Research Journal of Applied Sciences, Engineering and Technology 14(6): 221-226, 2017
DOI:10.19026/rjaset.14.4720
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2017 Maxwell Scientific Publication Corp.
Submitted: November 21, 2016 Accepted: February 13, 2017 Published: June 15, 2017

Research Article

Spatial Clustering Algorithm for Time Series Rainfall Data Using *X*-Means Data Splitting

[1]Noor Rasidah Ali and [2]Ku Ruhana Ku-Mahamud
[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kedah,
Kampus Sungai Petani, Malaysia
[2]School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, Malaysia

Abstract: The aim of this study is to present a new spatial clustering process for time series data. It has become an important and demanding application when the data involves chronological long time series and huge datasets. A great challenge in clustering is to achieve an optimal solution in searching similarity along the series. Furthermore, it also involves a very large-scale data analysis. Unfortunately, the existing clustering time series algorithms have become impractical since data do not scale properly for longer time series. The performance of the clustering algorithm gets even worse if it relies on actual data and many clustering algorithms are often faced with conflict in handling high dimensional data. In the case of spatial time series, the problem can be solved by unsupervised approaches rather than supervised classification, with appropriate preprocessing techniques to transform the actual data. The unsupervised solution using time series clustering algorithms is capable to extract valuable information and identify structure in complex and massive datasets as spatial time series. Therefore, a clustering algorithm by introducing data transformation using *X*-means data splitting is proposed to investigate the spatial homogeneity of time series rainfall data. The hierarchical clustering was used to demonstrate the similarity once the data was divided into training and testing sets. The proposed algorithm is compared with five types of data transformation techniques, namely mean and median in monthly data and the rest is in daily data such as binary, cumulative and actual values. Results indicate that data transformation using *X*-means data splitting in hierarchical clustering outperformed other transformation techniques and more consistent between training and testing datasets based on similarity measures.

Keywords: Clustering algorithm, similarity measures, spatial homogeneity, spatial time series, *X*-means data splitting

INTRODUCTION

Spatial time series data have gained a widespread popularity in most scientific researches. The application engages with measurements, in which the quantities vary in space and time. The main reason for the high demand of using this data is the abundance of applications, particularly in environmental and health science studies. This application arises in many other contexts, which include precipitation or regional ozone monitoring, disease mapping, remote sensing or satellite data, growth of population, moving of object, stock market fluctuation, real estate data in economic monitoring, electrocardiogram in medical data or seismic waves in seismology. Smoothing and predicting the time evolution of certain variables over spatial domain are the main interest in analyzing such data. The real challenge in spatial time series analysis is a large-scale analysis dealing with huge datasets.

A spatial time series data is defined as a time series measurement at each referencing location in a common spatial area. The relationship between time series data at different locations focuses on the spatial component in the data space, which is known as the spatial homogeneity of time series data. The process of grouping the objects based on their spatial and temporal similarity is called spatial time series clustering (Kisilevich *et al*., 2010). The outcome of the process is to find out the elements with high similarity within a group and widely differ from the elements in other groups. Thus, the similarity searching problem gives many benefits in exploring spatial time series databases (Keogh *et al*., 2001). Similarity search is a useful tool in many applications such as classification, clustering and mining of association rules.

It is crucial to improvise methods and algorithms in order to overcome data-driven problems as a wide range of spatial temporal problem at different data types. The explicit pattern can be extracted by establishing techniques for fast similarity searching and evaluation to perform clustering. A very intensive literature has been discussed on spatial variability purposely to analyze the existing spatial homogeneity of time series data. Montero and Vilar (2014)

Corresponding Author: Noor Rasidah Ali, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kedah, Kampus Sungai Petani, Malaysia

categorize clustering approaches into model-free, model-based, complexity-based and prediction-based measures. Exploring and understanding of the embedded information in the datasets with the quality of the data is important to fully exploit the richness of the data. There are numerous available analytical techniques such as statistical approach (Bierman *et al*., 2011) and data mining approach (Fu, 2011; Goler *et al*., 2012). These approaches offer various techniques to extract spatial patterns in order to provide full information hidden within spatial time series data. This information assists to identify regions or periods of time with different rainfall characteristics (wet or dry), whether rainfall data has significant difference at different regions, also to indicate which factors are responsible in rainfall variations. However, rainfall datasets are generally spatially and temporally comprehensive, large in volume and often contain internal correlation and redundancy.

