International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-3 Issue-5 September-2012

A Review of Density-Based clustering in Spatial Data

Pragati Shrivastava[1], Hitesh Gupta[2]
Department of Computer Science & Engineering, PCST, Bhopal[1,2]

Abstract

Data mining is a non-trivial process. That is identifying novel, valid and potentially useful patterns in data. Data mining supports automatic data exploration. That is extracting hidden information from the huge database. Data mining refers to search useful and relevant information from the database. Spatial mining is a branch of data mining. The spatial mining deals with the location or geo-referenced data. Spatial mining are based on the density based clustering. Density is covered area of any data.

Keywords

Data mining, Spatial data, Density-based clustering

1. Introduction

Data mining is a non-trivial process of identifying valid, novel, potential useful and ultimately understandable patterns in data. Data mining, the extraction of the hidden predictive information from large database, is a powerful new technology with potential to analyze important information in the data warehouse. The term 'data mining' refers to the finding of relevant and useful information from database.

Data mining [1] is used everywhere and large amounts of information are gathered: in business, to analyse client behavior or optimize production and sales [2]. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency networks, analyzing changes, and detecting anomalies. Data mining is the search for the relationship and global patterns that exist in large database but are hidden among vast amount of data, such as the relationship between patient data and medical diagnosis. This relationship represents valuable knowledge about the database, if the database is a faithful mirror of the real world registered by the database.

Spatial data mining is the branch of data mining that deals with spatial (location, or geo-referenced) data. The knowledge tasks involving spatial data include finding characteristic rules, discriminate rules, association rules; etc. A spatial characteristic rule is a general description of spatial data. A spatial discriminate rule is a general description of the features discriminating or contrasting a class of spatial data from other class. Spatial association rules describe the association between objects, based on spatial neighborhood relations. We can associate spatial attributes with spatial attributes, or spatial attributes with non-spatial attributes.

DBSCAN (Density Based Spatial Clustering of Application of noise).DBSCAN uses a density-based notion of clusters to discover clusters of arbitrary shapes. Density-based algorithm is another major clustering algorithm that has been purposed [3], [4]. The key idea of DBSCAN is that, for each object of a cluster, the neighborhood of a given radius has to contain at least a minimum number of data objects.

2. Related Work
Density-Grid Clustering Algorithms

In this section, we review the density-grid based clustering algorithms on data streams. In the density-grid based clustering algorithms, the data records are mapped into grid structure and cluster the grid based on density.

One of the common characteristics of reviewed algorithms is that the majority of them are based on CluStream framework The algorithms have online and offline phase for clustering. In the online phase, the algorithm record summary information about the data records and the offline phase perform clustering on synopsis information. Algorithms on clustering data streams can be categorized into two groups: one-pass approach and evolving approach. The one-pass approach clusters data stream by scanning the data stream only once, and under assumption that the data objects arrived in chunks such as DUCStream [5] in this paper. In the evolving approach the data streams are considered to be changing over time. There are three kinds of window models in evolving approach include landmark window, sliding window, and fading window [6]. Another common characteristic of reviewed algorithms is that they are based on

fading model which is the relevance of the data diminish over time.

A. DUCStream

Gao et al. [5] developed an incremental single pass clustering algorithm for data streams which is referred to as DUCStream. It has the ability to detect evolving clusters in limited memory and time. DUCStream is a single pass clustering algorithms in which the data objects arrive in chunks. Each chunk fits in the main memory and contains a number of data points. It partitions the data space into units and keeps only the units with large number of data points. The density of a grid is defined by the number of the data points in the grid and if it is higher than density threshold, it is considered as a dense unit.

Local dense unit concept is introduced for determining which unit should be maintained. The local dense unit is a candidate for dense unit which may become a dense unit. Therefore, DUCStream keeps the entire local dense unit and chooses the dense unit between them to do the clustering. The clustering results are shown by bits to reduce the memory requirements. "Clustering Bits" of a cluster shows the number of dense unit in the cluster by one and zero for dense and non-dense unit. For clustering data stream, DUCStream identifies the clusters as a connected component of a graph in which vertices represent the dense units and the edges are related to common attributes between two vertices. It uses depth first search algorithm in graph. The time complexity and memory space of the DUCStream is low due to utilizing the bitwise clustering.

