Binary Journal of Data Mining & Networking 5 (2015) 30-32
http://www.arjournals.org/index.php/bjdmn/index

Research Article

ISSN: 2229 -7170

Analysis And Implementation Of K-Mean And K-Medoids Algorithm For Large Dataset To Increase Scalability And Efficiency

Anjani Pandey[1], Mahima Shukla[2]

*Corresponding author:

Anjani Pandey

[1]Dept of Computer Engineering V.I.T.S, Satna (M.P)

[2]Dept of CSE (Software System) V.I.T.S, Satna (M.P)

Abstract

The experiments are pursued on both synthetic in data sets are real. The synthetic data sets which we used for our experiments were generated using the procedure. We refer to readers to it for more details to the generation of large data sets. We report experimental results on two synthetic more data sets in this data set; the average transaction of size and its average maximal potentially frequent item set its size are set, while the number of process in the large dataset is set. It is a sparse of dataset. The frequent item sets are short and also numerous data sets to cluster. The second synthetic data set we used is. The average transaction size and average maximal potentially frequent item set size of set to 30 and 32 respectively. There exist exponentially numerous frequent item data sets in this data set when the support based on threshold goes down. There are also pretty long frequent item sets as well as a large number of short frequent item sets in it. It process of contains abundant mixtures of short and long frequent data item sets.

Keywords: Clustring,Kmean,KMediod,Datamining

Introduction

A cluster is a collection of data objects are to be similar to one another within the same of cluster and are dissimilar to the objects in other in clusters. The process of based on grouping a set of abstract objects into classes of similar objects is called them clustering.

Clustering is a basically dynamic field in research in data mining. Many clustering algorithms have been developed. These can be categorized into partition method, hierarchical method, density based method, grid based method, and model based methods.

A partitioning method creates an initial set of number of partitions, where parameter k is the number of partitions to construct; then it uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. Typical partitioning methods include k-means, k-medoids, CLARANS, all so their improvements.

A hierarchical process creates a hierarchical decomposition of the given set of objects 0f data. The method can be classified as being either agglomerative (bottom up) or divisive (top down), based on how the hierarchical decomposition formed. To compensate for the rigidity of merge divide in the quality of hierarchical agglomeration can be its improved by analyzing object of linkages at each hierarchical partitioning (such as in CURE and Chameleon) or interesting other clustering techniques, such as iterative relocation (as in BIRCH)

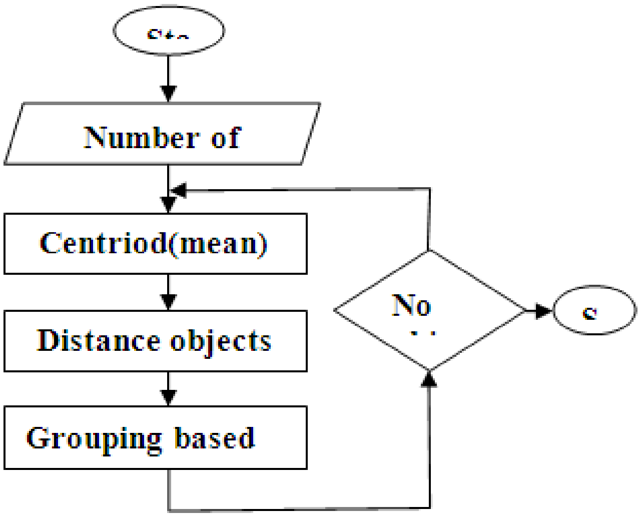
A density based method clusters objects based on the notation of density. It either grows cluster according to the density of neighborhood objects (such as DBSCAN) or according to some density function. OPTICS is a density based methods that generates an augmented ordering of the clustering structure of the data.

A grid based method first quantizes finite number of the object space into cells that form a grid structure, and then perform clustering on the grid structure. STING is a typical example of a grid-based method based on statistical information stored in grid cells. CLIQUE and Wave Cluster are two clustering algorithms that are both grid-based and density based.

A model-based method hypothesized a model for each of the clusters and finds the result of best fit in data to that model. Statistical Typical model based method involve approaches, neural network approaches [2].

