DISCOVERY OF TELECONNECTIONS USING DATA MINING TECHNOLOGIES IN GLOBAL CLIMATE DATASETS

Fan Lin, XingXing Jin, Cheng Hu, XiaoPing Gao*, Kunqing Xie, XiaoFeng Lei

Department of Intelligent Science, Peking University, Beijing 100871, China
**Email:* gaoxp@cis.pku.edu.cn

ABSTRACT

In this paper, we apply data mining technologies to a 100-year global land precipitation dataset and a 100-year Sea Surface Temperature (SST) dataset. Some interesting teleconnections are discovered, including well-known patterns and unknown patterns (to the best of our knowledge), such as teleconnections between the abnormally low temperature events of the North Atlantic and floods in Northern Bolivia, abnormally low temperatures of the Venezuelan Coast and floods in Northern Algeria and Tunisia, etc. In particular, we use a high dimensional clustering method and a method that mines episode association rules in event sequences. The former is used to cluster the original time series datasets into higher spatial granularity, and the later is used to discover teleconnection patterns among events sequences that are generated by the clustering method. In order to verify our method, we also do experiments on the SOI index and a 100-year global land precipitation dataset and find many well-known teleconnections, such as teleconnections between SOI lower events and drought events of Eastern Australia, South Africa, and North Brazil; SOI lower events and flood events of the middle-lower reaches of Yangtze River; etc. We also do explorative experiments to help domain scientists discover new knowledge.

KEYWORDS: Data mining, Global Climate Datasets, Teleconnection, Land precipitation, Sea surface temperature

1 INTRODUCTION

In recent years, because of the development of information technology, the amount of data has grown explosively. In particular, earth science data has been rapidly accumulating with the development of modern-day satellites, remote sensing technologies, and other data acquisition systems. Traditional analysis methods of earth science data are not good enough. The main statistical methods, such as RPCA (Rotated Principal Component Analysis) and SVD (Singular Value Decomposition), have been used to discover teleconnection patterns. However, they just do not fit the need of the data's growth. A data mining method that discovers episode association rules in a long event sequence can be applied to discovering relationships among the events.

This paper introduces an association data mining method that discovers teleconnection patterns and does experiments on real global earth science data to find many well-known and previously unknown patterns, such as an abnormally low sea surface temperature (SST) in the Eastern Pacific or an abnormally high SST in the Western Pacific coincides with abnormally high precipitation in Shanxi; and an abnormally low SST in the Northern Pacific coincides with abnormally low precipitation in Finland. The experimental results prove the feasibility and efficiency of this method.

2 DATA MINING METHOD

In this section, we describe the data mining method used to discover teleconnections in global climate datasets. Figure 1 illustrates the steps.

2.1 Preprocessing

It is necessary to preprocess earth science data because data of different grids is not compatible. Preprocessing earth science data means getting rid of periodicity and handling data of different time and space intervals.

The first preprocessing method is standardization, This allows variables to have the same weight (Han & Kamber, 2001). The standardization method we perform changes variables in every grid to have the same weight.

The second preprocessing method is getting rid of periodicity. Some patterns of earth science data are well known and may be important. For example, seasonal and yearly variations are very common; if we don't preprocess earth science data, we may get seasonal patterns after data mining instead of other interesting patterns, as the latter may be not as obvious as the former. Therefore we use a monthly z-score (Tan, et al, 2001) and moving average to get rid of periodicity.

The third preprocessing method is transforming data of different granularities. Because of different applications and different data collection methods, dissimilar space and time intervals in earth science datasets are common. Therefore, we must transform the data into uniform spatial and temporal intervals before we analyze them.

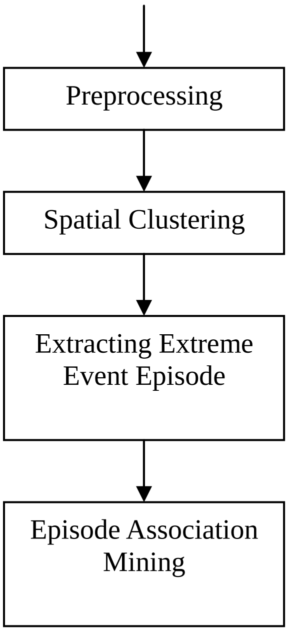

Figure 1. Data mining steps

2.2 Spatial clustering

Before extracting extreme events, we need to cluster the earth science data. First, as the amount of data is huge, the number of events is very large, making difficulties for analysis. After clustering, we can focus only on events of the same type. Second, spatial self-correlation will distort the results because it allows many similar association rules, most of which are not interesting.

