2024 IEEE International Performance, Computing, and Communications Conference (IPCCC) | 979-8-3503-6794-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IPCCC59868.2024.10850463

Spatiotemporal-DBSCAN: A Density-Based Clustering Method for Analyzing Spatiotemporal Ground-lightning Dataset

*

James Agresto
Department of Computer Science
James Madison University
Harrisonburg, USA
agrestjj@dukes.jmu.edu

Zhuojun Duan
Department of Computer Science
James Madison University
Harrisonburg, USA
duanzx@jmu.edu

Mace Bentley
ISAT Department
James Madison University
Harrisonburg, USA
bentleml@jmu.edu

Tobias Gerken
ISAT Department
James Madison University
Harrisonburg, USA
gerkentx@jmu.edu

Dudley Bonsal
ISAT Department
James Madison University
Harrisonburg, USA
bonsaldb@jmu.edu

***Abstract*—In recent years, the extraction of extensive spatiotemporal datasets from various sources has surged. One notable example is the cloud-to-ground (CG) lightning dataset in atmospheric area, where each entry records the flash location, including latitude and longitude, and the time the flash occurred.**

The flashes can be clustered into cells with a set of rules applied to the time-ordered sequence of flash locations, which is called as a thunderstorm cell. Analysis of the cell structure plays an important part in a description of a multicell thunderstorm system in atmospheric area. Clustering algorithms such as K-means and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) cannot be directly applied to these datasets due to their inability to handle spatial and temporal dimensions simultaneously. Our study aims to investigate a clustering method capable of handling large-scale spatiotemporal data in a reliable and efficient manner. The proposed method, called Spatiotemporal-DBSCAN, is applied to both idealized clusters and real datasets, such as the cloud-to-ground (CG) lightning dataset. Experiements demonstrates that the spatiotemporal-DBSCAN performs effectively on spatiotemporal datasets.

I. Introduction

In recent years, massive amounts of spatiotemporal datasets have been generated from different areas, such as social networks [1], biology [2], and meteorology [3]. These datasets have features of huge-volume, multi-columns, and complicated format, which brings about new opportunities and challenges for data analysis. Some datasets have spatial coordinates and time stamps simultaneously [3]. To cluster these spatiotemporal datasets, two popular algorithms can be used: K-means [5] and DBSCAN (Density Based Spatial Clustering of Applications with Noise) [4] algorithms. K-means algorithm first selects a number of clusters and assigns a data item to a clusters according based on the distance. All the clusters produced by k-means algorithm are spherical or convex in shape. On the other hand, DBSCAN performs very well to detect arbitrary cluster shapes. Moreover, none of them can be used to cluster spatiotemporal datasets directly.

In this study, our objective is to devise a practical clustering approach tailored for spatiotemporal data in tmospheric area, ensuring both efficiency and interpretability.

This work is funded by the NSF under grant number 2104299.

II. Ground-lightning dataset

From the U.S. National Lightning Detection Network (NLDN; Vaisala, Inc.), we obtained a lightning flashes dataset that include the warm season (May - September) in a 200km radius surrounding Washington, DC and for the years 2006–2020. The raw datasets consist of entries representing individual lightning flashes. Each entry includes essential information about a flash, such as the date and time, latitude and longitude coordinates, distance to the detected station in Washington, DC, and other relevant geographic details. Following initial data cleaning and reformatting, we retain only the *Date and Time*, *Latitude*, and *Longitude* columns from the datasets for use in our clustering algorithms.

Clustering lightning flashes based on their spatial and temporal attributes can investigate how urban

aerosols and pollution affect the magnitude and intensity of lightning flashes [6].

III. The Spatiotemporal-DBSCAN Method

Cluster analysis aims to uncover hidden groups and structures within datasets. It encompasses various methods that categorize data into homogeneous groups, emphasizing internal cohesion and external separation. Spatiotemporal clustering methods go beyond traditional approaches by explicitly incorporating both spatial and temporal dimensions, contrasting with methods that treat all attributes uniformly when calculating distances between data points.

Density-based clustering methods like DBSCAN are probably the most popular method in this clustering methods, and can also fit the natural structures of thunderstorms of flashes.

