Toward mining of spatiotemporal maximal frequent patterns

Luboš Popelínský and Jan Blaťák

KD Lab at Faculty of Informatics
Masaryk University in Brno, Czech Republic
{popel,xblatak}@fi.muni.cz

Abstract. We show that propositional spatiotemporal logic $PSTL$ is a powerful tool for mining in various spatiotemporal data including environmental and medical data, keystroke dynamics data or text. We introduce a refinement operator for a fragment of $PSTL$, ST_0 and describe the ILP system GRAPE for mining first-order frequent patterns in spatiotemporal data. We also show that in the classification task the use of frequent patterns as new features result in an accuracy increase.

1 Introduction

Inductive logic programming (ILP), or multirelational data mining [10, 16], has been proven to be a useful tool for mining in data of complex structure including temporal [18, 8, 11, 15] and spatial [13, 14, 17] data. Despite of many approaches to mining in particular forms of spatiotemporal data, a work that exploits not only the expressiveness of the first-order language but also other advantages of inductive logic programming seems be missing.

We are not pretending to solve the general problem, inductive inference in a spatiotemporal logic. We rather exploit our experience from mining various data with explicit or implicit spatial and temporal features and show that the spatiotemporal logic ST_0 is powerful enough for mining interesting patterns in spatiotemporal data.

In this paper we focus on mining of first-order maximal frequent patterns in spatiotemporal logic. A *first-order frequent pattern* is a conjunction of positive literals from domain knowledge that covers at least N examples. A frequent pattern F is *maximal* if there is no frequent pattern G such that F is a prefix of G. In other word, there is no specialization of the pattern F which is frequent. A *spatial (temporal) maximal frequent pattern* is then a maximal pattern which contain at least one spatial (temporal) predicate.

It is clear that the transformation of temporal data into a form appropriate for an ILP system can be easily performed. See e.g. [8] and for the case of mining episodes in temporal data. The same holds for spatial data [13, 17].

However, the problem of efficiency will arise if an ILP system – actually its refinement operator – cannot exploit information about semantics of temporal or spatial predicates.

Let us have e.g. a temporal logic with operators $\Box$ (always in future) and $\Diamond$ (sometimes in future), windstorms data where K is a unique identifier for a strong wind, and the frequent pattern

year(K,C),$1971 \leq$ C $\leq 1972, \Box$(K,K1,type(strong)&char(house_destruction))

"after a wind K in the period 1971-72 all the winds was strong"

it is useless to add the literal $\Diamond$(K,K2,X) if X does not refine the term type(strong) & char(house_destruction). Similarly, for spatial predicates po(X,Y) (X overlaps Y) and tpp(X,Y) (X is tangential partial part of Y) an appearance of one of them prevents from use of the other. In current ILP systems this is usually solved by defining constraints but it in fact does not solve the problem because candidate patterns have to be constructed anyway.

In this paper we make a small step in this direction and describe the improvement of ILP system RAP [5] that enables to mine first-order frequent patterns from spatiotemporal data efficiently. This work follows also the research lines started with *GeoMiner* [12] and ILP systems *GWiM* [17] and particularly *SPADA* [14]. SPADA [14] is the first system for first-order association rule mining in geographic data. Data from *ORACLE* database are first transformed with FEATEX into the form of deductive database which is then mined with *SPADA*. As *GeoMiner*, *SPADA* employs hierarchy of geographic concepts. For each level of hierarchy minimum support and confidence are defined. *SPADA* has been shown successful for mining real-world data [13].

The structure of this paper is as follows. In Section 2 we give a brief overview of spatial logic *RCC-8* and its extensions, spatiotemporal logics ST_i that we use in this paper. A new refinement operator is described in Section 5. Section 3 brings description of the spatiotemporal data model and the mining task. In Section 4 we introduce new version of RAP for efficient mining in spatiotemporal data. Section 6 brings information about the explored data. The patterns found are displayed in Section 7. Use of the patterns for classification is described in Section 8.

2 Spatial and spatiotemporal logics

2.1 The Region Connection Calculus

The language of *RCC-8* [7] consists of *region variables* $X_0, X_1, ...$ and the eight binary predicates

dc(X,Y) ... X is disconnected from B
ec(X,Y) ... X is in external contact with B
eq(X,Y) ... X is equal to Y

po(X,Y) ... X overlaps Y
tpp(X,Y) ... X is tangential partial part of Y
tppi(X,Y) ... its inverse (tpp(X,Y)=tppi(Y,X))
ntpp(X,Y) ... X is non-tangential partial part of Y
ntppi(X,Y) ... its inverse.