In this paper, we use the SNN algorithm (Ertoz, et al., 2001) to cluster earth science data. This algorithm uses a new definition of similarity. At first it finds the nearest neighbors of each data point and then redefines the similarity between pairs of points in terms of how many nearest neighbors the two points share. Using this definition of similarity, the SNN algorithm identifies core points and then builds clusters around them. SNN can handle data sets of different sizes, shapes, and densities and those with high dimensionality.

Because the run-time complexity of the SNN algorithm is $O(n^2)$, the computing cost becomes huge when the data sets are large. We focus on this and analyze the inclusion of nearest neighbors, which proves the spatial autocorrelation. We can take advantage of this spatial autocorrelation to reduce the time complexity.

2.3 Extracting an extreme event episode

After cluster the earth science data, we need to extract space-time extreme events and build the episodes. Earth scientists estimate whether an extreme event is happened by considering these factors: rarity, in other words, the frequency of the event; the numerical value magnitude; the spatial-temporal range influenced by the extreme event; its difference from the average value; and the influence on society.

2.4 Episode association mining

As we know, sequences of events are common forms of data that contain important knowledge. Episodes are patterns in event sequences, in other words, combinations of events with a partially specified order. The algorithm (Mannila & Toivonen, 1996) we used is based on minimal occurrences of episodes. First of all, we found a frequent simple episode of size k; then we formed candidate episodes of size k+1. Next we checked the candidate episodes for frequency, and repeated the steps until all frequent episodes were found. Using the above algorithm, we can form association rules and compute their confidence and support.

Considering the character of the earth science domain, we were able to make improvements. We found that it would take some time for one phenomenon to influence another. Therefore, we took time delay into consideration. For spatial self-correlation, we found many useless or uninteresting rules, which often inundated the interesting ones. In order to avoid this, we added spatial restrictions, so that our algorithm can focus on interesting rules.

3 EXPERIMENTAL RESULTS

In this section, we consider an application of our episode association rules data mining method to earth science data. The data referred in this paper contains CRU (Global Earth Science Data) TS 2.10, SOI (Southern Oscillation Index), and NINO3.4 (Sea Surface Temperature of El Niño Zone).

We did two kinds of experiments. One tries to prove the correctness and feasibility of our methods and the other looks for interesting patterns within the data.

3.1 Experiments to determine the method’s effectiveness

3.1.1 Land rainfall and the SOI exponent

This experiment tries to find abnormal rainfall areas involved with an abnormal SOI Exponent. We discovered the rules below:
Rule 1: Rainfall was abnormally low when the El Niño phenomenon occurred. (Red);
Rule 2: Rainfall was abnormally high when the El Niño phenomenon occurred. (Blue);
Rule 3: Rainfall was abnormally low when La Nina phenomenon occurred. (Yellow);
Rule 4: Rainfall was abnormally high when La Nina phenomenon occurred. (Green)

The table below shows the confidence and support of the rules we found, and the figure shows the results. In this figure, the different areas are shown in different colors. The circular regions are the ones found in earth science.

Table 1. Association rules between SOI and global events of abnormal rainfall

Rule	Confidence	Support	Rule	Confidence	Support
SOIL=>eap186H	0.77	24	SOIL=>sap30H	0.74	23
SOIL=>sap34H	0.74	23	SOIL=>eap188H	0.71	22
SOIL=>eap81H	0.71	22	SOIL=>nap91H	0.71	22
SOIL=>sap32H	0.71	22	SOIL=>eap141L	0.68	21
SOIL=>afp50L	0.68	21	SOIL=>eap56H	0.68	21

