBCA		
Bank Data	Simple K mean	513	12
	DBSCAN		
	HCA		
	MDBCA		
Reuter	Simple K mean	1554	1003
	DBSCAN		
	HCA		
	MDBCA		
SMS	Simple K mean	100	861
	DBSCAN		
	HCA		
	DBCA		
Webtk	Simple K mean	2010	1013
	DBSCAN		
	HCA		
	MDBCA		

Graphically we can also see the distribution of various datasets used. Figure 1 gives us the number of instances and the number of attributes used for each datasets.

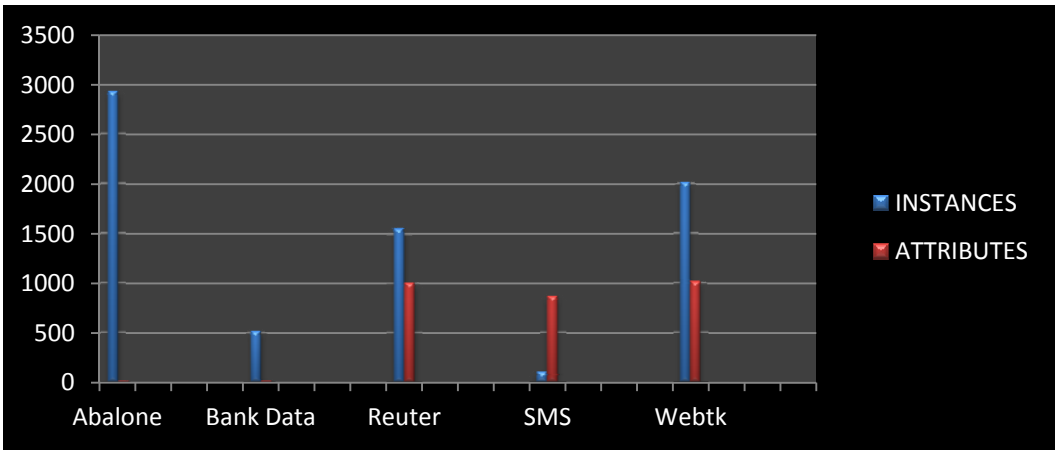

Figure 2 Graphical view of number of instances and attributes used

The Simple K-mean, DBSCAN, HCA and MDBCA are applied on the five different datasets and their results are compared on the basis of time complexity. The figure (3) shows the time taken by the cluster algorithms to make clusters when these datasets are deployed in the tool.

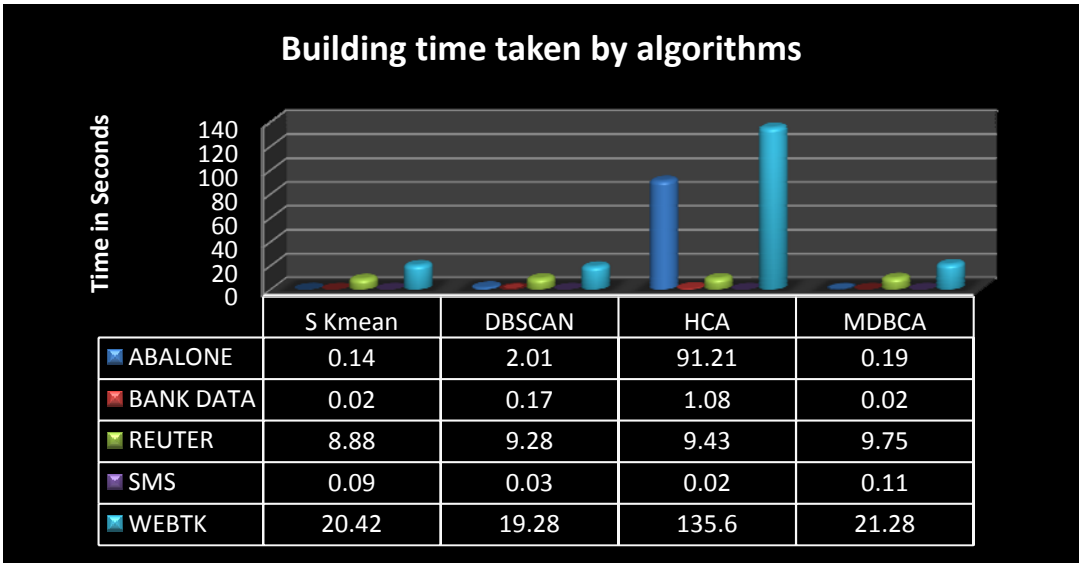

Figure 3 On the basis of building time taken by clustering algorithm to make cluster

With the help of analysis, it is shows that Simple K-mean has required minimum time to make cluster for the five datasets in comparison with other clustering algorithm. So, overall performance of Simple K-mean is higher.

