New Generation Computing, 25(2007)247-262
Ohmsha, Ltd. and Springer

Exploiting Temporal Relations in Mining Hepatitis Data

Tu-Bao HO, Canh-Hao NGUYEN, Saori KAWASAKI,
Si-Quang LE and
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, 923-1292 JAPAN
{bao, canhhao, skawasa, quang}@jaist.ac.jp
Katsuhiko TAKABAYASHI
Chiba University Hospital
Inohana, Chuo-ku, Chiba, 260-8677 JAPAN
takaba@ho.chiba-u.ac.jp

Received 20 January 2006
Revised manuscript received 27 March 2007

Abstract Various data mining methods have been developed last few years for hepatitis study using a large temporal and relational database given to the research community. In this work we introduce a novel temporal abstraction method to this study by detecting and exploiting temporal patterns and relations between events in viral hepatitis such as "event A slightly happened before event B and B simultaneously ended with event C". We developed algorithms to first detect significant temporal patterns in temporal sequences and then to identify temporal relations between these temporal patterns. Many findings by data mining methods applied to transactions/graphs of temporal relations shown to be significant by physician evaluation and matching with published in Medline.

Keywords: Temporal Patterns, Temporal Relations, 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 exact cause of HCC is still unknown – studies on viral hepatitis, specially on hepatitis type B and type C, are crucial in medicine.

Recently, a precious source for hepatitis study has been given by Chiba university hospital to the data mining community.[12] The hepatitis temporal database collected during twenty years (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. 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. In last few years, six problems P1-P6 posed by physicians in hepatitis study using the above database have attracted different research groups, for example.[8, 18, 21]

It is worth noting that methods for processing medical temporal data essentially aim to detect temporal patterns in temporal sequences,[6] and they can be viewed in two categories: methods for categorical time series with focus on discovering frequently occurring episodes in a sequence,[15, 16] and methods for numerical time series with focus on trend detection.[3, 4, 8, 11] Techniques in each category can be either supervised or unsupervised.

Temporal abstraction (TA) is an approach to temporal pattern detection that aims to derive an abstract description of temporal data by extracting their most relevant features over periods of time.[3, 6] Typical TA works in the literature deal with regular temporal data, says, temporal data of an individual measured on consecutive days in a short period,[4] diabetes data measured on consecutive days within two weeks; newborn infants regularly measured every minute.[11] Different from the regular data processed by the above mentioned TA methods, the hepatitis data was collected irregularly in long periods, and none of the above methods can be applied to.

We approach the hepatitis database by novel temporal abstraction methods aiming at *explaining the causes and mechanisms* of hepatitis diseases in a *comprehensible way* to physicians. Our early work [9] developed a supervised TA technique called *abstraction pattern extraction* (APE) whose task is to map (to abstract) a given fixed length sequence into one of predefined abstraction patterns. In this work we develop a unsupervised TA technique called *temporal relation extraction* (TRE) whose task is to find temporal relations in terms of temporal logic among detected temporal patterns, and use these relations together with abstraction patterns to solve problems P1-P2. Temporal logic was developed as a theory of action and time by Allen whose basis is relations between temporal events.[1, 2] The key idea that makes our TRE work efficient and different from other methods of temporal pattern detection is the domain-oriented temporal patterns are defined basing on properties of hepatitis disease but not in a formal manner.

This work contributes to methods of detecting temporal patterns from irregular temporal sequences and temporal relations among detected patterns, and more interestingly, various findings have reconfirmed reported medical knowledge and some are surprising to physicians. Section 2 of the paper presents the data, problem, and the framework. Section 3 describes the methods. Section 4 provides the obtained results and analysis. The last section gives discussion and

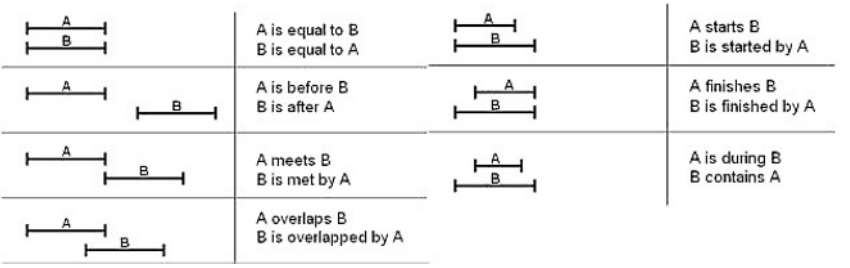


Fig. 1 Temporal Relations in Allen's Temporal Logic

conclusions.

