International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-3 Issue-11 September-2013

Comparative Analysis of Clustering Techniques for Real Dataset

Kishor T. Mane[1], Vandana G. Pujari[2]

Abstract

Clustering is a process of classification of data objects into similar groups or clusters. A clustering algorithm partitions the data set into several similar groups. By analyzing different algorithms, efficiency of clustering techniques has been calculated. System can define cluster quality by comparing different clustering algorithms. Data clustering techniques are used in a wide variety of scientific applications such as biology, pattern recognition, information systems etc. In this paper an attempt has been made to compare the different data clustering techniques such as K-Means, K-Medoid and Rough K-Means on the basis of various parameters like memory required, execution time and compactness of the cluster. These algorithms are applied on the real dataset.

Keywords

Data clustering, K-means, K-Medoid, Rough k-means, real dataset, clusters algorithms.

1. Introduction

Advances in sensing and storage technology and also dramatic growth in applications such as Internet search, digital imaging and video surveillance have created many high-volume, high-dimensional data sets. Most of this data is stored digitally in electronic media, thus providing huge potential for the development of automatic data analysis, classification, and retrieval techniques. Many of these data streams are unstructured, adding to the difficulty in analyzing them. This increase in both the volume and the variety of data requires advances in methodology to automatically understand process and summarize the data [1]. Different data clustering techniques are very useful to minimize this type of problems.

Kishor T. Mane, Assistant Professor, Information Technology Department, D. Y. Patil College of Engg. & Tech, Kolhapur, Maharashtra, India.

Vandana G. Pujari, Lecturer, E & TC Department, Dr. D. Y. PatilPolytechnic, Kolhapur, Maharashtra, India.

Data analysis techniques can be broadly classified into two major types as, first exploratory or descriptive, meaning that the investigator does not have pre-specified models or hypotheses but wants to understand the general characteristics or structure of the high dimensional data, and second confirmatory or inferential, means that the investigator wants to confirm the validity of a hypothesis or model or a set of assumptions given the available data. Data clustering is the assignment of a set of observations into subsets called as clusters so that observations in the same clusters are similar in some sense. The goal of data clustering or cluster analysis is to find the natural groupings of asset of patterns, points or objects. In this paper the comparison has been made on different clustering techniques.

2. Data Clustering Techniques

The Cluster analysis groups data objects based only on the information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in the other groups. The greater similarity (or homogeneity) of clustering is within a group, and the greater the difference between groups, the better or more distinct the clustering [1]. Given a representation of n objects, find K groups based on a measure of similarity such that objects within the same group are alike but objects in different are not alike. Here, the different clustering techniques are discussed as –

K-Means Clustering

K-means algorithm assigns each point to the cluster whose center called centroid is nearest to the value of point [2]. The center is the average of all the points in the cluster-that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. Let X= {xi, i=1 ….n} be the set of n d-dimensional points to be clustered into a set of K clusters, C= {ck, k=1, 2…k}.K-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized [3]. Let μ be the mean of cluster ck. The squared error between μk and the points in the cluster ck is defined as,

$J(c_k) = \sum_{x_i \in c_k} \|Xi - \mu k\|^2$ (1)

The goal of k-means is to minimize the sum of the squared error over all the K clusters.

$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|Xi - \mu k\|^2$ (2)

Minimizing this objective function is an NP-hard problem(even for K=2).Thus k-means is greedy algorithm converge to a local minimum .K-means starts with an initial partitions with k clusters and assign objects to clusters so as to reduce the squared error (distance from the centroid to the object). Since the squared error tends to decrease with an increase in number of clusters K (with J(C)=0 with K=n)(i.e. Every object is a centroid and it is closest from himself), it can be minimized only for fixed number of clusters. Main steps of algorithm are as follows:

1. Select an initial partition with K clusters; repeat this step2 and 3 clusters until membership stabilizes.
2. Generate a new partition by assigning each partition to its closest cluster i.e. centroid.
3. Compute new clusters centers.

The most critical is the choice of K (no. of clusters). For a given K, with several different initial partitions and choose the partitions with the smallest value of the squared error. K-means is used with Euclidean metric for computing the distance points and clusters. As a result, K-means finds spherical or ball shaped clusters in data.

K-Medoid Clustering

It is a clustering algorithm related to the k-means algorithm and the medoid shift algorithm. Both the k-means and k-medoid algorithms are partitioned (breaking the dataset up into groups) and both attempt to minimize squared error, the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoid chooses data points as centers as medoid [4]. The k-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters. It is more robust to noise and outliers as compared to k-means because it minimizes a sum of pair wise dissimilarities instead of a sum of squared Euclidean distances [5]. A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the cluster. The most common realization of K-Medoid clustering is the Partitioning around Medoids (PAM) algorithm and is as follows:

1. Initialize: randomly select k of the n data points as the medoid.
2. Associate each data point to the closest medoid.
3. Update medoid.
4. Repeat steps 2 to 3 until there is no change in the medoid.