B. D-Stream I

Chen et al. [7] proposed a framework for clustering data streams, which is referred to as D-Stream I. The framework is based on the observation that many clustering algorithms on data streams cannot find clusters of arbitrary shape or handle the outliers. The idea is using density-grid approach for clustering data streams. The algorithm procedure could be described as follows:

D-Stream I has online-offline components. The online component reads new data record, maps each input data record into a density grid and update the characteristic vector which records summary information about the grid. The offline component clusters the density grids by merging two dense neighboring grids. A grid cluster is a connected grid group which has higher density than the surrounding grids. For removing outliers, D-Stream I periodically detects sporadic grids mapped to by outliers. It also runs the offline component occasionally in order to adjust the clusters.

As D-Stream I is based on fading model [8] of data stream, it considers weight for each data record which is decreased as data record ages. The density of the grid is defined as sum of the weight of all data records in the grids. If no data record is added to this grid, the density of grid decrease over the time. Based on grid density, dense and sparse grid is introduced. Their differences referred to their density. Dense and sparse grid is defined as follows:

Definition 1: Dense Grid at time t, for a grid g, is defined as follows: $(g, t) \geq Cm/(1-\lambda) = Dm$, $Cm > 1$Definition 2: Sparse Grid at time t, for a grid g, is defined as follows: $(g, t) \geq Cl/(1-\lambda) = Dm$, $0 < Cl < 1$Where Cm and Cl controls the threshold because the density value could not be more than $1/(1-\lambda)$ according to [5]. The (g, t) is the density of the grid, it is defined as $(g, t) = \sum_{x \in E(g,t)} D(x, t)$ and $D(x, t) = \boldsymbol{\lambda}^{t-T(x)} = \boldsymbol{\lambda}^{t-T}$, where $\lambda \in (0, 1)$ is a constant called the decay factor (capture the dynamic changes of a data stream). N is the number of grids.

D-Stream I puts the grids under consideration on the grid list as hash table and checks the list in special time intervals. If the density value of a grid becomes lower than special density threshold, it will be removed from the grid list. Chen et al. showed that D-Stream I improves the time complexity and quality of clustering compared to CluStream. When a new data record arrives, D-Stream I needs to update the summary information of the grid which is the new data is mapped to it. Hence, the time complexity is O(1).

C. DD-Stream

Jia et al. in [9] proposed a framework called DD-Stream for density-based clustering of data streams in grids. They developed an algorithm, DCQ-means, for improving quality of clustering by considering the border points of the grids. The framework is online-offline phase in which the online phase reads the new data records and maps to the grids and the offline phase perform the clustering on the grids using DCQ-means algorithm. DCQ-means algorithm extracts the data points on the border of the grid and joins boundary data points in the grid before adjusting the cluster. In DCQ-algorithm if the data is located in the border of two or more grids, it uses the most direct distance from the center of these grids to determine which data point belongs to which grid. In order to determine the distance between the data points and the neighboring grid, the eigenvector of the grid is

defined for keeping a record of the central grid. If the distance of the grid is the same for more than one neighboring grids, the data point will be added to the grid with the higher density. If the neighboring grids have same density, the data point is added to the grids with the latest updates.

Jia et al. in [9] show that by extracting the border points, their algorithms has lower time complexity in comparison to CluStream and yet it has better scalability.

D. D-Stream II

In [10], Tu et al. improved the D-Stream II by considering the positional information about the data. They address this issue by introducing the grid attraction concept which shows to what extend the data in one neighbor is closer to another neighbor. It has attraction based mechanism to generate cluster boundaries. The clustering procedure of D-Stream II is the same as the D-Stream I; the only difference is that before merging two grids D-Stream II checks the grid attraction of two grids. If the grid attraction is higher than threshold, they are strongly correlated, and the grids will be merged.