A *spatial formula* is then constructed from the predicates above and the logical connectives $\wedge, \vee. \rightarrow, \neg$.

2.2 Extending RCC-8 with time dimension

In [1], Bennett et al. combines *RCC-8* with the propositional temporal logic *PTL*. Time is assuming to be isomorphic with the set of natural numbers and the relation $<$ is defined together with two temporal operators *Since* and *Until*. When allowing application of *Since* and *Until*, or other standard operators like $\bigcirc$ (next), $\Diamond$ (sometimes) or $\Box$ (always) to spatial formulas we have got the spatiotemporal language ST_0.

For example, the formula

$$\Diamond(change(diet, 7) \wedge \Diamond(no\_change(work\_char)))$$

"patients who came down with some disease and who changed their diet without any changes of work character in the period of the last two examinations"

is in ST_0.

3 Mining spatiotemporal patterns

The data model used here is a generalization of that one introduced in [19]. *Spatiotemporal data* are supposed to be a sequence of events. An *event* has a unique identifier and, unlike in [19], is connected with an *explicit time instant*. If the data contains no explicit time attribute, the temporal coordinate can be substituted with an order of an event in the sequence. At least one attribute has to be spatial. It can be x- and y-coordinates (e.g. in windstorm data coordinates of a place) or identifier of an area (e.g. the name of a district). In contrary to [19] there is no limit on the number of events with the equal time stamp. We also allow attributes of complex type: not only atomic like in [19] but also of the type of list.

A *domain knowledge* is a set of predicate definitions. A *spatiotemporal pattern* (or shortly pattern) is a conjunction of non-spatiotemporal atoms and at least one spatiotemporal atom. Negation is not allowed in a pattern. A non-spatiotemporal *atom* is either of the form `Attribute Operator Value`[1], or is defined by a predicate from domain knowledge that does not have a temporal attribute as its

[1] Operator is `'='` for categorical attributes and `'=', '=<', '<'` for numerical attribute.

argument. A *spatiotemporal atom* can be temporal – $\Diamond(X)$, $\bigcirc^n(X)$ or $\Box(X)$ – or one of the spatial atoms from *RCC-8*, e.g.. dc(X,Y) (X is disconnected from Y).

The problem of mining spatiotemporal maximal frequent patterns is finding, for a given M (M is usually called a *minimal support*), all frequent spatiotemporal patterns, i.e. those that cover at least M examples, and that cannot be further refined without decreasing support under M.

4 GRAPE

New refinement operator for mining spatiotemporal data has been implemented in GRAPE. GRAPE is an extension of the ILP system RAP. RAP [5, 4] is a system for mining of first-order maximal frequent patterns that employs different search strategies for mining long patterns. Frequent patterns learned with RAP has been successfully used as new features for knowledge discovery in mining medical data [4, 2] and in information extraction from biomedical text [3].

The downward refinement operator [16] in RAP consists of three operations:

1. Add a most general literal into the pattern;
2. Bind two distinct variables of the same type;
3. Split the range of a numeric variable.

We extended the refinement operator of RAP with two classes of operations, for descending in user-defined is-a hierarchy (not necessarily temporal or spatial) and for specialization of temporal formulas. These new specialization operations are explained in Section 5:

Minimal support in GRAPE can be defined as global (ignoring granularity level), separately for each level in a hierarchy (like in $SPADA$) or as a user defined predicate.

5 Specialization in spatiotemporal logic

Is-a hierarchy. We suppose that for each non-temporal attribute there is maximally one hierarchy defined by a predicate `is-a(Attribute,Node)`. Then the additional specialization operation in refinement operator ρ is

4. $\rho((P, is\_a(Attr, X_N), R)) = (P, is\_a(Attr, X_{N+1}), R)$ where X_N is a value in a level N in a hierarchy for the attribute $Attr$ and X_{N+1} is an ancestor of X_N in this hierarchy.

E.g. for a data with an attribute PLACE and a four-level hierarchy for PLACE with the usual semantics, CZECH-LANDS, REGION, DISTRICT and PLACE, the specialization of the pattern (Pref,is_a(X,'CZECH-LANDS'),Suff) is (Pref,is_a(X,'Southern-Moravia',Suff). Pref and Suff are the prefix and the suffix of the original pattern..

