DATE OF ACCEPTANCE: MAY 13, 2014
DATE OF PUBLICATION: MAY 18, 2014
ISSN: 2348-4098
VOLUME 02 ISSUE 04 APRIL-MAY 2014

A SURVEY ON CLUSTERING PROBLEM WITH OPTIMIZED K- MEDOID ALGORITHM

[1]SUNITA KUMARI, [2]ABHA KAUSHIK

[1]M.Tech, Computer Science & Engineering, Galgotia's University, Gr. Noida, UP, India ,
E-mail: sunita.rao86@gmail.com
[2]M.Tech, Computer Science & Engineering, Galgotia's University, Gr. Noida, UP, India ,
E-mail: abha15.1990@gmail.com

ABSTRACT

Clustering is the division of data into groups of similar objects. It disregards some details in exchange for data simplification. Informally, clustering can be viewed as data modeling concisely summarizing the data, therefore, it relates to many disciplines from statistics to numerical analysis. Such applications usually deal with large datasets and many attributes. Searching of such data is a subject of data mining. This survey based on clustering algorithms from a data mining viewpoint. The goal of this survey is to provide a comprehensive review of different clustering techniques in data mining. Clustering is a partition of data into groups of similar objects. Each and every group, called a cluster, consists of various objects that are similar to one another and dissimilar to objects of other groups.

KEYWORDS: K Means, K Medoid, Clustering, Partitional Algorithm,

1. INTRODUCTION

The goal of this survey is to provide a comprehensive review of different clustering techniques in data mining. Clustering, which aims at dividing a dataset into groups or clusters containing similar data, is a fundamental problem in unsupervised Learning and has many applications in various domains. In recent years, there has been significant interest in developing clustering algorithms to massive datasets. Clustering is useful technique for discover some or the entire hidden patterns. The discovery of data sharing and patterns in the underlying data. Cluster is a collection of data objects

- Similar to one another in similar cluster

ISSN: 2348-4098
VOLUME 02 ISSUE 04 APRIL-MAY 2014

- Different to the objects in other clusters
- A good clustering method will produce better quality clusters with
  - High intra class relationship
  - Low inter class relationship

The value of clustering result based on both the similarity measure used by the method and its implementation.

The value of clustering method is also measured by its capability to

DATA CLUSTERING ALGORITHMS: Data clustering algorithms can be divided into following categories. Some of these algorithms are given as follows:

PARTITIONING ALGORITHMS:

Build various partitions and then evaluate them by some measure.

HIERARCHY ALGORITHMS:

Create a hierarchical breakdown of the set of data (or objects) using some criterion.

DENSITY-BASED: built on connectivity and density functions.

GRID-BASED: Grid based clustering is depending on a multiple-level granularity structure.

MODEL-BASED: A model is offered for each of the clusters and the idea is to find the best fit of that model to each other.

2. PARTITIONING METHODS

2.1 K-Means Method

2.2 K –Medoids Method

2.3 CLARA

2.4 CLARANS

2.1 K-MEANS ALGORITHM

K-means is a widely used partitioned clustering method. While there are considerable exploration efforts to characterize the key features of K-means clustering, further exploration is needed to reveal whether the optimal number of clusters can be found on the run based on the cluster quality measure. It classifies a given set of n data objects in k clusters, where k is the number of preferred clusters and it is required in advance. A centroid is defined for all clusters. Each data objects are positioned in a cluster having centroid nearby to that data object. Later handling all data objects, k-means, or centroids, are rearranged, and the whole process is repeated. All data objects are certain to the clusters depend on the new centroids. In each repetition centroids change their location step by step. In other words, centroids move in each repetition. This process is sustained until no any centroid move. As a result, k clusters are found signifying a set of n

ISSN: 2348-4098
VOLUME 02 ISSUE 04 APRIL-MAY 2014

data objects. An algorithm for k-means method is given below.

Algorithm Input: 'k', is the number of clusters to be divided; 'n', is the number of objects. **Output**: A set of 'k' clusters based on given similarity function.

Steps:

i) Arbitrarily choose 'k' objects as the initial cluster centers;

ii) Repeat,

a. Reassign each object to the cluster to which the object is the most similar; based on the given similarity function;

b. Update the centroid (mean value of cluster), i.e., calculate the mean value of the objects for each cluster;

iii) Until no change.

