A Novel Spatio-temporal Clustering Approach by Process Similarity

Fan Lin, Kunqing Xie, Guojie Song*, Tianshu Wu
Key Laboratory of Machine Perception, Ministry of Education, Peking University, China
gjsong@cis.pku.edu.cn

Abstract

Spatio-temporal data in earth science is usually of huge volume and high dimensionality. Clustering is usually preparation work for many applications in this field. Traditional clustering methods are of high complexity when applied to spatial-temporal data. Traditional methods neglect the changing process of the temporal data by treating data with consecutive timestamps independently and do not consider objects' spatial proximity which is important in earth science. An effective spatial-temporal tight clustering approach with domain knowledge is proposed for this field. The similarity measurement for the cluster method named Value- Process (VP) measurement estimates similarity of two objects from the view of their attributes value and the value changing process. The computation of the measurement adopts a filter-and-refinement strategy with a growing search window to lift the efficiency and guaranty the spatial proximity. Based on the VP similarity measurement, a tight clustering approach, which has a more strict cluster rule, is applied to the global climatic dataset and the promising result shows that it was an effective clustering method for the spatial-temporal data in earth science.

1. Introduction

Clustering is usually a necessary preparation for many kinds of real life applications, and much research has been devoted to the problem of spatio-temporal sequence clustering in earth science data[3][4][5]. However, spatio-temporal data in earth science is massive and high dimensional, and brings great challenge for research. Besides that, spatio-temporal data has a regional character, and the time order of the observation implies the changing process of meteorological index, figure 1 shows the climate process lines of 4 cities in China(Jilin, in the northeast ; Kunming, in the southwest; Shantou, in the south; Urumqi, in the northwest). On the plan, the X co-ordinate value presents temperature and the Y co-ordinate value presents precipitation. Each point is the average value of that area in one month. The lines show how the climate of these 4 cities in China changed in the year and they are much more distinguished if we take the process in consideration when clustering.

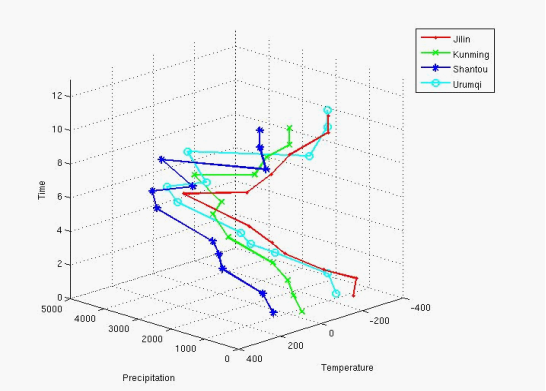

Figure 1. Process of Spatio-temporal data

The existing clustering methods do not take attribute value, spatial relationship and time sequences into account in their similarity measurement at the same time. Since now, most of the methods measure similarity of two objects only by their attributes values independently without considering their spatial distance nor their time order.

So we propose a tight clustering method based on our Value-Process (VP) measurement making use of the characters of spatio-temporal data. The Value-Process measurement estimates the similarity of two observation points both in the attributes value and the value changing process. The VP measurement is computed using a filter-and-refinement strategy with a growing search window. The strategy guaranties the spatial proximity and lifts the efficiency. After the similarity computation, the tight clustering method is proposed for clustering, which could insure that any two points in the same cluster are similar under the VP measurement.

This paper is organized as follows: Section 2 introduces the distance function based on process similarity measurement. Section 3 presents an effective tight cluster method

*Corresponding author

978-0-7695-3735-1/09 $25.00 © 2009 IEEE
DOI 10.1109/FSKD.2009.584

based on the above measurement. Section 4 presents experiments that show significant results, while Section 5 concludes the paper.

2. The Distance Function

The choice of distance function has great influence on the measuring of similarity, particularly important in process-based sequence clustering because of the computational complexity. Hence, we need an effective and efficient distance function to measure the similarity of sptiotemporal data in high dimensional space.

2.1. Brief Description of Process Data

We focused on the global climate dataset in form of raster data, which can be defined as a set of sequences. Each sequence presences the changing process of climate on some place of the earth.