The spatial and temporal elements are embedded in the rainfall data that is actually integrated with deterministic and stochastic (or random) components. The stochastic component consists of measurement error and systematic error that are due to a random nature in rainfall. The noise in rainfall data should be eliminated appropriately in order to obtain a true pattern. Moreover, it is difficult to evaluate similarity among time series data if the data has a noise that will affect the clustering accuracy. The performance can be achieved by considering suitable data preprocessing and transformation techniques (Wu *et al*., 2010; Wu and Chau, 2013). Hence, data preprocessing plays an important role and has the foremost contribution in extracting oblige information from spatial data systems (Mohan, 2014).

Dealing with large scale data, several preprocessing techniques have been applied for data dimensionality reduction and transformation to produce the best spatial clustering solution. There are many preprocessing techniques that have been studied to transform raw data into binary (Finch, 2005; Gaspar *et al*., 2012), mean or average (Koutroumanidis *et al*., 2009), median (Chambers, 2003) and cumulative (Liu and Liu, 2016). These transformation techniques can be used in clustering objects in order to optimize the performance.

To identify a suitable tool for the clustering algorithm, it is necessary to have a method that can be compared to other preprocessing techniques. Therefore, several tasks have been taken into consideration in designing the algorithm. This study proposes a clustering algorithm for the spatial homogeneity of spatial time series rainfall data by introducing the data splitting technique to transform the actual data. The algorithm does not only assign homogeneous objects into the same cluster, it also introduces similarity measures to evaluate the clustering performance between training and testing datasets. Thus, this study focuses on developing an algorithm to perform clustering using the hierarchical method.

MATERIALS AND METHODS

The procedures applied in this study were data preprocessing using *X*-means data splitting, hierarchical clustering for spatial homogeneity and preparing similarity measures to evaluate the clustering algorithm. The overall procedures undertaken in developing the spatial clustering for time series homogeneity is depicted in Fig. 1.

***X*-means data splitting:** One of the data preprocessing techniques is data transformation besides data cleaning and data reduction. In this study, *X*-means is employed to split the data into *k* distinct partitions or levels. The data will be partitioned in such a way that the within-level variation is minimized. By the extension of *K*-means, the algorithm in *X*-means is efficient in searching the space of segment positions using Bayesian Information Criterion (BIC):

$$BIC(C \mid X) = L(X \mid C) - \frac{p}{2}\log n$$

where, $L(X|C)$ is the log-likelihood of dataset X according to model C, $p = k(d + 1)$ is the number of parameters in the model C with dimensionality d and k cluster center and n is the number of points in the dataset. The algorithm searches and chooses the best k partitions over many values with the best BIC score. This statistical-based criterion is used to produce a local decision that minimizes the model's posterior probability. For the different number of partitions, the performance of *X*-means reveals the true number of partitions and it is much faster than repeatedly using accelerated *K*-means.

Hierarchical clustering: In the classical problem, correlation analysis is often used to identify spatial autocorrelation among different time series. This effect should filter out to perform clustering. However, the computational cost for the analysis is very high when the dimension of spatial time series data considered is large. The most common method in clustering is analyzing and exploring the association among the objects in groups, for which objects within group have common characteristics. It is considered unsupervised

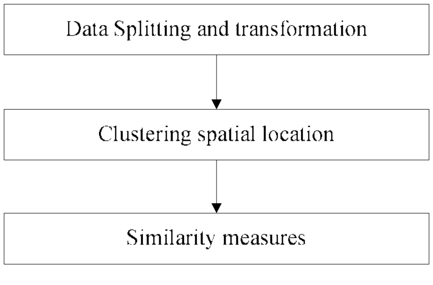

Data Splitting and transformation
Clustering spatial location
Similarity measures

Fig. 1: Clustering procedures

learning because the number of groups is delineated by the data itself that is derived from the proximity between time series (Ratanamahatana *et al.*, 2009).

