International Journal of Modern Computation, Information and Communication Technology 2020;3(1):12-18.
ISSN: 2581-5954
http://ijmcict.gjpublications.com

Review Article

Comparative Analysis of Various Spatio-Temporal Data Clustering Techniques in Spatio-Temporal Data Mining

Muhammad Haroon

Department of Computer Science & Information Technology, University of Gujrat Lahore Sub Campus, Lahore, Pakistan.

*Corresponding author's e-mail: haroon@uoglahore.edu.pk

Abstract

The data in digital world are increasing exponentially day by day. Spatio-temporal data hold data about an object in space over a period of time. Vast amount of spatiotemporal data is generated on daily basis in different application fields like weather forecasting, traffic flow, geo-tagging in social media etc. Spatiotemporal data mining (STDM) refers to the process of mining interesting and potentially useful patterns from a large spatiotemporal dataset. Clustering is one of the fundamental and important step in data mining process. The clustering is slightly different from trivial clustering technique in data mining. This paper emphasizes the temporal and spatial dimensions of data along with the techniques of spatiotemporal clustering techniques used and the comparative analysis of spatiotemporal algorithms with respect to methodology and complexity.

Keywords: Spatiotemporal data; Clustering; Data mining; Spatiotemporal data mining; Spatiotemporal clustering.

Introduction

Data in digital repositories are increasing on daily basis [1]. Management and handling of massive data does require extraordinary methods. Spatiotemporal data is a type of data with spatial and temporal characteristics. Spatial characteristics refer to the geometry and location of the data object whereas the temporal property refers to the time interval for which data object is valid. For example, weather forecasting data contains the spatial and temporal characteristics. Any phenomena occurs at specific time 't' and location 's' can be termed as spatiotemporal event. As with the increasing mass of spatiotemporal data sets, the spatiotemporal data mining (STDM) has become a hot research area in Data Mining field. The first step of STDM is preprocessing of data which usually includes cleaning of data. Then data is clustered and visualized and finally produced in suitable output pattern. The clustering is different from traditional data mining approach due to the different nature in data. Spatiotemporal data represents continuous space whereas traditional data set has discrete space [4]. Another difference in data sets is of statistical analysis. Trivial statistical analysis can be used for non-spatiotemporal data sets due to discrete nature but same statistical analysis cannot be used. The present paper targets to discuss the clustering techniques for STDM and their comparative analysis with respect to methodology, approach and complexity.

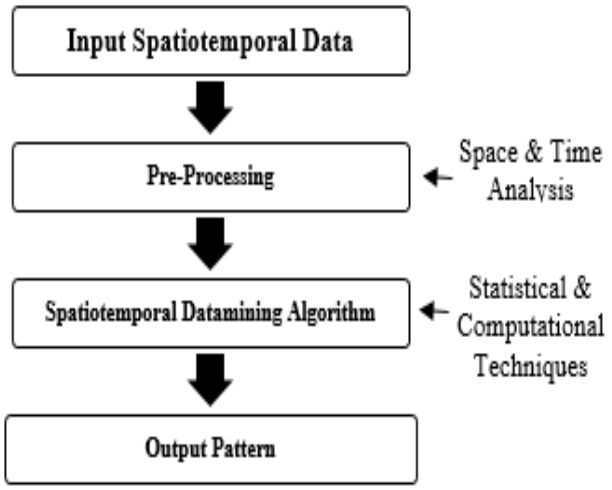

Fig. 1. Process of spatiotemporal data mining

Received: 03.02.2020; Received after Revision: 21.02.2020; Accepted: 21.02.2020

Brief introduction to spatio-temporal data

Spatialization and Temporalization are two properties that can be associated with some types of data which make data more complex to handle. A dataset of wireless communication networks that may exist only a particular period of time and on specific geographical area [2]. This type of data has an exponential growth graph. Another very common example could be related to tracking of moving objects over invariant geometrical regions [3]. The tracking of a moving car following invariant geometry paths associate specific temporal and spatial properties with it and create a large dataset of its points [21]. Other examples could include weather stations and geo-tagging data that are increasing in enormous manner and large data sets are building up on the World Wide Web.