Given

$$\left\{ \begin{array}{ccccccc} S^{(1,1)}, & S^{(1,2)}, & \cdots, & S^{(1,j)}, & \cdots, & S^{(1,n)} \\ S^{(2,1)}, & S^{(2,2)}, & \cdots, & S^{(2,j)}, & \cdots, & S^{(2,n)} \\ \vdots, & \vdots, & \ddots, & \vdots, & \ddots, & \vdots \\ S^{(i,1)}, & S^{(i,2)}, & \cdots, & S^{(i,j)}, & \cdots, & S^{(i,n)} \\ \vdots, & \vdots, & \ddots, & \vdots, & \ddots, & \vdots \\ S^{(m,1)}, & S^{(m,2)}, & \cdots, & S^{(m,j)}, & \cdots, & S^{(m,n)} \end{array} \right\}$$

is a sequences set, where (i, j) presents a location on the earth, while i and j presents the latitude and longitude individually, m and n are their max value. Each $S^{(i,j)}$ is a sequence in the location (i, j), which presences the process of climate changing. The sequence can be defined as:

$$S^{(i,j)} = S_1^{(i,j)} S_2^{(i,j)} \ldots S_k^{(i,j)} \ldots S_l^{(i,j)}$$

where $S_k^{(i,j)}$ is an event that occurs at the time point k, l is the total number of time points. In the global climate dataset, each time point presents one month, and the data covers from 1901 to 2002, so $l = 102 \times 12 = 1224$. An event at each time point contains nine attributes, such as temperature, precipitation, etc.

2.2. Process-based Similarity Measurement

In this paper we propose a similarity measurement named Value-Process(VP) measurement to calculate the similarity of observation points on the earth according to the climatological value, defined as

$$VP(X,Y) = \begin{cases} 1, & V(X,Y) \geq v \text{ and } P(X,Y) \geq p \\ 0, & \text{others} \end{cases} \quad (1)$$

where $VP(X, Y)$ present the relation of two spatial points X and Y , 1 means "they are similar" while 0 means "they are dissimilar". In this formulation, $V(X, Y)$ is the similarity measurement of climatological attributes value, $P(X, Y)$ is the similarity measurement of value changing process, while v, p are the threshold value respectively. The details are as follows.

2.2.1 The value-measurement

In our experiment, we use four attributes of the climatic dataset for similarity estimation, temperature, precipitation, annual amplitude of temperature, annual amplitude of precipitation. The four attributes are correlated to each other in one way so it is difficult to identify the respective weighings of them while calculating the result. Here we analyze the internal relation of the four attributes using principal component analysis[11] to obtain a new integrated indicator which involves most of the information contained in the primitive attributes. For a point (i, j), the matrix in formula (2) could be specified as a 4×1224 matrix and the principal component z_i is computed. As shown in formula (2).

We randomly chose an area on the earth which covers more than 10,000 points and compute their first principal components, and find that the first principal component contains 85% information of the four attributes. It implies that the new integrated indicator could be used to estimate the value-similarity of two points. So the value measurement is as follows:

$$\begin{aligned} V(X,Y) &= V(L_{x_i,x_j}, L_{y_i,y_j}) \quad (2) \\ &= \|PCA(L_{x_i,x_j}, z_1) - PCA(L_{y_i,y_j}, z_1)\| \end{aligned}$$

where $PCA(L_{x_i,x_j}, z_1)$ is the first principal component of L_{x_i,x_j}, a $1 \times T$ vector.

2.2.2 The process-measurement

Temperature and precipitation are time-varying, the time sequences imply the changing process and it is an important character of the climate. However, the existing clustering methods treat attributes value as uncorrelated discrete value, omitting the changing process. We use Pearson product-moment correlation coefficient as the process measurement.