Practically, the distinction form of the clustering analysis classifies into non-hierarchical and hierarchical. Non-hierarchical is a partitioning method that is known as *K*-means clustering. The main drawback of the *K*-means algorithm is highly depends on the choice of initial point (centroid). The accuracy of clustering has a direct impact from the different choice of centroids (Liu and Liu, 2016). Moreover, this point is initialized at random and is prone to local minima in its searching iteration. The algorithm is particularly centre-based clustering under Gaussian spherical assumption. It performs well if the cluster complies with unimodal distribution or is in spherical form.

Hierarchical clustering demonstrates differently than *K*-means. The procedure begins by merging the most similar time series into small clusters based on pairwise distance and the next step is to combine the most similar clusters until they become a large one. Without to provide the number of clusters, this bottom up clustering is called agglomerative. The hierarchical relationships between objects to perform clusters are commonly displayed graphically in a dendrogram or tree diagram. A basic concept in clustering is to minimize the dissimilarity within clusters and to maximize the dissimilarity between clusters. Superficially, proximity measures are required for clustering the time series data (Kleist, 2015).

Clustering similarity measures: Similarity measures are useful in solving problems related to pattern recognition, for instance, clustering and classification. Let us assume the results of the training and testing datasets produced two sets of clusters, such as $A = \{A_1, A_2,..., A_p\}$ and $B = \{B_1, B_2,..., B_q\}$ respectively. The similarities of the two sets of clusters denote $S(A_i, B_j)$ $for\ i = 1,2,..., p$ and $j = 1,2,..., q$ that must satisfy the following properties:

- $S(A_i, B_j) \geq 0$ (positivity)
- $S(A_i, B_j) = S(B_j, A_i)$ (symmetry)
- $S(A_i, A_i) \geq S(A_i, B_j)$ (Maximality)

A pair wise similarity cluster then to be computed to establish a similarity matrix ($p \times q$) before Jaccard's Similarity Coefficient (Jain and Dubes, 1988) can be applied. The matrix arrangement is as follows:

$$S(A,B) = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1j} & \cdots & S_{1q} \\ \vdots & \vdots & & \vdots & & \vdots \\ S_{i1} & S_{i2} & \cdots & S_{ij} & \cdots & S_{iq} \\ \vdots & \vdots & & \vdots & & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pj} & \cdots & S_{pq} \end{bmatrix}$$

The Jaccard's Similarity Coefficient is $S_{ij} = \frac{|A_i \cap B_j|}{|A_i \cup B_j|}$, where $|A_i \cap B_j|$ and $|A_i \cup B_j|$ represent the number of objects in the intersection and union of clusters A_i and B_j respectively. The similarity between cluster A and cluster B is defined as $\text{Sim}(A,B) = \sum_{\text{all}\, i,j} \frac{S_{ij}}{\max(p,q)}$, where Sim($A$,$B$) should be within [0,1].

Study area: Daily rainfall data series from 14 stations of Perlis are analyzed in this study with the latitude values ranging from 6.3944 to 6.6569 and longitude values from 100.1819 to 100.3528. The state of Perlis is located in the northern part of the west coast of Peninsular Malaysia. With the area of 821 km^2 and it is bounded by the state of Kedah to the south and Thailand on its northern side. The data was collected in a period of five years from January 2010 to December 2014 that was acquired from the Drainage and Irrigation Department (DID) Malaysia. The study area is presented in Fig. 2.

Generally, Malaysia is characterized as equatorial climate, which is warm and humid during the year with an average annual rainfall of more than 2500 mm. The climate in Malaysia is described as a monsoon type, which consists of the southwest monsoon, northeast monsoon and two shorter periods of inter monsoon seasons. Rainfall distribution is mainly due to these monsoon seasons and geographical regions. However, Perlis is considered an arid area or receives very minimum rainfall, particularly during February to April as compared to other states in Malaysia.