§2 Hepatitis Data and Temporal Basic Patterns

The hepatitis database consists of the following data tables:

T1. Basic information of patients (771 records)
T2. Results of biopsy (960 records)
T3. Information on interferon therapy (198 records)
T4. Information about measurements in in-hospital tests (459 records)
T5. Results of out-hospital medical tests (30,243 records)
T6. Results of in-hospital medical tests (1,565,877 records).

Our focus in this work is on problems P1-P2 among six problems posed by physicians to challenge the KDD community [12]:

P1. Discover the differences in temporal patterns between hepatitis B and C (HBV and HCV).
P2. Evaluate whether laboratory tests can be used to estimate the stage of liver fibrosis (LC (liver cirrhosis) vs. nonLC (non liver cirrhosis)).

For each patient O_k the measured values e_i on a medical test A_j over time are an event sequence $S_{jk} = (e_1, t_1), (e_2, t_2), ..., (e_n, t_n)$. In case of the hepatitis data, sequences S_{jk} can be long as observed during twenty years. The starting point of our work is the view on *temporal patterns*. It is worth noting that the temporal pattern is a rather broad concept and defined differently in temporal data mining, in particular the view on temporal patterns in terms of 13 kinds of temporal relations (Fig. 1) between two events A and B summarized by Allen in the temporal logic.[1,2] In [10] a temporal pattern is considered as a set of states together with their interval relationships described in the Allen's interval logic.[1,2] Kam and Fu [14] also use Allen interval operators to formulate patterns but restricted to the form with concatenation on the right hand side $(((\cdot\, A_1 rel_1 A_2) rel_2 A_3) \cdot el_{k-1} A_k)$.

In this work we consider a temporal pattern as a conjunction/relation of temporal basic patterns (hereafter called basic patterns). In the following we will define two kinds of basic patterns for a sequence of test values that are sensitive

to the type of tests. In the hepatitis study, we selected 24 typical tests from 983 tests based on the opinion of physicians and the preprocessing/analysis results of different research groups.[17] These tests are divided into two types:

1. *Short-term changed tests*: These include four tests GOT, GPT, TTT, and ZTT that characterize liver inflammation and their values can highly increase in short terms (within several days or weeks) when liver cells are destroyed by inflammation.
2. *Long-term changed tests*: The other twenty tests characterize the liver reserve capacity and change smoothly their values in long terms (within months or years) when their reserve capacity becomes exhausted. Among these tests, there are two subgroups with clear trends:
   - Going down: T-CHO, CHE, ALB, TP, PLT, WBC, and HGB.
   - Going up: D-BIL, I-BIL, T-BIL, and ICG-15.

The temporal abstraction process is based on temporal abstraction primitives viewed as abstraction units. In fact, each test value belongs to either the normal region or an anormal region, viewed as event state, such as high and low regions (which can be further divided into smaller regions) [*1]. Each subsequence of a given event sequence now can be abstracted using event states, concretely, assigning to it the label of the region where the majority of its events belong to.

We determine differently basic patterns for short-term and long-term changed tests using their detected abstraction states.

[1] Basic patterns in short-term changed test sequences

The abstraction states of short-term changed tests include N (normal region), H (high), VH(very high), XH (extreme high), L (low), VL (very low), and XL (extreme low). We call a *peak* the event that has its value suddenly much higher than that of its neighbors.

We define the *temporal basic patterns (BP)* of a short-term changed test the subsequence characterizing a inflammation period where the sequence suddenly has the high or very high state and with/without peaks. These basic patterns have the form:

$$< state\ of\ test > = high\_value \quad \text{or}$$
$$< state\ of\ test > = high\_value\ \&\ peaks$$

where $< state\ of\ test >$ denotes the abstraction state of the test sequence and the test name, and $high\_value$ is one value in $\{H, VH, XH\}$. For example, "$GOT = XH\ \&\ peak$" means "GOT is in extremely high state with peaks".

[2] Basic patterns in long-term changed test sequences

The abstraction states of short-term changed tests include N (normal), H (high), L(low).

[*1] The thresholds to distinguish the values regions of tests are given by physicians, for example, those to distinguish N, H, VH, XH of the short-term changed test TP (total protein) are 5.5, 6.5, 8.2, 9.2, respectively, where (5.5, 6.5) is the normal region.