Temporal predicates

5. $\rho(P) = (P, \Diamond(X))$ where X is a new variable and there is no other temporal predicate in P with a free variable (i.e. unused in P).
6. $\rho((P, \Diamond(T), S)) = (P, \Box(T), S)$ if there is no term T_1 θ-equivalent (in terms of θ-subsumption) with T in the rest of the pattern.
7. $\rho((P, \Diamond(T), S)) = (P, \bigcirc^n(T), S)$ if there is no term T_1 θ-equivalent with T in the rest of the pattern.
8. $\rho(P) = (P, \Diamond(T_1))$ if P contains $\Box(T)$, where T, T_1 are terms and T_1 is a proper specialization of T and T_1 does not appear elsewhere in the pattern.
9. $\rho(\star(X)) = \star(\rho(X))$ for $\star \in \{\Diamond, \bigcirc^k, \Box\}$

■

The second part of this paper concerns of experimental proof of usefulness of ST_0 logic and of the proposed refinement operator. We performed experiments with several datasets that cover different spatiotemporal domains. These datasets are described in Section 6 and the discovered patterns are discussed in Section 7. We also used frequent patterns as new features in a classification task. In Section 8 we bring results.

6 Data

Windstorms data [9] contains 4551 instances about strong winds in Czech lands since 16th century. Each example consists of attributes TIME, TYPE, DAMAGE, CHARACTERISTICS, X, Y, and PLACE. TIME is equal to the date of this event and can be of various granularity – year, a period of a year, months, days or a day. Three attributes describe a kind of a particular wind: TYPE describes the character of the wind, DAMAGE says whether there were no, moderate or serious damage caused by this strong wind and CHARACTERISTICS brings information about kind of damage. Attributes X, Y contains X and Y coordinates of the place, PLACE contains name of the village, town or region affected by this wind.

Keystroke dynamics data is a set of keystroke sequences, together 14483 records. Six persons (described with AGE, SEX, LEVEL of writing), have written repeatedly three different texts. Each keystroke record consists of TIMESTAMP (the moment of the event), TYPE (release or press) and CODE of the key pressed or released. For each key there are coordinates (its layout on the keyboard) and also its membership into spatial hierarchy on keys – FINGER-TO-WRITE (left thumb, right thumb, left forefinger, right forefinger etc.), and HAND-TO-WRITE (left, right).

STULONG data[2] consists of 10,572 records of long-term observations of 1,226 patients with atherosclerosis. There are 332 patients (each with a sequence of observations) who came down with a disease. The average number of observations before the disease occurs is 7.60. 136 of these patients belong to the RGI (*Risk Group Intervened*) group while 26 belong to RGC (*Risk Group Control*). In the data, there are 296 patients who never came down with any disease in the RGI and 89 in the RGC group.

7 Discovered patterns

We focused on searching for emerging patterns. For classified data a pattern is *emerging* if coverage on different classes differs significantly. Here we say that a pattern is emerging if difference between maximal and minimal coverage is greater or equal to 60%.

Windstorms. To classify data we took the DAMAGE attribute with values `NO, MODERATE` and `SERIOUS`. The pattern

key(K), year(K,X), $1650 \leq$ X ≤ 1716, type(K,4) , char(K,l)

("in the period 1650–1716 there were the windstorms with blow down") that has the support 45 (9.9%) corresponds with results from [6]. Coverage for the class SERIOUS is 43, for the class MODERATE 2, and for the class NO 0. The rule that has the support 19

$$key(K), year(K, C), 1971 \leq C \wedge C \leq 1972 \wedge \Diamond(type(B, 3) \wedge char(B, j))$$

says that "there was a wind in period 1971–72 and sometimes after that wind a strong wind of type TYPE = 'snow storm' and with type of damage CHARACTERISTICS = other".

Keystroke dynamics The LEVEL of a user - `NON-EXPERIENCED, ADVANCED` - has been used as the class attribute and the following patterns has been found.

$$\Diamond(\bigcirc(press(x_1) \wedge i(x_1,' v') \wedge \Diamond(press(x_2) \wedge I(x_2,' backspace')$$

"it is always true that, as the second event, the key 'v' was pressed and always in future the key 'bspace' was pressed". This patterns is frequent for non-experienced users.

In the next pattern the predicate d(Key1,Key2,Delay) computes the delay between press of two keys Key1 and Key2.

$$\Diamond(\bigcirc(P(x_1) \wedge I(x_1,' v') \wedge \Diamond(P(x_2) \wedge I(x_2,' h') \wedge \\ \bigcirc(P(x_3) \wedge I(x_3,' a') \wedge d(x_2, x_3, z) \wedge 162 \leq z \wedge z \leq 191))))$$

[2] The study (STULONG) was realized at the 2^{nd} Department of Medicine, 1^{st} Faculty of Medicine of Charles University and Charles University Hospital The data resource is on the web pages `http://euromise.vse.cz/challenge2004`.

