2004－ICS－138　(10)
2004／12／4

Temporal Logic in Mining Hepatitis Data

Tu Bao Ho, Si Quang Le, Canh Hao Nguyen, Saori Kawasaki

Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa, 923-1292 Japan

Hideto Yokoi, Katsuhiko Takabayashi

Chiba University Hospital
Inohana, Chuo-ku, Chiba, 260-8677 Japan

Abstract. In this paper we propose a method that uses Allen's temporal logic in mining hepatitis data. We first describe different temporal events for short-term changed tests using the base state and peaks and temporal patterns for long-term changed tests using the changes of states. We then develop algorithms to detect temporal relations between patterns found in terms of temporal logic such as "Event A happened before event B and B happened during event C". The early results shown the feasibility of the approach and they are worth to be pursued in hepatitis study.

1. Introduction

Viral hepatitis is a disease in which tissue of the liver is inflamed by the infection of hepatitis viruses. As viral hepatitis has a potential risk to liver cirrhosis and hepatocellular carcinoma (HCC) – which is the most common type of liver cancer and the fifth most common cancer and the exact cause of HCC is still unknown – studies on viral hepatitis, specially on hepatitis type B and type C, are crutial in medicine.

Recently, a precious source for hepatitis study has been given by Chiba university hospital to the research community [10]: the hepatitis temporal database collected during 1982-2001 containing results of 771 patients on 983 laboratory tests. It is a large temporal relational database consisting of six tables of which the biggest has 1.6 million records. However, collected during a long period with progress in test equipments, the database is un-cleansed and contains inconsistent measurements, many missing values, and a large number of non unified notations. Among targets in analyzing hepatitis data, the following received much attention from the community: (P1) Find differences in temporal patterns between hepatitis B and C? (P2) Evaluate whether laboratory tests can be used to estimate the stage of liver fibrosis (F0, F1, …, F4)? (P3) In which stage of viral hepatitis the interferon therapy can be effective?

Among various approaches to mining the hepatitis database [10], ours is essentially based on *temporal abstraction* (TA). TA methods aim to derive an abstract description of temporal data by extracting their most relevant features over periods of time [7], [10]. The fundamental problem here is how to transform a sequence of time-stamped data of each patient on each test into an abstracted statement such as "ZTT first increases to the high region then changes to the normal region and remains stable". In previous work [4], [5], [8], within the framework combining temporal abstraction with data mining, we developed appropriate TA techniques for irregular temporal data in hepatitis study and obtained encouraging results. Continuing the TA research direction, this paper presents a new trial to TA with temporal relations introduced in temporal logic [1], [2], [3].

Temporal logic was developed as a theory of action and time by Allen whose basis is relations between temporal events. Recently, there have been some works on finding association rules based on temporal relations, e.g., [6], [9]. Unlikely these works, we developed temporal relations techniques appropriately for hepatitis study. First, we specify different temporal events for short-term and long-term changed tests in hepatitis data, and develop algorithms to detect temporal relations between events found in terms of temporal logic. Second, we apply data

mining methods to abstracted hepatitis data and obtained considerable preliminary results.

2. Problems and the framework

Assume that for each object O_k, the observed values v_i on attributes A_j at absolute time t_i is a sequence S_{jk} of time-stamped data

$$S_{jk} = (v_1, t_1), (v_2, t_2), \ldots, (v_n, t_n)$$

We denote by (E, T) an event E that occurs in a time interval $T = (t_s, t_e)$ where $t_s, t_e \in \{t_1, t_2, \ldots, t_n\}$. By event E we mean any trend or property of interest such as "ALB decreases from normal to low state", "GOT has many peaks in very high state". In the context of temporal data, we can assume to consider only events happening in some period of time, and can implicitly write event E instead of (E, T). In temporal logic, Allen summarized 13 kinds of temporal relations between two events A and B as shown in Figure 1 [1], [2]. When linking such temporal relations we can have compound statements such as "Event A happened before event B and B happened during event C".