Limitations and problems: K-means attempts to minimize the squared or absolute error of points with respect to their cluster centroids. Although this is sometimes a reasonable criterion and leads to a simple algorithm, K-means has a number of limitations and problems.

Handling Empty Clusters: One of the problems with the basic K-means algorithm given earlier is that empty clusters can be obtained if no points are allocated to a cluster during the assignment step. If this happens, then an approach is needed to choose a replacement centroid, since otherwise, the squared error will be larger than necessary.

Reducing the SSE with Post processing: In k-means to get better clustering we have to reduce the SSE that is most difficult task. There are various types of clustering methods available which reduces the SSE [16].

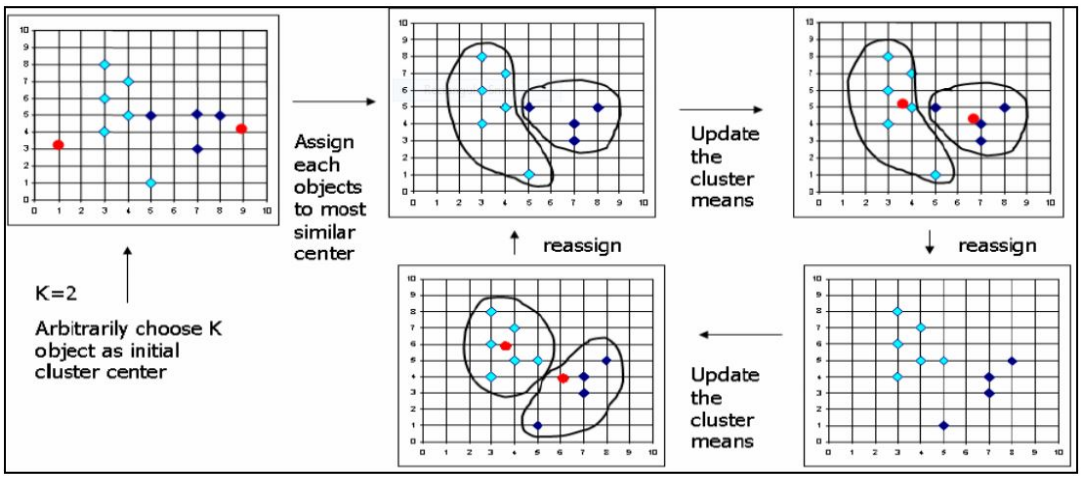

Figure1. Working of k means algorithm

ISSN: 2348-4098
VOLUME 02 ISSUE 04 APRIL-MAY 2014

2.2 K-MEDOID ALGORITHMS

The k-means method uses centroid to represent the cluster and it is sensitive to outliers. This means, a data object with an extremely large value may disrupt the distribution of data. K-Medoids method overcomes this problem by using Medoids to represent the cluster rather than centroid. A Medoids is the centrally positioned data object in a cluster. Here, k data objects are selected randomly as Medoid to represent k cluster and remaining all data objects are placed in a cluster having Medoids nearest (or most similar) to that data object. After handling all data objects, new Medoids is determined which can represent cluster in a better way and the whole process is repeated. Again all data objects are bound to the clusters depend on the new Medoids. In each repetition, Medoids change their location step by step. In other words, Medoids move in each repetition. This process is continued until no any Medoids change. [2]

Algorithm:

Input: 'k', is the number of clusters to be divided; 'n', the number of objects.

Output: A is the set of 'k' clusters that reduces the sum of the dissimilarities of all the objects to their neighboring Medoid.

Steps:

(i) Arbitrarily choose 'k' objects as the initial Medoid;

(ii) Repeat,

(a). Allot each remaining object to the cluster with the neighboring Medoid;

(b). Randomly select a non-Medoid object;

(c). Compute the total cost of swapping old Medoid object with a new selected non- Medoid object

(d). If the total cost of swapping is less than zero (< 0), then perform that swap operation to form the new set of k-Medoid.

(iii) Until no change.

