

DATA MINING FOR TELECONNECTIONS IN GLOBAL CLIMATE DATASETS*

Xiaoping Gao, Kunqing Xie, Xingxing Jin, Xiaofeng Lei*

Center of Information Science, Peking University, Beijing, 100871, China

*Email: xiaopinggao@gmail.com

ABSTRACT

Teleconnection is a linkage between two climate events that occur in widely separated regions of the globe on a monthly or longer timescale. In the past, statistical methods have been used to discover teleconnections. However, because of the overwhelming volume and high resolution of datasets acquired by modern data acquisition systems, these methods are not sufficient. In this paper, we propose a novel approach to finding teleconnections in global climate datasets using data mining technologies. We present experiments on real datasets and find some interesting teleconnections, including well-known ones such as ENSO. The experiments indicate that our method is usable and efficient.

Keywords: Global climate datasets, Data mining, Association rule, Teleconnections, Climate events

1 INTRODUCTION

Nowadays, with the development of geographic technologies such as remote sensing and the GCOS (global climate observing system), more and more geographic data have been acquired and collected. However, we cannot often take full advantage of these data because of the large data size. In geographic science, when analyzing earth science data, people often use traditional statistical methods, such as singular value decomposition (SVD), principal component analysis (PCA), EOF, etc. Widely used statistical methods have helped geography scientists find many geographic rules and models, but there are often difficulties when using them:

1. Data in geographical science may not be independent and identically distributed (iid). For example, because of spatial-autocorrelation, some climate events may occur closely. This can make statistical analysis results more difficult to explain.
2. There is a good deal of non-numeric data in geographical science, such as vegetation and earth type. Traditional statistical methods perform poorly when processing non-numeric data.

To deal with these problems, we use data mining methods to extract knowledge from large amounts of geographic data. In this paper, we use several data mining technologies to find the association rules between extreme climate events.

* This research is sponsored by the China National Advanced Science & Technology (863) Plan under Grant SQ2006AA12Z214047 and the Beijing Science Foundation under Grant 4062019.

First we use SNN, a high dimensional clustering method, to cluster original time series data, and then we extract extreme events from time series data according to the definition of IPCC at the clusters. Finally, we use MINEPS, a method for mining frequent episodes in event sequences, to help find association rules in these extreme events. As a result, we have found teleconnections in global climate datasets. Some of them are well-known in geographic science, such as ENSO, but there are other teleconnections that have not been affirmed yet. These teleconnections may be explained in geographic science in the future.

This paper is organized as follows: in section 2, we explain the basic concepts which we use in our experiment, including data mining, teleconnections, etc.; in section 3, we describe in greater detail the main algorithms for mining. We explain the experiment and results in section 4, and section 5 concludes the paper.

2 BASIC CONCEPTS

2.1 Data mining, spatial clustering, and association rules

Data mining means to “mine” interesting patterns in large amount of data. In geographic science, these data contain temporal data (often time-series data) and spatial data (often in raster format). Spatial-temporal data mining is a research direction to gain knowledge from these data.

Geographic events often happen in a region with a hard-to-locate edge, so it is not appropriate to define a threshold to determine the region. People often use the spatial clustering method to find the region of a spatial event. The principle is “to minimize variability within clusters, maximize variability between clusters.” Spatial clustering also has the advantage of reducing complexity. After clustering, we can analyze data at a higher level of granularity, so computation can be reduced when applying other mining methods.

Association rule mining finds interesting association rules or relationships among data items. Typically, they are used in “market basket analysis” to find what items are often purchased together, so the manager can adjust the layout of items in a store or optimize the price. Association rules are formulas such as “ $A \Rightarrow B(\text{sup}, \text{conf})$,” where A and B are item sets, and *sup* and *conf* are two measures of the association rule. For example, “ $\text{beer} \Rightarrow \text{bread}(20\%, 60\%)$ ” means “There are 20% of transactions which include beer and bread together and 60% of customers who brought beer also brought bread” (Agrawal, Imielinski, & Swami, 1993).