Importance and potential applications of spatio-temporal data mining

The importance of spatiotemporal data mining is growing with the increase in spatiotemporal data sets in many domains. Some are: Neurology: Analysis of neuroimaging data for estimating equations approach for spatially correlated data [5]. Ecology: Mining relationships in environmental changes around the globe and incidents related to environment. Security Agencies: Specific crime patterns are recorded from geographical map to allocate police and security forces effectively [7]. Transportation Industry: Analysis of taxi and bus trajectories based on GPS to engage more customers on specified routes. Oncology: Major cancer cells development over a specified period of time are monitored to dictate oncologists in suggesting medicines. Climatology: Climate changes over a specific period of time and space are recorded to use data for various healthful and eco-friendly purposes. Internet of Things (IOT): Many decision making systems requires lot of information of mining patterns for their future decision making [8].

Spatio-temporal clustering

Clustering is the process of combining large data set in the form of group based on some similarity attribute and the analysis of clusters falls under the umbrella of unsupervised learning because the data does not need to have the past knowledge of dataset [9]. Spatiotemporal clustering deals with the clustering of data which have spatial characteristics in terms of longitude & latitude and temporal characteristics with time dimensions [10].

There are various clustering techniques because of many spatiotemporal data types. Each technique has its own style of working and mechanism.

Geo-referenced time series clustering

Geo-referenced time series clustering data refers to the form of data that contains the complete past history of the evolving object over a specific period of time [12]. Hourly monitored Air Quality Index (AQI) value which records the hourly values of air pollution is an example of it [13].

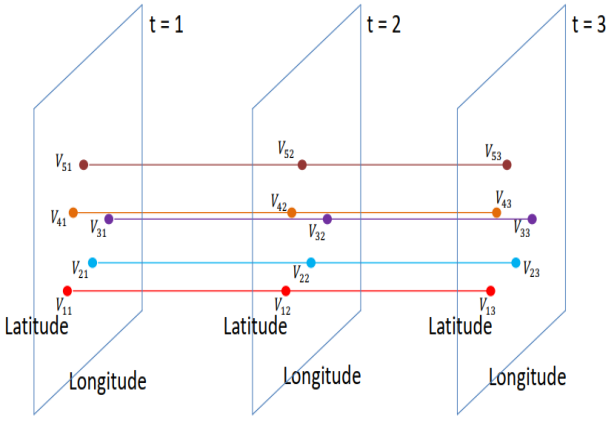

Fig. 2. Geo-referenced time series data

As the data is based on time series, it requires to compare time series of object with its spatial characteristics. Fuzzy logic based algorithm, Fuzzy C-Medoid (FCM) can be used to cluster this type of data by modifying its objective function to make it useful for this type of data. One latest algorithm named Correlation based Clustering of Big Spatiotemporal Datasets (CorClustST). It is based on correlation of neighbor points with respect to space over time. Its key feature is its unique behavior of comparing different type of clusters while placing data in specific cluster [13].

Geo-referenced data item clustering

It contains data which is gathered from a particular event or phenomena at a particular but fixed location over a certain period of time. Non-Spatial characteristics may also be associated with spatiotemporal data [14]. Typical example

could include the weather data collected from a fixed location over a specified period of time.

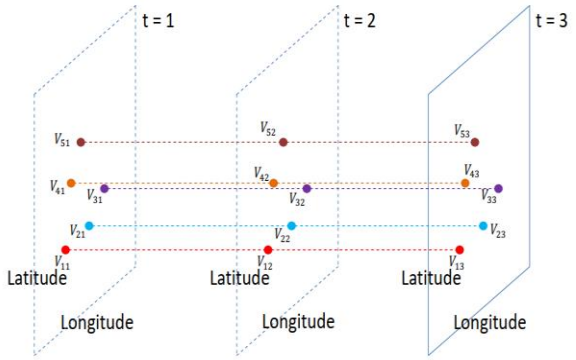

Fig. 3. Geo-referenced data items

Clustering defines a paradigm, Density Based Clustering which forms clusters by considering a density threshold of data points. Geo-referenced data items clustering works on the type of data which has similarities with respect to spatial attributes but on a given specific timestamp [11].

