# A Novel Frequent Patterns Mining Method of Unusual Climate Events in Data of East Asian Monsoon Zone

He Shan
Department of Machine Intelligence, Peking University, China
Email:hhh@pku.edu.cn

Lin Fan
Department of Machine Intelligence, Peking University, China
Email:linfanlf@gmail.com

Xie Kunqing
Key Laboratory of Machine Perception, Ministry of Education, Peking University, China.
Email:kunqing@cis.pku.edu.cn

***Abstract*—Unusual climate events which may cause disasters have great influence on both the natural environment and the human society. Finding association patterns among these events has great significance. Traditional data mining methods have several problems while applied to climate science data directly so we propose a novel method that mining frequent patterns among unusual events in climatic data, including spatial clustering algorithm based on tight clique, extracting unusual climate events algorithm and extended generalized sequential pattern (EGSP) algorithm. In order to verify our method, we do experiments on real climatic data (Climatic data of East Asian monsoon zone) and find lots of well-known and previously unknown patterns. It needs the climatic expert to judge whether the new patterns are significative. Overall, the experiment told us it was an effective and viable method for the climatic research.**

## I. Introduction

Climate system is very complex. Climate is different between different areas because of different conditions. However, climate system is a unitary system. In varying climate conditions there are interactions between climate events in a certain range that may form some association patterns. The disasters caused by unusual climate events often make a great deal of injuries and loses in life and the loses in wealth. To discover the spatio-temporal patterns of unusual climate events is very important. It can help us to understand the changing process of climate, find its rule and enhance the prediction capability.

With the developments of modern-day technologies, climate science data grow explosively. Due to the overwhelming volume and high resolution of datasets, traditional climate science data analysis methods are not enough. Much of the data acquired at great expense remains greatly under-exploited.

As a new way to discover hidden patterns in huge amount of data, data mining can help climatic experts to explore the data. There are lots of works using the data mining methods in climatic research. However, there are several problems while applied the method to climate science data directly. First, the traditional clustering method does not consider the spatio-temporal changes of the process climatic data. Second, ordinary events are not only complicated but also meaningless. Unusual events are important and useful for the research of climate. Third, Traditional sequence association rule mining method is suitable for mining common patterns in various regions, rather than mining the relationship between the regions. To solve the three problems above, we propose a method that mining frequent patterns among unusual events in climatic data. This paper is mainly about applying data mining method to climate science data and our work is as follows:

(1) **Spatial clustering based on tight clique**. We propose a clustering algorithm based on tight clique. The clustering algorithm considers the properties of the climate changing character as a similarity measure indicator, and form clusters based on this type of similarity measure. This part of work can be found in Ref. [1].

(2) **Extended generalized sequential pattern mining**. Considering unusual climatic events commonly take place in a huge area and a changefully spatio-temporal range, we merge the result of extraction in both spatio field and temporal field after extracting unusual climate events in each cluster. This step reduces the complexity of the frequent patterns mining. We extended the classic generalized sequential pattern (GSP) algorithm, and apply it to the field of frequent pattern mining.

(3) **Experiment with the climatic data of East Asian monsoon zone**. We focus upon the East Asian monsoon zone that China is located in. The monsoon system is very unstable and has a large interannual variation. The formation mechanism of drought, water-logging, cold, heat and other disasters under the influence of the East Asian monsoon was studied by many meteorologists. They found that there would be some interesting patterns brought by the unequal distribution of the east Asia monsoon. We do experiments on climatic data of East Asian monsoon zone and find lots of well-known and previously unknown patterns. It needs the climatic expert to judge whether the patterns are significative.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 is the unusual climate events extracting algorithm. Section 4 is the extended gener-

Xie Kunqing is the correspondent author.

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

978-1-4244-7739-5/10/$26.00 ©2010 IEEE

alized sequential pattern mining algorithm. Section 5 reports experimental results. We conclude our study in section 6.

## II. Related works

Statistical methods are widely used in climate science but they can not well deal with the overwhelming volume of data. Methods of data mining are used in the field of meteorological science. Ref.[2] used shared nearest neighbors (SNN) as a similarity measure indicator clustering the temperature data of surface seawater. Ref.[3] clustered trajectory of hurricane by expectation-maximization (EM) algorithm and composite regressive model. Ref.[4] proposed a hierarchical classification algorithm to solve a land cover classifications problem.

The co-location rule in Ref.[5] was used to discover non-sequential pattern between different regions and mainly to find the association rule among the events which happened in the same time and the location conform to some spatial relationships. Ref.[6] presented a Depth-First-Search algorithm (DFS-MINE) for fast mining of frequent spatiotemporal patterns in environmental data which is different with GSP. Ref.[7] got a traditional transactional database by pre-processing of remote sensing image, then used the method of Ref.[8] used P-tree to store transactions to deal with remote sensing data. Ref.[9] present an efficient method for mining frequent sequential association rules from multiple sequential data sets with a time lag between the occurrence of an antecedent sequence and the corresponding consequent sequence.