1. For each object O_k, from the event sequence Sj_k on each attribute A_j, find all possible significant abstracted temporal basic patterns BP on corresponding temporal intervals T.
2. Consider all temporal basic patterns found from all attributes for each object O_k and detect all significant temporal relations between those temporal basic patterns in terms of temporal logic. Represent each object O_k as a graph or a transaction of temporal relations.
3. Using data mining methods to find temporal rules from the collection of graphs or transactions.

Fig. 2 Framework of Mining Hepatitis Data by Temporal Relation Extraction (TRE)

We define the *temporal basic patterns (BP)* of a long-term changed test the subsequence characterizing the change of states between two state regions. These basic patterns have the form:

$$< state\ of\ test > = \ state_1 \ > \ state_2$$

where $state_1$ and $state_2$ are two different values in {N, H, L} and ">" stands for "change the state to". For example, "$ALB = N > H$", or more informally "$ALB = NormalToHigh$" means "ALB changes from normal to high state".

Denote by (BP, T) a temporal basic pattern BP that occurs in a time interval $T = (t_s, t_e)$ where $(t_s, t_e) = t_1, t_2, ..., t_n$. Examples of temporal basic patterns are "ALB decreases from normal to low state", "GOT has many peaks in very high state". In the context of temporal data, we consider only temporal patterns happening in some period of time, and can implicitly write patterns BP instead of (BP, T). As defined above, temporal patterns viewed as temporal relations between temporal basic patterns are compound statements such as "Pattern A happened before pattern B and B happened during pattern C" or the rule such as "If pattern A happened before pattern B and B happened during pattern C then hepatitis type B".

The problem of temporal abstraction using temporal relations in mining hepatitis data can be viewed as finding significant temporal patterns in hepatitis data to solve problems P1-P2, shown in Fig. 2.

§3 Finding Temporal Patterns

This section describes solutions to the problem of finding temporal basic patterns (step 1) and complex temporal patterns in form of temporal relations (step 2) in the framework. The key issue in these steps is that it is hard to determine exactly interval boundaries T in which temporal basic patterns BT occur while determining temporal relations between temporal basic patterns requires comparing their boundaries.

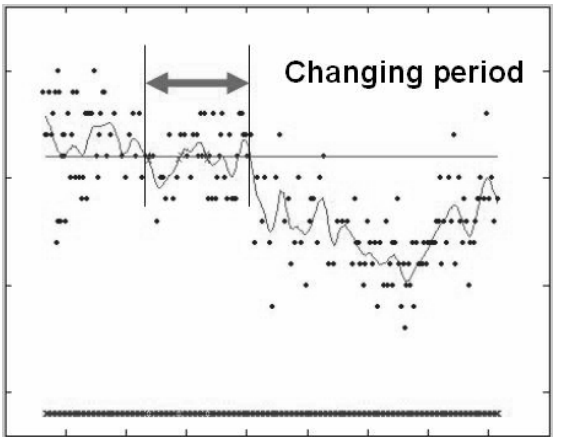


Fig. 3 Original, Smoothing Data and State Changing Period.

3.1 Smoothing Data

As hepatitis data was collected during a long period with progress in test equipments, the database is un-cleansed, besides various preprocessing works,[9, 17] in this work, we first do smoothing event sequences and use smoothed data instead. We employed a moving average filter to smooth data by replacing each data point with the average of the neighboring data points defined within the span. This process is equivalent to low-pass filtering with the response of the smoothing given by the difference equations. Given a data of one patient's test $(e_1, t_1), ..., (e_n, t_n)$ where e_i is the test result at time t_i. Then, the smoothed value e for at time t is predicted as

$$e = \frac{\sum_i F(t_i - t) \times e_i}{\sum_i F(t_i - t)}$$

where F(.) is an influent function. In our experiment, we choose $F(x) = 1/|x|$.

3.2 Finding Basic Patterns

After smoothing data, we detect periods of state changing for both short-term changed tests and long-term changed tests based on the following criteria:

- The first point and last point belong to different states.
- States of the first point and last point are stable for at least 6 months.
- Intervals between consecutive crossing pairs must less than parameter θ_1 or intervals between two crossing pairs are less than θ_3 and there are at least $MinPoint$ crossing pairs between them.
- The interval between two consecutive crossing pairs must be less than θ_2

By the statistics and visualization of the data, together with discussion with physicians, we choose $\theta_1 = 12 \times 4$ weeks, $\theta_2 = 3 \times 12 \times 4$ weeks and $\theta_3 = 5 \times 12 \times 4$ weeks. Figure 3 illustrates an original events sequence, its

Algorithm 1. Detecting basic patterns in short-term changed test sequences
Input: A sequence Sj_k of a test data from a short-term changed test A_j
Output: Basic patterns characterizing the inflammation in the short-term changed test.

