DCAD: a Dual Clustering Algorithm for Distributed Spatial Databases

ZHOU Jiaogen  GUAN Jihong  LI Pingxiang

Abstract  Spatial objects have two types of attributes: geometrical attributes and non-geometrical attributes, which belong to two different attribute domains (geometrical and non-geometrical domains). Although geometrically scattered in a geometrical domain, spatial objects may be similar to each other in a non-geometrical domain. Most existing clustering algorithms group spatial datasets into different compact regions in a geometrical domain without considering the aspect of a non-geometrical domain. However, many application scenarios require clustering results in which a cluster has not only high proximity in a geometrical domain, but also high similarity in a non-geometrical domain. This means constraints are imposed on the clustering goal from both geometrical and non-geometrical domains simultaneously. Such a clustering problem is called dual clustering. As distributed clustering applications become more and more popular, it is necessary to tackle the dual clustering problem in distributed databases. The DCAD algorithm is proposed to solve this problem. DCAD consists of two levels of clustering: local clustering and global clustering. First, clustering is conducted at each local site with a local clustering algorithm, and the features of local clusters are extracted. Second, local features from each site are sent to a central site where global clustering is obtained based on those features. Experiments on both artificial and real spatial datasets show that DCAD is effective and efficient.

Keywords  distributed clustering; dual clustering; distributed spatial database
CLC number  P208

Introduction

Clustering is one of the most important techniques in knowledge discovery and the data mining field, which is widely applied to fields that include biological information query\cite{1-2}, image processing\cite{3}, pattern recognition\cite{4}, trend analysis\cite{5}, Web mining, and document classification\cite{6}. Clustering is to partition the input data set to different groups according to some criteria, where intra-clusters are maximally similar, and inter-clusters are minimally similar\cite{7}.

In GIS application, spatial objects often have two types of attributes, i.e., geometrical attributes (e.g., geographical coordinate, latitude) and non-geometrical attributes (e.g., population density, economic index and pollution index).

Although spatial objects may geometrically scatter in a geometrical domain, they are possibly similar to each other in a non-geometrical domain. Most existing clustering algorithms group spatial datasets into different compact regions in a geometrical domain.
without considering the aspect of a non-geometrical domain. However, many applications require clustering results in which a cluster not only has high proximity in a geometrical domain, but also high similarity in a non-geometrical domain. That is, constraints are imposed on the clustering goal from both geometrical and non-geometrical domains simultaneously. Such a clustering problem is called dual clustering and is illustrated by Fig.1.

![Fig.1 Projections of a spatial dataset in geometrical and non-geometrical domains](image)

Distributed applications are becoming more and more popular, such as distributed mobile networks, sensor networks, and supermarket chains. How should dual clustering be conducted in distributed applications, where large amounts of heterogeneous, complex data reside on different, independently working computers that link to each other by local or wide area networks? Because of limited bandwidth and security concerns, we could not expect to conduct the clustering task on a central site when the whole dataset is amassed from the local sites. However, it is possible that we first cluster locally and extract some proper features from the local clusters, and then transfer those features to a central site to obtain the global clustering results.

In this paper, we explore a solution to dealing with dual clustering in distributed applications. We propose a new algorithm, DCAD (dual clustering algorithm for distributed databases) for solving this problem. We first cluster datasets residing on local sites with a local clustering algorithm, i.e., a density-based and statistical clustering algorithm that allows spatial datasets to form compact regions in geometrical domains and have maximal similarity in non-geometrical domains. We then extract proper features from the created local clusters and send these features to a central server site, where the local features are used to build global clusters based on a global clustering algorithm.

1 Related work

Roughly, we can classify the existing clustering algorithms into centralized clustering algorithms and distributed clustering algorithms. Centralized clustering algorithms run on a single computer, while distributed clustering algorithms work on different computers connected to each other by networks and finish the clustering task collaboratively.

There are four main types of centralized clustering algorithms: hierarchical clustering algorithms, partitioning clustering algorithms, density-based clustering algorithms and grid-based clustering algorithms.

A typical hierarchical algorithm yields a dendrogram representing the nested grouping of patterns underlying the datasets. According to the formation of the dendrogram, hierarchical algorithms are divided into agglomerative (down-top) and divisive (top-down) hierarchical clusters. BIRCH\textsuperscript{\cite{8}} is an integrated hierarchical clustering algorithm which uses clustering features and a clustering feature tree (CF tree) to represent clusters. This algorithm is relatively efficient in clustering large databases and effective for incremental and dynamic clustering of incoming objects. However, it can only find spherical clusters.

The partitioning clustering algorithms start with a random partition of data and then use an iterative and controllable strategy to optimize a pre-specified evaluation function. There are two types of partitioning algorithms: the $k$-means algorithm and $k$-medoids algorithm. CLARANS\textsuperscript{\cite{9}} is an improved $k$-medoids algorithm that overcomes most disadvantages of traditional clustering methods on large datasets. However, it is still inefficient with a computational complexity of $O(kN^2)$, and could not guarantee the quality of clustering results due to its use of data sampling.

Density-based clustering algorithms form clusters