In statistics, the Pearson product-moment correlation coefficient is a measure of the correlation of two sequences X and Y measured on the same object or organism, that is, a measure of the tendency of the sequences to increase or decrease together[6]. It is defined as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\delta_X \delta_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\delta_X \delta_Y} \quad (3)$$

where E is the expected value operator, cov means covariance, $\mu_X = E(X)$, $\delta_X^2 = E(X^2) - E(Y^2)$ and likewise for Y. The process-measurement $P(L_{x_i,x_j}, L_{y_i,y_j}) = \rho_{L_{x_i,x_j},L_{y_i,y_j}}$

Problem Statement: Our task is to find an efficient process-based sequence clustering method where the distance between two sequences is measured by Eq (1).

3 The Clustering Algorithm

To overcome the challenges caused by high dimensionality of spatio-temporal sequence data, filter-and-refinement strategy based on domain knowledge is introduced in section 3.1. To guarantee the quality of clustering, tight clustering method is present in section 3.2.

3.1 Filter-and-refinement strategy

For an arbitrary observation point (i, j), it is time-consuming to compare it to all the other observation points in the huge dataset to find all the similar points. On the other hand, according to the first law of geography[7], points in the same cluster are expected to have spatial proximity. So we propose a filter-and-refinement strategy with a growing search window to lift the efficiency and guaranty the spatial proximity.

(1) *The filter stage.* Compared to $P(X, Y)$, $V(X, Y)$ has a lower complexity. So in the filter stage, we use $V(X, Y)$ to estimate the similarity between the observation points to filter out some dissimilar points fail to satisfy the constraint of $V(X, Y)$. To further reduce the complexity and guaranty the spatial proximity, a growing search window is added to the filter stage. Bellows are some definitions.

Table 2. Parameter Description

Param	Description
$R_{i,j}(r)$	a sequence area centers on (i, j) where r is the side length
$D_{i,j}(r)$	the value-similar points of (i, j) in $R_{i,j}(r)$ $\{(e, f)\|V(L_{i,j}, L_{e,f}) \geq v, (e, f) \in R_{i,j}(r)\}$
Δ_r	the growing speed of r
$\Delta D_{i,j}(r)$	$\Delta D_{i,j}(r) = D_{i,j}(r) - D_{i,j}(r - \Delta r)$

In the filter stage, to search similar points of (i, j) by $V(X, Y)$, we first limit the searching range in $R_{i,j}(r)$, a square area centers on (i, j) where r is the side length. Considering all the observation points in the whole dataset as nodes in a graph named similarity graph, we draw a directed line from (i, j) to (e, f), if they are value-similar. Then the searching range is enlarged to $R_{i,j}(r+\Delta r)$, shown in figure 2. Search is repeated until the density of increased similar points in the newly enlarged area is smaller than a threshold. This process is describe in Algorithm 1.

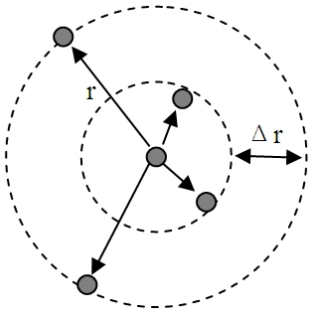

Figure 2. Searching in a growing window

Algorithm 1 Filtering process

Input: (i, j): the initial observation point for which the similar points will be searched for;
r and Δr: the same definition in Table 1;
Output: Directed links from (i, j) to its value-similar points;
Procedure:

```
1: r=0;
2: DO
3:    r = r + Δr;
4:    compute D_{i,j}(r) and ΔD_{i,j}(r);
5:    make all the points in ΔD_{i,j}(r) leaf nodes of (i, j);
6: While ΔD_{i,j}(r) >ε;
```

(2) The refine stage. When the filter stage is finished, every observation point has many links to its similar nodes. In the refine stage, the link from (i, j) to (e, f) is preserved only if there is also a link from (e, f) to (i, j) and after that, all the links in the similarity graph are bidirectional links. Then we examine the process similarity of the linked nodes by $P(X, Y)$ and remove the link between dissimilar points. When the filter and refine stages have finished, all the linked nodes in the similarity graph satisfy the VP measurement and have spatial proximity.

3.2. The Tight Clustering

A high quality cluster result with a strict similarity measurement would make applications based on it more effective.