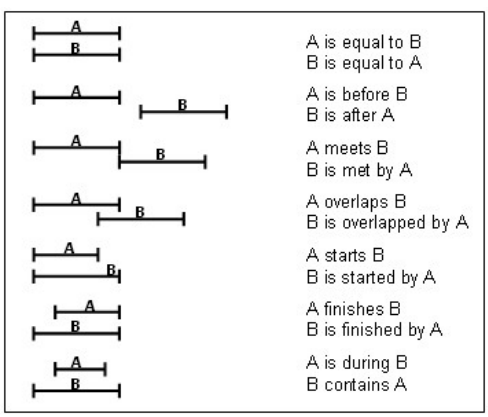

Figure 1. Temporal relations

The problem of temporal relations-based abstraction for mining hepatitis data can be formulated as follows: *Finding significant temporal relations for abstraction in hepatitis data to solve problems P1-P3.*

In this direction we propose the following framework for finding temporal relations as shown in Figure 2.

We started by a separation of two groups of 16 typical tests, one with values that can change in short terms and the other with values that can change in long terms when hepatitis B or C occur.

(1) *Tests with values that can change in short terms*: GOT, GPT, TTT, and ZTT. The tests in this group, in particular GOT and GPT, can rapidly change (within several days or weeks) their values to high or even very high values when liver cells were destroyed by inflammation.

(2) *Tests with values that can change in long terms*: The tests in the second group can slowly change (within months or years). Liver has the reserve capacity so that some products of liver (T-CHO, CHE, ALB, and TP) do not have low values until reserve capacity is exhaustive (the terminal state of chronic hepatitis, i.e., liver cirrhosis). Two main tendencies of change of tests in this group are:
   - Going down: T-CHO, CHE, ALB, TP, PLT, WBC, and HGB.
   - Going up: D-BIL, I-BIL, T-BIL, and ICG-15.

1. For each object O_k, from the data sequence S_{jk} on each attribute A_j, find all possible significant abstracted temporal events E on corresponding temporal intervals T, i.e., (E, T).
2. Consider all temporal events found from all attributes for each object O_k and detect all significant temporal relations between those events in terms of temporal logic.
3. Represent each object O_k as a graph or a transaction of temporal relations found. The collection of graphs or transactions is viewed as a abstracted data.
4. Using data mining methods to find temporal patterns from the abstracted data.

Figure 2. Framework of mining temporal relations

Temporal events

Based on visual analysis of various sequences by a created tool with MATLAB, we determined the following temporal abstraction primitives from which temporal events will be defined:

1. *State primitives*: N (normal), L (low), VL (very low), XL (extreme low), H (high), VH (very high), XH (extreme high).
2. *Peak primitives*: P (peaks occurred).

The thresholds to distinguish the state primitives of tests are given by physicians, for example, those to distinguish values N, H, VH, XH of TP are 5.5, 6.5, 8.2, 9.2 where (5.5, 6.5) is the normal region.

Currently, we limit our interests in considering two kinds of events:

(1) *Inflammation* characterized by the sudden occurrence of peaks in periods of high or very high state. Such events happen clearly in short-term changed tests and have the form:

E = <base state> & peaks

where <base state> can take any value of state primitives.

(2) *Change of states* between three state regions. Such events are typically for long-term changed tests and have the form:

E = <state> ">" <state>

where <state> can take any value from "N" (normal), "H" (high), and "L" (low), and ">" stands for "change the state to". For example, "ALB: N>L" means "ALB changes state from Normal to Low".

3. Abstracting data by temporal relations

The data abstraction process is carried out by two following algorithms according to the steps 1 and 2 described in our framework (Figure 2).

Algorithm 1. *Find significant abstracted temporal events*

Input: A sequence S_{jk} of a test data from a test A_j

Output: All abstracted temporal events from the sequence.

1. A_j is a short-term changed test

- Move from the left to the right of the sequence.
- A data point X is peak if $v(X) > v(Y) +$ threshold where Y is any neighbor of X.
- Find a peak X from the sequence. If there are no other peaks to the left of X within distance Δ_1 then the starting boundary of the interval is $t_s = X - \Delta_1$.
- Find other peaks Y to the right of X. If there is no other peak to the right of a peak Y farer than a distance Δ_1 then take the end boundary of the interval as $t_e = Y + \Delta_1$.
- Calculate the base state BS (without considering peaks) of the interval (t_s, t_e), and form the abstracted temporal event "BS&P" in this interval.