There is one key problem in all above researches: how to process the climate dataset into traditional transactional database? Then we may use the traditional methods of association rule mining. Ref.[10] reviewed four modes to process the earth science data into transactional database:

(1) **Non-sequential pattern in one region**.
(2) **Non-sequential pattern among different regions**.
(3) **Sequential pattern in one region**.
(4) **Sequential pattern among different regions**.

Ref.[8] and [11] belong to the 1st mode. Ref. [7] and [12] fall into the same category with the 2nd mode. To the 4th mode, Ref. [10] only presented a primary process: First, choosing a few interested areas and climate parameters. Second, add every climate parameter into all the points of the areas as a variable and get a new transactional database. Third, use the method of Ref. [13] in the new transactional database.

## III. Unusual Climate Events Extracting

### A. Extracting unusual climate events in each cluster

Climate disaster means the disaster brought on by large range, long time unusual climate. The unusual climate could bring on drought, flood, chilling injury, sandstorm and so on. In this paper, we use the data of temperature and precipitation to judge whether the unusual climate events take place. The judging standard includes the frequency and strength.

*Def 3.1:* **Frequency threshold of unusual climate event**. Frequency threshold of unusual climate event is the time interval of the unusual climate event taking place, denoted as $\lambda$. $\lambda$ is higher, the unusual climate event is more infrequent.

*Def 3.2:* **Unusual climate event**. To one area, under a given frequency threshold of unusual climate event $\lambda$, choose all time points which property value fulfills the threshold condition making up a dataset T. p is a threshold indicating the upper scale between maximum value MAX and average value. If the property value of a time point is in this range, we define this time point a climate unusual high event. q is a threshold indicating the lower scale between minimum value MIN and average value. If the property value of a time point is in this range, we define this time point a climate unusual low event. The climate unusual high event and climate unusual low event are generalized as climate unusual event.

*Def 3.3:* **Formal representation of unusual climate event**. An unusual climate event is denoted as C = $\langle P, E, T \rangle$ , where C is an unusual climate event, P is the location that unusual climate event takes place and could be a spatial extent or a set of spatial cluster labels, E is the type of unusual climate event including drought, flood, chilling injury, heat and so on, could be lone type or a combination of several types such as flood adding chilling injury. T is the duration of unusual climate event, a single number means unusual climate event happens in one month and a series of number means it continues for several months.

In this definition, the unusual climate events extracting algorithm is as follows:

(1) Convert frequency threshold of unusual climate event $\lambda$ into the first n items or last n items of climate property values in one location. According to domain knowledge, during a definite period of time, the corresponding time interval of the first or last n categories of climate property values expresses the frequency of unusual climate event. For example, during 100 years, if the maximum value of precipitation in one area is $v_1$ the unusual climate event $v_1$ is "once every 100 years". If the next to the highest one of precipitation in one area is $v_2$ the unusual climate event $v_2$ is "once every 50 years" without consideration of the time interval between $v_1$ and $v_2$.

(2) Record the corresponding unusual climate event C of the first or last n categories of property values in one area.

(3) Delete the event C that does not meet the threshold p and q then get the unusual events sequence of one area.

### B. Merge the result in both spatio field and temporal field

The neighborhood unusual events in time or space indicate the same event. If we do not merge the result for the next step of data mining there would be many reduplicate counts that may hide the real significative information.

*1) Merging in space:* (1) Put the unusual events of the same moth in the candidate set C and choose an event to the result set R. We use the cluster label A expresses the unusual event.

(2) Put the neighborhood clusters of A into a queue Q. With each step, we choose one cluster B from Q. If B have the same unusual climate with A we put B into R and put the neighborhood clusters of B into Q. Otherwise, choose the next cluster in Q for judgement until Q is empty.

(3) Clusters in the result R can be combined into one unit. Delete the clusters in R from the candidate set C and repeat

above steps until C is empty.

*2) Merging in time:* (1) In the temporal dimension, to each unusual event E, search forward for all the same type unusual climate event in the new data set after merging in space.

(2) If neighborhood unusual events in time is also neighborhood in space we merge the events into one event and let the duration of unusual climate event T a set of time.

## IV. Frequent patterns mining

The purpose of frequent patterns mining of unusual climate events is to find the relationship of unusual climate in different areas. These unusual climate events may take place in the same time or with some time delays. This kind of frequent patterns is interesting rules to meteorologists.