1. Call a data point (e_i, t_i) a peak if $e_i > e_j + threshold$ where (e_j, t_j) is any neighbor of (e_i, t_i).
2. Find the most left peak (e_i, t_i) from the sequence. Set the $CurrentPeak = (e_i, t_i)$, the starting and ending boundaries of the period are $t_s = t_i - 1$ and $t_e = t_i + 1$.
3. Find the closest peak on the right of $CurrentPeak$.
4. If $(t_j < t_e)$ then set $t_e = t_j + 1, CurrentPeak = (e_j, t_j)$ and return to step 3.
5. If $(t_j \geq t_e)$ then

   a. Calculate the base state BS (without considering peaks) of the interval (t_s, t_e),
   b. Form the abstracted temporal event "$BS\&P''$ in this interval,
   c. Set a new period with the starting and ending boundaries: $t_s = t_i - 1$ and $t_e = t_j + 1$. Set $CurrentPeak = (e_j, t_j)$ and Return to step 3.

Fig. 4 Algorithm for Finding Temporal Basic Patterns in a Short-term Changed Test Sequences

Algorithm 2. Detecting basic patterns in long-term changed test sequences
Input: A sequence Sj_k of a test data from a long-term changed test A_j
Output: Basic patterns characterizing the state change periods in the long-term changed test.

1. Detect crossing:
   If $state(f(t)) \neq state(f(t+1))$ then t is a crossing point.
2. Merging crossing points:

   - If $length(crossing\ point\ i,\ crossing\ point\ i+1) \leq \theta_1$ then merging i and $i+1$.
   - If $length(crossing\ point\ i,\ crossing\ point\ i+1) > \theta_2$ then separating i and $i+1$.
   - If $length(crossing\ point\ i,\ crossing\ point\ i+1) < \theta_3$ and $j - i > n$ then merging i and j.

3. Interval detecting: For each crossing point (an interval of merged crossing points), if it is stable for 6 months before and after, then this crossing point (the interval) is a change period.

Fig. 5 Algorithm for Detecting Basic Patterns in a Long-term Changed Test Sequences

smoothed sequence and state changing period. Algorithm 1 in Fig. 4 is for detecting basic patterns for short-term changed tests and Algorithm 2 in Fig. 5 is for long-term changed tests.

3.3 Finding Temporal Relations

The step 2 in our framework aims to build a graph or a transaction of

Algorithm 3. 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 as vertices and relations as edges.

Fig. 6 Algorithm for Finding a Transaction/Graph of Temporal Relations.

possible temporal relations from each object (patient) O_k starting from all of its detected events. 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 the same (time) 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 relations of such events in order to find all possible temporal relations.

§4 Mining Abstracted Data and Evaluation

In this section, we describe experiments and their results on the hepatitis data for the two problems of P1 and P2. Even though the primary purpose is to find the causes and explaining the mechanism of diseases, we carried out two studies: prediction study to see if the extracted data can be good for (even black box) classification; and description study to find comprehensible rules for the primary purpose.

4.1 Prediction Study

We would like to *evaluate the quality of extracted data* to see whether the proposed framework and preprocessing steps are appropriate. We studied whether the extracted data contain enough useful information for the two problems P1 and P2. We used WEKA [*2] as the experiment environment. For each problem:

1. Run the algorithms in Section 3 to generate a transaction of temporal patterns for each patient.
2. Converted the transactions with class label into table format.
3. Run feature selection techniques.
4. Run classification methods in WEKA.

In step 1, the algorithms in Section 3 returned 610 (out of 771) patients (372 HCV and 238 HBV) that more than one temporal relation was detected for P1. After converting into table format incorporating class labels, it resulted in a table with 1888 features. Similarly for problem P2, only 108 patients (71 non-LC and 37 LC) were returned with totally 403 features. Due to the large number of features, in step 3, we run the feature selection techniques. The feature selection techniques that gave the highest classification accuracy are Correlation Based Feature subset selection for P1 and Information Gain Filtering for P2. Step 3 resulted in 62 features for P1 and 20 features for P2. Filtering the patients without any event after feature selection, P1 data now contains 498 patients and P2 data contains 69 patients. From our observation, using feature selection improved significantly prediction accuracies. We run various classification methods and the best results were summarized as follows:

- For problem P1, Naive Bayes classifier gave an accuracy of 77.56% with 10 times 10-fold stratified cross-validation.
- For problem P2, Naive Bayes classifier gave an accuracy of 78.70% with leave-one-out cross-validation.