A modified version of a famous density based algorithm, Density-based spatial clustering of applications with noise, DBSCAN can be used to cluster spatiotemporal data which is ST-DBSCAN. It works by using Spatiotemporal Euclidean Distance for spatial characteristics of data point at given instance of time [22].

An example would be sufficient to understand this. A car moving on a highway passes through toll plazas at different instances of time. The recording of car movement data is of spatiotemporal nature.

Suppose at some instance of time, car moves from Toll Plaza 1 (A) to Toll Plaza 2 (B). It has geographical points on both toll plazas, A and B as $A(x_A, y_A, t_{1A}, t_{2A})$ and $B(x_B, y_B, t_{1B}, t_{2B})$.respectively. These points have spatial and temporal characteristics. Euclidian distance for spatial values can be calculated as

$$\sqrt{(x_A - x_B)^2 + (y_A - y_B)^2} \qquad (1)$$

which shows the closeness of points and temporal values are calculated as

$$\sqrt{(t_{1A} - t_{1B})^2 + (t_{2A} - t_{2B})^2} \qquad (2)$$

ST-DBSCAN also calculates average of values and compares these values with new ones to place them in clusters if any value is missing.

Event clustering

An event is simple capturing of record having spatial attributes at some instance of time. It basically stores three attributed i.e. Time (Temporal Attribute), Longitude and Latitude (Spatial Attributes) [15].

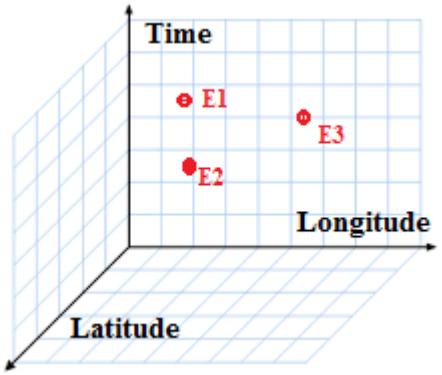

Fig. 4. Spatiotemporal data events

Fuzzy C-Means algorithm – a fuzzy Logic based algorithm can be modified to a new version, Spatiotemporal Extended Fuzzy C-Means (SEFCM) algorithm. It also uses the Euclidian distance as a part in its formula where one value is the data point value and other one is the center of cluster [15].

The given formula can be used to form clusters:

$$D_k = \sum_{i=1}^{M} \sum_{j=1}^{N} a_{ij}{}^{k} \left((x_i - c_j)^2 \times \partial \right) \qquad (3)$$

Where k is the number of data point, a_{ij} is the membership count of point x_i in cluster, c_j is the center of the cluster. Another multiplicative symbol ∂ denotes the temporal part of distance function.

Moving clusters

The monitoring of a fast moving car for security purpose covers many points with the passage of time. This car can be termed as moving object. This type of object has some unique type of nature. There is no need to store its previous history because at every point, it changes its dimension arbitrarily. We just need to have last immediate state of the moving object [16].

An algorithm based on DBSCAN named MOVCLUST to form such data type cluster has been devised which takes record of moving object at every specified timestamp t. This gives the position of the objet with respect to time. For

two moving clusters C1 and C2, the condition 1 > θ must be true where θ is some threshold for two clusters.

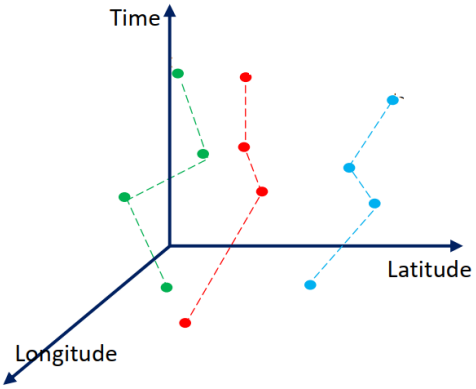

Fig. 5. Spatiotemporal moving object

Trajectory clustering

An object following trajectory has the sequence of spatial locations with timestamp associated with it. Its like the moving objects but the difference is to have complete history of moving object. This is used for scenarios where we need to use complete previous data for data points along a trajectory [17]. The example could include the tracking of a railway train. An algorithm used for this type of data is suggested as Trajectory Clustering (TRACLUS). It works in two phases: (i) Partitioning and (ii) Grouping [19].

