An adaptive density-based time series clustering algorithm

Yaolin Liu*[1], Xiaomi Wang[1], Yanfang Liu[1]
[1]School of Resource and Environment Science, Wuhan University, 129 Luoyu Road, Wuhan 430079, China;
*Email: yaolinliuwhu@163.com
806560802@qq.com
yfliu610@163.com

Abstract

Time series clustering algorithms have been widely used to mine the clustering distribution characteristics of real phenomena. However, these algorithms have several limitations to mine clustering characteristics in geography. First, current time series clustering algorithms fail to effectively mine clustering distribution characteristics of time series data without sufficient prior knowledge. Second, the algorithms ignore the spatial heterogeneity of geographical objects. Thirdly, these algorithms fail to simultaneously consider non-spatial time series attribute values and non-spatial time series attribute trends, which are all important similarity measurements. In view of these shortcomings, an adaptive density-based time series clustering (DTSC) algorithm has been proposed in this paper. DTSC algorithm simultaneously considers the spatial attributes, non-spatial time series attribute values, and non-spatial time series attribute trends. DTSC algorithm proceeds with two major parts. In the first part, the objects with spatial proximity relationship are considered as similar in the spatial domain. In the second part, an improved density-based clustering strategy is then adopted to detect clusters with similar non-spatial time series attribute values and time series attribute trends. The effectiveness and efficiency of the DTSC algorithm are validated by experiments on simulated datasets and real applications. In the applications of simulated datasets, the results indicate that the proposed DTSC algorithm effectively detects time series clusters with arbitrary shapes and similar attributes and densities while considering noises. In the real applications, both time series raining dataset and time series surface deformation dataset have been utilized, and several interesting patterns that cannot be effectively detected by other classical time series clustering algorithms have been found.

Keywords: Time series clustering, Adaptive, Data Mining.

1. Introduction

Time series data are very common in the real world and generally exhibit obvious spatial heterogeneity. Mining the spatial clustering characteristics of time series data is essential to exploring the potential distribution mechanism underlying this kind of data.

In the past few decades, many time series clustering algorithms have been developed. These algorithms can be roughly grouped into five classes, as follows: partitioning-based time series clustering algorithms (Guyet and Nicolas, 2016, Kaur, et al., 2016), hierarchical time series clustering algorithms (Yin, et al., 2006), density-based time series clustering algorithms (Uijlings, et al., 2014), graph-based time series clustering algorithms (YEANG and JAAKKOLA, 2003) and time series co-clustering algorithms(Xu, et al., 2013, Yan, et al., 2008). Although these algorithms can handle certain applications, they still suffer from several deficiencies and require improvement. For example, these algorithms generally consider either non-spatial attribute values or non-spatial attribute trends to measure the similarity between objects. However, time series data with similar non-spatial attribute trends and different attribute values or similar attribute values and different attribute trends exist in real applications. Therefore, the similarity of non-spatial time series attribute values and the similarity of non-spatial time series attribute trends should be considered simultaneously to correctly mine the clustering patterns. As another example, current algorithms ignore spatial heterogeneity and seldom consider spatial attributes. Non-spatial attributes of geographical objects in reality are generally similar with a short spatial distance. Meanwhile, if spatial attributes are neglected, objects in clusters will be dispersedly distributed in the spatial domain, and clusters with similar non-spatial attributes will overlap. This phenomenon partially violates the real conditions and affects the visualization effect. Hence, to obtain clusters non-overlapping to each other with arbitrary geometrical shapes, spatial attributes should be considered to construct the spatial proximity relationships between sequences during the clustering procedure. Furthermore, the results of existing clustering algorithms are largely affected by predefined parameters and depend on prior knowledge, which is generally unavailable in real applications.

To overcome the above-mentioned deficiencies, a novel density-based time series clustering (DBSC) algorithm, is proposed based on a density-based spatial clustering (DBSC) algorithm (Liu, et al., 2012). The proposed DTSC algorithm can adaptively detect clusters with similar spatial attributes, non-spatial time series attribute values and non-spatial time series attribute trends. In addition, the corresponding clusters under considering the spatial heterogeneity are non-overlapping for a clear visualization.

