ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

SPATIAL DATA MINING USING CLUSTERING TECHNIQUES

M.V.B.T. Santhi,

Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, India Email: santhi_ist@kluniversity.in

Abstract:

These days, there is no shortage of clustering approaches available, all of which are employed exclusively for the purpose of spatial data mining. A few of them use K-means, Clarans, DBscan, and numerous more techniques. Similar to this, there is a distributed dynamic clustering technique that uses local locations for clustering and then aggregates the local clusters that are obtained at these sites. Every local site produces local clusters, which are then sent to the global site for aggregation. In the aggregation step, local clusters that have been obtained from local sites are combined and connected to create global clusters. The current techniques cause great confusion in the aggregate step. The contour algorithm is used to determine the borders of the local clusters that are obtained during the parallel phase, which is followed by the aggregation phase. Global clusters are framed by the accumulation of local clusters. That being said, the aggregation process is incredibly intricate. This work modifies the technique to separate the aggregation and also does a comparison with conventional clustering algorithms.

Keywords: Spatial Data, clustering, data mining, distributed data.

1. Introduction

Spatial data mining using clustering techniques is a dynamic field that leverages advanced computational methods to uncover meaningful patterns and relationships within spatially referenced data sets. This interdisciplinary approach combines principles from data mining and geography to extract valuable insights from geospatial information. Clustering, a fundamental technique in spatial data mining, plays a pivotal role in identifying groups or clusters of spatial entities with similar characteristics, thereby facilitating the discovery of

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

hidden structures within geographic data. By employing sophisticated algorithms and statistical methods, spatial data mining with clustering techniques contributes to enhanced decision-making processes in diverse domains such as urban planning, environmental management, and public health, offering a powerful toolset for understanding the complex interplay of spatial attributes and uncovering valuable knowledge from geospatial datasets.

2. Literature Survey

Massive amounts of data are collected daily in this world and saved in databases. We use data mining techniques to retrieve useful information. We learned from [2] and [11] that the automated data gathering techniques have led to a massive build-up of data that is hidden in a number of databases or information repositories. The problem of the data explosion is inferred from this. We are in debt of knowledge even though we possess a wealth of facts. The original purpose of data mining was to address this issue of data explosion. When data mining was first established, there was no such thing as an idea of combining several methodologies. Data mining, which is defined as the process of uncovering patterns and obtaining knowledge from massive databases, is a crucial step in KDD and is included in [3].

A handful of the phases that make up the KDD process are listed below. The preprocessed data is obtained when the selection data is first extracted from flat files or data files and cleaned up to remove any missing values. Afterwards, by producing helpful patterns, the data is modified for data reduction and projections. To increase the knowledge, the patterns are subsequently subjected to interpretation or evaluation. From now on, data mining is done in this manner. Numerous applications and diverse domains can benefit from data mining. This paper's main goal is to use several hierarchical clustering approaches for geographical data mining.

2.1 Spatial Data Mining

The process of extracting knowledge from spatial databases is called spatial data mining. The information that has been obtained is used to distinguish between geographic and non-spatial data, examine their relationship through inspection [6], and evaluate additional data derived from the same. Information that represents the spatial data can be found in the spatial databases. Information about objects that can be numerically represented in a geographic

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

coordinate system is provided by spatial data. In addition to being represented by multiple topologies and coordinate systems, spatial data is used for mapping. The following are the few stated objectives for spatial data mining: to identify the geographical trends. to identify the spatial objects that are also the source of spatial patterns. to decide whether to reveal or explain the derived spatial pattern's information. to naturally convey the facts and also need to be enhanced for further analysis. Several tasks, including classification, clustering, characteristic rules, discriminant rules, and association rules, are involved in spatial data mining. GIS (Geographic Information System) and image processing both make use of spatial data. While clustering is an unsupervised learning process, classification is supervised learning. Class labels are defined in classification, but they are not defined in clustering. For the level under consideration, characteristic rules specify the attributes of the objects in relation to their closest counterparts. Contrasts between the items taken into account for the spatial data are defined by discriminate rules. To identify recurring patterns, correlations, and associations for the associated data, association rules are used.They have considered support, confidence, minimum support and minimum threshold keeping in mind the end goal of relating patterns to the objects considered for spatial data.