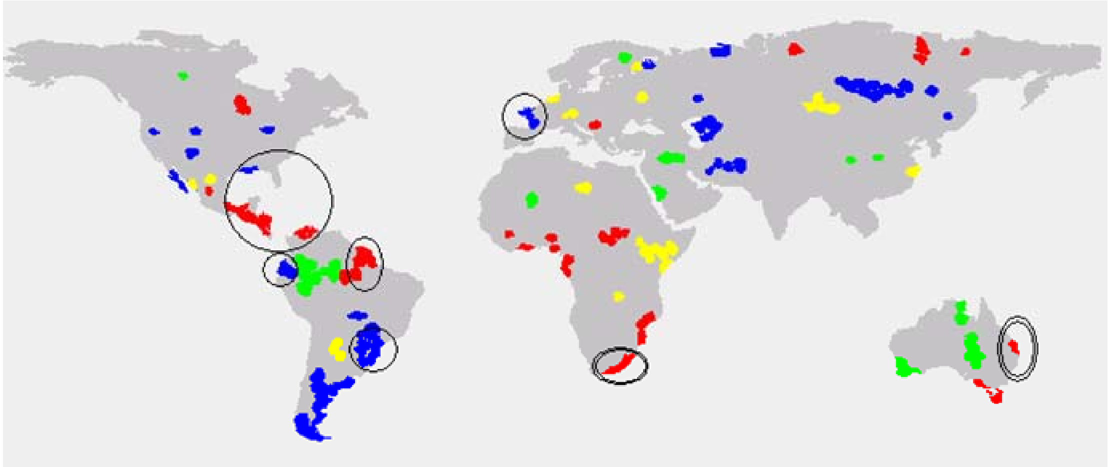

Figure 1. Areas with abnormal rainfall which relate to an abnormal SOI

3.1.2 Land temperature and the SOI exponent

This experiment tries to find an association between abnormal earth temperature and the SOI exponent. We discovered the results below:
Rule 1: Temperature was abnormally high when the El Niño phenomenon occurred (Red);
Rule 2: Temperature was abnormally low when the El Niño phenomenon occurred (Yellow);
Rule 3: Temperature was abnormally low when the La Nina phenomenon occurred (Blue).

Table 2. Association rules between SOI and global events of abnormal temperature

Rule	Confidence	Support	Rule	Confidence	Support
SOIL=>sat14H	0.77	24	SOIL=>nat56H	0.77	24
SOIL=>sat15H	0.74	23	SOIL=>sat2H	0.74	23
SOIL=>aft86H	0.74	23	SOIL=>sat14H,sat9H	0.71	22
SOIL=>sat14H,sat20H	0.71	22	SOIL=>sat14H,sat2H	0.71	22
SOIL=>sat9H	0.71	22	SOIL=>sat20H	0.71	22

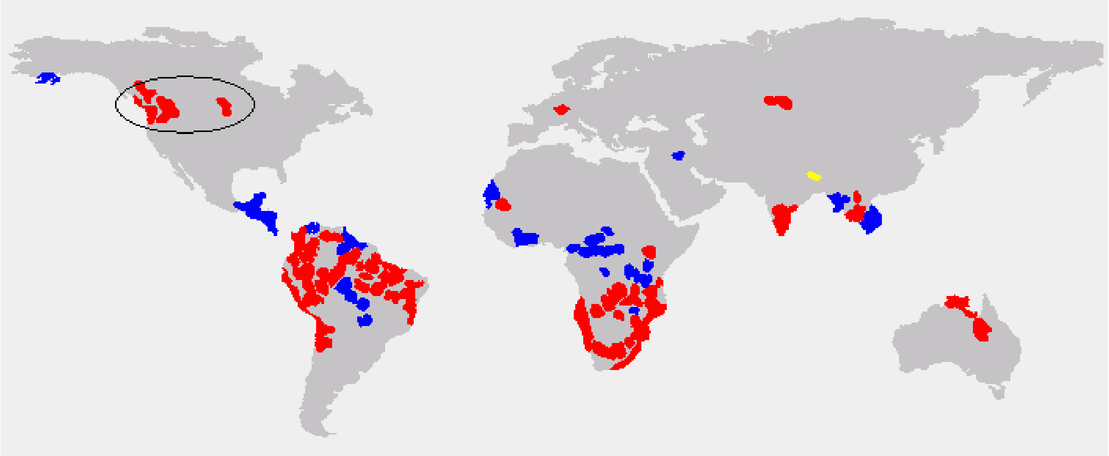

Figure 2. Areas with abnormal temperature which relate to an abnormal SOI

The table above shows the rules we found, and the figure shows the results. The circular result areas are proved by domain knowledge. Also the phenomenon of rule 2 only appears in Yunnan Province, China.

3.1.3 Land rainfall and NONI3.4 exponent

This experiment is similar to the first one and found similar rules. The table and figure below have the same meaning as the above ones.