DEFINITION 1(**Tight correlated cluster**). *A cluster in which every two objects are similar. For a tight cluster in our similarity graph, there is a bidirectional link between any two observation points.*

DEFINITION 2(**Maximal tight correlated cluster**). *The size of a tight correlated cluster is the number of points in it. Maximal tight correlated cluster is the tight correlated cluster with the maximum size in the graph.*

We want to find the maximal tight correlated cluster in the similarity graph and make it a cluster in the clustering result. Refreshing the graph by removing the clustered points, we identify the next cluster with the same approach. Repeat the process above until all the points are clustered. Maximal tight correlated clusters in our similarity graph is quite similar to cliques in a undirected graph. Finding the maximum clique in a graph is a NP-complete problem. It is impractical to exactly find the maximal tight correlated cluster in the similarity graph based on the huge climatic dataset. So we design a greedy algorithm for this problem.

Algorithm 2 Tight Clustering Algorithm

Input: The similarity graph G;
Output: Tight correlated clusters;
Procedure:

1: G'=G;
2: DO
3: the result set $R = \varnothing$;
4: put all the nodes in G' into candidate set C;
5: while $C \neq \varnothing$;
6: For each node in C, compute the number of other nodes linked with it in C, denoted as *C-degree*;
7: Move the node with the maximum *C-degree* from C to R;
8: Let candidate set $C = \varnothing$;
9: Put the nodes linked with all the nodes in R into the candidate set C;
10: R is the tight correlated cluster;
11: Remove all the nodes in R from G';
12: While $G' \neq \varnothing$;

The greedy algorithm is based on the assumption that the node with the maximum C-degree (see line 4 in Algorithm 2.2) is in the maximal tight correlated cluster.

4 Experiments and Results

4.1 Earth science data description

We use real global data for experiments. Global climate datasets CRU TS 2.10[8] contains nine variables including monthly average precipitation, monthly average temperature etc and we mainly use variables of monthly average precipitation, monthly average temperature.

Table 3. Description of datasets used in experiments

Dataset name	CRU TS 2.10
From	CRU
Space range	Global land
Space resolution	$0.5^o \times 0.5^o$
Time range	1901 - 2002
Time granularity	Month

4.2 PCA Value experiment

In this section, we discuss the average PCA value of points both in the specific circle and in the specific window. For each point of the datasets, calculate its distance to all points in the specific circle, and then figure out its average distance to these points. Finally, figure out the average of all points in dataset. For example, if i=1, then there are 8 points around it in the raw. The calculation of the specific windows is the same with the circle one, while such method add up all circle with the window.

Figure 3 shows the result, which shows one rule: the closer the points are, the smaller the distance of their PCA value is. Such rule is correspond to the spatial autocorrelation.

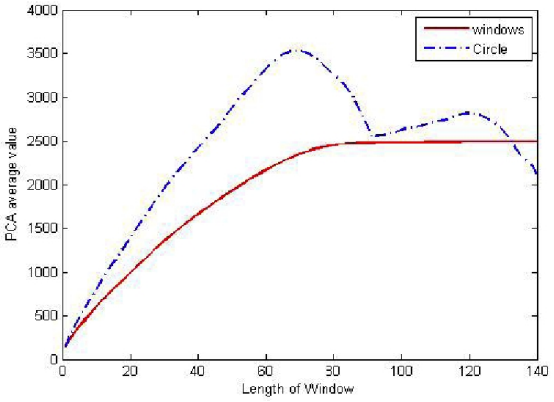

Figure 3. Searching in a growing window

4.3 Tight Clustering result of Global Climate Sets

In this section, we consider an application of our tight clustering method to the earth science data. In particular, we choose the CRU TS 2.10, which was introduced above. These time series were preprocessed to remove seasonal variation. Figure 4 shows the result, the data points with the same color is in the same tight clique.