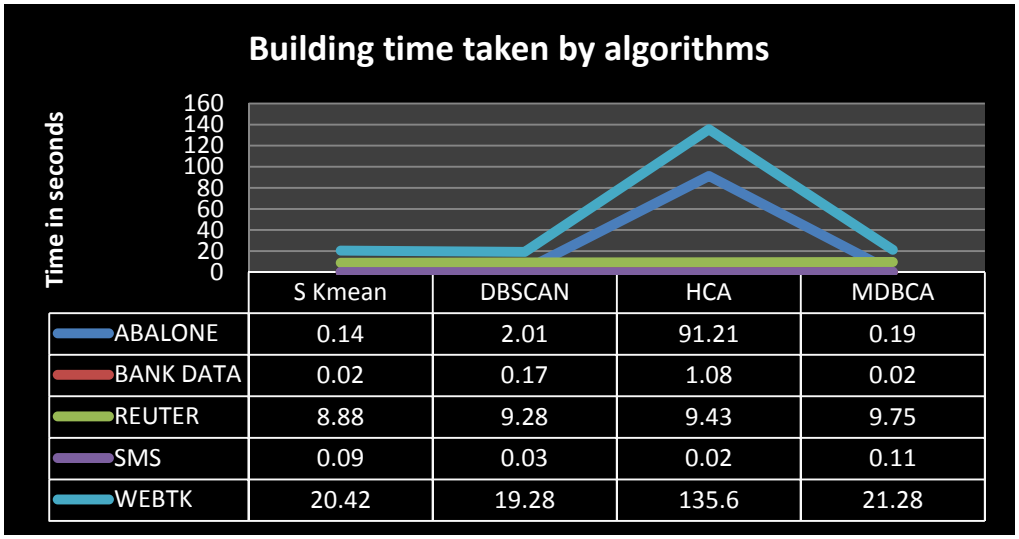

Figure (4) Building time comparison among clustering algorithms with used datasets

The performance of five datasets is compared for Simple K-mean, DBSCAN, HCA and MDBCA is shown graphically in figure (4). An analysis shown in this figure shows that Simple K-mean clustering algorithm has higher overall performance for five datasets whereas MDBCA has almost equal to Simple K-mean. The other two clustering algorithms DBSCAN and HCA has lowest performance.

V. Conclusion And Future Work

The main conclusion of this paper is to make a comparative performance analysis of Simple K-means, DBSCAN, HCA and MDBCA. It is important to remember that cluster analysis is an exploratory tool. While hundreds of clustering algorithms are available and new ones continue to appear, we compare only four of them. All the algorithms have some ambiguity in some (noisy) data when clustered. Simple K-means make clusters with minimum amount of time. Whereas MDBCA shows slightly equal performance to Simple K-mean in making clusters. HCA is more sensitive for noisy data and shows much variation with time complexity. DBSCAN is not suitable for data with high variance in density. In terms of time complexity and dataset used, K-means produces better results in comparison to all explained algorithms.

This can be seen as the base for future work. Evaluations with the parameters show that none of the parameter can provide an overall rating of quality of cluster. Some parameters give contacting quality rating on some datasets. Such effects show us that further research should be done in this area. However, further work has to be done to collect a bigger test set of high dimensional datasets. On such a benchmarking set one could collect best parameter settings of various algorithms and best quality results of clustering results on these datasets. The aim of an overall evaluation will then lead to a more mature clustering research field in which one easily judge the quality of algorithm by comparing it with approved results of competing approaches.

References

1. Bhoopender Singh, Gaurav Dubey, “A comparative analysis of different data mining using WEKA”, International Journal of Innovative Research and Studies, ISSN: 2319-9725, Volume 2, Issue 5, Page 380-391 and May 2013.
2. Dr.Naveeta Mehta, Shilpa Dang, "A Review of Clustering Techiques in various Applications for Effective Data Mining",International Journal of Research in IT & Management, ISSN 2231-4334,Volume 1, Issue 2, Page 50-66 and June 2011.
3. Sunila Godara, Amita Verma, "Analysis of Various Clustering Algorithms", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-3, Issue-1, Page 186-189 and June 2013,.
4. Dr Shilpa Dang and Peerzada Hamid Ahmad, “A comparative study on text mining techniques”, International Journal of Science and Research, ISSN (online): 2319-7064, Volume 2, Issue 12, Page 2222-2226 and Dec 2014.
5. Glory H. Shah, C. K. Bhensdadia, Amit P. Ganatra, "An Empirical Evaluation of Density-Based Clustering Techniques", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-2, Issue-1and March 2012
6. Bandana Parida, Dr. Rakesh Chandra Balabantaray, "A Comparative Study of Document Clustering", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Volume 2, Issue 6, Pages 2750-2757 and June – 2013.
7. Bhoj Raj Sharmaa, Aman Paula "Clustering Algorithms: Study and Performance Evaluation Using Weka Tool" International Journal of Current Engineering and Technology, ISSN 2277 - 4106, Volume 3, Issue 3, Page 1094-1098 and August 2013.
8. Slava Kisilevich, Florian Mansmann, Daniel Keim, “FIP-DBSCAN:A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos”, University of Konstanz.
9. Rui Xu, Donald Wunsch II, "Survey of Clustering Algorithms", IEEE transactions on neural networks, Volume16, NO. 3, Page 645-678 and MAY 2005.
10. Ranjini K and Rajalinngum N, “Performance Analysis of Hierarchical Clustering Algorithm” International Journal of Advanced Networking and Applications, ISSN: 1006-1011 and Year 2011.
11. Yiling Yang, Xud ong Guan, Jinyuan You, “CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data”, ISSN: 58113-567-X/02/0007 and Year 2002
12. Sharmila, R.C Mishra, "Performance Evaluation of Clustering Algorithms", International Journal of Engineering Trends and Technology (IJETT), ISSN: 2231-5381, Volume4 Issue7,Page 3113-3116 and July 2013.
13. Garima Sehgal, Dr. Kanwal Garg "Comparison of Various Clustering Algorithms", International Journal of Computer Science and Information Technologies, Volume 5 Issue 3, Page 3074-307 and Year 2014.
14. Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya "Comparison the various clustering algorithms of weka tools", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 5 and May 2012.
15. UCI machine learning repository, archive.ics.uci.edu/ml.