The primary distinction between Spatiotemporal-DBSCAN and DBSCAN lies in how the search boundary is delineated. Unlike DBSCAN, Spatiotemporal-DBSCAN employs two epsilon values, Eps_D and Eps_T: Eps_D functions similarly to Eps in DBSCAN, while Eps_T defines the threshold for temporal differences. By specifying these parameters separately, Spatiotemporal-DBSCAN can explicitly incorporate temporal similarities into the clustering process. This methodology can be implemented as follows:

1) The spatiotemporal data is initially arranged in chronological order, and the three parameters, Eps_D, Eps_T, and $MinPts$, are set.
2) Begin with the first flash in the time-ordered data and count the number of other flashes within both Eps_D and Eps_T.
3) Form a cluster (thunderstorm) if the number of flashes exceeds $MinPts$, and mark all these flashes as part of the same cluster. Otherwise, mark the flash as noise.
4) Expand the clusters by recursively repeating the neighborhood calculation for each neighboring point.
5) Repeat from Step 2 until all flashes in the data have been visited.

The complexity of this algorithm is $O(n^2)$, where n is the number of flashes in the data.

IV. Initial Experiments and results

We made some experiments of the Spatiotemporal-DBSCAN method using using the Washington DC lightning flash dataset in May of 2017, which includes more than 39,000 flashes.

Table I shows the parameters used in each analysis, along with the number of clusters and noise points derived from these parameters. Eps_D represents the spatial threshold, and Eps_T represents the temporal threshold. The $MinPts$ parameter is set to 10 by default. Additionally, the running time is presented in seconds.

TABLE I
Effects of Parameter Changes on Cluster Formation.

Run #	Eps_D	Eps_T	clusters	Noise	Running time
0	15	15	17023	5574	5851.53
1	9	15	14974	9214	5862.36
2	12	15	16302	6863	5981.29
3	15	9	15829	7352	3799.50
4	15	12	16571	6218	4949.53
5	9	9	13858	10911	3750.05
6	12	12	15826	7538	4997.34
7	10	15	15523	8270	5923.33
8	15	10	16127	6909	4067.94

As the parameter range increases, the amount of noise decreases and the number of clusters also decreases, while the running time increases. This indicates that although data attributes within a cluster are similar, slight differences can cause them to belong to different clusters.

V. Conclusion and Future work

Our ongoing work aims to improve clustering of the Ground-lightning Dataset using spatial and temporal attributes. To achieve this goal, we have developed a practical Spatiotemporal-DBSCAN method tailored for analyzing the dataset. Initial experimental results demonstrate how parameters like epsilon values impact cluster formation and running time. Moving forward, we will conduct additional experiments with various parameters and employ statistical and visualization methods to comprehensively evaluate the results. Additionally, further investigation is needed to enhance the method's efficiency.

References

[1] Curiskis, Stephan A., Barry Drake, Thomas R. Osborn, and Paul J. Kennedy. "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit." Information Processing & Management 57, no. 2 (2020): 102034.

[2] Mahmud, Mufti, M. Shamim Kaiser, T. Martin McGinnity, and Amir Hussain. "Deep learning in mining biological data." Cognitive computation 13 (2021): 1-33.

[3] Khan, Nida, and Radu State. "Lightning network: A comparative review of transaction fees and data analysis." In Blockchain and Applications: International Congress, pp. 11-18. Springer International Publishing, 2020.

[4] Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In kdd, vol. 96, no. 34, pp. 226-231. 1996.

[5] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in 5th Berkeley Symposium on Mathematical Statistics and Probability, U. of California Press, Ed., 1967, pp. 281–297.

[6] Stallins, J. Anthony, James Carpenter, Mace L. Bentley, Walker S. Ashley, and James A. Mulholland. "Weekend–weekday aerosols and geographic variability in cloud-to-ground lightning for the urban region of Atlanta, Georgia, USA." Regional Environmental Change 13 (2013): 137-151.

[7] Řezanková, H. A. N. A. "Different approaches to the silhouette coefficient calculation in cluster evaluation." In 21st international scientific conference AMSE applications of mathematics and statistics in economics, pp. 1-10. 2018.