"Always the second key was 'v' and after there were always the sequence of 'h' and 'a' and the delay between pressing these two keys was in the interval [162,191]". In opposite to the previous pattern, this one is frequent for advanced users.

$$\Diamond(p(x_1) \wedge r(x_1, r) \wedge \bigcirc(p(x_2) \wedge r(x_2, l) \wedge \bigcirc(p(x_3) \wedge r(x_3, l) \wedge \\ \Diamond(p(x_4) \wedge r(x_4, r) \wedge d(x_3, x_4, z) \wedge 818 \leq z \wedge z \leq 1071))))$$

"sometimes there is a key pressed with right hand followed by sequence left-hand → left-hand → right-hand, and the delay between that last two keystrokes is in the interval [818,1071] milliseconds".

STULONG The goal was to find out relationship of observed men to groups RGI (*Risk Group Intervened*) and RGC (*Risk Group Control*) with regarding their health and changes in their behavior. Each record consists of eight input attributes refer-ed to changes of the character of occupation, of physical activity, diet, etc.

An example of the difference between RGI and RGC

$$\Diamond(no\_change(phys\_act) \wedge \bigcirc^5 change(diet, 7))$$

"a patient did not change his physical activity in the period of five examinations before he changed his diet to the value 7[3]". Among patients who never came down with any disease this pattern covers 9 patients, 8 (89%) of them belong to the RGI group and only one to RGC. On the other hand, for patients who suffer from any disease, the pattern holds for 4 (80%) from RGC and only 1 (20%) to RGI. The pattern

$$\Diamond(no\_change(work\_char)) \wedge \Diamond(change(phys\_act, 30))$$

says that 73% (8 of 11) of patients who came down with some disease, who left their work to the partial retirement (the value 30) and who did not change their diet belong to the RGC group. On the other side, there is only 48% (10 of 21) patients from RGC who never came down with any disease and who satisfy this condition. Besides many patterns with a small support we have found the rule

$$\Diamond(change(diet, 7))$$

which says that 29% (2 of 7) of patients belonging to the group RGC who changed their diet to value "he take medicines for reduction of serum cholesterol" never came down with any disease. For the group RGI it is true for 71% (15 of 21) of men.

[3] 7 means that he takes medicines for reduction of serum cholesterol.

8 Frequent patterns as new features

For windstorm data, we also exploit the frequent patterns as new features and used them for learning classifiers. We combined also the original data with these new features and compared the results in terms of accuracy. We employed three classifiers from Weka package [21], Naive Bayes, decision tree classifier J48 and support vector machines SMO. 10-cross validation has been used so that for each fold the frequent patterns has been generated and then used as new or additional features.

Table 1. Windstorm: Frequent patterns as new features for classification

	Naive Bayes	J4.8	SMO
original	58.8	80.2	79.7
only frequent	81.3	81.1	84.1
orig+frequent	73.1	82.1	84.1
orig+max	64.3	76.2	84.1

As can bee seen in Table 8, adding this kind of new features - either frequent or maximal - always results in an accuracy increase. The highest accuracy has been reached when all frequent patterns was used as new attributes and has not depend on the learning algorithm used. The accuracy always overcomes 80%: SMO 84.1%, J48 81.1% Naive Bayes 81.3%. Adding original attributes has not affected accuracy significantly and an increase of accuracy has been observed only for J48.

9 Future work and conclusion

In future work we want to answer the following questions: Is the refinement complete for ST_0 logic? What are refinement operators for ST_1 and ST_2 logics? reasonable addition of spatial predicates

This work describes the initial experiments with a refinement operator for spatiotemporal data. In a long term perspective we aim at building universal framework based on ILP and basic domain knowledge that together can be exploited for building a task-oriented systems for mining in various spatiotemporal data.

This approach can be also useful for mining in text, e.g. for morphologically tagged data. For a word W and all words in the left and the right contexts we have set of all possible tags. The time dimension is given by the order of a word in a context (e.g. the left neighbor of W has the time stamp -1). Spatial attributes can describe particular grammatical categories – case, gender, number etc. Then we can e.g. explore dependency between cases in a context

like $numbers(W1, C1), \bigcirc(W2), numbers(W2, C2), po(X, Y)$ ("for two adjacent words W1, W2 the regions of possible values of grammatical number overlap"). This is an emerging patterns for verb phrases in inflectional languages like Czech [20].