For the same problem of P1, a completely different approach [18] also reported a comparable accuracy of 77.60%. However, the key difference in our work is that we are able to extract data for 576 patients in comparison with 193 patients in theirs. This means that our approach give similarly reliable information from a much larger number of patients. For problem P2, our accuracy was lower than that reported in [21] (88.2%).

4.2 Description Study

As it is crucial that physicians need to evaluate results of hepatitis data mining, the main target of this work is to describe the diseases in a comprehensible form. Therefore, we used rule learning algorithms to generate rules from the extracted data. Our work follows four steps:

1. Created a transactional database for each hepatitis problem by

[*2] http://www.cs.waikato.ac.nz/ml/weka/

proposed algorithms described in Section 3.
2. Used software CBA [*3], our LUPC [*4] and See5 to find rules from the transactional database with default parameters.
3. Filtered statistically significant rules by hypothesis testing.
4. Analyzed the findings with/by physicians.

The key difference from prediction study is that the set of description rules was not meant to cover the whole data set. Instead, each rule itself should be of high precision or high coverage. Another difference is that for interpretable reason, we eliminated the rules containing any condition like "*if NOT ALB changes from normal to low etc then ...*". Such a condition is considered not to make any medical sense. For the above reasons, the set of description rules might not perform well on the training data in terms of accuracy. However, the rule set should be able to explain some part of the data in a comprehensible way.

[1] Rules for hepatitis types: HBV and HCV

Using CBA, we were able to generate a set of 238 rules, in which 20 rules for HBV and 218 rules for HCV. The overall accuracy of the prediction rule sets on the training data is 89.34%. Contingency table of the rule set on training data is as follows.

Predicted		HBV	HCV
Correct	HBV	208	30
	HCV	35	337

Table 1 shows the set of typical rules for describing HBV and HCV. The first column is the rule identification number generated by the classifier. Next, "Class" is the predicted class of the rule. "Cov." means the number of patients the rule covers, and "Conf." is the confidence of the rule. The "Rule Conditions" is a conjunction of temporal patterns playing the role of condition for the rule. We can observe the component test items in the temporal events exhibit different temporal patterns for each of HBV and HCV as follows:

Observation 1: Even when there are temporal relations between GOT and GPT, even both GOT and GPT have peaks in High region, the rules in which ALP oscillate between Normal and Low are for HBV while the ones in which ALP oscillate between High and Normal are for HCV.

Some rules support this observation are the numbers: 145 (ALP changes from Low to Normal etc., class HBV), 206 (ALP changes from Normal to Low etc., class HBV), 20 (ALP changes from High to Normal, class HCV) and 202 (ALP changes from Normal to High etc., class HCV).

Observation 2: Among patients who have peaks on both GPT and TTT in High regions, T-BIL decreases from High to Normal in HBV patients, while T-BIL decreases Normal to Low in HCV patients. Some rules support

[*3] http://www.comp.nus.edu.sg/∼dm2

[*4] http://www.jaist.ac.jp/ks/labs/ho/Projects.htm

Table 1 Some Typical Rules for HBV and HCV

RID	Class	Cov.	Conf.	Rule Conditions
145	B	3	100.0%	ALP=LowToNormal & GOT=Normal
206	B	20	80.0%	ALP=NormalToLow & GOT=High *Ends* GPT=High
20	C	13	100.0%	ALP=HighToNormal & GOT=High *Starts* GPT=High
202	C	56	82.1%	ALP=NormalToHigh *Before* GOT=High & GPT=High *Before* GOT=High
196	B	12	83.3%	T-BIL=HighToNormal & GPT=High *Ends* TTT=High
185	C	7	85.7%	T-BIL=NormalToHigh & GPT=Normal
167	C	25	92.0%	T-BIL=NormalToLow *Before* TTT=High & GPT=High *Before* TTT=High
25	C	12	100.0%	T-BIL=NormalToLow *Before* TTT=High & T-BIL=NormalToLow *Before* GPT=High
203	C	28	82.1%	T-BIL=NormalToLow *Before* GPT=High & TTT=High
188	B	13	84.6%	GPT=High *Ends* TTT=High & GPT=High *Ends* ZTT=High
217	C	139	77.0%	GPT=High *Before* TTT=High & TTT=High *Before* ZTT=High
176	C	10	90.0%	GPT=Normal & TTT=High *Starts* ZTT=High
151	B	3	100.0%	TP=NormalToLow *Before* ZTT=High & TTT=High *Starts* ZTT=High
8	C	18	100.0%	TP=NormalToHigh & TTT=High *Before* ZTT=High
2	C	23	100.0%	TP=HighToNormal & TTT=High *Before* ZTT=High
219	B	56	75.0%	TTT=Normal & ZTT=Normal
227	B	34	70.6%	TTT=Normal & CHE=HighToNormal
226	C	78	70.5%	TTT=High *Before* GOT=High & GPT=High Start GOT=High
193	C	49	83.67.3%	TTT=High *Before* ZTT=High & F-A1.GL=NormalToLow