ISSN: 2348-4098
VOLUME 02 ISSUE 04 APRIL-MAY 2014

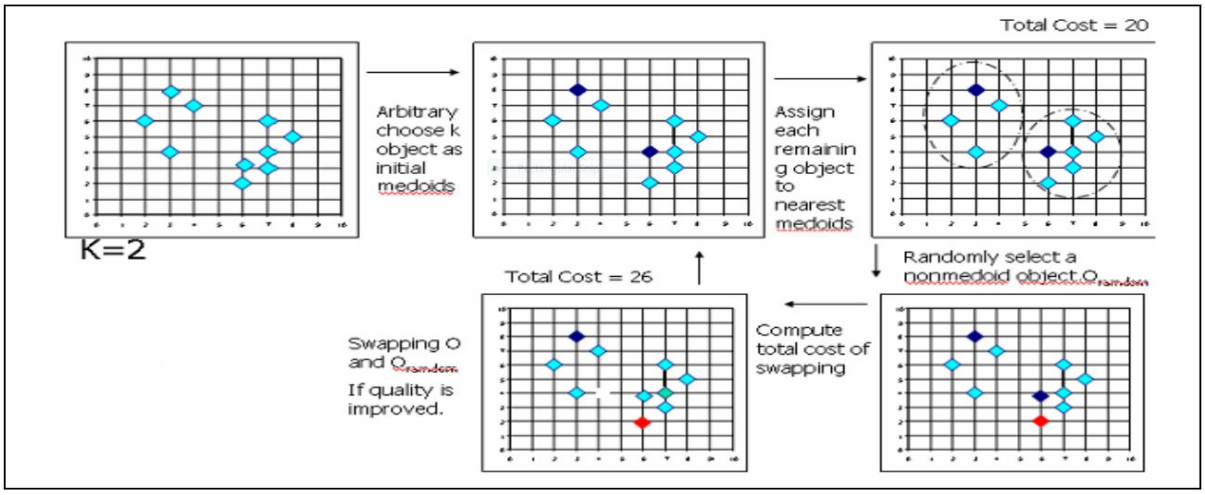

Figure2. Working of k- Medoid algorithm

Features of K-Medoid Algorithm

K-Medoid works on the dissimilarity matrix of the given data set or when it is offered with data matrix, the algorithm first computes a dissimilarity matrix. It is more robust; because it minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances it provides a novel graphical representation, the outline plot, which permits the user to select the optimal number of clusters. However, PAM lacks in scalability for very large databases and it present high time and space complexity It is understood that the average time for normal distribution is greater than the average time for the uniform distribution. This is true for both the algorithms K-Means and K-Medoids. When the number of data points is less than the K-Means algorithm takes less execution time as compare to k-Medoid algorithm. But when the data points are increased to maximum the K-Means algorithm takes maximum time and the K-Medoids algorithm performs reasonably better than the K-Means algorithm. The specific feature of k-Medoid algorithm is that it needs the distance among every pair of objects only once and uses this distance at every stage of repetition.[3]

2.3 CLARA

CLARA (Clustering large Applications) algorithm is designed by Kaufman and Rousseeuw to handle large data sets, It depend on sampling. Instead of finding representative objects for the whole data set, CLARA draws a model of the data set,

ISSN: 2348-4098
VOLUME 02 ISSUE 04 APRIL-MAY 2014

applies PAM on the model, and finds the Medoid of the sample. The point is that, if the model is drawn in a sufficiently random way, the Medoid of the model would approximate the Medoid of the whole data set. For finding the better approximations, CLARA draws several models and provides the best clustering as the result. Here, for perfection, the quality of a clustering is measured based on the average dissimilarity of all objects in the whole data set, and not only of those objects in the models. [4]

2.4 CLARANS

CLARANS is very efficient and effective. Second, we study how CLARANS can handle not only point's objects, but also polygon objects efficiently. One of the approaches measured, called the IR-approximation, is very efficient in clustering convex and nonconvex polygon objects. CLARANS is a main-memory clustering technique, while many of the above-mentioned techniques are designed for out-of-core clustering applications. We admit that whenever wide I/O operations are difficult, CLARANS is not effective as the others. However, we argue that CLARANS still has considerable applicability.

CLARANS uses a randomized search approach to improve on both CLARA and PAM. Conceptually CLARANS does the following.

1) Randomly pick K candidate Medoids.

2) Randomly consider a swap of one of the selected points for a non-selected point.