The rainfall spatial time series data is a time series measurement collected at specific spatial locations taken at constant intervals of time at each sampling location. Each rainfall station is located at a specific latitude and longitude. Rainfall is measured in millimeters (mm) and observed at different locations as presented in Table 1. The original data has been organized into a data matrix, which is categorized as S-mode (time versus locations) that is common in atmospheric sciences (Richman, 1986). The sampling

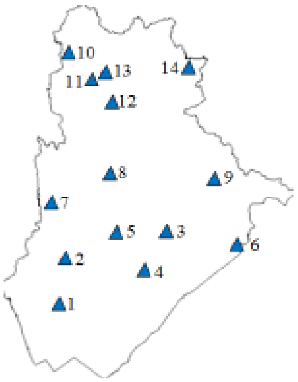
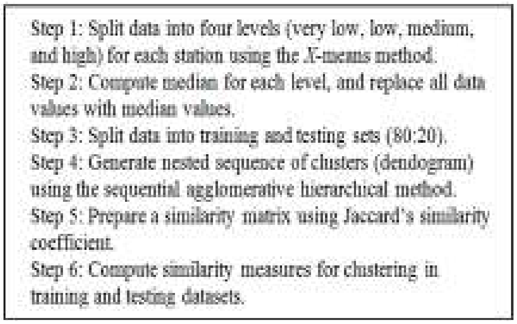


Fig. 2: Rainfall stations in Perlis, Malaysia

Table 1: Time series rainfall data (in mm)

Time (Day)		Location (Rainfall station)				
Year	Day	S1	S2	S3	...	S14
2010	1	0.0	0.0	0.0	...	0.0
	2	0.0	0.5	0.0	...	0.5
⋮	⋮	⋮	⋮	⋮		⋮
	365					
	⋮					
2014	1					
	2					
	⋮					
	365					



Step 1: Split data into four levels (very low, low, medium, and high) for each station using the *X*-means method.
Step 2: Compute median for each level, and replace all data values with median values.
Step 3: Split data into training and testing sets (80:20).
Step 4: Generate nested sequence of clusters (dendogram) using the sequential agglomerative hierarchical method.
Step 5: Prepare a similarity matrix using Jaccard's similarity coefficient.
Step 6: Compute similarity measures for clustering in training and testing datasets.

Fig. 3: Algorithm for spatial homogeneity clustering

location represents the attributes and the sampling time is the data elements.

Spatial homogeneity clustering algorithm: This algorithm is combined with data splitting technique and hierarchical clustering and the overall procedures described in Fig. 3. The procedure begins with data cleaning task, if missing values and inconsistent data values are present in the dataset. The average value from same station, month and day for other different years will replace it.

When the data is considered clean, *X*-means is performed to split the data into four different levels, which are Very Low (VL), Low (L), Medium (M) and High (H) using WEKA (Waikato Environment for Knowledge Analysis) for each rainfall station in step 1. Find the median value in each level and transform all the original values within the level by their own median value. The median value is somehow appropriate because in most of the levels, the data is normally distributed and only the VL level is skewed due to too many zero values.

In order to cross validate the data splitting and data transformation techniques in the previous step, the data has to be splitted into 80% of training (*A*) and 20% of testing (*B*) sets. A dendogram was performed using sequential agglomerative hierarchical clustering from hclust package (Montero and Vilar, 2014) in R programming for both datasets. The clustering is based on Ward's algorithm. To retain the consistency of clustering performance, the number of clusters chosen in this study is four. From the dendogram, Jaccard's Similarity Coefficient was computed between clustering in the training and testing datasets. Finally, similarity matrix using 4 by 4 contingency table was prepared and the similarity of clustering in both the datasets was determined.

In experimenting with the proposed *X*-means splitting data integrated with hierarchical clustering, other data transformation techniques such as mean, median, cumulative, binary and actual data were studied. The mean and the median were calculated in monthly form, while the others were in daily form.

RESULTS AND DISCUSSION

The dendograms were generated from both training and testing datasets by applying the clustering algorithm in above section. The assignments of stations into clusters are shown in Fig. 4 and 5. All stations have been assigned into the same cluster, namely clusters 1, 2, 3 and 4 and are summarized in Table 2.

The clustering is almost consistent between the training and testing datasets except for station 5 (S5) that was assigned in cluster 2 for training data but in cluster 1 for testing data. This is because the position of S5 itself is close to both cluster 1 and cluster 2.