this observation are the numbers: 196 (class HBV), 185 (class HCV), 203 (class HCV), 167 (class HCV) and 25 (class HCV).

Observation 3: Patients who have temporal relations of peaks in both TTT and ZTT have different state change on TP. In case of HCV , TP moves from High to Normal, meanwhile it changes from Normal to Low for HBV. Some rules support this observation are the numbers: 151 (class HBV), 8 (class HCV) and 2 (class HCV).

Matching with Medline abstracts: We looked for some reported results from medical researches to find evidences for and against our findings. We developed a simple search program integrating both keywords and synonyms in the query.

Murawaki et al.[19] showed that the main difference between HBV and HCV is that the base state of TTT in HBV is normal, while that of HCV is high. We examined the rule sets and found that our rules are more complicated than that as they also include various temporal relations. However, there are many rules of very high coverage and high confidence, TTT appeared to be mostly in High state for HCV but in Normal state for HBV. We showed some

rules support this finding in the table with numbers: 219, 227 226 and 193. This means that even though our rules are not exactly identical to reported knowledge of medical research, such knowledge is confirmed true in our rule set under certain condition.

[2] Rules for liver cirrhosis: LC and non-LC

Using CBA, we were able to generate a set of 61 rules, in which 21 rules for LC and 40 rules for non-LC. The overall accuracy of the prediction rule sets on the training data is 96.30%. Contingency table of the rule set on training data is as follows.

Predicted		LC	non-LC
Correct	LC	37	0
	non-LC	4	67

Some rules in the set can be seen from the Table 2. Notions in the table are identical to that in Table 1. From the rule sets, we observed the following phenomena:

Observation 1: There are more rules for non-LC patients and most of them are of higher precision and coverage. This conforms to the common knowledge of experts that LC is harder to detect.

Observation 2: There were some long term changed test items that appeared mostly in LC patients. They are I-BIL and ALB. The following rules for LC patients support this observation:

- Rule 15: I-BIL changes from normal to low (coverage: 5 patients, precision: 100%).
- Rule 26: I-BIL changes from high to normal and ALB changes from low to normal (coverage: 4 patients, precision: 100%).
- Rule 27: ALB changes from low to normal and TTT has peaks in normal state (coverage: 4 patients, precision 100%).

From this, we may induce that I-BIL and ALP change their states mostly in LC patients, not in non-LC ones. They can be good indicators for predicting liver cirrhosis patients.

Observation 3: There were some long term changed test items that appeared mostly in non-LC patients. They are LDH, CRE, T-BIL and ALP. The following rules for non-Lc patients support this observation:

- Rule 1: CRE changes from normal to low and ZTT has peaks in normal (coverage: 10 patients, precision 100%).
- Rule 2: T-BIL and LDH change from normal to Llow, GOT and TTT have peaks in high (coverage: 10 patients, precision: 100%).
- Rule 5: T-BIL changes from normal to low, ALP changes from normal to high, GOT and TTT have peaks in high (coverage: 8 patients, precision: 100%).
- Rule 8: ALP changes from normal to high Before TTT has peaks in high and ZTT has peaks in high (coverage: 7 patients, precision: 100%).