2.2 Cluster Analysis

The technique of determining whether or not an object is related to a certain cluster is known as cluster analysis. Examined from [7], [8], it can be shown that the process of clustering consists of putting together objects with comparable attributes. There are no class labels on the items in this unsupervised learning process. Among the collection of objects that are taken into consideration for grouping of objects based on attributes, clustering is used to identify objects. We discuss two ways to cluster the objects. For these clustering techniques, there are unique algorithms: K-means, PAM, CLARA, and CLARANS, which are part of the partitioned clustering technique. In contrast, the hierarchical clustering method includes the CURE, BIRCH, CHAMELEON, and ROCK algorithms. The algorithms DBSCAN, OPTICS, and DENCLUE belong to the density-based clustering technique. Grid-based clustering techniques include the CLIQUE and STING algorithms. Applications for clustering can be found in the fields of marketing, insurance, land use, city planning, earthquake research, and other areas. The purpose of the paper we studied from [5], [10] was to cluster spatial data that is dispersed throughout the space acquired in certain geographic databases using the

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

Chameleon algorithm of hierarchical clustering approach. There are two approaches in the Chameleon algorithm: the Agglomerative method and the Devise method.

While the Devise method has a top-down perspective, the Agglomerative method adheres to a bottom-up approach [8]. In the agglomerative technique, each object in the set of objects to create a cluster is treated as a cluster in and of itself. For the objects acting as clusters, the nearest distance between them is determined, and then the objects are grouped together to form a cluster. Using the divide and conquer strategy, the devise method separates a set of items into groups based on similar qualities after first treating the objects as a single cluster.

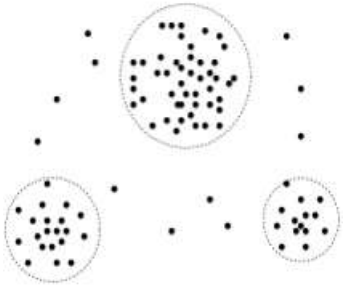


Figure 1 Clusters grouped from data

2.3 Distributed Data mining

This area of data mining organizes dispersed data by paying close attention to the data and taking resources into account. It is used to glean insights and patterns from large, dispersed information. For sophisticated and complicated data kinds, including geographical temporal data from [13], focal point is intended to be predominant. The goal of distributed mining is to improve performance and scalability in situations when there is a bottleneck. It also aims to integrate non-trivial data, which was previously deemed unrealistic. It offers the special ability to divide large datasets into smaller ones for improved scalability while making use of processing resources. When data from multiple geographic places needed to be preprocessed, evaluated, surveyed from [4] and so on, distributed mining's expertise became apparent. The majority of the methods use a two-phase methodology. Local computation and global aggregation are the two stages. The local computation step is the first one. Every site is involved in the creation of unique outcomes. The results of the several phases are relayed to the central site or to every other site connected to the network as they are produced at locally autonomous locations, with the goal of computing the ultimate result. After obtaining the

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

local findings from different individual sites, the global site aggregates these results to get a final result. The methods that different websites employ to obtain the local results may differ from one another. Since the two sites are independent of one another, one local site may employ an unsupervised algorithm and the other a supervised learning technique. It is preferable for productivity reasons if all local sites employ the same learning strategies. To obtain the final result, the global site may combine the local findings using an algorithm of its own.