2.2 Extreme climate events

In geographic science, there are events that happened abnormally or suddenly. Some of these we call “extreme climate events.” These are some conditions we can consider to help us define “extreme” are:

1. **Infrequency:** IPCC made a definition of “extreme” according to the infrequency of the events: “An extreme weather event is an event that is rare within its statistical reference distribution at a particular place. Definitions of ‘rare’ vary, but an extreme weather event would normally be as rare as or rarer than the 10th or 90th percentile” (Gitay, Suarez, Watson, & Dokken, 2002)
2. **Intensity:** For example, we can consider a precipitation of 300mm as an extreme climate event. This can give us an intuitional definition of “extreme,” but it depends on the knowledge of experts.

Beside infrequency and intensity, we can also consider the difference from an average value and the influence on the society as additional measures of extremes.

2.3 Teleconnections

According to the American Meteorological Society's definition, teleconnection is "a linkage between weather changes occurring in widely separated regions of the globe and a significant positive or negative correlation in the fluctuations of a field at widely separated points" (American Meteorological Society, 2000). For example, El Nino in winter can cause high precipitation in spring in Europe (van Oldenborgh, Burgers, & Klein Tank, 2000).

At first, people discovered teleconnections by visual observation of data, but now, it is impossible to do this because data size is exploding. Now we always use statistical methods such as RPCA (Rotated Principal Component Analysis) and SVD to analyze the relationship among climate events. However, these methods are not sufficient because of the overwhelming volume and high resolution of datasets acquired by modern-day satellites and other data acquisition systems. In this paper, we introduce a new method of data mining to cope with this problem.

3 ALGORITHMS

We use three steps to mine teleconnections among extreme climate events. First, the high dimensional clustering method SNN is used to cluster original time-series data. This helps us reduce the complexity of the large data size. Then we extract extreme climate events from the time-series data of the clusters according to the "extreme" definition of IPCC (Gitay, et al., 2002) Finally, we use the MINEPS algorithm to find frequent episodes in these event sequences to help find association rules (teleconnections) among the extreme climate events.

3.1 SNN clustering methods

In geographic science, when clustering time series data, we need to measure the similarity of two time series, so clustering on time series data is often of high dimension because time series always have a very large length. For example, the data in our experiment has 12x100(12months multiplied by 100 years) dimensions. The traditional Euler distance measure is not very good at such high dimensions because as the dimension increases, the difference reflected by the Euler distance gets smaller. In this paper, we use a high-dimensional technique more suitable as a measure, SNN, to measure the similarity of two time series. SNN is the shortened form of Shared Nearest Neighbor, which was first introduced by Jarvis and Patrick (1973) and was improved to cope with different sizes, shapes and densities data by Ertoz, Steinbach, and Kumar (2003).

Here we introduce some definitions:

1. **Similarity** according to SNN: given two objects p and q in dataset D , the similarity of p and q is:

$$\text{similarity}(p, q) = |NN(p) \cap NN(q)|$$

For an arbitrary k , $NN(p)$ means the k nearest neighbors of p . The similarity measure means the size of the Shared nearest neighbors of p and q .

2. **Density** of an object p : given a threshold Eps ,

$$density(p) = |\{q : similarity(p, q) \geq Eps\}|$$

If $density(p) > minEps$, we can conclude that p is in a high density region, and we call p a *Core Point*.

With these definitions, we describe the clustering algorithm as below:

input : $k, Eps, minEps$

1. Calculate the distance matrix M . Here we use the Euler distance of the Manhattan distance.
2. With distance matrix M , find the top k nearest neighbours of each object and save the result to a binary matrix B .
3. According to B , calculate the SNN matrix and density of each object.
4. Find the core points with respect to $minEps$.
5. Cluster the core points. If the SNN distance of two core points is less than Eps , then merge them into the same cluster.
6. Abandon the noisy points. Noisy points are points outside of the Eps distance region for any core points
7. Merge the non-noisy and non-core points into the core points clusters according to SNN similarity.

3.1.1 Optimization using auto-correlation

In geographic science, there is a observation, "Everything is related to everything else, but near things are more related than distant things." This is called "Tobler's First Law" (Tobler, 1979). According to this, we can search the k -nearest neighbors within a limited region to avoid searching the entire data set. In our experiment, we choose a circle with a proper radius as the region and find that the efficiency was evidently improved without affecting the result of the cluster.