2. Methods for Time Series Clustering

Performing the adaptive time series clustering depends on two aspects. One is the measurement of the similarity between time series objects, and the other involves the adaptive strategy of the time series clustering.

2.1 Similarity measurements

Spatial and non-spatial similarities are considered interdependently to eliminate the need to determine suitable weightings for the similarity between objects in the spatial and non-spatial domains. In the spatial domain, Euclidean distance is generally adopted to measure the similarity degree. In the non-spatial domain, the similarity of non-spatial attribute values is generally defined as the mean value of the difference of attribute values of every time interval. The similarity of non-spatial attribute trends can be measured by correlation coefficients such as the Pearson coefficient and Spearman coefficient (Fu, 2011).

2.2 A Strategy for Adaptive Time Series Clustering

In the proposed DTSC algorithm, a strategy of separately detecting clusters in the spatial and non-spatial domains is proposed to adaptively identify the clusters with similar spatial and non-spatial attributes. The objects with spatial proximity relationship are considered as similar in the spatial domain. If the dataset is a vector data, spatial proximity relationships between time series objects are adaptively obtained by removing inconsistent edges in the constructed Delaunay triangulation of objects by merging the particle swarm optimization (PSO) algorithm (Liu, et al., 2016); If the dataset is a raster data, the eight-connected pixels can be considered proximity pixels. Then, based on the spatial proximity relationships, clusters with neighboring objects having similar non-spatial time series attributes are adaptively clustered by using an improved density-based time series clustering method in the non-spatial domain. The density based time series clustering method is improved by integrating the proposed similarity measurements and the strategy of density indicator of a dual density-based clustering method (DBSC) (Liu, et al., 2012).

3. Selected Results and Discussion

The DTSC algorithm is applied to the annual rainfall data to mine the clustering pattern of rainfall. These data are provided by the China Meteorological Bureau and comprise the annual average rainfall monitoring data of 599 rainfall stations in mainland China from 1960 to 2009. The results (Figures 1 and 2) show that 15 interesting clusters were obtained. The proximity clusters are significantly different in terms of non-spatial attributes. The results also indicate that rainfall gradually increased from the northwestern to southeastern areas. Furthermore, Hou, et al. (2002) stated that separating line between C1 and C2, C3, C8 and C10 is consistent with the line of semi-humid and semi-arid regions; these results are consistent with actual conditions.

