AN EFFICIENT DENSITY BASED IMPROVED K-MEDOIDS CLUSTERING ALGORITHM

ABSTRACT:

Clustering is the process of classifying objects into different groups by partitioning sets of data into a series of subsets called clusters. Clustering has taken its roots from algorithms like k-medoids and k-medoids. However conventional k-medoids clustering algorithm suffers from many limitations. Firstly, it needs to have prior knowledge about the number of cluster parameter k. Secondly, it also initially needs to make random selection of k representative objects and if these initial k medoids are not selected properly then natural cluster may not be obtained. Thirdly, it is also sensitive to the order of input dataset.

Mining knowledge from large amounts of spatial data is known as spatial data mining. It becomes a highly demanding field because huge amounts of spatial data have been collected in various applications ranging from geo-spatial data to bio-medical knowledge. The database can be clustered in many ways depending on the clustering algorithm employed, parameter settings used, and other factors. Multiple clustering can be combined so that the final partitioning of data provides better clustering. In this paper, an efficient density based k-medoids clustering algorithm has been proposed to overcome the drawbacks of DBSCAN and k-medoids clustering algorithms. The result will be an improved version of k-medoids clustering algorithm. This algorithm will perform better than DBSCAN while handling clusters of circularly distributed data points and slightly overlapped clusters.

INTRODUCTION:

Numerous applications require the management of spatial data, i.e. data related to space. Spatial Database Systems (SDBS) (Gueting 1994) are database systems for the management of spatial data. Increasingly large amounts of data are obtained from satellite images, X-ray crystallography or other automatic equipment. Therefore, automated knowledge discovery becomes more and more important in spatial databases. Clustering algorithms are attractive for the task of class identification. However, the application to large spatial databases raises the following requirements for clustering algorithms:

(1) Minimal requirements of domain knowledge to determine the input parameters, because appropriate values are often not known in advance when dealing with large databases.

(2) Discovery of clusters with arbitrary shape, because the shape of clusters in spatial databases may be spherical, drawn-out, linear, elongated etc.

(3) Good efficiency on large databases, i.e. on databases of significantly more than just a few thousand objects. Clustering is considered as one of the important techniques in data mining and is an active research topic for the researchers.

The objective of clustering is to partition a set of objects into clusters such that objects within a group are more similar to one another than patterns in different clusters. So far, numerous useful clustering algorithms have been developed for large databases, such as K-MEDOIDS, CLARANS, BIRCH, CURE, DBSCA, OPTICS, STING and CLIQUE. These algorithms can be divided into several categories. Three prominent categories are partitioning, hierarchical and density-based. All these algorithms try to challenge the clustering problems treating huge amount of data in large databases. However, none of them are the most effective. In density-based clustering algorithms, which are designed to discover clusters of arbitrary shape in databases with noise, a cluster is defined as a high-density region partitioned by low-density regions in data space. DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a typical Density-based clustering algorithm. In this paper, we present a new algorithm which overcomes the drawbacks of DBSCAN and k-medoids clust

EXISTING SYSTEM:

The objective of clustering is to partition a set of objects into clusters such that objects within a group are more similar to one another than patterns in different clusters. So far, numerous useful clustering algorithms have been developed for large databases, such as K-MEDOIDS, CLARANS, BIRCH, CURE, DBSCAN, OPTICS, STING and CLIQUE. These algorithms can be divided into several categories. Three prominent categories are partitioning, hierarchical and density-based. All these algorithms try to challenge the clustering problems treating huge amount of data in large databases. However, none of them are the most effective. In density-based clustering algorithms, which are designed to discover clusters of arbitrary shape in databases with noise, a cluster is defined as a high-density region partitioned by low-density regions in data space. DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a typical Density-based clustering algorithm.

PROPOSED SYSTEM:

The proposed clustering and outlier detection system has been implemented using Weka and tested with the proteins data base created by Gaussian distribution function. The data will form circular or spherical clusters in space.

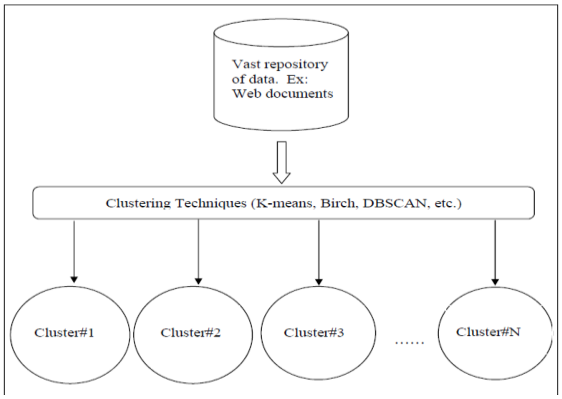
DIFFERENT TYPES OF CLUSTERING ALGORITHMS:

Clustering can be done in many different ways; each clustering technique produces different types of clusters. Some take input parameters from the user like number clusters to be formed etc, but some decide on the type and amount of data given. The main developments have been the introduction to density based and grid based clustering methods. Clustering algorithms can be classified into five distinct types:

- Partitioning methods;
- Hierarchical methods;
- Model-based methods;
- Density based methods; and
- Grid based methods.

MODULES:

- DBSCAN
- Optics
- K-means
- K-Medoids


Vast repository of data. Ex: Web documents
Clustering Techniques (K-means, Birch, DBSCAN, etc.)
Cluster#1
Cluster#2
Cluster#3
……
Cluster#N

Architecture of clusters

CONCLUSION:

The main objective of this thesis was to survey the most important clustering algorithms and determine which of them can be used for clustering large datasets. Extending or improving basic models of clustering as discussed in chapter 3 can help in some ways to deal with large datasets but the most successful clustering methods stored summary statistics in trees. Building a tree requires only single scan of data and inserting a new object into an existing tree is usually very simple. By limiting the amount of memory available in the tree building process, it is possible for the tree to adapt to fit into main memory. This thesis focuses on inspection of most important clustering algorithms and further we have discussed the key concepts that allow the current clustering methods to manage very large datasets. Determining clusters of arbitrary shape, identifying outliers as sparse regions and providing computational speed-ups through ignoring sparse regions of the data space were the essential steps found in most of the current clustering methods.

An optimally efficient tree-based data structure should be ascertained for clustering problems. Multi-resolution clustering techniques (i.e. ability to detect clusters with in a cluster) need to be formalized. The ability to cluster data arriving in a constant stream should be considered. Tree-based data structures within the online systems should be explored as they are likely to be very effective. The below is a list of all the clustering methods and their corresponding run times along with other specifications.

REFERENCES:

1. Clustering Large Datasets by D. P. Mercer

http://www.stats.ox.ac.uk/~mercer/documents/Transfer.pdf

2. Amit Singhal, ‘Modern Information Retrieval: A Brief Overview’, IEEE Data Engineering Bulletin, Volume 24, pages 35-43, 2001.

3. C. J. Van Rijsbergen, ‘Information Retrieval’, Second Edition, Chapters 1, 6 & 7, Information Retrieval Group, University of Glassgow.

TEAM MEMBERS:

GOWRI.M

KALYANA SUNDARI.R

KALEESHWARI.R

VASAVIANANDHA LAKSHME.S.S