Rough K-means clustering

Rough set algorithm has two boundaries i.e. upper boundary and lower boundary hence we can cluster the points having same value also [5]. Threshold value is defined which is used for the clustering. Therefore in the rough set approach with any rough set pair of crisp sets- called the lower and the upper approximation of the Rough set is associated. The lower approximation consists of all objects which surely belong to the set and upper approximation contains all objects which possibly belong to the set. The difference between the upper and the lower approximation constitute the boundary region of the rough set. Let U denote a finite universal set and let R be subset of U * U be an equivalence relation on U. the pair A = (U, R) is called an approximation space. the equivalence partitions the set U into disjoint subsets denoted by U/R =E1,E2,…En where Ei is equivalence class of R .to elements u and v are indistinguishable if they belong to same equivalence class E. the equivalence classes of R are called the elementary sets in the approximation space A=(U, R). The union of one or more elementary sets is called a composed set in A. Since it is not possible to differentiate the elements within the same equivalence class, one may not be able to obtain precise representation for an arbitrary sets X subset of U in terms of elementary sets in A. instead, any X may be represented by its lower and upper bound. The lower bound A(X) is the union of the elementary sets which are subset of X, and the upper bound (X) is the union of all elementary sets which have a non-empty intersection with X. The pair (A(X), (X)) is the representation of the ordinary set in the approximation space A = (U, R). Or simply the rough set of X, while elements in upper bound may or may not belong to X.

3. Implementation

The previous section describes the theoretical part of the clustering techniques. Now, in this section the focus goes on the implementation of three clustering techniques used for clustering of real data. The implementation has been done using JAVA language with windows platform. The real data set has been considered for comparison. The figure1, figure 2, figure 3 shows the graphical representation of cluster

techniques as K-means, K- Medoid, and Rough K-Means resp.

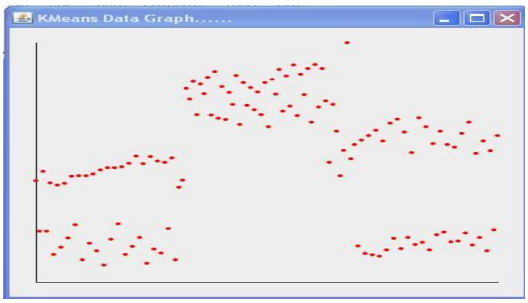

Figure 1: K-Means cluster graphical representation

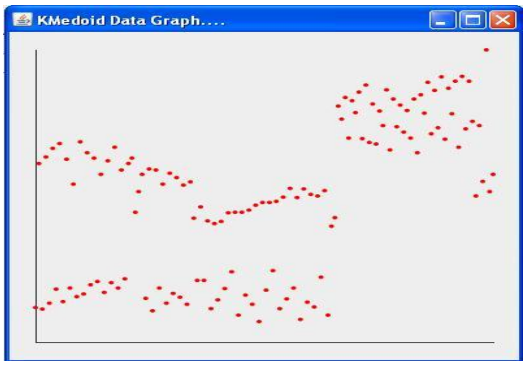

Figure 2: K– Medoid cluster graphical representation

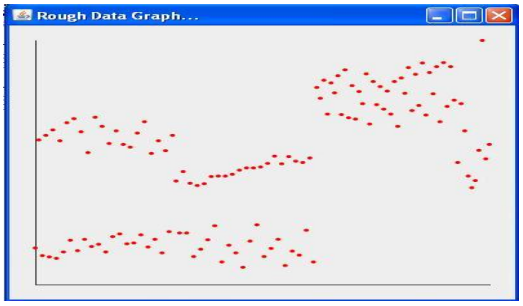

Figure 3: Rough K-means Cluster graphical representation

4. Experimental Results

The Cluster techniques are compared on the basis of following factors:

1. Memory Required.
2. Execution Time.
3. Compactness: Compactness of each cluster is measured on the basis of radius of the cluster.

Table 1 shows the comparison between three algorithms with the help of execution time and memory requirement.

Table 1: Comparison between algorithms

Algorithm	Execution Time	Memory
K-Means	31	322 KB
K – Medoid	16	355 KB
Rough K-means	47	314 KB

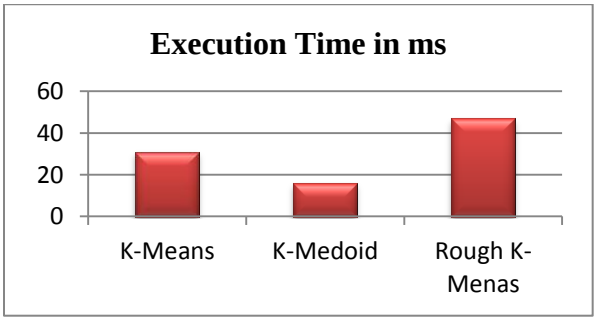

Figure 4: Execution time

The Figure 4 shows the comparative results of the clustering techniques in the form of execution time. In this K-Medoid is efficient than other two because it takes less execution time. The Figure 5 shows the comparative results of the clustering techniques in the form of memory requirement. In this Rough K-Means is efficient than other two because it takes less memory.