Partitioning phase

The complete trajectory of an object containing several points in between them is partitioned into collection of line segments. Fig.5 has three trajectories showing in green, red and blue. Each having three line segments S1, S2 and S3. These line segments are divided and stored in a collection of S1, S2 and S3 for every colored trajectory. Principle of minimum description length is followed [18].

Grouping phase

Grouping phase again works on a principle of density based clustering by using DBSCAN [22]. A line segment can be decomposed into three components to compute distance.

- Perpendicular Distance
- Parallel Distance
- Angular Distance

Following figure (Fig. 6) can be helpful to understand this computing phenomena.

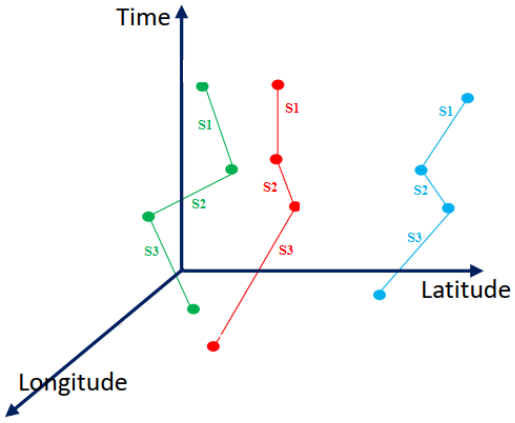

Fig. 6. Spatiotemporal moving object

Here, we have two line segments, L_a and L_b making an angle θ among them with respective lengths shown in fig. 7.

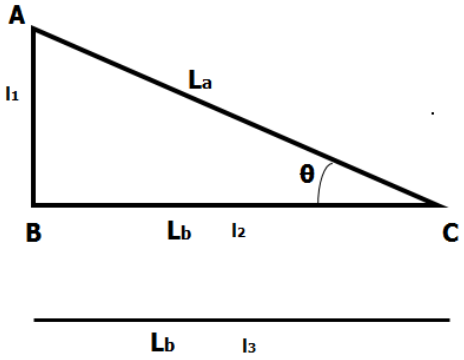

Fig. 7. Components of a trajectory

The perpendicular distance between L_a and L_b can be computed using following formula:

$$d_{\perp}(L_a, L_b) = \frac{l_1^2 + l_2^2}{l_1 + l_2} \quad (4)$$

For the parallel distance:

$$d_{\parallel}(L_a, L_b) = Min(l_1, l_2) \quad (5)$$

And for the angular distance:

$$d_{\theta}(L_a, L_b) = |L_a| \sin\theta \quad (6)$$

Some other related algorithms

A technique called, Clustering-Based Stops and Moves of Trajectories (CB-SMoT) which works on trajectories paths by considering a parameter, velocity of moving object along the trajectory. It is useful for the dataset which have velocity parameter associated with it hence it does not perform well if there is some other semantic attached with trajectory [23]. It first works on

actual points in trajectory and continue expanding these points to form a complete cluster [23].

Hypertext Induced Topic Search (HITS) algorithm mines relevant site spots and moving sequence of an object. It can be of the good choice to cluster ST data but it is unable to group location histories [20]. Such type of data in which trajectory points are missing or there is the task to locate points on moving object's trajectories, there's another algorithm works better called Direction Based Stops and Moves of Trajectories (DB-SMoT) which is purely direction finding oriented. It runs perfectly on the data types which have only spatial and direction based characteristics.