K-Means Clustering Method

Given *k*, the *k-means* algorithm is implemented in 4 steps:
Partition objects into *k* non-empty subsets
Arbitrarily choose *k* points as initial centers.
Assign each object to the cluster with the nearest seed point (center).
All so calculate the mean of the any cluster and update its seed point.
Repeat back to Step 3, stop when no more new assignment



Figurre. 1

K-medoids algorithm

The K-Medoids algorithm is a clustering algorithm related to the K-means algorithm is that k-medoid shift algorithm. Both the K-means and K-medoids algorithms are partitioned (breaking the dataset up into groups) and both attempt to minimize squared error of 0.75 to calculate distance between points every point of labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the K-means algorithm K-medoids chooses data points as centers (medoids or exemplars). (Hae-Sang, Jong-Seok Lee and Chi-Hyuck Jun, [17]) K-Mediod is a classical partitioning technique of clustering that clusters the data set of n objects into k of clusters known as apriori.

It is more robust to noise and outliers as compared to k-means

A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal it is a most centrally located point in the given data set.

K-medoid clustering algorithm is as follows

1) The algorithm begins with arbitrary selection of the k objects in medoid points n data points (n>k)

2) The similarity here is defined using distance measure that can be Euclidean distance, when after selection of the k medoid points, associate each data object in the given data set to most similar medoid. Based on Manhattan distance [4].

3) Randomly select nonmedoid object O'

4) Compute total cost, S of swapping initial medoid object to O'

5) If S<0, then swap initial medoid with the new one (S<0)then there will be new set of medoids

6) Repeat steps 2 to 5 until there is no change in the medoid.