2. A_j is a long-term changed test

- Move from the left to the right of the sequence.
- Detect a set of relative consecutive pairs of points of different statuses and with Δ_2 points with the same status to the left of the starting point of the first pair t_s and to the right of the right point t_e of the last pair.
- Form the abstracted interval with the change of state from the above set of pairs in the interval (t_s, t_e).

Algorithm 2. *Find a transaction or a graph of temporal relations*

Input: The set of all associated events to one object O_k.

Output: A transaction or graph of temporal relations.

1. To build a transaction
   - Initialize the transaction as an empty set.
   - Check all pairs of events for each temporal relation type. If a pair matches the relation, add this relation to the transaction.
2. To build a graph
   - Build the transaction of relations as in the previous step.
   - Build the graph by adding each existing temporal relation to the graph when considering the events a vertices and relations as edges.

The step 2 in our framework aims to build a graph or a transaction of possible temporal relations from each object O_k starting from all of its events found. A basic algorithm to do this task was originally given in [1] using constraint propagation technique (the transitive property of temporal events). In this work on hepatitis data, due to the specific

features of the data, we develop an appropriate technique based on:

- *Soft matching*: at the boundaries of intervals for relations "equal ", "meet", "start", "finish", and "overlap". The boundary points of two events are considered as one if their absolute difference is smaller than a given threshold, or considered as different in "overlap" relation if their absolute difference is greater than a given threshold.
- "*Slightly*" is a key constraint for the "before" relation, i.e., we consider only relations of the form "A slightly before B" viewed by some threshold.

Noting that the constraint propagation in [1] causes a great number of induced relations usually when applied to the relation "before" to, and the set of events associated to each object (patient) has size up to several hundreds, we propose an exhaustive and direct examination of all such events against the relations in order to find possible temporal relations.

4. Results on hepatitis data

According to the medical background knowledge, we focus on exploiting the 15 most frequent tests. Unlike our previous work on temporal abstraction [3], [5] that requires fixing episodes on which abstracted patterns are generated, this work on temporal relations considers dynamically intervals on which events are detected.

In this paper we report the preliminary results of applying abstraction techniques with temporal relations to problem P1-P3. For problem P1 we consider events on the whole data sequence; for P2 we consider events on a acceptable neighborhood of the day doing biopsy and before the days having interferon therapy. For problem P3 episodes are backwardly taken from last day before the treatment with interferon. We have to separate the patients into four groups by response to interferon (IFN) therapy based on the domain knowledge of doctors:

(1) *Response*: GPT data turned into the normal region within 6 months after IFN therapy finished, and keep this level for more than 6 months.

(2) *Partial response*: GPT data turned into twice as high as the normal region within 6 months after IFN therapy finished, and kept this level for more than 6 months.

(3) *Aggravation*: GPT data changed remarkably higher than the level before IFN therapy within 6 months after IFN therapy finished.

(4) *No change*: GPT data does not show any change.

We began with 197 patients who are treated with IFN. Among them, we removed one patient who has no GPT test data and six others who are with many missing values. By using one set of parameters, we came to a final dataset with 190 instances with a distribution as follows {response: 121, partial-response: 35, aggravation: 5, no-response: 29}. In this current version of the paper we reported only the results for P1.

4.1 Finding temporal relation associations

We used tool CBA (http://www.comp.nus.edu.sg/~dm2/) to find associations from the transactional database created by algorithms described in section 3. Some potentially interesting result from description rules for each class: B and C, here the rule format is

rule (Cover% Conf% CoverCount SupCount Sup%)

We use the notation of *B*, *O*, *E*, … for relations "Before", "Overlap", "Equal", etc.