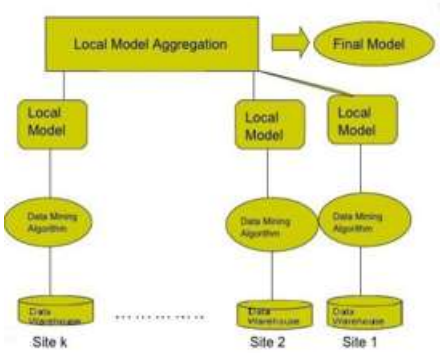


Figure 2 Distributed DataminingFramework

2.4 Spatial Distributed Clustering

The objective of this sample challenge is to determine the homogenous group of objects based on attribute values. Explicit placement and implicit bonding of spatial objects in their locale are generalized in this type of grouping. Like other distributed data mining methods, the spatial distributed clustering also uses a two-step process.

There are two parts in this process:

i) First, the sub dataset that is present on each site is used to create local clusters.

ii) Local clusters created at each specific site provide the basis for the formation of global clusters, which are analyzed from [1].

Density-based, partition-based, or hierarchical clustering techniques are the three types of algorithms that a local site can employ to generate clusters. Let's assume that the K-means clustering technique, which is a partition-based clustering algorithm, is being used at the local sites.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

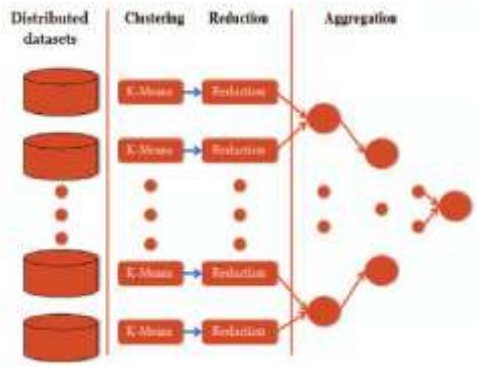


Figure 3 Overview of the approach

3. Implementation

The K-means computation examined through and [14] can be used to do local bunching or clustering in the first step, also referred to as the parallel stage. To deliver Ki adjacent bunches, each hub or node (di) applies K-means on its native data. We determine their limits once all neighboring groupings have been determined. These borders will function as related groups' agents. The exchange of each node's local boundary with its surrounding hubs is the next step in the process. This makes it possible to determine whether or not there are any overlapping bunches, or clusters. Each leader attempts to consolidate the covering clusters of their group in the third stage. Every class's encircled hubs are where the leaders are picked. Following the resolution of the leaders, each leader forms new clusters, or bunches. Until the root hub is reached, the second and third steps are repeated. After the sub-results are aggregated, a tree is created, with the final results being located at the root node. The typical significant variability in cluster forms and densities is a problem, as it is in all cluster computations.

1) Every node will be assigned as a part of the dataset or of the general dataset.
2) The existing leaf hub (ni) carries out the K- means function with Ki variable on its local dataset.
3) The adjacent or the nearest hubs have to allocate their clusters or bunches in order to frame considerably bigger clusters utilizing theoverlay strategy.
4) The outcomes are made to be stayed in the parent hub (also known as precursor).
5) Repeat the bottom two steps until and unless the root node is acquired.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

4. Experimental Results

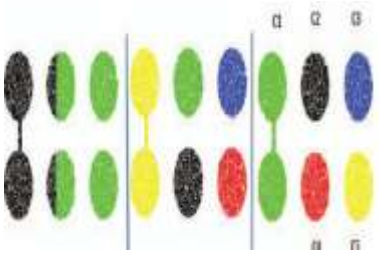
This section discusses the outcomes and compares them with more conventional clustering techniques, such as k-mediods, cure, and birch. It also contains a performance study of the distributed dynamic clustering technique.

It creates several partitions for K-MEDIODS and determines how many clusters to produce by calculating the centroid based on a set of criteria. Each data point is assigned to the cluster with the closest mediod value once the mediods are first computed. Following the data point assignment, the cluster's new median is once more determined and updated appropriately. Until the convergence requirements are satisfied, this process is repeated.

This method uses archetypal points for CURE, where the data points shrink toward the mean. Rather of taking into account every point in the combined groups, the characteristic points for the current combined group are selected from one of the two unique groups when two groups are converged in every phase of the algorithm.