To measure the clustering similarity between two data sets, Jaccard's Similarity Coefficient was used to prepare the similarity index as tabulated in Table 3. Thus, the similarity between clustering using training data (*A*) and testing data (*B*) is Sim(*A*, *B*) = (0.5+0.1667+0.75+1+1)/4 = 0.8542. It gives 85.42% similar clusters which is higher similarity, since the value is close to 1.

Table 2: Cluster membership for training and testing datasets

Station	Station name	Training data (A)	Testing data (B)
S1	Kg Behor Lateh	1	1
S2	Padang Katong	1	1
S3	Guar Nangka	2	2
S4	Arau	2	2
S5	Ngolang	2	1
S6	Ulu Pauh	2	2
S7	Abi Kg Bahru	3	3
S8	Bukit Temiang	3	3
S9	Ladang Perlis selatan	4	4
S10	Wang Kelian	4	4
S11	Kaki Bukit	4	4
S12	Tasoh	4	4
S13	Lubok Sireh	4	4
S14	Padang Besar	4	4

Table 3: Clustering similarity index between training and testing datasets

Training data	Testing data			
	B1	B2	B3	B4
A1	0.5	0	0	0
A2	0.1667	0.75	0	0
A3	0	0	1	0
A4	0	0	0	1

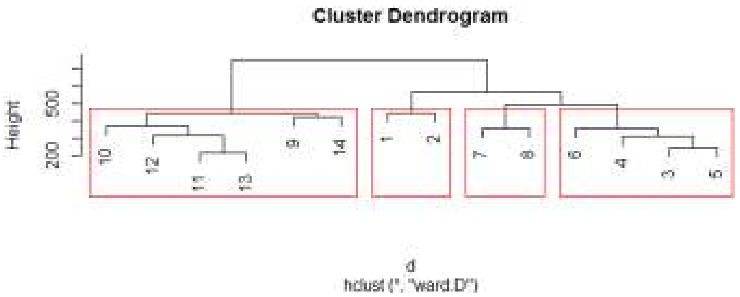


Fig. 4: Dendogram for clustering the rainfall stations using training data

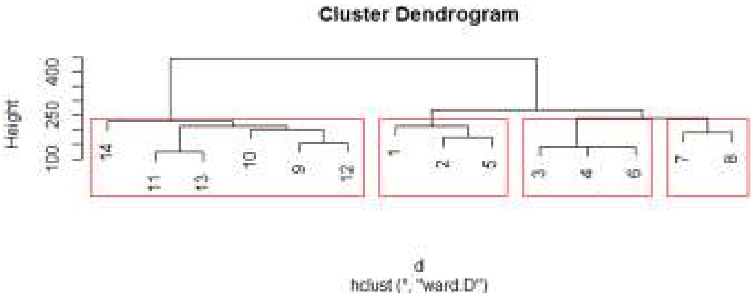


Fig. 5: Dendogram for clustering the rainfall stations using testing data

Table 4: The similarity measures for various data transformations

Type	Transformation	Similarity	Percentage
Month	Mean	0.8081	80.81
	Median	0.6033	60.33
Day	Binary	0.6155	61.55
	Cumulative	0.6834	68.34
	Raw	0.7691	76.91
	Split	**0.8542**	**85.42**

The comparison of results between the proposed clustering algorithm and other data transformation techniques is shown in Table 4. The clustering using mean values for monthly data is better than median with 80.81% similarity. However, the highest similar cluster for daily data is *X*-means data splitting with 85.42%, concurrently outperforming other transformation techniques for all types of data.

The results reveal that the proposed clustering algorithm increases the performance of clustering. The spatial homogeneity of time series clustering is more consistent between training and testing datasets based on the partitioning data and median transformation. These techniques are purposely applied to reduce noise from the series in order to obtain the optimal cluster solution using hierarchical clustering. The same techniques have been applied for all stations, which contain time series rainfall data and then the stations are clustered corresponding to their similarity and homogeneity pattern. The experimental results exhibit that the proposed algorithm outperforms the other five different data transformation techniques for both daily and monthly.

