# Events-coverage based spatio-temporal association rules mining method

**LI Guangqiang, DENG Min, ZHANG Weiling, CHEN Yi**

*Department of Surveying and Geo-informatics, Central South University, Hu'nan Changsha* 410083, *China*

**Abstract:** Spatio-temporal association rules mining is a key technology and a hot issue in the field of spatio-temporal data mining. The classical Apriori algorithm is usually utilized to detect the spatio-temporal association rules from the spatio-temporal transaction table, which is derived from the original spatio-temporal data. In most existing approaches to generate the spatio-temporal transaction table, many defects, such as data redundancy, further affects the efficiency of spatio-temporal association rules mining. This paper proposes an events-coverage based spatio-temporal association rules mining (ECSTAR for short) to overcome these limitations. ECSTAR employs the event's coverage to divide the researching spatio-temporal domain into some cells to generate a spatio-temporal transaction. Among each cell, spatio-temporal relationship predications are utilized to present the spatio-temporal relationship between the events and spatio-temporal objects. Thus, the spatio-temporal transaction table is built and spatio-temporal association rules are mined by the Apriori algorithm. Moreover, many concepts about ECSTAR are expounded and its algorithm is narrated in detail. Finally, a practical experiment demonstrates the feasibility and validity of the ECSTAR.

**Key words:** spatio-temporal association rules, spatio-temporal event, event coverage, spatio-temporal transaction table

**CLC number:** TP751.1 **Document code:** A

**Citation format:** Li G Q, Deng M, Zhang W L and Chen Y. 2010. Events-coverage based spatio-temporal association rules mining method. *Journal of Remote Sensing*. **14**(3): 468—481

## 1 INTRODUCTION

Spatio-temporal data mining (STDM for short), an important tool for spatio-temporal analysis, is utilized to discover potential information and knowledge hided within magnanimous spatio-temporal data (Hsu *et al.*, 2008). STDM includes five sub-fields: spatio-temporal association rules (STARs for short) mining, spatio-temporal sequential pattern mining, spatio-temporal data prediction, spatio-temporal clustering, and spatial discriminate rules mining. As a key technology in the field of STDM, STARs mining could discover potential spatio- temporal association relationships from spatio-temporal data set. The most existing association rules (ARs for short) mining methods only focus on temporal or spatial data mining, and do not consider time and space integration. Currently, a few researchers employ spatio-temporal transaction table to detect the STARs. These methods include: (1) spatio-temporal relationship based method, and (2) spatio-temporal division based method. The former usually employs the spatio-temporal relationship among spatio-temporal objects to construct the spatio-temporal transaction table, and then to detect STARs. Celik *et al.* (2006) studied continuous emerging spatio-temporal co-occurrence patterns algorithm (SECOP); however, the algorithm dissever the nature spatial and temporal property of a spatio-temporal object, and its computation is very intricate. Su *et al.* (2004) proposed a spatio-temporal configuration model for environmental changing. The model is suitable for spatio-temporal database which changes continuously. Meng (2005) and Li (2001) developed the concept of ARs credibility, being relevant to the time, and proposed an ARs mining algorithm for temporal data. Li *et al.* (2003) proposed calendar-based temporal association rules algorithm. From the above overview, it may be found that these methods only focus on the temporal association rules mining. However, the spatial relationship among the entities is not considered. Presently, Ren *et al.* (2003) and Verhein and Chawla (2006) utilized the object - relationship database to establish mobile spatio-temporal database, and then mined mobile objects routing STARs relying on the semantic-based spatio-temporal relationship expression. But this method does not suit other non-mobile spatio-temporal database.

The spatio-temporal division based STARs mining method divides spatio-temporal domain into lots of fixed cells, then constructs transaction table by means of these cells. Mennis and Liu (2005)

**Received**: 2009-01-08; **Accepted**: 2009-05-11

**Foundation**: National High Technology Research and Development Program of China (863 Program) (No. 2009AA12Z206), Natural Science Foundation of Hu'nan Province of China (No. 09JJ6061), Foundation of Jiangsu Key Laboratory of Resources and Environment Information Engineering (No. 20080101) and Foundation of Key Laboratory of State Bureau of Surveying and Mapping of Geo-Spatial Information Engineering (No. 200805).

**First author biography**: LI Guangqiang (1972— ), male, Doctor. He received the doctor degree in the Central Southern University, China. His research interests include spatial data mining and spatio-temporal outlier detection. E-mail: ligq168@163.com

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

firstly proposed a method which divides the researching area into a number of spatial fixed cells and constructed mining table based on spatio-temporal relationship in each cell, then mined association rules in mining table. Yu *et al.* (2008) reseached a spatio-temporal division based method via the expansion of cellular automata, and utilized genetic algorithm to find the transition rules of cellular state (Hu, 2006). Wang *et al.* (2008) proposed a time series prediction and association rules mining method for spatio-temporal snapshot database. They assumed to improve the ARs mining efficiency via clustering spatio-temporal sequence. For these spatio-temporal division based methods, the size of cell impacts severely the result and computational efficiency. Moreover, the attribute of cell should be interpolated sometimes, so the uncertainty of interpolation also affects the reliability of mining results.

As mentioned above, it plays a magnificent role to build spatio-temporal transaction table in the process of mining STARs and directly impact the efficiency of computation and mining results. To overcome these limitations, an Events-Coverage based Spatio-Temporal Association Rules mining (ECSTAR for short) is proposed in this paper. The event's coverage is employed to divide the researching spatio-temporal region into cells firstly, and then construct spatio-temporal transaction relying on these cells. This method may remove the data redundancy, and the efficiency of calculation and the reliability of the results in the process of mining could be improved.

## 2 SPATIO-TEMPORAL EVENT AND EVENT COVERAGE

In the real world, the evolution of spatio-temporal object may be mainly controlled by some important events, and these events only affect a certain spatio-temporal scope. Namely, the event only affects some spatio-temporal objects in the temporal and spatial scope. Spatio-Temporal Event (STE for short) is an object which could affect a certain spatio-temporal scope. Generally, the appearance, disappearance or change of spatio-temporal object may be a STE, which could affect the other objects nearby. Obviously, the STE is related to the research area and is relative. In an application field, a thing may be a STE, but may not be one in another application field. The raining, for example, is a very important STE in the field of geological disaster monitoring, while it is not a meaningful event in the field of LBS (location-based services).

Spatio-temporal coverage (STC for short) is the coverage domain of a STE. The time projection of STC is temporal coverage length (denoted by $w$), and the space projection of STC is spatial coverage domain (denoted by $s$). The time coverage of STE is one-way, which is range from the beginning of the event to the extinction of it. For simplicity, this paper assumes that the STC is homogeneous in different direction, namely, STE's effect is related only to the distance, and has nothing on with the direction. Therefore, the STC is a circle area, where the spot of STE is the center and coverage distance is the radius. In Fig. 1, the temporal coverage length of $STE_1$, whose $w_1=3$, and the spatial coverage domain of the $STE_1$ includes $O_1$~$O_5$. That is, $STE_1$ affects the attributes of $O_1$~$O_5$ in the period from $t_p$ to $t_{p+3}$. Similarly, $STE_2$ affects the attributes of $O_6$~$O_9$ in $t_p$~ $t_{p+5}$.

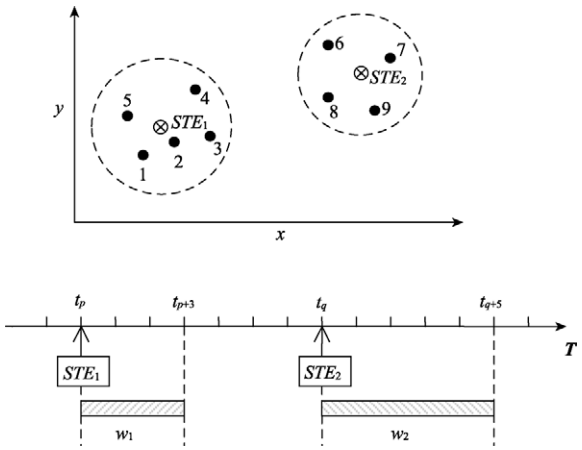


Fig. 1 Illustration of spatio-temporal event coverage

In the process of spatio-temporal evolution, the STE may closely impact the attributes of other spatio-temporal objects in the STC. Spatio-temporal association rules mining relying on events coverage is equivalent to mine association rules between spatio-temporal events and the attributes of other spatio-temporal objects in the STC. Thus, the STC may be used to build spatio-temporal transaction table and mine STARs, which is the theoretical basis of the method proposed in this paper.