The algorithm operates in two stages for BIRCH. Pre-clustering is the first phase, while hierarchical clustering based on centroids is the second. The total number of points obtained following the pre-clustering phase, or step 1, determines the time and space complexity.

Additional extensions can be carried out by comparing the time complexity of the methods or the cluster morphologies of different algorithms to determine which is more practical while mining large datasets. These criteria can be further expanded by evaluating their scalability and performance across a range of current platforms.



Birch Cure DDCA

Figure 4 Clusters obtained for dataset

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

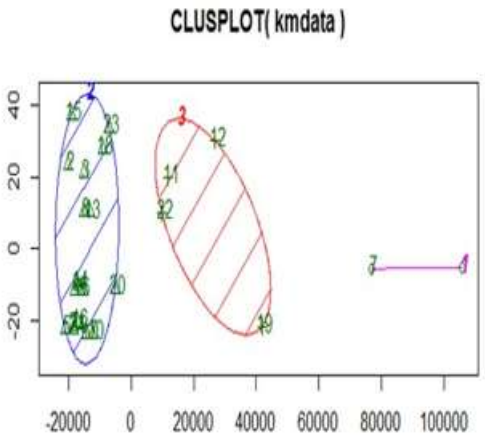


Figure 5 Clusters obtained from K-mediods(k=3)

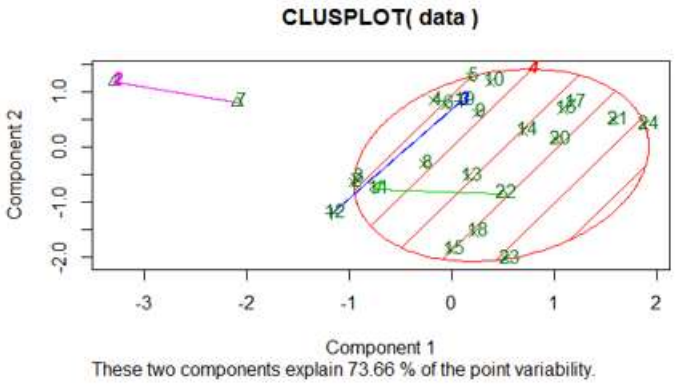


Figure 6 Clusters obtained from K-mediods(k=4)