Table 3. Association rules between NONI3.4 and global events of abnormal rainfall

Rule	Confidence	Support	Rule	Confidence	Support
NINOH=>sap34H	0.84	26	NINOH=>afp44L	0.81	25
NINOL=>aup4H	0.81	25	NINOH=>afp46L	0.77	24
NINOL=>aup10H	0.77	24	NINOL=>aup1H	0.77	24
NINOH=>eap150H	0.77	24	NINOH=>sap30H	0.74	23
NINOH=>sap32H	0.74	23	NINOH=>sap6L	0.74	23

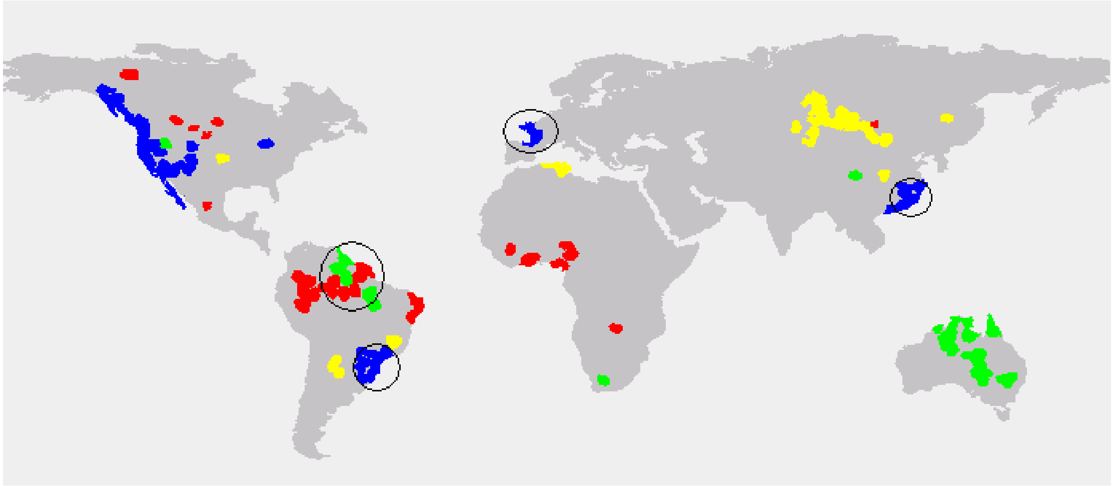

Figure 3. Areas with abnormal rainfall which relate to an abnormal NONI3.4

3.1.4 Land temperature and the NONI3.4 exponent

This experiment is similar to the second one and found similar rules. The table and figure below are also the same as the second one.

Table 4. Association rules between NONI3.4 and global events of abnormal temperatures

Rule	Confidence	Support	Rule	Confidence	Support
NINOH=>sap34H	0.84	26	NINOH=>afp44L	0.81	25
NINOL=>aup4H	0.81	25	NINOH=>afp46L	0.77	24
NINOL=>aup10H	0.77	24	NINOL=>aup1H	0.77	24
NINOH=>eap150H	0.77	24	NINOH=>sap30H	0.74	23
NINOH=>sap32H	0.74	23	NINOH=>sap6L	0.74	23

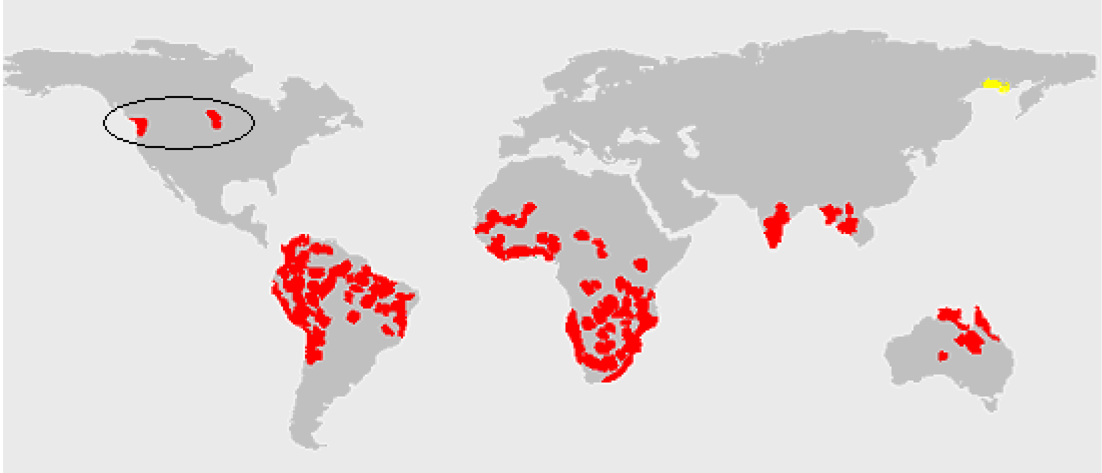

Figure 4. Areas with abnormally temperatures which relate to an abnormal NONI3.4

As the four experiments above illustrate, we have found results similar to those in the domain knowledge. In other words, our method is reasonable and effective. While the association rules appear correct, the mechanism behind the association is unclear.