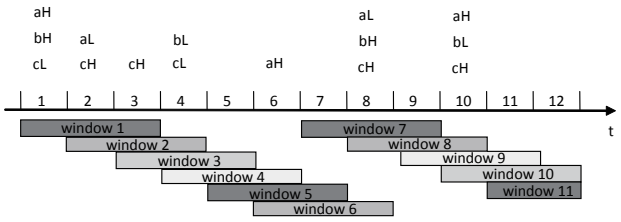



Fig. 1. Frequent patterns mining of climate

Figure 1 shows a unusual climate queue of an area during 12 months. There are some locations with the label a,b,c. H is unusual high event. L is unusual low event. In this paper, our goal is to find all frequent patterns among these different locations. For example, the rule expressions in the form $aL \Rightarrow cH$, which means after the unusual low event take place in location a, an unusual high event will happen in location c.

### A. Sliding time window

Since the annual cycle climate changing, the time interval of the relationship between two locations could not exceed one year. We add a time window to the unusual climate sequeue within 12 months. As shown in Figure 1, the window length is 3. For each month, item $I_p$ is the unusual climate taking place at the same time and all these items combining to event $e_l$. The events in a certain window make a sequeue s. There are 11 window that length is more than 1, so there are 11 different queues. We pay attention to find the frequent subsequences of all queues. The frequent sub-sequences that we found cover all locations, so it could express the relationship among different locations.

### B. Time aligning

Using the traditional GSP algorithm, there would be many redundancy and reduplicate counts. For example, to the rule $aH \Rightarrow cH$, aH appears in the 2nd time point of window 1 and cH in 3rd, count plus 1. In window 2, they appear in the 1st time point and 2nd, so count plus 1 again. To avoid this kind of reduplicate counts, we add a limit of time aligning to the original GSP algorithm: counting only when the first event of the rule aligning to the first time point of current window. With the limit of time aligning the count would not duplicate with the sliding window.

### C. EGSP algorithm

EGSP (Extended Generalized Sequential Pattern) is aimed at climate mining, but it also can be used for general sequential pattern mining:

(1) Add a time sliding window when reading the sequeue.

(2) Add a limit of time aligning in the algorithms of association rule mining.

The function of candidacy generation is same with GSP algorithm.

### D. Scenarios of EGSP algorithm

(1) Without division of region. That means to mine patterns base on the original clustering result neglecting to merge in space and time.

(2) Pattern mining base on division of region. This method forms a new set of unusual climate event with merging the original result in space and time. We can divide the East Asia Monsoon Zone into several parts and mine patterns in the division. The standard of division may base on domain knowledge. In this paper, we divide the range into two parts: South and North.

## V. Experiments and Results

We use real global data for experiments. Global climate datasets CRU TS 2.10[14] contains nine variables including monthly average precipitation, monthly average temperature etc and we mainly use variables of monthly average precipitation, monthly average temperature.

### A. Unusual Climate Events Extracting

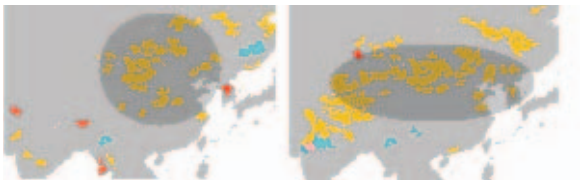
TABLE I
Chinese climate disaster in 1937 and 1962

| Time | Location | Type | Detail |
|---|---|---|---|
| 1937 | Anhui, Shanxi, Shandong, Sichuan, Henan, Guizhou, Guangxi, Ningxia, Gansu | Great drought | Affected population hit 30 million |
| 1962 | North China, Westnorth | Great drought | Drought acrossing Spring and Summer |

Table I is taken out randomly from Chinese climate disaster record chart. Set the frequency of unusual climate event as "once every 10-years ", p,q as $1\%$, the result shows in Figure 2. The large scale drought in north China checks up with the facts in the two years.

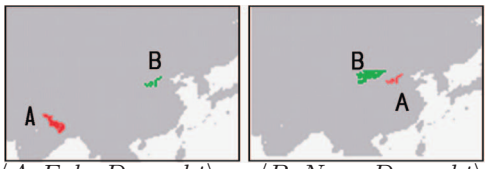
### B. Frequent patterns mining

*1) Without division of region:* In this experiment, frequency of unusual climate is "once every 10-years ", confidence threshold is 0.5, support threshold is 8, the result shows in Figure 3 (B is latter than A).



The shaded part is the range of drought

Fig. 2. Result of 1937 and 1962




$\langle A, Feb., Drought\rangle \Rightarrow \langle B, Nov., Drought\rangle$,
confidence=51.8%,support=9

Fig. 3. Result without division region

*2) Pattern mining base on division of region:* We divide the East Asia Monsoon Zone into south and north (Figure 4). If the disaster area is larger than 3% of the part we consider that there is a disaster of the part. The frequency of unusual climate event is "once every 10-years ", Confidence threshold is 0.5, support threshold is 7, the result shows in table II.

