Adaptive Threshold Based Clustering
A Deterministic Partitioning Approach

Mamta Mittal, Department of Computer Science and Engineering, G. B. Pant Govt. Engineering College, New Delhi, India
Rajendra Kumar Sharma, Thapar University, Patiala, India
Varinder Pal Singh, Thapar University, Patiala, India
Raghvendra Kumar, Department of Computer Science and Engineering, LNCT College, Indore, India

ABSTRACT
Partitioning-based clustering methods have various challenges especially user-defined parameters and sensitivity to initial seed selections. K-means is most popular partitioning based method while it is sensitive to outlier, generate non-overlap cluster and non-deterministic in nature due to its sensitivity to initial seed selection. These limitations are regarded as promising research directions. In this study, a deterministic approach which do not requires user defined parameters during clustering; can generate overlapped and non-overlapped clusters and detect outliers has been proposed. Here, a minimum support value has been adopted from association rule mining to improve the clustering results. Further, the improved approach has been analysed on artificial and real datasets. The results demonstrated that datasets are well clustered with this approach too and it achieved success to generate almost same number of clusters as present in real datasets.

KEYWORDS
Algorithm, Clustering, Data Mining, Deterministic, K-Means

1. INTRODUCTION
Exponential development of information assets has required new procedures that can change over it into helpful data. Clustering techniques of data mining investigate these information assets for concealed examples. Numerous clustering techniques exist in the literature. The work presented in this study accentuates on partitioning-based clustering techniques and is an endeavor for proposing new technique which can effectively recognize clusters in real databases. The K-means clustering algorithm, is widely accepted in all application areas. But, this algorithm is non-deterministic as it requires user specified initial centroids; it is sensitive to outliers; and it generates non-overlapped clusters. However, these aspects of k-means have not restricted the use of this algorithm; rather, motivated researchers for improving its functionality. K-means, k-medoid and single pass clustering algorithms are dependent on random determination of initial centroids or data objects. Due to this, these algorithms produce varying results on successive runs, so these are considered as non-deterministic algorithms.

There is a need of such algorithms in which random selection of initial data objects or centroids can be avoided. Besides this, for user defined parameters like threshold value or the number of clusters, there is always need of domain knowledge especially when it has to be done for the first time. These user defined parameters are generally a constant value, so there is the requirement of adaptive approaches.
in which parameters changes as algorithm executes. To overcome these limitations a deterministic partitioning-based algorithm, adaptive threshold based clustering (ATC) algorithm has been proposed.

This algorithm is deterministic in nature as it does not select the objects randomly and; rather it is based on selecting the farthest data objects. Moreover, changes in the order of selection of data objects do not affect the clustering results. It uses a parameter, Neighborhood distance which is not specified by the user, rather, it is calculated automatically and moreover, it is an adaptive parameter. This algorithm can generate both overlapped and non-overlapped clusters; and can detect outliers too. Another parameter used in ATC algorithm is the minimum support value which prunes the insignificant clusters. This parameter does not play any rule in clustering. Performance of the ATC algorithm is also assessed on artificial and real datasets and it has been observed that it is capable of detecting outliers and generates overlapped and non-overlapped clusters. It has an advantage over the single pass and modified single pass clustering algorithms as it does not rely upon the request of the choice of the data on objects. Here, it has also been observed from the experiments that ATC algorithm generates clusters of different sizes whereas k-means algorithm generates clusters of almost same sizes.

This paper has been systematized into five segments. Brief overview is presented in the first part. The second segment discusses the literature work regarding this field. The third segment presents assumption and prerequisite, parameters used, adaptive property, its illustration and the proposed algorithm. The fourth segment describes the experimental work carried out on real datasets for k-means and proposed technique and provides key observations. The last section presents the concluding remarks with future directions of the work in hand.

2. LITERATURE REVIEW

In 1960s, analysts utilized “Data Fishing” or Data Digging” for getting information from the repositories. They developed numerous methods to parcel it into various groups where objects in each group is similar to one another. For this purpose, MacQueen (1967) developed k-means method for acquiring k clusters, where each cluster is separated from other cluster based on the distance from the centroid and data objects. From that point onward, Salton and Wong (1971) acquainted single pass clustering for parcel it into disjoint sets. Duda and Hart (1973) portrayed how parceling the datasets is important in pattern recognition and acknowledgment. Dubes and Jain (1976) emphasized that the user should be aware of intrinsic properties of a clustering technique before using them. Lloyd (1982) presented k-means method for apportioning the datasets. In this, separation between the data objects and centroids is least as they are at the minimum distance. On the other hand, in 1987, Kaufman presented k-medoid method, depends on the hunt of k delegates called medoids for the provided databases.

Agarwal & Srikant (1994) initiated the field of association rule mining. They proposed a well-known apriori algorithm to generate frequent itemsets. Jain et al. (1999) have displayed a thorough overview on the clustering techniques. The k-means algorithm is widely accepted in the research community owing to its simplicity. But, analysts experienced with the accompanying issues in this approach: How many clusters are present in the dataset? How to decide what should be the seed values for the centroids? How to beat the impact of exceptions (noise) on the quality of clustering?

Keeping in mind the end goal to decide the quantity of groups, in 2001, Tibshirani et al. presented a gap statistic-based method. This method tells the distinction between the detected intra-group distance and predictable intra-cluster distance and decides ideal estimation of k. Further, issue of initialization for the centroids has been taken in 1994 by Katsavounidis et al. In this, the specific first centroid is in one corner of the dataset having most astounding standard. In 2003, Likas et al. displayed a global k-means approach, in that, initial centroids are acquired progressively in turn until the point where centroids becomes equivalent for the estimated value of k. Khan and Ahmad (2004) presented an approach for deciding preliminary centroids by concentrated upon estimations about discrete qualities.
Related Content

Managing the Quality of UML Models in Practice
www.igi-global.com/chapter/managing-quality-uml-models-practice/26823?camid=4v1a

www.igi-global.com/article/towards-ontological-approach-to-security-risk-analysis-of-information-system/160710?camid=4v1a
A Comparison of Stochastic Models for the Interarrival Times of Packets in a Computer Network
Dennis Guster, Semyon Litvinov, Mary Richardson and David Robinson (2002).
Optimal Information Modeling Techniques (pp. 248-257).
www.igi-global.com/chapter/comparison-stochastic-models-interarrival-times/27842?camid=4v1a

Information Systems and Software Development
www.igi-global.com/chapter/information-systems-software-development/75746?camid=4v1a