CONCLUSION

A novel spatial clustering for time series using *X*-means data splitting and data transformation techniques was proposed to produce an optimal unsupervised classification among time series rainfall datasets at different locations. The capability of the proposed algorithm to capture a similar pattern along the series between the training and testing datasets are verified with achieving the highest similarity index. To study the effect of data transformation techniques in time series clustering, comprehensive experiments on five datasets with different transformation techniques and using raw data as control are to be compared. It can be concluded that for this, model-free approaches give more consistency and are able to improve the clustering performance. The aim of future work is to extend this method using model-based approaches.

ACKNOWLEDGMENT

The authors wish to thank the Ministry of Higher Education Malaysia for funding this study under Long Term Research Grant Scheme (LRGS/b-u/2012UUM/ Teknologi Komunikasi dan Informasi) and the Department of Irrigation and Drainage Malaysia for supplying hydrology and reservoir operation data.

REFERENCES

Bierman, P., M. Lewis, B. Ostendorf and J. Tanner, 2011. A review of methods for analysing spatial and temporal patterns in coastal water quality. Ecol. Indic., 11(1): 103-114.

Chambers, L.E., 2003. South Australian rainfall variability and trends. BMRC Research Report NO. 92. Bureau of Meteorology Research Centre, Melbourne, pp: 33-34.

Finch, H., 2005. Comparison of distance measures in cluster analysis with dichotomous data. J. Data Sci., 3: 85-100.

Fu, T.C., 2011. A review on time series data mining. Eng. Appl. Artif. Intel., 24(1): 164-181.

Gaspar, P., J. Carbonell and J.L. Oliveira, 2012. On the parameter optimization of support vector machines for binary classification. J. Integr. Bioinform., 9(3): 201-211.

Goler, I., P. Senkul and A. Yazici, 2012. Spatio-temporal Pattern and Trend Extraction on Turkish Meteorological Data. In: Gelenbe, E., R. Lent and G. Sakellari (Eds.), Computer and Information Sciences II. Springer, London, pp: 505-510.

Jain, A.K. and R.C. Dubes, 1988. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, New Jersey, pp: 320.

Keogh, E., K. Chakrabarti, M. Pazzani and S. Mehrotra, 2001. Dimensionality reduction for fast similarity search in large time series databases. Knowl. Inf. Syst., 3(3): 263-286.

Kisilevich, S., F. Mansmann, M. Nanni and S. Rinzivillo, 2010. Spatio-temporal Clustering. In: Data Mining and Knowledge and Discovery Handbook. Springer, US, pp: 855-874.

Kleist, C., 2015. Time series data mining methods: A review. M.S. Thesis, Department of Statistics, School of Business and Economics, Humboldt-Universitat zu Berlin, Berlin.

Koutroumanidis, T., G. Sylaios, E. Zafeiriou and V.A. Tsihrintzis, 2009. Genetic modeling for the optimal forecasting of hydrologic time-series: Application in Nestos river. J. Hydrol., 368(1-4): 156-164.

Liu, Y. and L. Liu, 2016. Rainfall Feature Extraction using Cluster Analysis and its Application on Displacement Prediction for a Cleavage-parallel Landslide in the Three-Gorges Reservoir Area. Natural Hazards and Earth System Sciences Discussions Papers (January), pp: 1-15.

Mohan, A., 2014. A new spatio-temporal data mining method and its application to reservoir system operation. M.S. Thesis, University of Nebraska, Lincoln.

Montero, P. and J.A. Vilar, 2014. TSclust: An R package for time series clustering. J. Stat. Softw., 62(1): 1-43.

Ratanamahatana, C.A., J. Lin, D. Gunopulos, E. Keogh, M. Vlachos and G. Das, 2009. Mining Time Series Data. In: Data Mining and Knowledge Discovery Handbook. Springer, US, pp: 1069-1103.

Richman, M.B., 1986. Rotation of principal components. Int. J. Climatol., 6(3): 293-335.

Wu, C.L. and K.W. Chau, 2013. Prediction of rainfall time series using modular soft computing methods. Eng. Appl. Artif. Intel., 26(3): 997-1007.

Wu, C.L., K.W. Chau and C. Fan, 2010. Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. J. Hydrol., 389(1-2): 146-167.