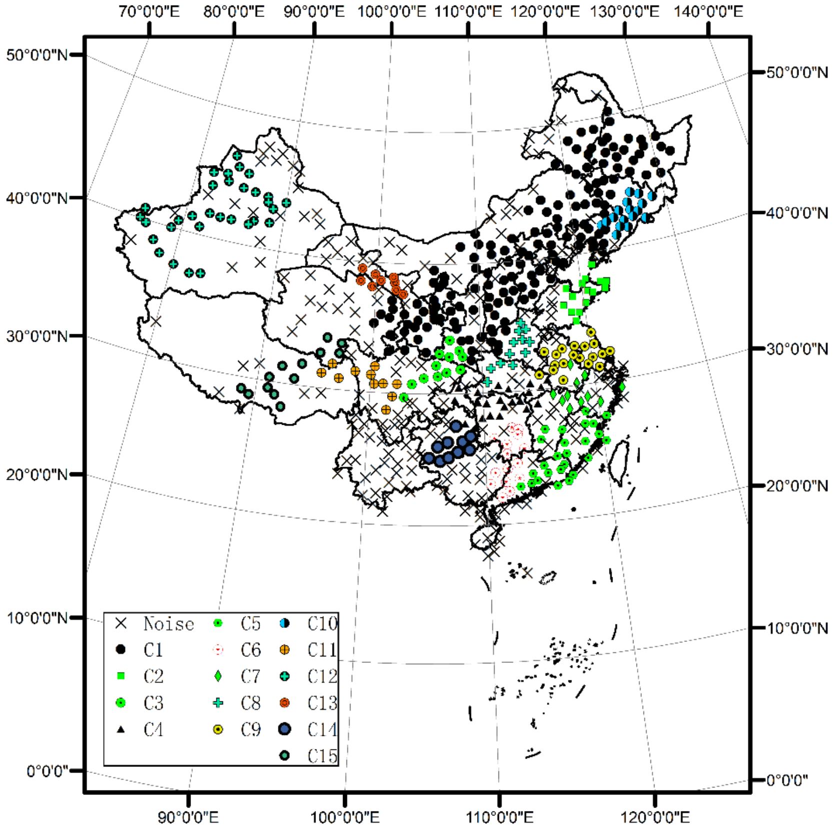

Figure 1: Clustering result of rainfall data using DTSC algorithm.

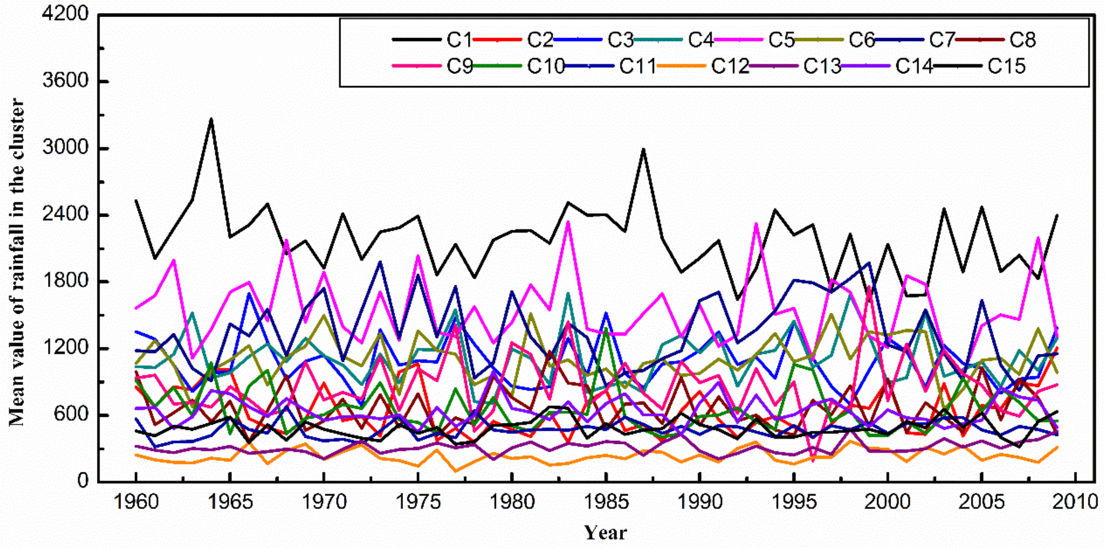

Figure 2: Mean values of clusters of rainfall in Figure 1.

The DTSC algorithm is also conducted on the time series surface deformation data provided by Ningbo Bureau of Surveying and Mapping in China. Deformation values between two neighboring time points are shown in Figure 3 and the end time point of the time interval is labeled above the image. The result is shown in Figure 4, and the statistic values of accumulative deformation of clusters in Figure 5. According to the results, the following rules can be observed: (1) The DTSC algorithm can detect clusters with arbitrary shapes under the interference of uneven deformation areas; (2) Most of the constructed areas in 20 years continue to have subsidence; (3) Several districts constructed more than two centuries slightly uplifted due to ground rebound; (4) The surface deformation in the reclamation area (areas in the red boundaries in Figure 4(a)) in Ningbo city remains unstable.