Class B rules

Rule 8: che:N>H*O*t-cho:H>N
(1.751% 100.000% 10 10 1.751%)

Rule 16: che:N>H*O*t-bil:H>N
(1.401% 100.000% 8 8 1.401%)

Rule 29: che:H>N*O*d-bil:N>H AND gpt:VH*E*got:H
(1.226% 100.000% 7 7 1.226%)

Rule 30: alb:L>N*O*tp:N>L AND gpt:VH*E*got:H
(1.226% 100.000% 7 7 1.226%)

Rule 35: ztt:N*E*ttt:H
(1.051% 100.000% 6 6 1.051%)

Rule 45: t-bil:N>H*O*che:H>N
(1.051% 100.000% 6 6 1.051%)

Class C rules

Rule 2: ttt:XH*B*ztt:VH
(2.277% 100.000% 13 13 2.277%)

Rule 3: alb:N>L*O*ztt:VH
(2.277% 100.000% 13 13 2.277%)

Rule 5: alb:N>L*B*ztt:H
(2.102% 100.000% 12 12 2.102%)

Rule 10: ttt:XH*O*alb:N>L
(1.751% 100.000% 10 10 1.751%)

Rule 11: alb:N>L*O*tp:H>N
(1.751% 100.000% 10 10 1.751%)

Rule 14: ttt:H*O*ztt:VH
(1.576% 100.000% 9 9 1.576%)

Rule 18: i-bil:N>H*B*ztt:VH
(1.401% 100.000% 8 8 1.401%)

Rule 19: ztt:VH*B*ttt:H
(1.401% 100.000% 8 8 1.401%)

Rule 22: d-bil:N>H*O*tp:N>L
(1.226% 100.000% 7 7 1.226%)

Rule 26: ttt:XH*O*ztt:VH
(1.226% 100.000% 7 7 1.226%)

Some interpretations:

- Class B: “che increases (N>H) overlaps with t-cho or t-bil decreases (H>N)”
- Class B: “che decreases (H>N) around the time of t-bil or d-bil increases”
- “alb:L>N*O*tp:N>L” occurs in both classes, but when “gpt:VH end got:H”, it is B.
- ztt and ttt have peaks at almost the same time (R #35, #2, #14, #19, #26). In class C, both ztt and ttt tend to have much higher base states than class C
- Class C: alb decreases (N>L) followed by some events.

Comments from physicians:

We can find a highly frequency of bilirubins (t-bil, d-bil, i-bil) in the class B rules (3 cases / 6 rules), compared with class C rules (2 cases / 10 rules). The movement of bilirubins seem to be correlated with t-cho (total cholesterol), che (cholinesterase). All the movements of the substances mean hepatitis aggravation and recovery. Bilirubins are considered to seldom move in most part of the natural course of hepatitis. These rules suggest there is a difference between natural course of hepatitis B and C.

4.2 Finding temporal relation rules for prediction

We used the program See5 to find prediction rules from the created database. Taking classifier number 1, most rules are to describe class B. The classifier either gives rule for class B, and default rule for C and vice verse. The reason is that if there rule for both classes, there is a branch of “if A=y then else (A=N) then”. This seems not meaningful and statistically significant.

Only a subset of rules predicting class B are selected.

Format of the rule is as below, where id is just a rule identification.

Rule id: condition -> class B [estimated accuracy]

Rule 1: che:N>H*O*t-cho:H>N -> class B [0.892]

Rule 9: che:N>H*O*t-cho:N>H -> class B [0.820]

Rule 24: che:L>N*O*ztt:H -> class B [0.763]

Rule 29: che:N>L*O*t-cho:H>N -> class B [0.654]

Rule 31: che:N>L*F*tp:N>L -> class B [0.654]

Rule 32: che:N>L*S*alb:N>L -> class B [0.654]

Rule 5: che:H>N*O*alb:H>N -> class B [0.853]

Rule 4: d-bil:H>N*O*ztt:N -> class B [0.853]

Rule 11: alb:H>N*S*d-bil:H>N -> class B [0.808]

Rule 14: tp:L>N*S*d-bil:H>N -> class B [0.808]

Rule 16: d-bil:H>N*O*che:N>H -> class B [0.805]

Rule 12: i-bil:N>H*B*got:XH -> class B [0.808]

Rule 15: d-bil:N>H*B*t-cho:H>N -> class B [0.808]

Rule 19: i-bil:H>N*E*alb:L>N -> class B [0.776]

Rule 25: t-bil:N>H*O*alb:N>H -> class B [0.763]

Rule 33: ztt:H*E*ttt:XH -> class B [0.618]

Rule 2: ztt:N*F*ttt:N -> class B [0.882]

Rule 3: ztt:N*E*ttt:H -> class B [0.853]

Rule 21: ttt:XH*F*ztt:H -> class B [0.773]

Rule 23 ttt:N*S*ztt:N -> class B [0.773]

Some interpretations:

- “che decreases” overlaping “t-cho or tp or alb decreases”
- “d-bil decreases around tp, che increases”
- ztt and ttt also tend to be smaller in class B.