## 3 STC BASED SPATIO-TEMPORAL TRANSACTION TABLE

According to the STC, the spatio-temporal domain could be divided into lots of spatio-temporal transaction cells. In each cell, a spatio-temporal transaction record is built depending on the concept generalization method, which is commonly utilized in data mining to discretize the continuous numerical data. The concept generalization method mainly includes the iso-deep method and the iso-wide method. The former method separates the samples into a few groups and each group have equal samples. The latter method separates the samples into some groups with the equal interval. In specific application areas, we should choose the appropriate principles of concept generalization method referring to the background knowledge. The rainfall, for example, should be grouped referring to the classification methods in the meteorological field.

In this paper, the spatio-temporal objects, having a significant impact on its spatio-temporal adjacent domain, are named as the spatio-temporal events. The tables, recording spatio-temporal events, are named as spatio-temporal event table (***STET*** for short). The other spatio-temporal objects impacted by the spatio-temporal events are named as spatio-temporal objects (STO for short). Thus, the tables recording spatio-temporal

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

objects are named as spatio-temporal table (***STDT*** for short). Then, the spatio-temporal data sets include the ***STET*** and ***STDT***, *etc*. The concept generalized procession for the spatio-temporal data includes: (1) to generalize the ***STET***, (2) to generalize the ***STDT***, (3) to Generalize the spatial relationships between STEs and STOs, and (4) to generalize the temporal relationships between STEs and STOs.

According to single or multi-items, the concept generalization method is employed to generalize the ***STET*** and ***STDT***, and to divide the STEs and the spatio-temporal data into a few corresponding concept-lattices. The generalized table of the STEs is named as events generalized table (***EGT*** for short), and the generalized table of spatio-temporal data is data generalized table (***DGT*** for short). The concept generalization of the spatio-temporal relationship between the STEs and the STOs, which requires the algebra operations among tables (i.e. join and project), is relatively complicated. The process of the concept generalization of spatial relationship between the STEs and the STOs includes: (1) to join the ***STET*** to the ***STDT*** and compute their spatial distance, and (2) to present the spatial distance via the spatial predictions. These predictions includes *IsFarTo*, *IsMediumTo* , *IsCloseTo* and so on. Supposed DG (•) to denote the operator of distance generalization, the spatial generalized table between STEs and STOs can be expressed as follows:

$$\boldsymbol{SGT}=\mathrm{DG}(\sigma_{\mathrm{Dists}(ste,sto)\leqslant r}(\boldsymbol{STET}\bowtie\boldsymbol{STDT})) \tag{1}$$

In Eq. (1), $\sigma$ is selecting operator, one of the relational algebras, and Dists(ste, sto) aims to compute spatial distance between STEs and STOs, Dists(ste, sto)$\leqslant r$ is the requirement of the selection, which aims to filter all spatial objects with the condition, namely spatial distance is less than r, $\bowtie$ denotes Cartesian product.

This paper also utilizes temporal predication such as *Concurrent*, *Follow*, $Behind_i$ to represent temporal relationship between STEs and STOs when generalizing temporal relationship. Being similar to the process of spatial concept generalization, the process of temporal concept generalization includes: (1) to compute temporal distance via joining ***STET*** to ***STDT***, (2) to utilize temporal distance generalization to generate time generalized table (***TGT*** for short). If used TG (•) to denote the operator of temporal generalization, time generalized table can be expressed as follows:

$$\boldsymbol{TGT}=\mathrm{TG}(\sigma_{\mathrm{Distt}(ste,sto)\leqslant w}(\boldsymbol{STET}\bowtie\boldsymbol{STDT})) \tag{2}$$

In Eq. (2), Distt(ste, sto)$\leqslant w$ aims to filter all STOs, whose temporal distance is less than *w*.

Events-coverage based spatio-temporal transaction table (***STT*** for short) is the result of the operations of join and projection among ***EGT***, ***DGT***, ***SGT*** and ***TGT***. It can be depicted as follows:

$$\boldsymbol{STT}=\pi_{Express}(\boldsymbol{EGT}\bowtie\boldsymbol{DGT}\bowtie\boldsymbol{SGT}\bowtie\boldsymbol{TGT}) \tag{3}$$

In Eq. (3), $\pi$ is the projecting operator, the *Express* is the expression of data items, which contains at least one spatial relationship item and one spatio-temporal relationship item. $\pi_{Express}$ is to select data items in the *express* via joining ***EGT***, ***DGT***, ***SGT***, ***TGT***.

## 4 EVENTS-COVERAGE BASED SPATIO-TEMPORAL ASSOCIATION RULES

For the data items of spatio-temporal transaction table, this paper uses $\boldsymbol{\tau}$ to denote the temporal relationship items and spatial relationship items, and uses $\boldsymbol{I}$ to denote all the items among spatio-temporal table, then the pre-incident $\boldsymbol{X}\subseteq\boldsymbol{I}\cup\boldsymbol{\tau}$, post-incident $\boldsymbol{Y}\subseteq\boldsymbol{I}(\boldsymbol{X}\cap\boldsymbol{Y}=\varnothing)$, where $\boldsymbol{X}$ contains at least one temporal relationship item and one spatial relationship item. The temporal relationship items and the spatial relationship items among $\boldsymbol{X}$ are named as spatio-temporal items (denoted as $\boldsymbol{\varphi}$**)**. Thus, $\boldsymbol{\varphi}=\boldsymbol{X}\cap\boldsymbol{\tau}$**,** the records of $\boldsymbol{\varphi}$ is denoted as $\boldsymbol{EGT}<\boldsymbol{\varphi}>$ in spatio-temporal events generalized table. The records of $\boldsymbol{\varphi}$, $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{X}\cup\boldsymbol{Y}$ is respectively denoted as $\boldsymbol{STT}<\boldsymbol{\varphi}>$, $\boldsymbol{STT}<\boldsymbol{X}>$, $\boldsymbol{STT}<\boldsymbol{Y}>$, $\boldsymbol{STT}<\boldsymbol{X}\cup\boldsymbol{Y}>$, where $\boldsymbol{STT}<\boldsymbol{X}\cup\boldsymbol{Y}>$ denotes the records for $\boldsymbol{X}$ and $\boldsymbol{Y}$ happening at the same time.

The events-coverage based STARs mining is dandified as follows:

**Definition 1** An spatio-temporal association rule is an form like $\boldsymbol{X}\Rightarrow\boldsymbol{Y}$(*p*%, *s*%, *c*%), where *p*% is the event probability, *s*% is the support, *c*% is the confidence .

**Definition 2** The event probability, the ratio of the number of records including $\boldsymbol{\varphi}$ to all the records of ***EGT***, is expressed as follows:

$$probability(\boldsymbol{\varphi})=\frac{\|\boldsymbol{EGT}<\boldsymbol{\varphi}>\|}{\|\boldsymbol{EGT}\|} \tag{4}$$

In Eq. (4), the event probability *p*% denotes the probability of occurrence of $\boldsymbol{\varphi}$, In order to prevent some small-probability events generating the greater support and confidence, this paper utilizes a given minimum threshold *minPr* of *p*% to weed the small probability events from spatio-temporal transaction table. Therefore, the efficiency and the credibility of spatio-temporal association rules could be improved.

**Definition 3** The support is the ratio of the number of transactions, which records $\boldsymbol{X}$ and $\boldsymbol{Y}$ happening concurrently, to the number of transactions including $\boldsymbol{\varphi}$ in the ***STT***. The support could be expressed as follows:

$$Support(\boldsymbol{X}\Rightarrow\boldsymbol{Y})=\frac{\|\boldsymbol{STT}<\boldsymbol{X}\cup\boldsymbol{Y}>\|}{\|\boldsymbol{STT}<\boldsymbol{\varphi}>\|} \tag{5}$$

The support denotes the ratio of the number of transactions recording $\boldsymbol{X}$ and $\boldsymbol{Y}$ happening concurrently to the number of transactions including all the spatio-temporal items.