3.2 Define the extreme event sequence

After clustering the entire climate data set into a set of clusters, we normalize all the event sequences in a cluster into one sequence. According to the definition of extreme climate events in section 2, we formally define the extreme climate event as below:

Given p and q ($0 < p < q < 1$), and a time series data with length n , we sort the data in ascending order; the first $p*n$ and the last $(1-q)*n$ data items are extreme events.

Here p and q are thresholds provided by experts, According to IPCC's definition (Gitay, et al., 2002) p is 10% while q is 90%.

3.3 Association rules mining

Before we discuss association rules mining, we introduce some concepts in mining sequence data.

1. **Event**: a data item e happening at a specific time. Let E be an event set, $e \in E$. $\langle e, t \rangle$ means an event

that happened at time t .

2. **Event sequence:** a triple $\langle s, Ts, Te \rangle$, s means the event set $\langle e1, t1 \rangle \dots \langle en, tn \rangle$; and Ts and Te mean the start and end time of s . For example: Figure 1 shows an event sequence, where $s = \{ \langle B, 10 \rangle, \langle B, 11 \rangle, \langle A, 15 \rangle, \dots, \langle B, 35 \rangle \}$, and $Ts = 10, Te = 35$

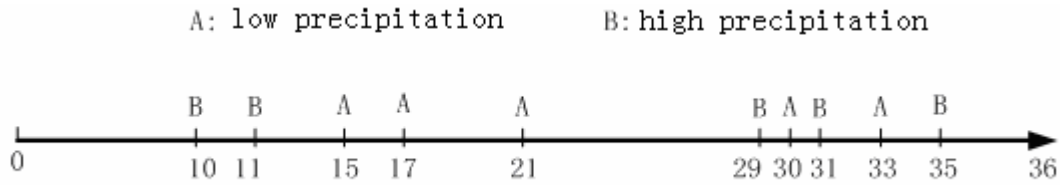


Figure 1. An event sequence

3. **Episode:** a triple $\langle V, \leq, g \rangle$, where V is a set of nodes, \leq is a partial order on V , and g is a mapping from V to E . There are three kinds of episodes according to \leq , serial episodes, parallel episodes, and composite episodes. For example: α is serial, β is parallel and γ is composite episode.

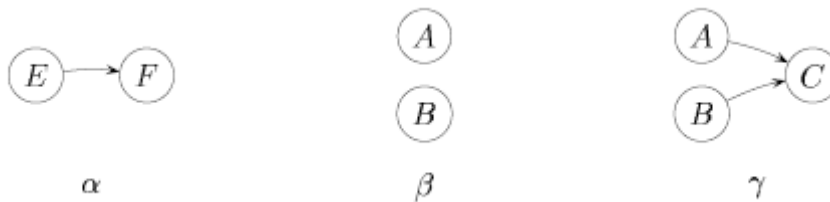


Figure 2. Serial, parallel, and composite episodes

MINEPI is based on the concept of minimal occurrence. A minimal occurrence $[ts, te)$ of episode a in a sequence $S = \langle s, Ts, Te \rangle$ means: a occurs in $[ts, te)$ in S , and there is not any $[ts', te')$ in $[ts, te)$, which is also an occurrence of a . For example, $[11, 15)$ is a minimal occurrence of (B, A) . We denote all minimal occurrences of episode a in S by $mo_{win}(a)$, that is $mo_{win}(a) = \{ [ts, te) \mid [ts, te) \text{ is a minimal occurrence of } a \}$.

With a predefined win and a $threshold$, if the times of minimal occurrence of episode a are larger than or equal to $threshold$, we call a a frequent episode. We use MINEPI (Mannila & Toivonen, 1996) to find all frequent episodes. According to this, the time complexity is $O(|mo(a)| + |mo(b)|)$.

Next we describe the algorithm of episode association rules finding. Episode association rules are rules such as $a[win1] \geq b[win2]$, where a and b are episodes and $a \leq b$. This rule means that if there is a minimal occurrence of a , there is a minimal occurrence of b . The support and confidence of this rule are defined below:

$$sup = |\{ [t_s, t_e) \in mo_{win1}(a) \mid [t_s, t_s + win2) \text{ is a occurrence of } b \}|$$

$$conf = \frac{|\{[t_s, t_e) \in mo_{win1}(a) \mid [t_s, t_s + win2) \text{ is a occurrence of } b\}|}{|mo_{win1}(a)|}$$

We provide an algorithm for mining these association rules. Given minimal confidence min_conf , minimal support min_sup , maximal window $win1$ of the first episode, and $win2$ of the second episode, we use MINEPI to find all the frequent episodes in set F . Now we use the algorithm below to find the association rules and the corresponding support and confidence.