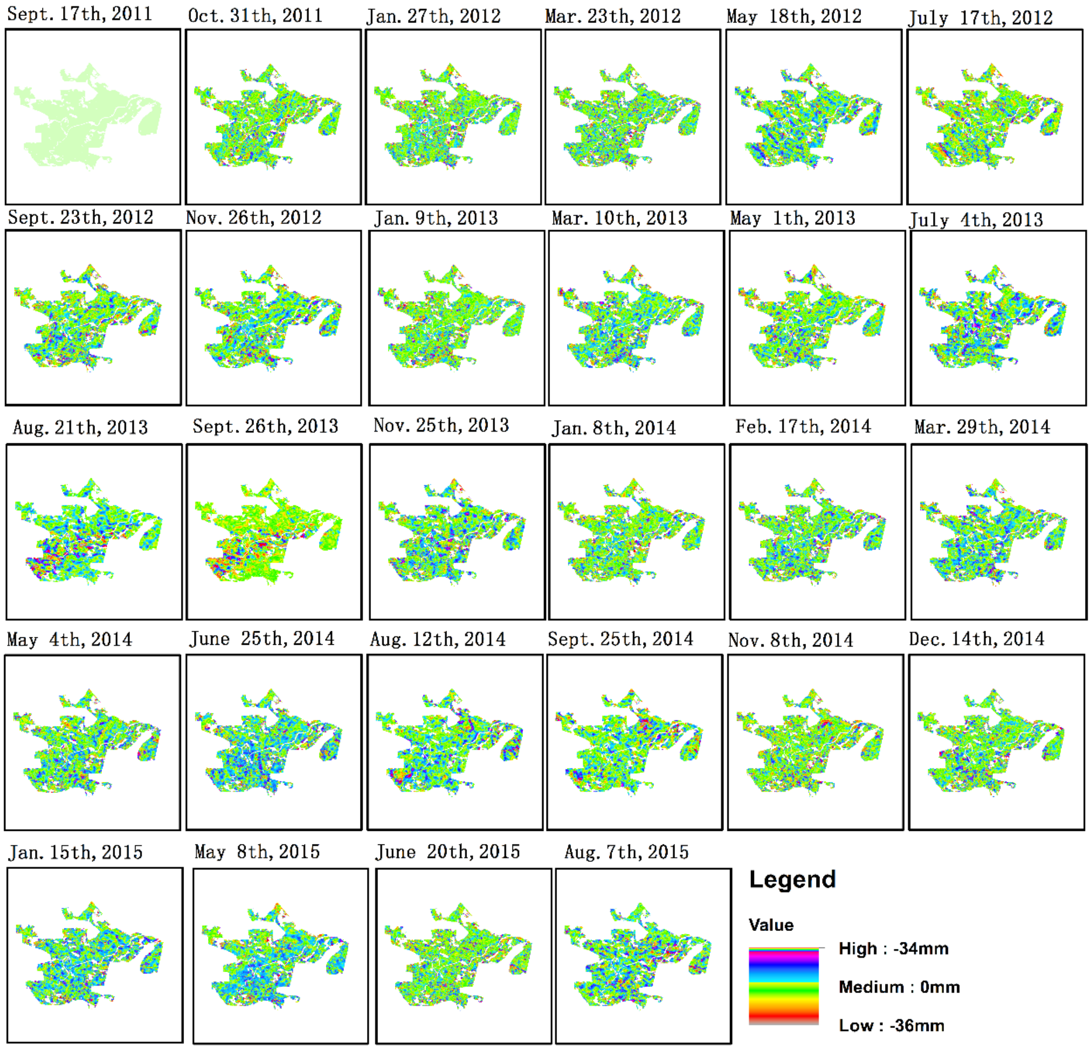

Figure 3: Image time series data of surface deformation detection in Ningbo City.