Table 2 Some typical rules for (non-) liver cirrhosis

RID	Class	Cov.	Conf.	Rule Conditions
58	NonLC	12	91.7%	GOT=High *Ends* GPT=High & TTT=Normal
1	NonLC	10	100.0%	CRE=NormalToLow & TTT=Normal
2	NonLC	10	100.0%	T-BIL=NormalToLow & LDH=NormalToLow & GOT=High & TTT=High
3	NonLC	10	100.0%	T-BIL=NormalToLow & ZTT=High & LDH=NormalToLow & TTT=High
5	NonLC	8	100.0%	T-BIL=NormalToLow & ALP=NormalToHigh & GOT=High & TTT=High
9	NonLC	7	100.0%	ZTT=High *Before* GPT=High & ALP=NormalToHigh & TTT=High *Before* GPT=High
8	NonLC	7	100.0%	ALP=NormalToHigh *Before* TTT=High & ZTT=High
13	NonLC	6	100.0%	ZTT=High & T-BIL=HighToNormal & GOT=High & TTT=High
11	NonLC	6	100.0%	GPT=High *Before* ZTT=High & TTT=Normal & TTT=High
12	NonLC	6	100.0%	ZTT=High & LDH=NormalToLow & T-BIL=HighToNormal
17	NonLC	5	100.0%	CRE=NormalToLow & ALP=HighToNormal
14	NonLC	5	100.0%	GPT=High *Before* ALP=NormalToHigh
15	LC	5	100.0%	I-BIL=NormalToHigh
26	LC	4	100.0%	I-BIL=HighToNormal & ALB=LowToNormal
27	LC	4	100.0%	TTT=Normal & ALB=LowToNormal
37	LC	3	100.0%	ALB=NormalToLow & LDH=NormalToLow
38	LC	3	100.0%	T-BIL=LowToNormal & ALP=NormalToHigh & TTT=High *Before* GPT=High

- Rule 12: LDH changes from normal to low, T-BIL changes from high to normal and ZTT has peaks in high (coverage: 6 patients, precision: 100%).
- Rule 61: LDH changes from normal to low, ZTT and TTT have peaks in high (coverage: 36 patients, precision 100%).

From this, we may induce that LDH, CRE, T-BIL and ALB change their states mostly in non-LC patients, not in LC ones. They can be good indicators for predicting non-liver cirrhosis patients.

§5 Discussion and Conclusion

We have presented a temporal relation approach to mining the temporal hepatitis data. The early findings in our on-going project present some interesting temporal patterns to physicians. The main contribution of the work is temporal relations allows us to find a kind of temporal relations that well describe hepatitis. Some findings are either quantitatively reconfirmation of medical observations or providing insight, some time contrast, to the medical general knowledge. In short, the approach is well evaluated by hepatitis experts.

In our opinion, the main advantage of temporal abstraction techniques is their generalization and summarization power for the description task from

temporal data. Even though temporal abstraction techniques are not developed for prediction task as the abstraction process may discard many details, they still give encouraging prediction accuracies. It is natural to think that temporal abstraction techniques, when combining appropriately with numerical conditions or domain knowledge represented in other formalisms can be well applied to the prediction task.[13] 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.

Acknowledgements

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., "Maintaining Knowledge About Temporal Intervals," *Communications of the ACM, 26(11)*, pp. 832–843, 1983.
2) Allen, J., "Time and Time Again: The Many Ways to Represent Time," *Int. J. Intelligent Systems, 6(4)*, pp. 1–14, 1991.
3) Balaban, M., Boaz, D. and Shahar, Y., "Applying temporal abstraction in medical information systems," *Annals of mathematics, computing and teleinformatics 1(1)*, pp. 56-64, 2003.
4) Bellazzi, R., Larizza, C., Magni, P., Monntani, S. and Stefanelli, M., "Intelligent Analysis of Clinic Time Series: An Application in the Diabetes Mellitus Domain", *Intelligence in Medicine, 20*, pp. 37–57, 2000.
5) Bruzzese, D. and Davino, C., "Statistical Pruning of Discovered Association Rules", *Computational Statistics 16 (3)*, pp. 387–398, 2001.
6) Chittaro, L., Montanari, A., "Temporal representation and reasoning in artificial intelligence: Issues and Approaches", *Annals of Mathematics and Artificial Intelligence 28*, pp. 47–106, 2000.
7) Das, G., Lin, K.I., Mannila, H., Renganathan, G., Smyth, P., "Rule discovery from time series" *ACM Int. Conf. on Knowledge Discovery and Data Mining KDD'98*, pp. 16–22, 1998.
8) Hirano, S., Tsumoto, S., "Mining Similar Temporal Patterns in Long Time-series Data and Its Application to Medicine", *IEEE Int. Conf. on Data Mining ICDM'02*, pp. 219–226, 2002.
9) Ho, T.B., Nguyen, T.D., Kawasaki, S., Le, S.Q., Nguyen, D.D., Yokoi, H., Takabayashi, K. , "Mining Hepatitis Data with Temporal Abstraction," *ACM Int. Conf. on Knowledge Discovery and Data Mining KDD'03*, pp. 369–377, 2003.
10) Hoppner, F., "Learning dependencies in multivariate times series," *the ECAI'02 Workshop on Knowledge Discovery in (Spatio)-Temporal Data*, pp. 25–31, 2002.