**Definition 4** The confidence is the ratio of the number of transactions recording $\boldsymbol{X}$ and $\boldsymbol{Y}$ happening concurrently to the number of transactions recording $\boldsymbol{X}$ only happening in the ***STT***. The confidence could be expressed as follows:

$$Confidence(\boldsymbol{X}\Rightarrow\boldsymbol{Y})=\frac{\|\boldsymbol{STT}<\boldsymbol{X}\cup\boldsymbol{Y}>\|}{\|\boldsymbol{STT}<\boldsymbol{X}>\|} \tag{6}$$

**Definition 5** The expected confidence, namely the ratio of the number of transactions including $\boldsymbol{X}$ to the number of transactions including $\boldsymbol{\varphi}$**,** could be expressed as follows:

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$EC(\boldsymbol{X} \Rightarrow \boldsymbol{Y}) = \frac{\|\boldsymbol{STT} < \boldsymbol{X} >\|}{\|\boldsymbol{STT} < \boldsymbol{\varphi} >\|} \quad (7)$$

**Definition 6** The lift, namely the ratio of confidence to expected confidence, could be expressed as follows:

$$Lift(\boldsymbol{X} \Rightarrow \boldsymbol{Y}) = \frac{Confidence(\boldsymbol{X} \Rightarrow \boldsymbol{Y})}{EC(\boldsymbol{X} \Rightarrow \boldsymbol{Y})} \quad (8)$$

If *minSup* denotes the minimum support threshold, and *minConf* denotes the minimum confidence threshold, all rules, whose $p\% \geqslant minPr$, $s\% \geqslant minSup$, $c\% \geqslant minConf$, and $Lift > 1$, are called strong STARs, which are potential interesting and useful.

## 5 ECSTAR ALGORITHM

The mining process of ECSTARs includes: (1) to generalize data, (2) to build spatio-temporal transaction table, and (3) to mine association rules from the ***STT***. Thus, ECSTAR algorithm could be described as follows:

Input:

(1) spatio-temporal events table ***STET***

(2) spatio-temporal table ***STDT***

(3) *r* and *w*

(4) *minPr*, *minSup*, *minConf*

Output: the set of spatio-temporal association rules

```
EADSTAR (STET, STDT, r, w, minPr, minSup, minConf)
{
   EGT = STET_Generalize(STET) ;                  //to generalize STET
   EGT = Prune (EGT, minPr);                      // to utilize minPr to prune EGT
   DGT = STDT_Generalize(STDT);                   //to generalize the spatio-temporal data
   SGT = DG (σ_Dists(stest,o)≤r (STET ⋈ STDT)); //to generalize the spatial relationship
   TGT =TG (σ_Distt(stest,o)≤w (STET ⋈ STDT)); //to generalize the temporal relationship
   STT = π_Express (EGT ⋈ DGT ⋈ SGT ⋈ TGT);    // to build SIT table
   STARs = GetSTARs (STT, minSup, minConf ) ;   // mine association rules
   Return STARs;
 }
// The process to prune the STARs
Procedure Prune (EGT, minPr)
{
   For each e ⊆ X∩τ
   {
      if ||EGT<e>|| / ||EGT|| ≤ minPr then
      {
            Delete all EGT<e> from EGT;          // to delete all the records of small probability events from EGT
      }
 }
 Return EGT;
}
// The process to mine spatio-temporal association rules from spatio-temporal transaction table.
GetSTARs (STT, minSup, minConf )
```

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

{

***STARs***=∅; // to initialize the set of spatio-temporal relationship rules

for each $i \subseteq \boldsymbol{X}$

for each $j \subseteq \boldsymbol{Y}$

for each $e \subseteq i \cap \boldsymbol{\tau}$

{

*Sup*= ||***STT***<$i \cup j$>|| / ||***STT***<$e$>||; // to compute *Support*

*Conf*= ||***STT***<$i \cup j$>|| / ||***STT***<$i$>||; // to compute *Confidence*

*Lift*= (||***STT***<$i \cup j$>||×||***STT***<$e$>||)/ ||***STT***<$i$>||$^2$; //to compute *Lift*

If *Sup*≥*minSup*∧ *Conf*≥*minConf* ∧ *Lift*≥1

***STARs***=∪ {"i⇒*j*"};

}

Return ***STARs***;

}

In order to reduce the computational complexity and improve the efficiency of algorithm, EADSTAR should prune ***EGT*** to reduce the number of records, which are involved in later computing steps, after generalized ***STET***.

## 6 EXPERIMENTS

The data, including three air quality monitoring stations and two rainfall stations in a city locating in southern china from 2004 to 2005, are employed to verify the feasibility of the ECSTAR. All the stations are shown in Fig. 2.

The data items of the air quality monitored table include station number, monitoring date, average daily concentration of $SO_2$, average daily concentration of $PM_{10}$, and average daily concentration of $NO_x$. Some data is shown in Table 1. The data items of the rainfall monitored table include station number, monitoring date, rainfall, and so on. Some data is shown in Table 2.

In this experiment, the rainfall is considered as the STE, and the association rules between rainfall and average daily concentration of $SO_2$, $PM_{10}$ and $NO_x$ are mined. Therefore, the rainfall monitored table is considered as ***STET***, and air quality monitoring table is considered as ***STDT***. In order to present the change of air quality, this experiment firstly computes the difference between each monitor values and the previous, and then generalizes the difference values.

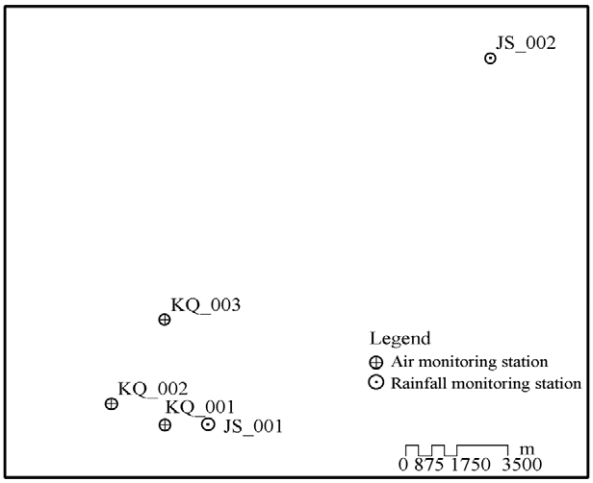


Fig. 2 Distribution of all monitor station

**Table 1 Air quality monitored data**

| Station number | Date | Average daily concentration of $SO_2$ /(mg/m$^3$) | Average daily concentration of $PM_{10}$ /(mg/m$^3$) | Average daily concentration of $NO_x$ /(mg/m$^3$) |
|---|---|---|---|---|
| KQ_001 | 2004-01-01 | 0.071 | 0.147 | 0.020 |
| KQ_002 | 2004-01-01 | 0.019 | 0.141 | 0.023 |
| KQ_003 | 2004-01-01 | 0.019 | 0.140 | 0.053 |
| KQ_001 | 2004-01-02 | 0.044 | 0.142 | 0.020 |
| KQ_002 | 2004-01-02 | 0.013 | 0.134 | 0.021 |
| KQ_003 | 2004-01-02 | 0.015 | 0.137 | 0.051 |
| … | … | … | … | … |

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

**Table 2 Rainfall monitored data**

| Station number | Date | Rainfall/mm |
|---|---|---|
| JS_001 | 2004-01-18 | 6.4 |
| JS_001 | 2004-01-19 | 4.8 |
| JS_001 | 2004-02-07 | 80 |
| JS_001 | 2004-03-28 | 16 |
| JS_001 | 2004-03-30 | 32 |
| JS_001 | 2004-04-01 | 48 |
| JS_002 | 2004-05-04 | 13.2 |
| … | … | … |

It is assumed that each of the rainfall stations could impact the whole area, namely $r$=42km. And the length of temporal coverage $w$=4. The ***STT*** is shown in Table 3.

Supposed *minPr* = 5%, *minSupp* = 5% and *minConf* = 55%, the STARs mined by the Apriori algorithm from ***STT*** are shown in Table 4.

According to the background knowledge and logical reasoning, all the mined ***STARs*** should be pruned, that is, to delete unreasonable rules, and to merge logically consistent rules. The pruned ***STARs*** are shown in Table 5.

The results in the table 5 show that the $PM_{10}$ concentration is impacted more significantly by the rainfall than the concentration of $SO_2$ and $NO_x$. For example, the R3, R5 and R8 denote that the $PM_{10}$ concentration increases slowly if the air monitoring station *is far to* the rainfall monitoring station in the third day after raining. The R1, R4, R7 and R11 denote that the $PM_{10}$ concentration declining slowly if the air monitoring station *is*

**Table 3 Events-coverage based spatio-temporal transaction table**

| TID | Rainfall grade | Temporal relationship | Spatial relationship | The change of $SO_2$ average daily concentration | The change of $PM_{10}$ average daily concentration | The change of $NO_x$ average daily concentration |
|---|---|---|---|---|---|---|
| 1 | *Light* | *Concurrent* | *IsCloSTETo* | *L*+ | *L*+ | *L*+ |
| 2 | *Light* | *Concurrent* | *IsMediumTo* | *L*+ | *L*+ | *L*+ |
| 3 | *Light* | *Concurrent* | *IsFarTo* | *L*+ | *L*+ | *L*+ |
| 4 | *Moderate* | *Concurrent* | *IsFarTo* | *L*- | *L*- | *L*- |
| 5 | *Light* | *Concurrent* | *IsCloSTETo* | *L*- | *M*- | *H*- |
| 6 | *Light* | *Concurrent* | *IsMediumTo* | *L*- | *L*- | *L*- |
| 7 | *Heavy* | *Follow* | *IsMediumTo* | / | / | *L*- |
| 8 | *Light* | *Follow* | *IsCloSTETo* | *L*+ | *L*+ | *L*+ |
| 9 | *Light* | *Follow* | *IsFarTo* | *L*- | *L*+ | *L*- |
| … | … | … | … | … | … | … |

Note: *L* denotes changing slowly, *M* denotes changing medially, *H* denotes changing highly of slow, + denotes rising, – denotes falling and / denotes no change.