4.4 Intra- and Inter- Cluster Distance Experiment

In order to estimate the performance of our algorithm, we use intra- and inter- cluster distances. Given a set of clusters $C^1, C^2, \ldots, C^i, \ldots, C^V$ presents the clustering result of N objects, where $C_i = c_1^i, c_2^i, \ldots, c_j^i, \ldots, c_{m_i}^i$ is the i-th cluster in which c_j^i is an object in it. V is the total number of clusters, m_i is the number of objects in the i-th cluster. The inter- and intra- cluster distances can be defined below:

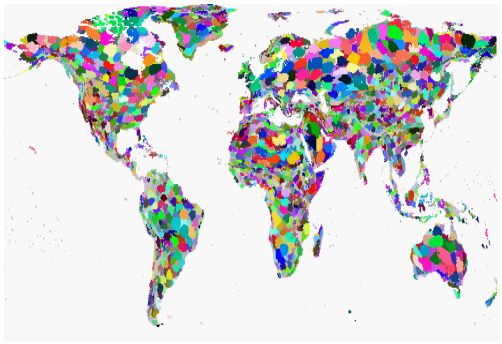

Figure 4. Clustering result of global climate sets

$$D_{inter} = \sum_{i=1}^{V} \frac{m_i}{N} \sqrt{\frac{1}{n} \sum_{j=1}^{m_i} (s_j^i - \overline{s^i})^2} \quad (4)$$

$$D_{intra} = \sqrt{\frac{1}{n(n-1)} \sum_{i \neq j}^{V} (\overline{s^i} - \overline{s^j})^2} \quad (5)$$

where s_j^i is the sequence of j-th point in the i-th cluster, $\overline{s^i}$ is the average value of the i-th cluster.

If the inter cluster distance is smaller and the intra cluster distance is larger, we can say that the performance of the algorithm is better. We do this experiment compare with SNN algorithm[5] and K-means[3] algorithm, the result shows our method gets good result.

Table 4. Inter- and Intra- Cluster Distance Experiment

	Inter			Intra		
Dataset	TC	Kmeans	SNN	TC	Kmeans	SNN
Africa	48	52	211	279	222	179
America	65	75	210	304	285	87
Asia	50	56	140	369	260	124
Australia	33	40	158	227	219	150
Europe	51	49	76	325	231	117

5 Conclusions

Spatio-temporal data in earth science is usually of huge volume and high dimensionality. Clustering is usually preparation work for many applications in this field. Traditional clustering methods are of high complexity when applied to spatio-temporal data. Traditional methods neglect the changing process of the temporal data by treating data with consecutive timestamps independently and do not consider objects' spatial proximity which is important in earth science. An effective spatio-temporal tight clustering approach with domain knowledge is proposed for this field. The similarity measurement for the cluster method named Value-Process (VP) measurement estimates similarity of two objects from the view of their attributes value and the value changing process. The computation of the measurement adopts a filter-and-refinement strategy with a growing search window to lift the efficiency and guaranty the spatial proximity. Based on the VP similarity measurement, a tight clustering approach, which had a more strict cluster rule, was applied to the global climatic dataset and the promising result told us it was a effective clustering method for the spatio-temporal data in earth science.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No.60703066 and No.60874082.

References

[1] Clustering Algorithms for Spatial Databases: A Survey. Erica Kolatch.http://www.cs.umd.edu/ kolatch/papers/SpatialClustering.pdf

[2] Lizheng Jiang. Study on the Methods of Correlation Pattern Mining(In Chinese). Peking University Doctoral Dissertation, 2007.

[3] A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, New Jersey, March 1988.

[4] Ester, Martin, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, Oregon, August 1996.

[5] Levent Ertoz, Michael Steinbach, Vipin Kumar. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In Proceedings of Third SIAM International Conference on Data Mining, San Francisco, CA, USA, May 2003.

[6] Rafael C. Gonzalez and Richard E. Woods. Digital image processing. Addison Wessley Publishing Company, 1992.

[7] W. R. Tobler. Cellular Geography, Philosophy in Geography. Gale and Olsson, Eds., Dordrecht, Reidel, 1979.

[8] Mitchell, T.D., Carter, T.R., Jones, P.D., Hulme,M., New, M., 2003: A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: the observed record (1901-2000) and 16 scenarios (2001-2100). Journal of Climate: submitted.