Algorithm 1: *generateRules*($F, min_sup, min_conf, win1, win2$)

For all b in F

If $|mowin2(b)| \geq min_sup$

For every sub episode a in b

If $|mowin1(a)| \geq min_sup$

(sup, conf) = *compute_sup_conf*($a[win1] \Rightarrow b[win2]$)

If sup > min_sup, conf >= min_conf

Return rule $a[win1] \Rightarrow b[win2]$

Algorithm 2: *compute_sup_conf*($a[win1] \Rightarrow b[win2]$)

$sup = 0$

For all $[ts, te)$ in $min_{win1}(a)$

For all $[us, ue)$ in $min_{win2}(b)$ and $us > ts$

If $u_e - t_s \leq win2$

$sup++$

$conf = sup / |mo_{win1}(a)|$;

Return (sup, conf)

3.3.1 An extension to the association rule

1. Association rules with time-lags

In geographic science, teleconnections often happen with a time lag of from one to several months. For example: El Nino in the winter can cause high precipitation in spring in Europe; the time lag is about 3 months (van Oldenborgh, et al., 2000). We introduce the association rule with time lags

as: $a[win1]_{lag1, lag2} \Rightarrow b[win2]$, which means if there is a minimal occurrence of a , then with a time lag of minimum lag1 and maximum lag2, there will be a minimal occurrence of b .

When mining these association rules with time lags, we change the algorithm of support and confidence: *compute_sup_conf*($a[win1] \Rightarrow b[win2]$). Let $us-ts$ change from $lag1$ to $lag2$ when

countering the support of a rule. The other parts of the algorithms remain the same.

2. Association rules with constraints

Although we have used clustering to avoid some auto-correlation in the data source, there are still rules with high support and confidence because of auto-correlation, for example the precipitation of two neighboring regions has a strong association. To avoid this, we can constrain the second part of the association rule from different data sources. For example, one is temperature while the other is precipitation. Experiments showed that the result was improved after introducing these constraints.

4 EXPERIMENTS AND RESULTS

4.1 Datasets for experimentation

We have applied our methods to geographic datasets. These datasets include global climate datasets CRU TS 2.10 from CRU, HadISST Sea Surface Temperature data, land surface NDVI data from AVHRR Land Pathfinder, etc. The properties of these data are described in Table 1.

Table 1. Data sets and properties

Dataset name	From	Space range	Space resolution	Time range	Time granularity
CRU TS 2.10	CRU	Global land	0.5°×0.5°	1901~2002	month
HadISST	Hadley	Global Ocean	1°×1°	1870~2003	month
NDVI	NOAA	Global land	1°×1°	1981~2001	Month
SOI	CRU			1866~2004	Month
NINO3.4	NOAA	NINO3.4 region		1871~Now	Month

4.2 Results of experiments

We summarize some abnormal climate events of the global land in Table 2.

Table 2. Some teleconnections

EL NINO/LA NINA	Region	Extreme events
EL NINO	Northeast Brazil, middle of America, East Australia, South Africa	drought
EL NINO	Ecuador, Peru, Chile South Brazil, East Africa, West Europe, South USA, Middle and end reaches of ChangJiang River	flood
LA NINA	Northeast Brazil, India, South Africa	flood
LA NINA	Equatorial Africa, Southeast USA	drought

With our methods and datasets, we found about 80 association rules in the datasets. These rules can be divided

into 4 groups:

1. EL NINO -> Low precipitation (drought)
2. EL NINO -> High precipitation (flood)
3. LA NINA -> Low precipitation (drought)
4. LA NINA -> High precipitation (flood)

We use red, blue, yellow, and green to mark regions where the association rules 1,2,3,4 occurred. We can see that the regions in which extreme events happened (marked in circles) are consistent with the regions with extreme events in Table 2.