**Table 4 Spatio-temporal association rules**

| Number | Spatio-temporal association rules(*minSupp*=5%, *minConf*=55%, *Lift*>1) |
|---|---|
| R1 | *torrential rain*∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}$(*L*-) (*Supp*=8.19% *Conf*=66.67% *Lift*=84.39) |
| R2 | *heavy rain*∧*Behind*$_1$∧*IsFarTo* ⇒ $PM_{10}$(*L*+) (*Supp*=4.35% *Conf*=66.67% *Lift*=98.04) |
| R3 | *light rain*∧*Follow*∧*IsFarTo* ⇒ $PM_{10}$(*L*+) (*Supp*=9.17% *Conf*=64.71% *Lift*=14.91) |
| R4 | *torrential rain*∧*Behind*$_1$∧*IsFarTo* ⇒ $PM_{10}$(*L*+) (*Supp*=4.09% *Conf*=63.64% *Lift*=163.18) |
| R5 | *light rain*∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}$(*L*-) (*Supp*=5.00% *Conf*=61.76% *Lift*=26.06) |
| R6 | *light rain*∧*Concurrent*∧*IsMediumTo* ⇒ $SO_2$(*L*-) (*Supp*=4.88% *Conf*=60.29% *Lift*=26.10) |
| R7 | *torrential rain*∧*Follow*∧*IsFarTo* ⇒ $SO_2$(*L*-) (*Supp*=5.26% *Conf*=60.00% *Lift*=117.65) |
| R8 | *torrential rain*∧*Concurrent*∧*IsFarTo* ⇒ $PM_{10}$(*L*-) (*Supp*=20.47% *Conf*=58.33% *Lift*=29.61) |
| R9 | *torrential rain*∧*Concurrent*∧*IsCloSTETo* ⇒ $PM_{10}$(*L*-) (*Supp*=7.02% *Conf*=57.14% *Lift*=84.03) |
| R10 | *moderate rain*∧*Behind*$_1$∧*IsFarTo* ⇒ $PM_{10}$(*L*+) (*Supp*=4.19% *Conf*=57.14% *Lift*=50.57) |
| R11 | *light rain*∧*Concurrent*∧*IsCloSTETo* ⇒ $PM_{10}$(*L*-) (*Supp*=4.52% *Conf*=55.88% *Lift*=26.11) |
| R12 | *moderate rain*∧*Concurrent*∧*IsFarTo* ⇒ $PM_{10}$(*L*-) (*Supp*=21.17% *Conf*=55.80% *Lift*=9.79) |
| R13 | *light rain*∧*Follow*∧*IsFarTo* ⇒ $NO_x$(*L*+) (*Supp*=7.86% *Conf*=55.46% *Lift*=14.91) |
| R14 | *light rain*∧*Concurrent*∧*IsFarTo* ⇒ $PM_{10}$(*L*-) (*Supp*=16.90% *Conf*=55.25% *Lift*=6.90) |
| R15 | *heavy rain*∧*Concurrent*∧*IsMediumTo* ⇒ $NO_x$(*L*+) (*Supp*=5.80% *Conf*=55.17% *Lift*=61.30) |
| R16 | *heavy rain*∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}$(*L*-) (*Supp*=5.80% *Conf*=55.17% *Lift*=61.30) |
| R17 | *moderate rain*∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}$(*L*-) (*Supp*=5.66% *Conf*=55.10% *Lift*=36.25) |

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

**Table 5 Pruned spatio-temporal association rules**

| Number | spatio-temporal association rules (*minSupp*=5%, *minConf*=55%, *Lift*>1) |
|---|---|
| R1 | *torrential rain*∧*Concurrent* ∧*IsMediumTo*⇒ $PM_{10}(L\text{-})$ |
| R2 | *torrential rain*∧*Follow*∧*IsFarTo* ⇒ $SO_2(L\text{-})$ |
| R3 | *torrential rain*∧$Behind_1$∧*IsFarTo* ⇒ $PM_{10}(L+)$ |
| R4 | *heavy rain*∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}(L\text{-})$ ∧$NO_x(L+)$ |
| R5 | *heavy rain*∧$Behind_1$∧*IsFarTo* ⇒ $PM_{10}(L+)$ |
| R6 | *moderate rain*∧*Concurrent*∧*IsFarTo* ⇒ $PM_{10}(L\text{-})$ |
| R7 | *moderate rain*∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}(L\text{-})$ |
| R8 | *moderate rain*∧$Behind_1$∧*IsFarTo* ⇒ $PM_{10}(L+)$ |
| R9 | *light rain*∧*Concurrent*∧*IsCloseTo* ⇒ $PM_{10}(L\text{-})$ |
| R10 | *light rain*∧*Concurrent*∧*IsFarTo* ⇒ $PM_{10}(L\text{-})$ |
| R11 | *light rain*∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}(L\text{-})$ ∧$SO_2(L\text{-})$ |
| R12 | *light rain*∧*Follow*∧*IsFarTo* ⇒ $PM_{10}(L+)$ ∧$NO_x(L+)$ |

*medium to* rainfall monitoring station on the same day of raining. The R4 denotes that the $NO_x$ concentration is rising slowly, the R11 denotes that the $SO_2$ concentration declining slowly. Additionally, to present the STARs, this experiment introduces spatial distance relationship predications and temporal distance relationship predications to express the relationship between the rainfall events and the change of air quality. The results are useful to the knowledge of environmental protection and are important to the environmental protection. The results also can prove the feasibility and practicality of the ECSTAR algorithm.

## 7 CONCLUSIONS

The method to mine spatio-temporal association rules based on the events coverage utilizes it to partition the spatio-temporal domain, and build the spatio-temporal transaction table via using the spatio-temporal relationship predications to express spatio-temporal relationship between the events and the objects. Thus, the method could ignore all the spatio-temporal objects unaffected by events and improve the efficiency of mining spatio-temporal association rules via removing a large number of redundant information. And, the interest and credibility of the association rules are improved, too. However, the method proposed in this paper is affected by the parameters *r* and *w*. So far, there is no the effective evaluating method about the impact of the two parameters, which is further research of this paper.

## REFERENCES

Celik M, Shekhar S, Rogers J P and Shine J A. 2006. Sustained emerging spatio-temporal co-occurrence pattern mining: a summary of results. Proceedings of the ICTAI

Hsu W, Lee M L and Wang J M. 2008.Temporal and spatio-temporal data mining. Hershey: IGI Publishing

Hu G Z. 2006. An extended cellular automata model for data mining of land development data. Proceeding of the 5th IEEE/ACIS International Conference on Computer and Information Science, Honolulu

Li X J and Meng Z Q. 2005. Data Mining of Temporal Sequence Patterns in a temporal space. *Microelectronics & Computer*, **22**(9): 29—35

Li Y J, Ning P and Wang X Y. 2003. Discovering calendar-based temporal association rules. *Data Knowledge Engine*, **44**(2): 193—218

Meng Z Q. 2001. Some properties for the associations rule of temporal data mining. *Computer Engineering and Application*, **10**: 86—87

Mennis J and Liu J W. 2005. Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change. *Transactions in GIS*, **9**(1): 13—18

Ren J D, Bao J and Huang H Y. 2003. The research on Spatio-temporal data model and related data Mining. Proceedings of the Second International Conference on Machine Learning and Cybernetics, Xi'an, 2—5 November

Su F Z, Du Y Y, Yang X M and Liu B Y. 2004. Geo-association rule with spatiotemporal reasoning .*Geo-Information Science*, **6**(4): 66—69

Verhein F and Chawla S. 2006. Mining spatio-temporal association rules, sources, sinks, stationary regions and thoroughfares in object mobility databases. Proceeding of the 11th International Conference on Database Systems for Advanced Applications (DASFAA'06)