Comments from physicians:

The temporal relation gave us the rules' position on time sequence. We note that bilirubins move in advance of other attributes on the occasion of aggravation (rule 12, 15). The rules are useful when we consider fine changes of the blood test data of hepatitis. But there are some rules which seem to be contradictory to each other (rule 1, 9). We need to improve the tool so that more exact information can be checked.

5. Discussion and Conclusion

We have presented a temporal relation approach to mining the time-series hepatitis data, and obtained some preliminary results that are interesting.

In our opinion, the main advantage of temporal abstraction techniques is their generalization and summarization power for the *description task* from temporal data. However, TA techniques may not be appropriate in several *prediction task* because of the abstraction process may discard many details that are necessary in prediction. It is natural to think that TA techniques, when combining appropriately with numerical conditions or domain knowledge represented in other formalisms can be well applied to the prediction task. Our future work consists of the continuation of making temporal relations feasible and useful in mining temporal data, in particular hepatitis data, and the integration of data mining methods with text mining and expert knowledge.

Acknowledgments

This research is supported by the project “Realization of Active Mining in the Era of Information Flood”, Grant-in-aid for scientific research on priority areas (B), and project “Discovery of Hepatitis Knowledge by Data Mining Methods with Multi-Sources”.

References

1. Allen, J. (1983) “Maintaining Knowledge About Temporal Intervals”, *Communications of the ACM* 26(11), 832-843.
2. Allen, J. (1984) “Towards a General Theory of Action and Time”, *Artificial Intelligence* 23(2), 123-15.
3. Allen, J. (1991) “Time and Time Again: The Many Ways to Represent Time”, *Int. J. Intelligent Systems*, 6(4), 1-14.
4. Ho, T.B., Nguyen, T.D., Kawasaki, S., Le, S.Q., Nguyen, D.D., Yokoi, H., Takabayashi, K. (2003) “Mining Hepatitis Data with Temporal Abstraction”, *ACM International Conference on Knowledge Discovery and Data Mining KDD-03*, Washington DC, 24-27 August, 369-377.
5. Ho, T.B., Nguyen, T.D., Kawasaki, S., Le, S.Q. (2004) “Combining Temporal Abstraction and Data Mining Methods in Medical Data Mining, *Chapter 7, Intelligent Knowledge-Based Systems, Vol. 3*, T. Leondes (Ed.), Kluwer Academic Press, 198-222.
6. Hoppner, F. (2001) “Discovery of Temporal Patterns”, *IJCAI Workshop on Learning from Temporal and Spatial Data*, Seattle, 25-31.
7. Horn, W., Miksch, S., Egghart, G., Popow, C., Paky, F. (1997) “Effective Data Validation of High-Frequency Data: Time-Point-, Time-Interval-, and Trend-Based Methods”, *Computer in Biology and Medicine, Special Issue: Time-Oriented Systems in Medicine*, 27(5), 389-409.
8. Kawasaki, S., Nguyen, T.D., Ho, T.B. (2003) “Temporal Abstraction for Long-Term Changed Tests in the Hepatitis Domain”, *Journal of Advanced Computational Intelligence & Intelligent Informatics*, Vol. 17, No. 3, 348-354.
9. Lee, J.W., Lee, Y.J., Kim, H.K., Hwang, B.H., Ryu, K.H. (2002) “Discovering Temporal Relation Rules from Interval Data”, *First EurAsian Conference on Advances in Information and Communication Technology*, 57-66, Springer.
10. http://lisp.vse.cz/challenge/ecmlpkdd2004/
11. Shahar, Y. (1997) “A Framework for Knowledge-based Temporal Abstraction”, *Artificial Intelligence,* 90, 79-133.