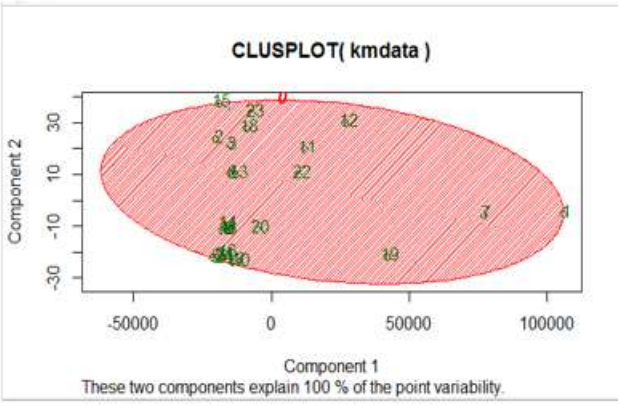


Figure 7 Cluster obtained from DBScan forthe same dataset

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

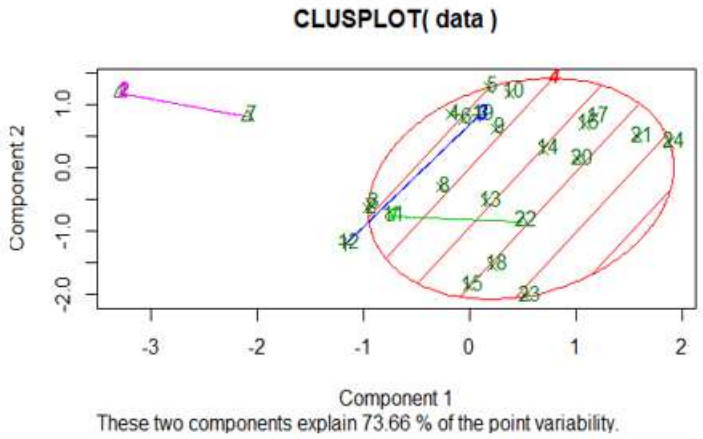


Figure 7.5 Cluster obtained from DDCA

In the case of distributed dynamic clustering algorithm, k-means algorithm is used on every individual site to get local clusters. Then these results are passed onto neighboring sites and integrated to obtain the global clusters as the output. It has been observed that the time complexities to obtain the clusters vary in different algorithms. The time complexity in DDCA algorithm is relatively less compared to the existing algorithms because the task is subdivided to be performed at individual sites and then aggregated. So the sub dataset would take less timeto group themselves as clusters.

5. Conclusion

The distributed dynamic clustering method we employed in this paper makes it easier to aggregate local clusters into a global cluster. We have taken into consideration a dataset obtained in accordance with a survey carried out in London, which was based on the sports activities that different people from the cities were participating in on that specific day. The dataset is transformed into a format with comma-separated values, which is then used as an input by the distributed dynamic clustering algorithm and conventional clustering algorithms like k-means, k-mediods, birch, cure, and dbscan. These algorithms produce clusters as their output, and the resulting results are compared to one another for performance analysis in terms of both space and temporal complexity. Every research result is enumerated and discussed. Moreover, it is shown that the distributed dynamic clustering algorithm preserves the quality of the clusters being formed while still outperforming the current clustering techniques. A thorough investigation is still pending. We plan to explore further with alternative techniques and explore the possibilities of extending to many large-scale distributed datasets.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

6. References

[1] Bendechache,Malika,andM- TaharKechadi"Distributed clustering algorithm for spatial data mining", 2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2015.

[2] Nedunchezhian, Raju Anbumani, Kalirajan. "Post mining-discovering valid rules from different sized data sources." International Journal of Information Tec, Jan 2006 Issue.

[4] Moez Ben Haj Hmida. "Meta-learning in grid- based data mining systems", International Journal of Communication Networks and Distributed Systems, 2010.

[5] Birant, D.. "ST-DBSCAN: An algorithm for clustering spatial-temporal data", Data & Knowledge Engineering, 200701.

[6] Wang, Shuliang, Deren Li, Qing Zhu, Yaolin Liu, and Shuliang Wang. "", MIPPR 2005 Geospatial Information Data Mining and Applications, 2005.

[9] Rafal A. Angryk. "Distributed Document Clustering Using Word-clusters", 2007 IEEE Symposium on Computational Intelligence and Data Mining, 04/2007.

[10] Gallego, Guillermo, Anthony Yezzi, Francesco Fedele, and AlviseBenetazzo. "Two Variational Stereo Methods for Space-Time Measurements of Ocean Waves", Volume 5 Ocean Engineering, 2013.

[11] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge discovery and data mining:Towards a unifying framework," in Proc. KDD-96, 1996, pp. 82–88,.

[12] M. Bertolotto, S. Di Martino, F.Ferrucci, and M-T. Kechadi, "Towards a framework for mining and analysingspatio-temporal datasets," International Journal of Geographical Information Science – Geovisual Analytics for Spatial Decision Support, vol. 21, pp. 895-906, January 2007.

[14] A. A. Freitas and S. H. Lavington, Mining ssvery large databases with parallel processing. 1st edition, Springer; 2000 edition, 30 November 2007.

[15] L-M. Aouad, N-A. Le-Khac, and M-T. Kechadi, "Grid-based approaches for distributed data mining applications," algorithms Computational Technology, vol. 3, pp. 517–534, 10 Dec. 2009.

[16] M.Sreemaa, S.Rama, A.Sivaranjini, "Data Embedding-Without Distortion In Video File", International Innovative Research Journal Of Enginerring And Technology, vol. 2, pp.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 1, 2019

34-38,September 2016.