Wang D W, Wang R J, Li Y and Wei B Z. 2008. Based on space-time snapshot of database time series prediction. *Computer Engineering and Application*, **44**(3): 180—182

Yu Y P, Chen X Y, Liu J N and Du J. 2008. Application of extended state cellular automata to spatiotemporal data mining . *Geomatics and Information Science of Wuhan University*, **33**(6): 592—595

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

# 利用事件影响域挖掘时空关联规则

李光强 邓 敏 张维玲 陈 翼

中南大学 测绘与国土信息工程系，湖南 长沙 410083

摘 要： 首先发展了基于事件影响域的时空事务表构建策略，提出了基于事件影响域的时空关联规则挖掘方法，给出了相应的挖掘算法(简称 ECSTAR 算法)。通过一个实际算例验证了所提方法的可行性和有效性。

关键词： 时空关联规则，时空事件，事件影响域，时空事务表

中图分类号： TP751.1 文献标识码： A

**引用格式**: 李光强，邓 敏，张维玲，陈 翼. 2010. 利用事件影响域挖掘时空关联规则. 遥感学报, **14**(3): 468—481
Li G Q, Deng M, Zhang W L and Chen Y. 2010. Events-coverage based spatio-temporal association rules mining method. *Journal of Remote Sensing*. **14**(3): 468—481

## 1 引 言

时空数据挖掘是从海量时空数据中获取隐藏的、有用的信息和知识的过程，是空间信息领域研究的重要方向之一。时空数据挖掘主要包括时空关联规则挖掘、时空序列模式挖掘、时空数据预测、时空聚类和区分规则挖掘等内容(Hsu 等, 2008)。其中，时空关联规则挖掘旨在时空数据库中发现数据项集之间潜在有用的时空关联关系，是时空数据挖掘领域里最为关键的技术难点之一。目前，针对时态或空间数据挖掘的研究文献众多，但是在综合考虑时态和空间(即时空耦合情况)下研究时空关联规则挖掘尚没有经典成果。现有的时空关联规则的挖掘方法是通过构造事务表(又称挖掘表)，并在事务表中挖掘时空关联规则。构造事务表的方法大致可以分为基于时空关系和基于时空划分的方法。

基于时空关系的方法主要是考虑时空目标间的时空关系来构建事务表。例如, Celik 等(2006)研究提出了时空共位规则挖掘算法(SECOP)，但 SECOP 算法割裂了时空耦合的特性，而且计算复杂。苏奋振等(2004)针对渔场环境变化提出了时空配置模型(STAMM)，该模型适用于连续变化的时空数据库。孟志青(2001)和李向军(2005)研究了基于时态(时间段)数据的关联规则挖掘方法，并提出了与时间段有关的关联规则可信度概念，进而提出了时态序列模式的挖掘方法。Li 等(2003)通过引入日历表达式，提出日历时间约束的关联规则挖掘算法。这些研究成果主要用于时态型关联规则挖掘，没有顾及空间关系。Ren(2003)、Verhein 等(2006)依据移动目标的时空位置，利用对象-关系数据库建立移动时空数据库，借助基于语义的时空关系表示方法，研究移动目标时空路线的关联规则，但该方法不适用于其他非移动类型的时空数据库。

基于时空划分的方法是将时空区域划分为若干相同大小的单元格，然后在每个单元格里构建事务表。例如, Mennis 等(2005)将研究区域划分为若干个相同大小的空间网格，以每个网格单元作为挖掘表构造单元，然后在每个单元中，利用时空关系概念层次表达方法构造基于时空关系概念的挖掘表，最后在挖掘表中挖掘时空关联规则。喻永平等(2008)研究了基于状态扩展元胞自动机的时空研究区域划分方法，然后利用遗传算法寻找元胞状态转换规则(Hu, 2006)。王大为等(2008)提出了基于时空快照数据库的时间序列预测和关联规则的挖掘方法，将时空快照序列聚集为若干簇，然后在簇内挖掘关联规则，提高挖掘效率。基于时空划分的方法不仅受单元格大小影响，而且单元格的属性数据有时需用内插方法计算，计算效率较低，且内插方法的不确定

收稿日期：2009-01-08；修订日期：2009-05-11
基金项目：863 计划资助(编号: 2009AA12Z206)；湖南省自然科学基金项目(编号: 09JJ6061)；地理空间信息工程国家测绘局重点实验室开放基金项目(编号: 200805)和江苏省资源环境信息工程重点实验室开放基金项目(编号: 20080101)
第一作者简介：李光强(1972— )，男，江苏徐州人，博士，现主要从事数据挖掘研究及教学工作。E-mail: ligq168@163.com。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

性影响了挖掘结果的可靠性。

综上所述，时空事务表的构造方法在时空关联规则挖掘过程中起到关键作用，直接影响挖掘效率及其挖掘结果。为此，本文发展一种基于事件影响域的时空事务表构建方法，该方法利用事件影响域划分时空域，并在每个影响域中构建事务记录，从而可以去除冗余数据，提高挖掘效率和挖掘结果的可靠性。进而，提出一种基于事件影响域的时空关联规则挖掘方法(event-coverage based Spatio-Temporal Association Rules Mining，简称 ECSTAR)。

## 2 时空事件及其影响域

在现实世界里，时空目标的演变主要受到一些重要事件的控制，而任何事件都具有一定的时间和空间影响范围，即事件仅影响一定时间、空间范围内的时空目标。时空事件(spatio-temporal event，简称 *STE*)是指具有一定时间和空间影响范围的事物，可以是时空目标的产生、消亡，或者是时空目标状态的改变，而且会在一定时间内对空间邻近区域的其他目标产生一定的影响。显然，时空事件具有相对性，并且与研究领域密切相关。同一个事物在一个应用领域中可能是时空事件，但在另一个应用领域中可能不是时空事件。例如在地质灾害监测中，暴雨就是一个非常重要的时空事件，但是在位置服务领域中，则不是有意义的时空事件。

时空事件影响的时间、空间范围，称为时空影响域(spatio-temporal coverage)。时空影响域在时间上的投影称为时间影响长度(记为 $w$)，在空间上的投影称为空间影响区(记为 $s$)。时空事件在时间上的影响是单向的，即从事件发生时刻起，开始向后延续，直到事件的影响彻底消除为止。为了简化计算，本文假定时空事件在空间上的影响是各向同质的，即时空事件对周围的影响是距离的函数，与方向无关。因此，时空事件的空间影响区域是以时空事件发生位置为中心，以影响距离(记为 $r$)为半径的圆。图 1 中，时空事件 $STE_1$ 的时间影响长度 $w_1$=3，空间上影响 $O_1$—$O_5$ 目标，即 $STE_1$ 在 $t_p$—$t_{p+3}$ 时间内对 $O_1$—$O_5$ 的属性变化产生影响。同样，$STE_2$ 在 $t_q$—$t_{q+5}$ 时间内对 $O_6$—$O_9$ 的属性变化产生影响。

在时空演化过程中，由于时空事件与时空影响域中的其他时空目标的属性值之间存在相关性，基于事件影响域的时空关联规则挖掘方法也就等价于时空事件与其影响域中其他时空目标属性数据变化之间的关联规则的挖掘。因此，可以利用时空事件

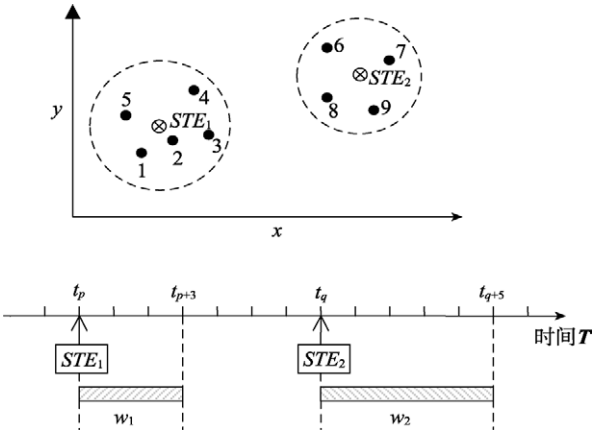


图 1 时空事件影响域示意图

的影响域来构建时空事务表，进而挖掘时空关联规则，这亦是本文研究的理论依据。

## 3 基于事件影响域的时空事务表

时空事件影响域将时空研究区域划分出若干个时空事务单元，然后在影响域中使用数据概念泛化方法构建时空事务记录。数据概念泛化是数据挖掘中常用的方法，是将连续数值型数据项进行离散，并使用概念表达的一种方法。数据概念泛化常用的方法有等深和等宽两种，前者是将样本平均分配到每个概念分组中，后者是依据相同的宽度将样本值域平均分成若干个分组。在具体的应用领域中，经常参考领域背景知识制定适当的概念泛化原则。如在对降雨量进行泛化时，可以参考气象部门常用的降雨分级方法进行泛化。