Acknowledgment This work has been supported by the Grant Agency of the Czech Republic under the Grant No. MSM0021622418. We thank to Petr Dobrovolný for permission to use windstorm data, to Jan Dušek for keystroke dynamics data. Thanks also to Karel Bařina and Radim Štampach for experiments with SPADA.

References

1. B. Bennett, A.G. Cohn, F. Wolter, and M. Zakharyaschev. Multi-dimensional modal logic as a framework for spatio-temporal reasoning. *Applied Intelligence*, 17(3):239–251, 2002.
2. J. Blaťák. Mining first-order frequent patterns in the STULONG database. In *Proceedings of the ECML/PKDD 2004 Challenge*.
3. J. Blaťák and Popelínský. Learning genic interactions without expert domain knowledge: Comparison of different ilp algorithms. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*.
4. J. Blaťák and L. Popelínský. Feature construction with RAP. In T. Horváth and A. Yamamota, editors, *Proceedings of the Work-in-Progress Track at the 13th International Conference on ILP*, pages 1–11. University of Szeged, 2003.
5. J. Blaťák and L. Popelínský. Mining maximal frequent patterns in first-order logic. *Neuroworld*, 5(4):381–390, 2004.
6. R. Brázdil, P. Dobrovolný, J. Štekl, O. Kotyza, H. Valášek, and J. Jež. *History of weather and climate in the Czech lands VI: Strong winds*. Masaryk University in Brno, 2004.
7. A.G. Cohn, B. Bennett, J.M. Gooday, and N. Gotts. RCC: a calculus for region based qualitative spatial reasoning. *GeoInformatica*, 1:275–316, 1997.
8. L. Dehaspe and H. Toivonen. Frequent query discovery: a unifying ILP approach to association rule mining. Technical Report CW 258, Katholieke Univesiteit Leuven, Departmen of Computer Science, Celestijnenlaan 200A – B-3001 Heverlee (Belgium), March 1998.
9. P. Dobrovolný and R. Brázdil. Documentary evidence on strong winds related to convective storms in the czech republic since ad 1500. In J. T Snow, editor, *Atmospheric Research*, pages 95–116. ELSEVIER, Jul - Sep 2003.
10. S. Džeroski and N. Lavrač, editors. *Relational Data Mining*. Springer-Verlag, Berlin, September 2001.
11. A. Fern, R. Givan, and J.M. Siskind. Specific-to-general learning for temporal events. In *Eighteenth national conference on Artificial intelligence*, pages 152–158, Menlo Park, CA, USA, 2002. American Association for Artificial Intelligence.
12. J. Han, K. Koperski, and N. Stefanovic. Geominer: A system prototype for spatial data mining. In *Proc. 1997 ACM-SIGMOD Int'l Conf. on Management of Data(SIGMOD'97)*, Tucson, Arizona, 1997.

13. D. Malerba and F.A. Lisi. Discovering associations between spatial objects: An ilp application. In *ILP '01: Proceedings of the 11th International Conference on Inductive Logic Programming*, volume 2157 of *Lecture Notes in Computer Science*, pages 156–16, London, UK, 2001. Springer-Verlag.
14. D. Malerba and F.A. Lisi. An ILP method for spatial association rule mining. In *Working notes of the First Workshop on Multi-Relational Data Mining*, pages 18–29. Albert Ludwigs Universitaet Freiburg, 2001.
15. S. Moyle and S. Muggleton. Learning programs in the event calculus. In *ILP 1997*, pages 205–212.
16. S. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19/20:629–679, 1994.
17. L. Popelínský. Knowledge discovery in spatial data by means of ilp. In *Zytkow J.M., Quafafou M.(Eds.): Proc. of 2nd European Symposium PKDD'98*, volume 1510 of *Lecture Notes in Computer Science*, Nantes, France, 1998. Springer-Verlag.
18. J.J. Rodríguez, C.J. Alonso, and H. Boström. Learning first order logic time series classifiers. In *Proceedings of the Work-in-Progress Track at the 10th International Conference on Inductive Logic Programming*, pages 260–275, 2000.
19. I. Tsoukatos and D. Gunopulos. Efficient mining of spatiotemporal patterns. In *SSTD '01: Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases*, pages 425–442, London, UK, 2001. Springer-Verlag.
20. E. Žáčková, M. Nepil, and L. Popelínský. Automatic tagging of compound verb groups in Czech corpora. In *Text, Speech and Dialogue: Proceedings of TSD'2000 Workshop, LNAI*. Springer-Verlag, 2000.
21. I.H. Witten and E. Frank. *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufman, 1999.