Comparison and analysis of studied algorithms

We have discussed many algorithms for spatiotemporal data mining for different types of clusters data. Some works well for a specific kind of data whereas other one works well for some other kind of nature. Here is the brief analysis and comparison of every algorithm.

Table 1. Comparison & Analysis of Algorithms

Algorithm	Comparison and Analysis
FCM	Uses Fuzzy Logic for clustering of ST Data. Good for the geo-referenced time series data.
CorClustST	Also works well for geo-referenced time series data but it is majorly used for big data sets because it uses the concept of correlation of points with respect to space over time. Unlike FCM, it compares data from different clusters and then place it in correct cluster.
ST-DBSCAN	Used for geo-referenced data items. It works on the basis of density of data points in a cluster. It does use the spatial and temporal properties.
SEFCM	Used for Event-Based Clustering. Used Euclidian distance with a modification that one point of data and other one is the center of cluster.
MOVCLUST	When the data is in the form of moving object which changes direction with passage of time, here this algorithm devised method to capture record at a time interval t. This is somehow lightweight algorithms in terms of memory as it does not store previous history of points.
TRACLUS	Unlike MOVCLUST, it stores complete history of trajectory points. This makes this algorithm different from previous one.
CB-SMoT	Used for trajectory based data of having velocity in terms of spatial attribute. It gets points on trajectories first then explores these points to form clusters.
DB-SMoT	It is same like CB-SMoT with one exception that it works with the direction attribute instead of velocity as in previous one.
HITS	Works good for moving points towards some specific location. Does not store group history.

Time complexities

Table 2. Time Complexities of Algorithms

Algorithm	Time Complexity
FCM	$O(ndic^2)$ Where: n = Number of objects d = Number of dimensions i = Number of iterations c = Number of Clusters
ST-DBSCAN	$O(n \lg n)$ Where: n = Number of objects
SEFCM	$O(ndic^2)$ Where: n = Number of objects d = Number of dimensions i = Number of iterations

	c = Number of Clusters
MOVCLUST	$O((n+u)\log(n+u)\log n+n\log_3 n)$ Where: n = Number of objects d = Number of object update
TRACLUS	$O(n \lg n)$ Where: n = Number of line segments
CB-SMoT	$O(n+n\log S)$ Where: n = Number of points in trajectories S = Number of defined stops
DB-SMoT	$O(n+n\log S)$ Where: n = Number of points in trajectories S = Number of defined stops
HITS	$O(n^2)$ Where: n = Number of points

Challenges handling spatiotemporal data

This section throws light on the challenges one can face while working with spatiotemporal data. Due to advent of latest and affordable technological gadgets and service, data can be reported at very small time intervals [24]. As the data became more complex with additional of spatial and temporal attributes, hence working on it also need some special care to handle it properly and use for further processes. Geographic Information System (GIS) aids the spatiotemporal clustering as the functions used in it are powerful enough to work for spatiotemporal data. Like distance function in GIS is not like the simple Euclidian distance.

One challenge which commonly arises related to Heterogeneity of data with spatial and temporal properties. In this type data, clusters have different type of values which have same feature values. To overcome this problem, spatiotemporal algorithms can be used [25].

Conclusions

The present paper provides the discussion of spatiotemporal data. It is explained that how spatiotemporal data is important in many fields of life like Neurology, Ecology, Security, Transportation, Climatology and Internet of Things. A detailed overview of different type of spatiotemporal data is given along with the various clustering techniques used to cluster data and the comparative analysis of described algorithms with respect to working methodology, approach and the time complexity analysis. Finally, some challenges in terms of future advancements of technology have also been explained.

Conflicts of interest

Authors declare no conflict of interest.

References

[1] Li S, Dragicevic S, Castro FA, Sester M, Winter S, Coltekin A, Pettit C, Jiang B, Haworth J, Stein A, Cheng T. Geospatial big data handling theory and methods: a review and research challenges. ISPRS J Photogramm Remote Sens 2016;115:119-33

[2] Brent Hall, Michael G, Leahy. Open Source Approaches in Spatial Data Handling. Springer. Advances in Geographic Information Science 2008; 126–128.