为叙述方便，本文把所有对时空邻近域产生显著影响的时空目标称为时空事件，并把记录时空事件的数据表称为时空事件表；把所有受时空事件影响的其他时空目标称为时空目标，并将记录时空目标的数据表称为时空数据表。因此，一个时空数据集至少包括时空事件表和时空数据表两类数据，时空数据概念泛化的步骤则包括：(1)时空事件表概念泛化；(2)时空数据表概念泛化；(3)时空事件与其他时空目标的空间关系概念泛化；(4)时空事件与其他时空目标的时间关系概念泛化。

时空事件数据表(spatio-temporal events table，记为 ***STET***)和时空数据表(spatio-temporal data table，记为 ***STDT***)只需要依据数据表中的 1 个或多个数据项，使用概念分层泛化技术，将时空事件或时空数据归入相应概念格，实现数据概念泛化。***STET*** 泛化表记为 ***EGT***(events generalized table)，***STDT*** 泛化表

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

记为 ***DGT***(data generalized table)。时空事件与时空目标的空间关系概念泛化、时间关系概念泛化较为复杂，需要数据表的连接和选择等操作。

时空事件与时空目标的空间关系概念泛化过程是先通过 ***STET*** 与 ***STDT*** 的连接操作，计算事件与时空目标间的空间距离，再使用空间谓词表示空间距离关系，实现空间概念泛化。本文将事件与时空目标间的空间距离泛化为空间谓词 *IsFarTo*(远)、*IsMediumTo*(中等)和 *IsCloseTo*(近)等概念。若距离泛化算子用 DG(·)表示，则时空事件与时空目标的空间概念泛化表(spatial generalized table，记为 ***SGT***)用关系代数表达为:

$$\boldsymbol{SGT}=\mathrm{DG}(\sigma_{\mathrm{Dists}(ste,sto)\le r}(\boldsymbol{STET}\bowtie\boldsymbol{STDT})) \quad (1)$$

式中，$\sigma$是关系代数中的选择运算，Dists(*ste*, *sto*)是时空事件与时空目标的空间距离计算公式，Dists(*ste*, *sto*)≤$r$ 为选择条件，即过滤所有距时空事件空间距离 *ste*≤$r$ 的时空目标，$\bowtie$ 表示笛卡尔积运算符。

时空事件与时空目标的时间关系概念泛化时，本文使用时间谓词 *Concurrent*(同时)、*Follow*(相邻或随后)和 $Behind_i$(间隔，其中 $i$ 表示间隔时刻数)表示时空事件与时空目标间的时间关系。时间概念泛化过程与空间概念泛化过程类似，首先将 ***STET*** 和 ***STDT*** 进行连接并计算时间距离，根据时间距离泛化后得到时间泛化表(time generalized table，记为 ***TGT***)。若时间泛化算子用 TG(·)表示，则时空泛化表用关系代数表达为:

$$\boldsymbol{TGT}=\mathrm{TG}(\sigma_{\mathrm{Distt}(ste,sto)\le w}(\boldsymbol{STET}\bowtie\boldsymbol{STDT})) \quad (2)$$

式中: Distt(*ste*, *sto*)≤$w$ 是从时空数据表中选择出所有与事件时间距离 *ste*≤$w$ 的时空目标。

基于时空事件影响域的时空事务数据表(Events-Coverage based Spatio-Temporal transaction table，简记为 ***STT***)是 ***EGT*** 表、***DGT*** 表、***SGT*** 表和 ***TGT*** 表连接和投影操作的结果，用关系代数表达为:

$$\boldsymbol{STT}=\pi_{Express}(\boldsymbol{EGT}\bowtie\boldsymbol{DGT}\bowtie\boldsymbol{SGT}\bowtie\boldsymbol{TGT}) \quad (3)$$

式中，$\pi$是投影操作，*Express* 是数据项选择表达式，其中至少包含一个空间关系和一个时空关系数据项。$\pi_{Express}$ 是从 ***EGT***, ***DGT***, ***SGT***, ***TGT*** 4 个表连接操作结果集中选择 *Express* 包括的数据项的操作，并组成 ***STT*** 表。

## 4 基于事件影响域的时空关联规则

若在时空事务表的数据项集中，时间关系、空间关系项集记为 $\boldsymbol{\tau}$，时空数据表项集记为 $\boldsymbol{I}$，前置件 $\boldsymbol{X}\subseteq\boldsymbol{I}\cup\boldsymbol{\tau}$ 和后置件 $\boldsymbol{Y}\subseteq\boldsymbol{I}(\boldsymbol{X}\cap\boldsymbol{Y}=\varnothing)$，且 $\boldsymbol{X}$ 中至少包含一个时间关系和一个空间关系数据项。$\boldsymbol{X}$ 中的时间关系项集和空间关系项集的统称为时空项集，并记为 $\boldsymbol{\varphi}$，则有 $\boldsymbol{\varphi}=\boldsymbol{X}\cap\boldsymbol{\tau}$。在时空事件泛化表中，$\boldsymbol{\varphi}$ 的记录集记为 $\boldsymbol{EGT}<\boldsymbol{\varphi}>$；在时空事务数据表中，$\boldsymbol{\varphi},\boldsymbol{X},\boldsymbol{Y},\boldsymbol{X}\cup\boldsymbol{Y}$ 的记录集分别记为 $\boldsymbol{STT}<\boldsymbol{\varphi}>$，$\boldsymbol{STT}<\boldsymbol{X}>$，$\boldsymbol{STT}<\boldsymbol{Y}>$，$\boldsymbol{STT}<\boldsymbol{X}\cup\boldsymbol{Y}>$；其中 $\boldsymbol{STT}<\boldsymbol{X}\cup\boldsymbol{Y}>$表示 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 同时发生的记录。

基于事件影响域的时空关联规则定义如下:

**定义 1** 时空关联规则是形如 $\boldsymbol{X}\Rightarrow\boldsymbol{Y}(p\%, s\%, c\%)$ 的规则，其中 $p$%是事件概率，$s$%是支持度，$c$%是置信度。

**定义 2** 事件概率是指时空事件泛化表中包括 $\boldsymbol{\varphi}$ 的记录数与所有记录数的比值，表达为:

$$probability(\boldsymbol{\varphi})=\frac{\|\boldsymbol{EGT}<\boldsymbol{\varphi}>\|}{\|\boldsymbol{EGT}\|} \quad (4)$$

式中，事件概率 $p$%表示 $\boldsymbol{\varphi}$ 发生的概率。为了避免小概率事件产生较大的支持度和置信度，在给定 $p$% 的最小阈值 *minPr* 时，可以从时空事务数据表中剔除小于 *minPr* 的小概率事件的记录，从而可以提高时空关联规则挖掘的效率和关联规则的可信度。

**定义 3** 支持度是指时空事务数据表中 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 同时发生的事务数与所有包含 $\boldsymbol{\varphi}$ 的事务数的比值，表达为:

$$Support(\boldsymbol{X}\Rightarrow\boldsymbol{Y})=\frac{\|\boldsymbol{STT}<\boldsymbol{X}\cup\boldsymbol{Y}>\|}{\|\boldsymbol{STT}<\boldsymbol{\varphi}>\|} \quad (5)$$

支持度表示在包含所有时空项集的事务数中，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 同时发生的事务数所占的比例。

**定义 4** 置信度是指在时空事务数据表中，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 同时发生的事务数和 $\boldsymbol{X}$ 的事务数的比值，表达为:

$$Confidence(\boldsymbol{X}\Rightarrow\boldsymbol{Y})=\frac{\|\boldsymbol{STT}<\boldsymbol{X}\cup\boldsymbol{Y}>\|}{\|\boldsymbol{STT}<\boldsymbol{X}>\|} \quad (6)$$

**定义 5** 期望置信度是指在时空事务数据表中，$\boldsymbol{X}$ 的事务数与 $\boldsymbol{\varphi}$ 的事务数的比值，即

$$EC(\boldsymbol{X}\Rightarrow\boldsymbol{Y})=\frac{\|\boldsymbol{STT}<\boldsymbol{X}>\|}{\|\boldsymbol{STT}<\boldsymbol{\varphi}>\|} \quad (7)$$

**定义 6** 作用度是指置信度与期望置信度的比值，即

$$Lift(\boldsymbol{X}\Rightarrow\boldsymbol{Y})=\frac{Confidence(\boldsymbol{X}\Rightarrow\boldsymbol{Y})}{EC(\boldsymbol{X}\Rightarrow\boldsymbol{Y})} \quad (8)$$