By considering both density and attraction they generate better result for clustering. D-Stream II keeps the grid list in black red tree which improves the running time for lookup and update. The space complexity is $O(\log_{1/\lambda} N)$, and yet time complexity is $O(\log \log_{1/\lambda} N)$ for looking up in the grid list (N the total number of grids and λ is the decay factor).

E. PKS-Stream

Ren et al. in [11] proposed an algorithm for clustering data streams based on grid density for high dimensional data streams. Most of existing density-grid clustering algorithms cannot handle high dimensional data stream efficiently, therefore Ren et al. in [11] proposed PKS-Stream algorithm based on grid density and Pks-tree. By using Pks-tree for clustering, the efficiency of storage and indexing are improved.

In the grid-based clustering approach, there are a lot of empty cells especially for high dimensional data. If all the grids are saved, it has an easy computation with high time complexity. If only non-empty grids are saved, the algorithm loses the relation between grids. So Pks-tree is used for recording not only the non-empty cells, but also the relation between grids. It is online-offline algorithm. In the online phase of PKS-Stream algorithm, the new data record in the data stream are continuously read and mapped to the related grid cells in the Pks-tree at all levels. If there is a grid cell for the data record, the data record is inserted. Otherwise, a new grid cell is created in the tree. In the offline phase, the clustering is started with non-empty cells in the leaf node level of Pks-tree. Firstly, it hacks the density of the grid, if it is higher than a threshold, a new cluster is created. After that, the neighboring grids are checked if their density is higher than the threshold; the grids will put in the same cluster as the first grid.

The sporadic grid is omitted in two situations: if the grid receives a few records over the time, or if many data records mapped to it but the density is reduced and is less than threshold. For improving the efficiency of the algorithms the empty grid cell is omitted using K-cover concept periodically-cover shows that the number of non-empty grids in the neighboring of leaf node grids. The average computational complexity of PKS-Stream is $(\log N)$ and in the worst case.

3. Conclusions

In this paper, we represent the density based clustering. That is uses to reduced core points, outliers and noise. When reduces this points than increase the efficiency of clustering. Core points are basically related to the center at any single tone problem and noise is the combination of outlier and core point.

References

[1] M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2003.

[2] Jean-Francois Laloux, Nhien-An Le-Khac, M-Tahar Kechadi," Efficient Distributed Approach for Density-Based Clustering" in 20th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises in 2011.

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of 2nd International Conference on Knowledge Discoveryand Data Mining (KDD). AAAI Press, 1996, pp. 226–231.

[4] Amineh Amini, Teh Ying Wah, Mahmoud Reza Saybani, Saeed Reza Aghabozorgi Sahaf Yazdi," A Study of Density-Grid based Clustering Algorithms on Data Streams' in Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) in 2011.

[5] J. Gao, J. Li, Z. Zhang, and P.-N. Tan, "An incremental data stream clustering algorithm based on dense units detection," Lecture Notes in Computer Science, vol. 3518, 2005.

[6] A. Zhou, F. Cao, W. Qian, and C. Jin, "Tracking clusters in evolving data streams over sliding windows," Knowledge and Information Systems, vol. 15, pp. 181–214, May 2008.

[7] Y. Chen and L. Tu, "Density-based clustering for real-time stream data," in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD '07. New York, NY, USA: ACM, 2007, pp. 133–142.

[8] W. Ng and M. Dash, "Discovery of frequent patterns in transactional data streams," in Transactions on Large-Scale Data- and Knowledge-Centered Systems II, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2010, vol. 6380, pp. 1–30.

[9] C. Jia, C. Tan, and A. Yong, "A grid and density-based clustering algorithm for processing data stream," in Proceedings of the Second International Conference on Genetic and Evolutionary Computing. Washington, DC, USA: IEEE Computer Society, 2008, pp. 517–521.

[10] L. Tu and Y. Chen, "Stream data clustering based on grid density and attraction," ACM Transactions on Knowledge Discovery Data, vol. 3, no. 3, pp. 1–27, 2009.

[11] C. H. Jiadong Ren, Binlei Cai, "Clustering over data streams based on grid density and index tree," Journal of Convergence Information Technology, vol. 6, pp. 83 – 93, 2011.