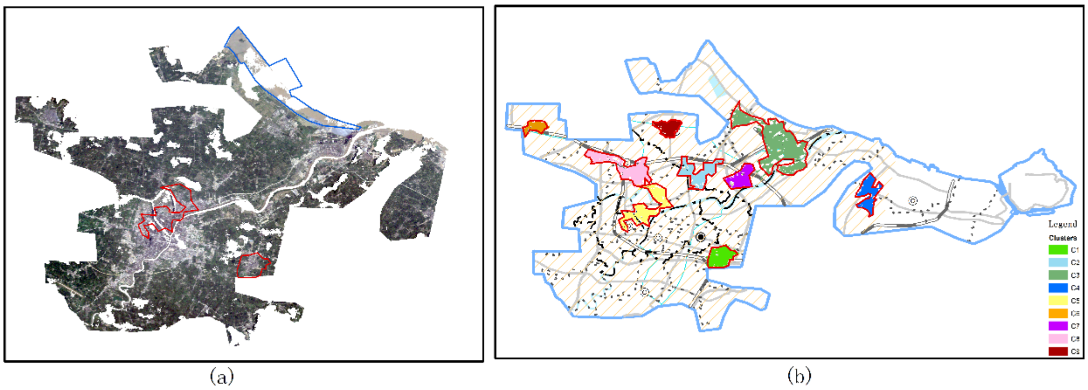

Figure 4: (a) Landsat image of 1995 in deformation detecting area; (b) Clustering result of surface deformation using DBTSC-IR.

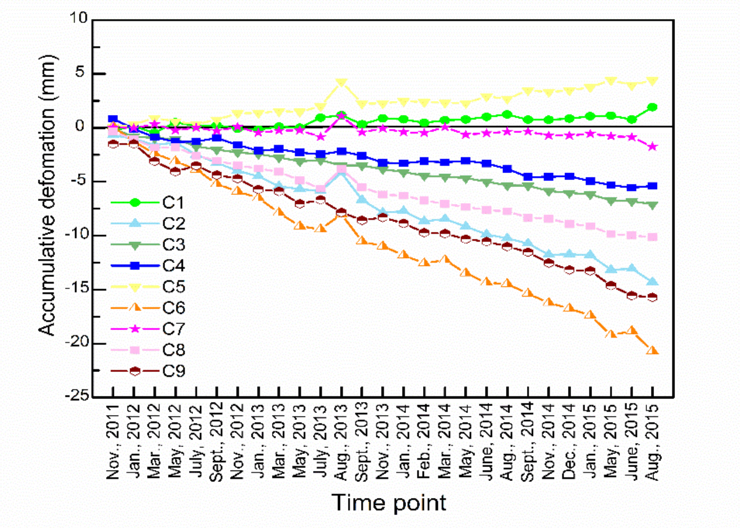

Figure 5: Accumulative deformation of clusters in Figure 4 (b).

4. Acknowledgements

The research reported in this paper was supported by the Natural Science Foundation of China (NO. 41371429).

Reference

T. Guyet and H. Nicolas. 2016. Long term analysis of time series of satellite images[J]. Pattern Recognition Letters, 70:17-23.

G. Kaur, J. Dhar and R. K. Guha. 2016. Minimal variability OWA operator combining ANFIS and fuzzy c-means for forecasting BSE index[J]. Mathematics and Computers in Simulation, 122:69-80.

J. R. R. Uijlings, I. C. Duta, N. Rostamzadeh and N. Sebe. 2014. Realtime Video Classification using Dense HOF/HOG[J].145-152.

C. YEANG and T. JAAKKOLA. 2003. TIME SERIES ANALYSIS OF GENE EXPRESSION AND LOCATION DATA[J]. International Journal of Artificial Intelligence Tools, 14(5):305-312.

T. Xu, X. Shang, M. Yang and M. Wang. 2013. Bicluster algorithm on discrete time-series gene expression data[J]. Application research of computers, 30(12):3552-3557.

L. Yan, Z. Kong, Y. Wu and B. Zhang. 2008. Biclustering Nonl inearly Correlated Time Series Gene Expression Data[J]. Journal of Computer Research and Development, 45(11):1865-1873.

Q. Liu, M. Deng, Y. Shi and J. Wang. 2012. A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity[J]. Computers & Geosciences, 46:296-309.

T.-c. Fu. 2011. A review on time series data mining[J]. Engineering Applications of Artificial Intelligence, 24(1):164-181.

Y. Liu, X. Wang, D. Liu and L. Liu. 2016. An adaptive dual clustering algorithm based on hierarchical structure: A case study of settlement zoning[J]. Transactions in GIS.

G. Hou, J. Wang, Q. Guo and X. Yan. 2002. A Study on the Cumulative Distributions of Rainfall Rate R1 (0. 01) Over China[J]. Journal o f Beijing Institute of Technology, 22(2):262.