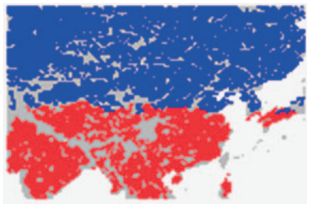


Fig. 4. Division of South and North

TABLE II
ASSOCIAION RULES OF UNUSUAL CLIMATE

| |
|---|
| $\langle South, Sept., Cool\rangle \Rightarrow \langle North, Dec., Drought\rangle$ confidence=81.8%, support=9 |
| $\langle North, June, Heat\rangle \Rightarrow \langle South, Dec., Drought\rangle$ confidence=88.9%, support=8 |
| $\langle North, Sept., Cool\rangle \Rightarrow \langle South, NextJan., Drought\rangle$ confidence=91.7%, support=11 |
| $\langle North, Mar., Flood\rangle \Rightarrow \langle South, Dec., Drought\rangle$ confidence=82.3%, support=14 |
| $\langle South, July, Flood\rangle \Rightarrow \langle North, Nov., Drought\rangle$ confidence=62.5%, support=10 |

Some of the mining results are in accordance with existing knowledge: if the Pacific Monsoon is weak, the rainfall belt staying at south in the long term, there will be the phenomena as "drought in north and flood in south". Otherwise, if the Pacific Monsoon is power, the rainfall belt staying at north in the long term, there will be the phenomena as "drought in south and flood in north". The results indicate the correctness and the feasibilities of our methods.

## VI. CONCLUSION

Climate system is a unitary system. With the interaction among climate elements there are many different climate phenomena so there are some relationship among the climate phenomena. Mining the spatio-temporal patterns we can delve more deeply into climate system and discover the law behind the climate phenomena. In this paper, we mine frequent Patterns of Unusual Climatic Events in Climatic data of East Asian monsoon zone. We propose tight clustering algorithm, Unusual Climate Extracting Algorithm and Extended generalized sequential pattern algorithm. In the experiments, we got some interesting climate rule.

## REFERENCES


[1] G. S. F. Lin, K. Xie and T. Wu, "A novel spatio-temporal clustering approach by process similarity," in *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, 2009, pp. 150–154.

[2] V. K. Levent Ertz, Michael Steinbach, "A new shared nearest neighbor clustering algorithm and its applications," in *Proceedings of the Workshop on Clustering High Dimensional Data and its Applications*. Second SIAM International Conference on Data Mining, April 2002.

[3] S. Gaffney and P. Smyth., "Trajectory clustering with mixtures of regression models," in *Fifth ACM SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining*, August 1999, pp. 63–72.

[4] J. G. S.Kumar and M. Crawford, "Hierarchical fusion of multiple classifiers for hyperspectral data analysis," *Pattern Analysis and Applications, spl. Issue on Fusion of Multiple Classifiers*, vol. 5, no. 2, pp. 210–220, 2002.

[5] S. S. Yan Huang and H. Xiong, "Discovering co-location patterns from spatial datasets: A general approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 12, pp. 1472–1485, December 2004.

[6] I. Tsoukatos and D. Gunopulos, "Efficient mining of spatiotemporal patterns," in *the 7th International Symposium on Advances in Spatial and Temporal Databases*, 2001.

[7] T. H. H. S. J. G. John A. Rushing, Heggere S. Ranganath, "Using association rules as texture features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 845–858, August 2001.

[8] Q. Ding and W. Perrizo, "Association rule mining on remotely sensed images using p-trees," in *the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2002.

[9] J. D. S. Harms and T. Tadesse, "Discovering sequential association rules with constraints and time lags in multiple sequences," in *the 2002 International Symposium on Methodologies for Intelligent Systems*, June 2002, pp. 432–442.

[10] V. K. S. K. C. P. P. N. Tan, M. Steinbach and A. Torregrosa, "Finding spatio-termporal patterns in earth science data," in *KDD Temporal Data Mining Workshop*, August 2001.

[11] R. H. S. G. S. J. Hinke T., Rushing J., "Techniques and experience in mining remotely sensed satellite data," *Artificial Intelligence Review : Issues on the Application of Data Mining*, vol. 14, no. 6, pp. 503–531, Dec. 2000.

[12] V. L. Y. D. Fenzhen Su, Chenghu Zhou and W. Shi, "A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution," *Ecological Modeling*, vol. 174, pp. 421–431, 2004.

[13] R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements," in *5th International Conference on Extending Database Technology*, March 1996.

[14] C. T. J. P. H. M. N. M. Mitchell, T.D., "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)," *ournal of Climate: submitted*, 2003.