若用 *minSup* 和 *minConf* 表示支持度和置信度的最小阈值，则所有 $p\%\ge minPr$、$s\%\ge minSup$、$c\%\ge minConf$ 且 $Lift>1$ 的时空关联规则称为强时空关联规则，强时空关联规则是用户感兴趣、潜在有用的规则。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

## 5 ECSTAR 算法

基于事件影响域的时空关联规则挖掘过程大致包括：(1) 数据概念泛化；(2) 时空事务表构建；(3) 从时空事务表 ***STT*** 中挖掘关联规则。ECSTAR 算法描述如下：

输入:
(1)时空事件表 ***STET***
(2)时空数据表 ***STDT***
(3)*r* 和 *w*
(4)*minPr*, *minSup*, *minConf*
输出: 时空关联规则集 ***STARs***

```
EADSTAR(STET, STDT, r, w, minPr, minSup, minConf )
{
  EGT = STET_Generalize(STET) ;                    //对 STET 进行概念泛化
  EGT = Prune(EGT , minPr);                        //利用 minPr 修剪 EGT 表
  DGT = STDT_Generalize(STDT);                     //时空数据概念泛化
  SGT = DG(σ_Dists(stest,o)≤r(STET ⋈ STDT));       //空间关系概念泛化
  TGT =TG(σ_Distt(stest,o)≤w(STET ⋈ STDT));        //时间关系概念泛化
  STT = π_Express(EGT ⋈ DGT ⋈ SGT ⋈ TGT);          //构建 STT 表
  STARs = GetSTARs (STT, minSup, minConf ) ;       //挖掘关联规则
  Return STARs ;
}

//修剪过程
Procedure Prune(EGT , minPr)
{
  for each e ⊆ X∩τ
  {
    if ||EGT<e>||／||EGT|| ≤ minPr then
    {
        Delete all EGT<e> from EGT;                //从 EGT 中删除所有小概率事件的记录
    }
  }
return EGT;
}

//从时空事务表中挖掘时空关联规则过程
GetSTARs (STT, minSup, minConf )
{
    STARs=∅; //初始化时空关系规则集合
    for each i⊆X
     for each j⊆Y
       for each e ⊆ i∩τ
       {
         Sup= ||STT<i∪j>|| / ||STT<e>||; //计算支持度
         Conf= ||STT<i∪j>|| / ||STT<i>||; //计算置信度
         Lift= (||STT<i∪j>||×||STT<e>||)/ ||STT<i>||^2 ; //计算作用度
         if Sup≥minSup∧ Conf≥minConf ∧ Lift≥1
           STARs=∪{“ i⇒j ”};
       }
    Return STARs;
}
```

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

为了降低计算复杂度，EADSTAR 算法在 ***STET*** 表泛化后，接着进行 ***EGT*** 修剪，减少在后面各运算步骤中参与运算的记录数，以提高算法执行效率。

## 6 实验算例

本文选用华南某市 2004—2005 年间的 3 个空气质量监测站的监测数据与 2 个降水监测站的监测数据进行实验，验证 ECSTAR 算法的可行性，各监测站点的分布如图 2。

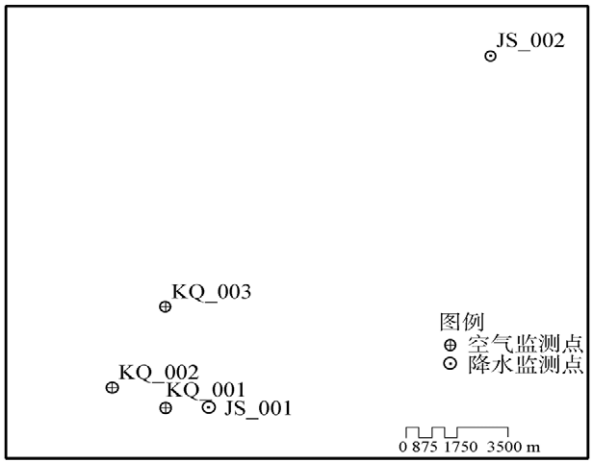


图 2 监测站空间分布图

2004~2005 年间的空气质量监测数据表的数据项包括站号、日期、$SO_2$ 日均浓度、$PM_{10}$ 日均浓度、$NO_x$ 日均浓度，部分数据如表 1；降水监测站的监测数据表的数据项包括测站号、日期、降水量等，部分数据如表 2。

本实验以降水发生作为时空事件，挖掘降水与空气中 $SO_2$、$PM_{10}$ 和 $NO_x$ 浓度的关联规则，因此降水监测数据表为时空事件数据表(***STET***)，空气质量监测数据表为时空数据表(***STDT***)。为了表达空气质量的变化情况，首先针对空气质量监测数据，计算各站的每个观测值与前一时刻观测值的差值，然后再对变化量进行概念泛化。

本实验中，假定每个降水监测站均对整个研究区域产生影响，即 $r$=42km，取时间影响长度 $w$=4，构建时空事务数据表如表 3。

若给定 *minPr*=5%，*minSupp*=5%和 *minConf*=55%；使用 Apriori 算法从 ***STT*** 中挖掘出的时空关联规则如表 4。

**表 1** 空气质量监测数据

| 站号 | 日期 | $SO_2$ 日均浓度 /(mg/m$^3$) | $PM_{10}$ 日均浓度 /(mg/m$^3$) | $NO_x$ 日均浓度 /(mg/m$^3$) |
|---|---|---|---|---|
| KQ_001 | 2004-01-01 | 0.071 | 0.147 | 0.020 |
| KQ_002 | 2004-01-01 | 0.019 | 0.141 | 0.023 |
| KQ_003 | 2004-01-01 | 0.019 | 0.140 | 0.053 |
| KQ_001 | 2004-01-02 | 0.044 | 0.142 | 0.020 |
| KQ_002 | 2004-01-02 | 0.013 | 0.134 | 0.021 |
| KQ_003 | 2004-01-02 | 0.015 | 0.137 | 0.051 |
| … | … | … | … | … |

**表 2 2004-2005 年**降水监测数据表

| 站号 | 日期 | 降水量/mm |
|---|---|---|
| JS_001 | 2004-01-18 | 6.4 |
| JS_001 | 2004-01-19 | 4.8 |
| JS_001 | 2004-02-07 | 80 |
| JS_001 | 2004-03-28 | 16 |
| JS_001 | 2004-03-30 | 32 |
| JS_001 | 2004-04-01 | 48 |
| JS_002 | 2004-05-04 | 13.2 |
| … | … | … |

**表 3** 时空事务数据表(***STT***)

| TID | 降水等级 | 时间关系 | 空间关系 | $SO_2$ 日均浓度变化 | $PM_{10}$ 日均浓度变化 | $NO_x$ 日均浓度变化 |
|---|---|---|---|---|---|---|
| 1 | 小雨 | *Concurrent* | *IsCloSTETo* | *L*+ | *L*+ | *L*+ |
| 2 | 小雨 | *Concurrent* | *IsMediumTo* | *L*+ | *L*+ | *L*+ |
| 3 | 小雨 | *Concurrent* | *IsFarTo* | *L*+ | *L*+ | *L*+ |
| 4 | 中雨 | *Concurrent* | *IsFarTo* | *L*- | *L*- | *L*- |
| 5 | 小雨 | *Concurrent* | *IsCloSTETo* | *L*- | *M*- | *H*- |
| 6 | 小雨 | *Concurrent* | *IsMediumTo* | *L*- | *L*- | *L*- |
| 7 | 大雨 | *Follow* | *IsMediumTo* | / | / | *L*- |
| 8 | 小雨 | *Follow* | *IsCloSTETo* | *L*+ | *L*+ | *L*+ |
| 9 | 小雨 | *Follow* | *IsFarTo* | *L*- | *L*+ | *L*- |
| … | … | … | … | … | … | … |