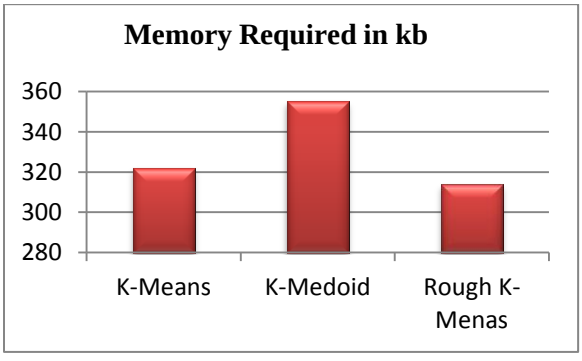

Figure 5: Memory required

5. Applications of Cluster Analysis

There are different field where the clustering analysis plays vital role. Some of them are described as follows –

1. Social network analysis–

In the study of social networks, clustering may be used to recognize communities within large groups of people.

2. Recommender systems –

Recommender systems are designed to recommend new items based on a user's tasks. They sometimes use clustering algorithms to predict a user's preferences based on the preferences based on the user's in the user's cluster.

3. Biology -

In biology clustering has many applications. In the fields of plant and animal ecology, clustering is used to describe and to make spatial and temporal comparisons of communities of organisms in heterogeneous environments; it is also used in plant systematic to generate artificial phylogenies or clusters of organisms at the species, genus or higher level that share no. of attributes.

4. Medicine -

In medical imaging, such as PET scans, cluster analysis can be used to differentiate between different types of tissue and blood in a three dimensional image. In this application, actual position does not matter, but the vexed intensity is considered as a vector, with a dimension for each image that was taken over time. This technique allows, for example, accurate measurements of the rate a radioactive tracer is delivered to the area of interest, without separate sampling of arterial blood, an intrusive technique that is most common today.

6. Conclusion

Different clustering techniques like k-means, k-medoid, rough k-means, and algorithms are used to cluster the data. To get the efficient technique of clustering this different techniques are compared on the basis of execution time, memory required and compactness of clusters. K-Medoid is efficient on the basis of execution time because it required only 16milliseconds to give the output. Rough K-Means is efficient because it required only 314KB memory and K-Medoid is worst because it required more memory than other techniques. In summary, clustering is an interesting, useful, and challenging problem. It has great potential in applications like object recognition, image segmentation, and information filtering and retrieval. However, it is possible to exploit this potential only after making several design choices carefully. These algorithms can be used in conjunction with other neural or fuzzy systems for further refinement of the overall system performance.

References

[1] Bruce Moxon "Defining Data Mining, TheHows and Whys of Data Mining, and How It Differs From Other Analytical Techniques" Online Addition of DBMS Data Warehouse Supplement, 200-208.

[2] Hartigan, J.A. and Wong, M.A. (1979). Algorithm AS136: A K-Means Clustering Algorithm. Applied Statistics, 28, 100–108.

[3] P. Arabie, L. Hubert, Advanced Methods in Marketing Research, in:Cluster Analysis in arketing Research, Blackwell, Oxford, 1994, pp. 160–189.

[4] Marek, V.W. and Truszczy_nski, M.: Contributions to the theory of rough sets, Fundamenta Informaticae 39, 389-409 (1999).

[5] Lingras, P.: Rough K-Medoids clustering using GAs. Proceedings of the 8th IEEE International Conference on Cognitive Informatics, 315-319 (2009).

[6] L. D. Baker and A. McCallum. Distributional clustering of words for text classification. In SIGIR '98: Proceedings of the21st Annual International ACM SIGIR, pages 96–103. ACM, August1998.

[7] A.K. Jain, M.N. Murty, P.J. Flynn," Data Clustering: A Review" Journal ACM computing Surveys. Volme 31 Issue 3, Sept. 1999, Pages 264-323.

[8] Pawan Lingras, C.West," Interval Set Clustering of Web Users with Rough K-Means" Journal of Intelligent Information Systems. July 2004, Volume 23, Issue 1, pp 5-16.

Kishor T. Mane received the BE in Computer Sci. &Engg. from Dr. J. J. Magdum college of Engg., Jaysingpur in 2007 and ME in computer sci. & engg. from D.Y. Patil college of engg. & tech., Kolhapur. He is working as Asst. Prof. in D.Y Patil college of engg. & tech., Kolhapur.

Vandana G. Pujari received the BE in Electronics from D.K.T.E, Ichalkaranji in 2011 and is currently pursuing ME degree in E&TC and also working as lecturer in E & TC dept, Dr. D. Y. Patil Polytechnic, Kolhapur.