[3] Ralf Hartmut Güting, Markus Schneider. Moving Objects Databases. Academic Press 2005.

[4] Shekhar S, EvansMR, J. M. Kang and P. Mohan, Wiley. Identifying patterns in spatial information: a survey of methods 2011;193-214.

[5] Albert P, McShane L. A generalized estimating equations approach for spatially correlated binary data: Applications to the analysis of neuroimaging data. Biometrics 1995;51:627-38.

[6] Scally R. GIS for Environmental Management. ESRI Press 2006.

[7] Leipnik MR, Albert DP. GIS in Law Enforcement: Implementation Issues and Case Studies. CRC Press, Sacramento 2002.

[8] Gubbi J, Buyya R, Marusic S, Palaniswami M. Internet of Things (IoT): A vision, architectural elements, and future directions. Future Gener Comput Syst 2013;29:1645-60.

[9] Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data 2007;63:503-27.

[10] Tork HF. Spatio-Temporal Clustering Methods Classification. Proceedings of the Doctoral Symposium on Informatics Engineering 2012;1-12.

[11] Kisilevieh S, Mansmann F, Nanni M, Rinzivillo S. Spatio-temporal c1ustering. In Data mining and Knowledge Discovery Handbook 2010;855-874.

[12] Zhang D, Lee K, Lee I. Hierarchical trajectory clustering for spatio-temporal periodic pattern mining. Expert Systems With Applications 2018; 92:1-11.

[13] Hüsch M, Schyska B, Bremen L. CorClustST—Correlation-based clustering of big spatio-temporal datasets. Future Generation Computer Systems 2018.

[14] Kalyani D, Chaturvedi SK. A Survey on Spatio-Temporal Data Mining. International Journal of Computer Science and Network 2012;1(4).

[15] Martino FD, Pedrycz W, Sessa S. Spatiotemporal extended fuzzy C-means clustering algorithm for hotspots detection and prediction. Fuzzy Sets and Systems 2018;340:109-26.

[16] Kalnis P, Mamoulis N, Bakiras S. On Discovering Moving Clusters in Spatio-temporal Data. Advances in Spatial and Temporal Databases 2005;3633:364-81.

[17] Lee J, Han J, Whang K-Y. Trajectory clustering: a partition-and-group framework. Proceedings of the ACM SIGMOD international conference on Management of data - SIGMOD 2007. doi:10.1145/1247480.1247546

[18] Gafiney S, Robertson A, Smyth P, Camargo S, and Ghil M. Probabilistic Clustering of Extratropical Cyclones Using Regression Mixture Models, Technical Report UCI-ICS 06-02, University of California, Irvine. 2006.

[19] Lee JG, Han J, Whang KY. Trajectory clustering: A partition-and-group framework. In Proceedings of the ACM SIGMOD Conference on Management of Data. ACM 2007;593-604.

[20] Zheng Y, Zhang L, Xie X, Ma W-Y. Mining interesting locations and travel sequences from GPS trajectories. Proceedings of the 18th international conference on World wide web - WWW 2009.

[21] Lee K, Zhang D, Lee I. Hierarchical trajectory clustering for spatio-temporal periodic pattern mining. Expert Systems With Applications 2018:92:1-11.

[22] ElKaffas S, Zaghlool E, Saad, A. A Density-Based Clustering of Spatio-Temporal Data. New Contributions in Information Systems and Technologies 2015;41-50.

[23] Palma AT, Bogorny V, Kuijpers B, Alvares LO. A clustering based approach for discovering interesting places in trajectories," in ACMSAC. New York, NY, USA: ACM Press 2008;863-68.

[24] Yao X. Research issues in spatio-temporal data mining. A white paper submitted to the University Consortium for Geographic Information Science (UCGIS) workshop on Geospatial Visualization and Knowledge Discovery, Lansdowne, Virginia 2003;18-20.

[25] Faghmous JH, Kumar V. A big data guide to understanding climate change: The case for theory-guided data science. Big data 2014;2(3):155-63. doi:10.1089/big.2014.0026

****