注: *L*, *M*, *H* 分别表示浓度变化慢，中等，快; +和-分别表示上升和下降, / 表示没有变化。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 4 时空关联规则

| 编号 | 时空关联规则(*minSupp*=5%, *minConf*=55%, *Lift*>1) |
|---|---|
| R1 | 暴雨∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}(L-)$ (*Supp*=8.19% *Conf*=66.67% *Lift*=84.39) |
| R2 | 大雨∧$Behind_1$∧*IsFarTo* ⇒ $PM_{10}(L+)$ (*Supp*=4.35% *Conf*=66.67% *Lift*=98.04) |
| R3 | 小雨∧*Follow*∧*IsFarTo* ⇒ $PM_{10}(L+)$ (*Supp*=9.17% *Conf*=64.71% *Lift*=14.91) |
| R4 | 暴雨∧$Behind_1$∧*IsFarTo* ⇒ $PM_{10}(L+)$ (*Supp*=4.09% *Conf*=63.64% *Lift*=163.18) |
| R5 | 小雨∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}(L-)$ (*Supp*=5.00% *Conf*=61.76% *Lift*=26.06) |
| R6 | 小雨∧*Concurrent*∧*IsMediumTo* ⇒ $SO_2(L-)$ (*Supp*=4.88% *Conf*=60.29% *Lift*=26.10) |
| R7 | 暴雨∧*Follow*∧*IsFarTo* ⇒ $SO_2(L-)$ (*Supp*=5.26% *Conf*=60.00% *Lift*=117.65) |
| R8 | 暴雨∧*Concurrent*∧*IsFarTo* ⇒ $PM_{10}(L-)$ (*Supp*=20.47% *Conf*=58.33% *Lift*=29.61) |
| R9 | 暴雨∧*Concurrent*∧*IsCloSTETo* ⇒ $PM_{10}(L-)$ (*Supp*=7.02% *Conf*=57.14% *Lift*=84.03) |
| R10 | 中雨∧$Behind_1$∧*IsFarTo* ⇒ $PM_{10}(L+)$ (*Supp*=4.19% *Conf*=57.14% *Lift*=50.57) |
| R11 | 小雨∧*Concurrent*∧*IsCloSTETo* ⇒ $PM_{10}(L-)$ (*Supp*=4.52% *Conf*=55.88% *Lift*=26.11) |
| R12 | 中雨∧*Concurrent*∧*IsFarTo* ⇒ $PM_{10}(L-)$ (*Supp*=21.17% *Conf*=55.80% *Lift*=9.79) |
| R13 | 小雨∧*Follow*∧*IsFarTo* ⇒ $NO_x(L+)$ (*Supp*=7.86% *Conf*=55.46% *Lift*=14.91) |
| R14 | 小雨∧*Concurrent*∧*IsFarTo* ⇒ $PM_{10}(L-)$ (*Supp*=16.90% *Conf*=55.25% *Lift*=6.90) |
| R15 | 大雨∧*Concurrent*∧*IsMediumTo* ⇒ $NO_x(L+)$ (*Supp*=5.80% *Conf*=55.17% *Lift*=61.30) |
| R16 | 大雨∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}(L-)$ (*Supp*=5.80% *Conf*=55.17% *Lift*=61.30) |
| R17 | 中雨∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}(L-)$ (*Supp*=5.66% *Conf*=55.10% *Lift*=36.25) |

根据专家背景知识和逻辑推理、归纳等方法，修剪挖掘出的关联规则，删除不合理的规则，合并逻辑上一致的规则，修剪后的时空关联规则如表 5 所示。

表 5 修剪后的时空关联规则

| 编号 | 时空关联规则(*minSupp*=5%, *minConf*=55%, *Lift*>1) |
|---|---|
| R1 | 暴雨∧*Concurrent* ∧*IsMediumTo*⇒ $PM_{10}(L-)$ |
| R2 | 暴雨∧*Follow*∧*IsFarTo* ⇒ $SO_2(L-)$ |
| R3 | 暴雨∧$Behind_1$∧*IsFarTo* ⇒ $PM_{10}(L+)$ |
| R4 | 大雨∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}(L-)$ ∧$NO_x(L+)$ |
| R5 | 大雨∧$Behind_1$∧*IsFarTo* ⇒ $PM_{10}(L+)$ |
| R6 | 中雨∧*Concurrent*∧*IsFarTo* ⇒ $PM_{10}(L-)$ |
| R7 | 中雨∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}(L-)$ |
| R8 | 中雨∧$Behind_1$∧*IsFarTo* ⇒ $PM_{10}(L+)$ |
| R9 | 小雨∧*Concurrent*∧*IsCloseTo* ⇒ $PM_{10}(L-)$ |
| R10 | 小雨∧*Concurrent*∧*IsFarTo* ⇒ $PM_{10}(L-)$ |
| R11 | 小雨∧*Concurrent*∧*IsMediumTo* ⇒ $PM_{10}(L-)$ ∧$SO_2(L-)$ |
| R12 | 小雨∧*Follow*∧*IsFarTo* ⇒ $PM_{10}(L+)$ ∧$NO_x(L+)$ |

在表 5 的实验结果中，很多关联规则涉及 $PM_{10}$ 浓度，说明降水对 $PM_{10}$ 浓度的影响较为显著，相较而言，降水对 $SO_2$ 和 $NO_x$ 浓度的影响相对较弱。例如, R3、R5 和 R8 规则均表示降水的第 3 天，在距离降水监测站较远处，$PM_{10}$ 浓度缓慢上升；R1、R4、R7 和 R11 均表示降水当天，在距离降水监测站中等距离处，$PM_{10}$ 浓度缓慢下降，R4 说明 $NO_x$ 浓度呈现缓慢上升, R11 还说明 $SO_2$ 浓度缓慢下降。此外，在表达时空关联规则时，本实验通过引入时间、空间距离的关系谓词来表达降水事件与空气质量变化的关系，是对环保领域知识的有益补充，对于环保工作具有较为重要的意义，本实验可以证明 ECSTAR 算法的可行性和实用性。

## 7 结 论

基于事件影响域的时空关联规则挖掘方法利用事件的影响域划分时空研究区域，进而利用时空关系谓词建立事件与影响域中目标间的时空关系，构建时空事务数据表，从而忽略所有不受事件影响的时空目标，去除大量冗余信息，能够提高时空关联规则的挖掘效率和结果规则的兴趣度和可信度。此外，通过引入事件概率过滤掉小概率事件，保证了时空关联规则的可靠性。但是，本文研究的方法受到时空事件影响域特征值 $r$ 和 $w$ 的影响，这两个参数的确定及其对挖掘结果的影响尚缺乏有效的评价手段，这也是本文有待进一步延伸的研究工作。

## REFERENCES

Celik M, Shekhar S, Rogers J P and Shine J A. 2006. Sustained

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

emerging spatio-temporal co-occurrence pattern mining: a summary of results. Proceedings of the ICTAI

Hsu W, Lee M L and Wang J M. 2008.Temporal and spatio-temporal data mining. Hershey: IGI Publishing

Hu G Z. 2006. An extended cellular automata model for data mining of land development data. Proceeding of the 5th IEEE/ACIS International Conference on Computer and Information Science, Honolulu

Li X J and Meng Z Q. 2005. Data Mining of Temporal Sequence Patterns in a temporal space. *Microelectronics & Computer*, **22**(9):29—35

Li Y J, Ning P and Wang X Y. 2003. Discovering calendar-based temporal association rules. *Data Knowledge Engine*, **44**(2): 193—218

Meng Z Q. 2001. Some properties for the associations rule of temporal data mining. *Computer Engineering and Application*, **10**: 86—87

Mennis J and Liu J W. 2005. Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change. *Transactions in GIS*, **9**(1): 13—18

Ren J D, Bao J and Huang H Y. 2003. The research on Spatio-temporal data model and related data Mining. Proceedings of the Second International Conference on Machine Learning and Cybernetics, Xi'an, 2—5 November

Su F Z, Du Y Y, Yang X M and Liu B Y. 2004. Geo-association rule with spatiotemporal reasoning .*Geo-Information Science*, **6**(4): 66—69

Verhein F and Chawla S. 2006. Mining spatio-temporal association rules, sources, sinks, stationary regions and thoroughfares in object mobility databases. Proceeding of the 11th International Conference on Database Systems for Advanced Applications (DASFAA'06)

Wang D W, Wang R J, Li Y and Wei B Z. 2008. Based on space-time snapshot of database time series prediction. *Computer Engineering and Application*, **44**(3): 180—182

Yu Y P, Chen X Y, Liu J N and Du J. 2008. Application of extended state cellular automata to spatiotemporal data mining . *Geomatics and Information Science of Wuhan University*, **33**(6): 592—595

## 附中文参考文献

孟志青. 2001. 时态关联规则采掘的若干性质.计算机工程与应用, **10**: 86—87

苏奋振, 杜云艳, 杨晓梅, 刘宝银. 2004. 地学关联规则与时空推理的渔业分析应用.地球信息科学, **6**(4): 66—69

王大为, 王儒敬, 李营, 魏保子. 2008. 时空快照数据库的时间序列预测.计算机工程与应用, **44**(3): 180—182

喻永平, 陈晓勇, 刘经南, 都洁. 2008. 自动机模型在时空数据挖掘中的应用.武汉大学学报(信息科学版), **33**(6): 592—595

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net