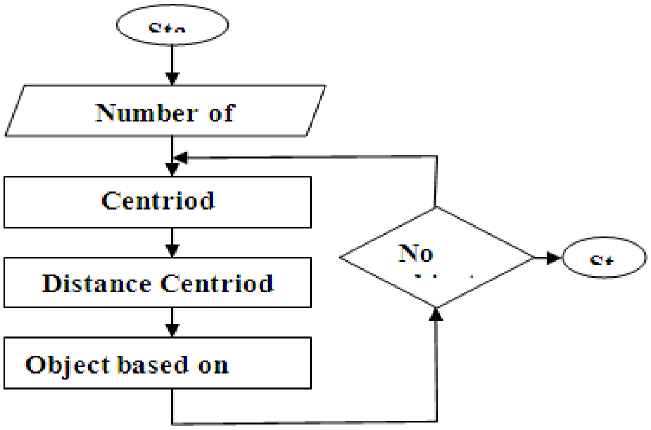


Figure: 2

Result

I .Comparison on the basis of varying number of Records

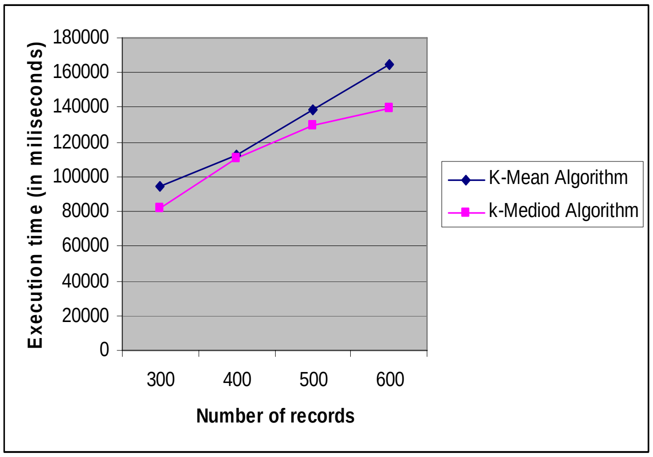


Figure: 3

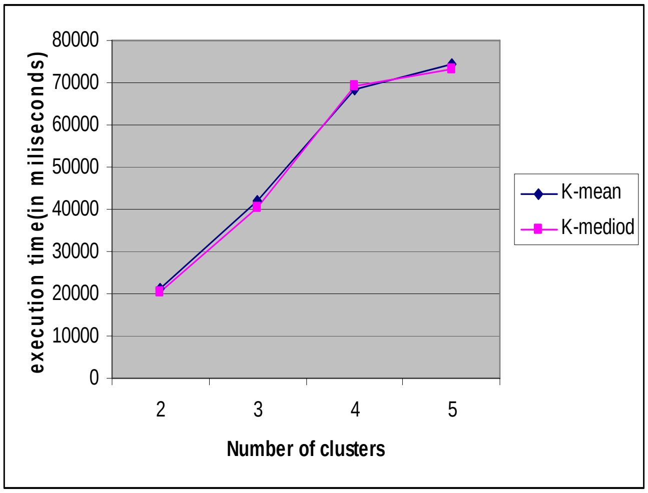


Figure.: 4

Conclusions

The result of their experiments illustrate that the random and the when initialization process of outperforms the rest of the compared methods as they make the K-means and K-mediod both are more effective independent on to the initial clustering and on instance of order. In this paper we present a new algorithm for *K*-means and k-mediod is based on optimization formulation and a novel method iterative. The rest of the paper is organized as follows. We present our proposed modified *K*-means clustering algorithm. Precise clustering centers in some sense are obtained main conditions and iteration methods. We present some simulation results with a group of randomly constructed sets data.

References

[1]. Dechang Pi, Xiaolin Qin and Qiang Wang. "Fuzzy Clustering Algorithm Based on Tree for Association Rules", International Journal of Information Technology, 2006;12, 3.

[2]. Fahim AM, Salem AM. "Efficient enhanced k-means clustering algorithm", Journal of Zhejiang University Science, 2006; 1626 – 1633,

[3]. Fang Yuag, Zeng Hui Meng. "A New Algorithm to get initial centroid", Third International Conference on Machine Learning and cybernetics, Shanghai, 26-29 August, 2004;1191 – 1193.

[4]. Friedrich Leisch1 and Bettina Gr un2. "Extending Standard Cluster Algorithms to Allow for Group Constraints", Compstat 2006, Proceeding in Computational Statistics, Physica verlag, Heidelberg, Germany.

[5]. MacQueen J. "Some method for classification and analysis of multi varite observation", University of California, Los Angeles, 281 – 297.

[6]. Maria Camila N, Barioni Humberto L, Razente Agma J, Traina M. "An efficient approach to scale up k-medoid based algorithms in large databases", 265 – 279.

[7]. Michel Steinbach, Levent Ertoz and Vipin Kumar, "Challenges in high dimensional data set", International Conference of Data management, 2005; 2, (3).

[8]. Parsons L, Haque E, and Liu H. "Subspace clustering for high dimensional data: A review", SIGKDD, Explor, Newsletter 6, 90 -105, 2004.

[9]. Rui Xu, Donlad Wunsch . "Survey of Clustering Algorithm", IEEE Transactions on Neural Networks,2005; 16, 3.

[10]. Sanjay garg, Ramesh Chandra Jain. "Variation of k-mean Algorithm: A study for High Dimensional Large data sets", Information Technology Journal 2006; 5 (6); 1132 – 1135,

[11]. Vance Febre, "Clustering and Continues k-mean algorithm", Los Alamos Science, Georgain Electonics Scientific Journal: Computer Science and Telecommunication,1994 ; 4, 3; 1.

[12]. Zhexue Huang. "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining".

[13]. Prof. Brian D. Ripley, "Study of the pure interaction dataset with CART algorithm", Professor of Applied Statistics

[14]. Yinmei Huang. "Classification and regression tree (CART) analysis: methodological review and its application", Ph.D. Student, The Department of Sociology, The University of Akron Olin Hall 247, Akron, OH 44325-1905,

[15]. Nathan Rountree. "Further Data Mining: Building Decision Trees", first presented 28 July 1999.

[16]. Yang liu. "Introduction to Rough Set Theory and Its Application in Decision Suppot System"

[17]. Wei-YIn loh. "Regression trees with unbiased variable selection and interaction detection", University of Wisconsin–Madison.

[18]. Rasoul Safavian S and David Landgrebe. "A Survey of Decision Tree Classifier Methodology", School of Electrical Engineering ,Purdue University, West Lafayette, IN 47907.

[19]. David S. Vogel, Ognian Asparouhov and Tobias Scheffer, "Scalable Look-Ahead Linear Regression Trees" .

[20]. Alin Dobra, "Classification and Regression Tree Construction", Thesis Proposal, Department of Computer Science, Cornell university, Ithaca NY, November 25, 2002.