3.2 Exploratory experiments

The experiments below try to find interesting rules from an even larger group of data. Perhaps these rules will be part of the domain knowledge in the future.

(1) Experiment 1: Sea Surface Temperature and Land Rainfall

Besides the association between the abnormal sea surface temperature of the Equatorial Eastern Pacific (the area where the El Niño phenomenon occurs) and the abnormal rainfall of the middle and lower reaches of the Yangtze River, Guangdong Province and Western Europe, we found the following rules: a high sea surface temperature in the Equatorial Eastern Pacific or Western Pacific is associated with high rainfall in the Shanxi Province, China; a high sea surface temperature in the Southern Pacific is associated with high rainfall in the middle and lower reaches of the Yangtze River; a high temperature in the Equatorial Eastern Pacific is associated with high rainfall in the north of Russia and Afghanistan; and so on. The mechanisms that cause these associations remain to be discovered.

(2) Experiment 2: Sea Surface Temperature and Land Temperature

This experiment yields the following rules: a high temperature in the Indian Ocean causes high temperatures in Africa’s west coast, such as in Ghana, Benin and Togo; a high temperature in the Indian Ocean results in high temperatures in Africa’s east coast, such as in Tanzania and Kenya; a warming in the Equatorial Eastern Pacific or Southern Pacific’s sea surface is associated with high temperatures in Southeast Africa. The mechanisms that cause these associations remain to be discovered.

(3) Experiment 3: Land Rainfall and NDVI

The most frequent associations we found are between abnormal rainfall in South America and an abnormal NDVI exponent, but these rules are not reasonable. Thus, this experiment did not find meaningful rules.

(4) Experiment 4: Land Temperature and NDVI

This experiment did not find meaningful rules.

(5) Experiment 5: Land Rainfall and Temperature

We found an interesting phenomenon that if there is abnormally low rainfall in an area, in the same or the next month, the average temperature in adjacent areas rises abnormally.

These exploratory experiments found some interesting patterns, which remain to be proven by domain experts.

4 CONCLUSIONS AND FUTURE WORK

This paper introduces a data mining method for earth science data. We extracted extreme events before associations; used an improved SNN algorithm to solve high-dimension problems; and did research on a data mining method of episode association rules that fits the character of the earth science domain. We did some experiments and discovered several rules that can be proven by further earth science domain research. This shows that our method is feasible and effective. Also we did some exploratory experiments and discovered interesting rules, which we think will lead domain experts to discover new knowledge.

However, there are still many problems remaining: our method cannot predict the future, which would be very helpful for earth science experts in their research. There are situations when it appears that the improved SNN algorithm is not effective enough; we need to design a more powerful episode association data mining algorithm.

5 REFERENCES

Ertoz, L., Steinbach, M., & Kumar, V. (2001) Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach. *Text Mine’ 01, Workshop on Text Mining, First SIAM International Conference on Data*

Mining, Chicago, IL.

Ertoz, L., Steinbach, M., & Kumar, V. (2002) A new shared nearest neighbor clustering algorithm and its applications. *In Workshop on Clustering High Dimensional Data and its Applications, SIAM Data Mining*.

Ertoz, L., Steinbach, M., & Kumar, V. (2003) 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.

Han, J. & Kamber, M. (2001) *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Mitchell T., Carter T., Jones P., 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*.

Mannila, H. & Toivonen, H. (1996) Discovering generalized episodes using minimal occurrences. *In Proceedings of the Second Int'l Conference on Knowledge Discovery and Data Mining (KDD'96)*, pp. 146 – 151. Portland, Oregon.

Rayner, N., Parker, D., Horton, E., Folland, C., Alexander, L., Rowell, D., Kent, E., & Kaplan, A. (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res. 108 (D14).*

Tan, P., Steinbach, M., Kumar, V., Klooster, S., Potter, C., & Torregrosa, A. (2001) Finding spatio-termporal patterns in earth science data. *KDD Temporal Data Mining Workshop.* San Francisco, California, USA.