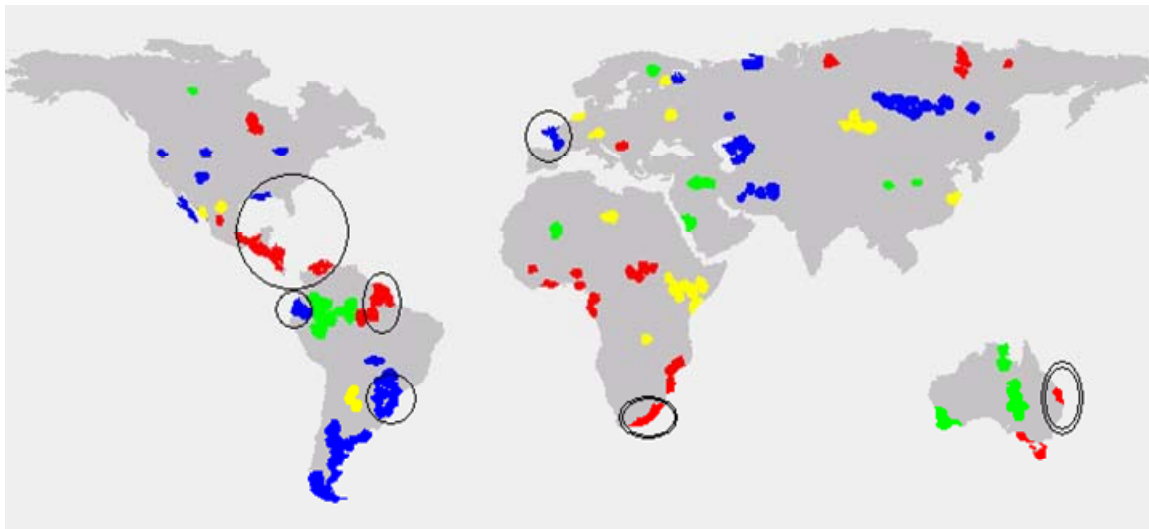


Figure 3. Data mining results

There are also some regions (not marked in circles) where association rules have occurred but have not been confirmed in geographic science. Further work is needed to confirm these associations, which would show that our method is valid and can actually be used in geographic science.

5 CONCLUSION

In this paper, we introduced a method of finding teleconnections in global climate data. Unlike the traditional statistical methods, we use technologies from data mining, so we can avoid the inefficiency of traditional statistical methods like SVD or RPCA. Also, we can take spatial auto-correlation into consideration to reduce the complexity of data size.

We did experiments on real data sets, found some well-known teleconnections such as ENSO, and also found some unknown ones to be confirmed. These experiments show that our method has potential to be usable and efficient and does not have the disadvantages of the traditional methods.

6 REFERENCES

Agrawal, R., Imielinski, T. & Swami, A. (1993) Mining Association Rules between Sets of Items in Large

Databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, D.C.

American Meteorological Society (2000) *Glossary of Meteorology, 2nd Edition*. Cambridge, Massachusetts. Retrieved from the WWW, September 20, 2007:<http://amsglossary.allenpress.com>

Ertoz, L., Steinbach, M., & Kumar, V. (2003) Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. *Proceedings of Third SIAM International Conference on Data Mining*, San Francisco, CA, USA.

Gitay, H., Suarez, A., Watson, R., & Dokken, D. (2002) Climate Change and Biodiversity. *IPCC Technical Paper V*.

Jarvis, R. & Patrick, E. (1973) Clustering Using a Similarity Measure Based on Shared Nearest Neighbors. *IEEE Transactions on Computers C-22 (11)*.

Mannila, H. & Toivonen, H. (1996) Discovering generalized episodes using minimal occurrences. *Proceedings of the Second Int'l Conference on Knowledge Discovery and Data Mining*. Portland, Oregon.

Tobler, W. (1979) Cellular Geography. In Gale, S. & Olsson, G. (Eds.) *Philosophy in Geography*. D. Reidel Publishing Company: Dordrecht, Holland.

Van Oldenborgh, G., Burgers, G., & Klein Tank, A. (2000) On the El-Nino Teleconnection to Spring Precipitation in Europe. *Int. J. Climatology* 20: 565-574.