3) If the new configuration is better, i.e., has lower cost, then repeat step 2 with the new configuration.

4) Otherwise, repeat step 2 with the current configuration unless a parameterized limit has been exceeded. (This limit was set to max (250, K *(m - K)).

5) Compare the current solution with any previous solutions and keep track of the best.

6) Return to step 1 unless a parameterized limit has been exceeded. (This limit was set to 2.)

3. CONCLUSION

This survey starts with a brief introduction about clustering in data mining. Since measuring similarity between data objects is simpler than mapping data objects to data points in feature space, these pairwise similarity

ISSN: 2348-4098
VOLUME 02 ISSUE 04 APRIL-MAY 2014

based clustering algorithms can greatly reduce the difficulty in developing clustering based pattern recognition applications. The advantage of the K means algorithm is its favorable execution time. Its disadvantage is that the user has to know in advance how many Clusters are searched for. It is observed that K means algorithm is efficient for smaller data sets and K-Medoids seems to perform better for large datasets. [3]

REFERENCES

[1] Weiguo Fan, Linda Wallace, Stephanie Rich, Zhongju Zhang "Tapping into the Power of Text Mining" February 16, 2005.

[2] Singh s s , Chauhan N C " K-means v/s K-medoids: A Comparative Study" National Conference on Recent Trends in Engineering & Technology, 13-14 May 2011.

[3] Batra A. " Analysis and Approach: K-Means and K-Medoids Data Mining Algorithms" 5th IEEE International Conference on Advanced Computing & Communication Technologies [ICACCT-2011] ISBN 81-87885-03-3

[4] Raymond T. Ng and Jiawei H. "CLARANS: A Method for Clustering Objects for Spatial Data Mining" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 14, NO.5, SEPTEMBER/OCTOBER 2002

[5] An Introduction to Cluster Analysis for Data Mining (ebook).

[6] Soni T. M. "AN OVERVIEW ON CLUSTERING METHODS" IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-725

[7] Berkhin P. "Survey of Clustering Data Mining Techniques" Accrue Software, 1045 Forest Knoll Dr., San Jose, CA, 95129

[8] STEFANOWSKI J. delivered Lecturer on "Data Mining-Clustering" at Institute of Computing Sciences Poznan University of Technology Poznan, Poland Lecture 7 SE Master Course 2008/2009

[9] Han J. and Kamber M. "Data Mining: Concepts and Techniques"

[10] Alsabti K., Ranka S., and Singh V." An Efficient K-Means Clustering Algorithm.", 1997

[11] Das S., Abraham A. and Konar A. "Document Clustering Using Differential Evolution" link.springer.com/content/pdf/bfm%3A978-3-540-93964-1%2F1.pdf

[12] Das S., Abraham A . and konar A., " Automatic Clustering Using an Improved

ISSN: 2348-4098
VOLUME 02 ISSUE 04 APRIL-MAY 2014

Differential Evolution Algorithm" IEEE VOL. 38, NO. 1, JANUARY 2008.

[13] Jusoh Shaidah and Hejab M. Alfawareh "Techniques, Applications and Challenging Issue in Text Mining" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012 ISSN.

[14] Xindong Wu, Kumar V. J. Ross Quinlan, Ghosh J, Yang. Q ,"Top 10 algorithms in data mining",Knowl Inf Syst (2008) 14:1–37 DOI 10.1007/s10115-007-0114,December 2007

[15] Malik. S. , Sharma N ., and Singh S. , "Evolving limitations in K-means algorithm in data mining and their removal" IJCEM (International Journal of Computational Engineering & Management), Vol. 12, April 2011

[16] Monz.C, "Delivered a lecture on Machine Learning for Data Mining", Week 6: Clustering at Queen mary university of London.

[17] Velmurugan T. and Santhanam T. "Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points" Journal of Computer Science 6 (3): 363-368, 2010 ISSN 1549-3636 © 2010 Science Publications.

[18] Pratap A ,R, Suvarna V.K., Devi R.J, Rao N.K. " An Efficient Density based Improved K- Medoids Clustering algorithm" International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[19] Polinati S.G, Babu V., Satapathy C. S, Pradhan G. "Effective Image clustering with Differential Evolution Technique", International Journal of computer and communication Technology, Vol. 2,No. 1.2010.