11) Horn, W., Miksch, S., Egghart, G., Popow, C., Paky, F., "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)*, pp. 389–409, 1997.

12) http://lisp.vse.cz/challenge/ecmlpkdd2004/

13) Jonsson P., Backstrom, C., "A unifying approach to temporal constraint reasoning," *Artificial Intelligence 102*, pp. 143–155, 1998.

14) Kam, P.S., Fu, A.W.C., "Discovering Temporal Patterns for Interval-Based Events," *Second International Conference on Data Warehousing and Knowledge Discovery DaWaK'00, LNAI 1874*, Springer, pp. 317–326, 2000.

15) Liu, B., Hsu, W., Ma, Y., "Pruning and Summarizing the Discovered Associations," *ACM Int. Conf. on Knowledge Discovery and Data Mining KDD'99*, pp. 125–134, 1999.

16) Mannila H., Toivonen, H., Verkamo, A.I., "Discovery of Frequent Episodes in Event Sequences," *Data Mining and Knowledge Discovery*, pp. 259–289, 1997.

17) Motoda, H., *Active Mining: New directions of data mining* (Ed.), IOS Press, 2002.

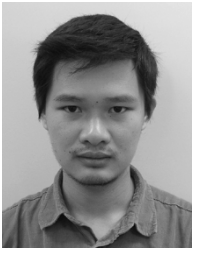
18) Ohara, K., Yoshida, T., Geamsakul, W., Motoda, H., Washio, T., Yokoi, H., Takabayashi, K., "Analysis of Hepatistis Dataset by Decision Tree Graph-Based Induction," *Discovery Challenge 2004*, (Berka, P. and Cremillieux, B. Eds.), ECML/PKDD'04, 173–184, 2004.

19) Murawaki Y., Ikuta Y., Koda M., Kawasaki H., "Comparison of clinical laboratory liver tests between asymptomatic HBV and HCV carriers with persistently normal amino-transferase serum levels," *Hepatol Research 21(1)*, pp. 67–75, 2001.

20) Sakai H., Horinouchi H., Masada Y., Takeoka S., Ikeda E., Takaori M., Kobayashi K., Tsuchida E., "Metabolism of hemoglobin-vesicles (artificial oxygen carriers) and their influence on organ functions in a rat model," *Biomaterials 25(18)*, pp. 4317–4325, 2004.

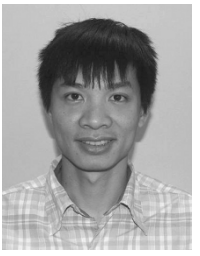
21) Yamada, Y., Suzuki, E., Yokoi, H., Takabayashi, K., "Experimental evaluation of time-series decision tree," *Twentieth Int. Conf. on Machine Learning ICML'03*, pp. 840–847, 2003.

Tu-Bao Ho: He is a professor at School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST), Japan. He received his M.S. and Ph.D. from Marie and Pierre Curie University in 1984 and 1987, respectively. His research interest include knowledge-based systems, machine learning, data mining, medical informatics and bioinformatics.



Canh-Hao Nguyen: He is a PhD student at School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST), Japan. He received his B.S. in Computer Science (with Honours) from the University of New South Wales in 2002, and M.S. from JAIST in 2005. His research interests lie in statistical machine learning, ranging from theoretical to application studies.

Saori Kawasaki: She is an assistant professor at School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST), Japan. She received her M.S. and Ph.D. from JAIST in 2000 and 2003, respectively. Her research interest includes data mining, knowledge evaluation and medical informatics.



Si-Quang Le: He is currently a postdoctoral scientist at LIRMM in Montpellier, France. He received his M.S. from College of Technology, Vietnam National University, Hanoi in 2003, and Ph.D. from Advanced Institute of Science and Technology (JAIST) in 2005.

Katsuhiko Takabayashi: He is MD, FACP and professor at Division for Medical Informatics and Management, Chiba University Hospital, 1-8-1 Inohana, Chuo-ku, Chiba, 260-8677, Japan.