Discovery of trends and states in irregular medical temporal data

Trong Dung Nguyen and Tu Bao Ho

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

Abstract. Discovery of trends and states is one of the main tasks in mining medical temporal data. Temporal abstraction has been known as a powerful approach of data abstraction by converting temporal data into interval with abstracted values including trends and states. Most temporal abstraction methods, however, has been developed for regular temporal data, and they cannot be used when temporal data are collected irregularly. In this paper we introduced a temporal abstraction approach to irregular temporal data inspired from a real-life application of a large database in hepatitis domain.

1 Introduction

Extensive amounts of data gathered in medical databases require specialized tools for data analysis and effective use of data. It can be seen that medical databases open many opportunities and challenges to the new interdisciplinary field of knowledge discovery and data mining (KDD).

Hepatocellular carcinoma (HCC) is the most common type of liver cancer and the fifth most common cancer in the world. The exact cause of HCC is unknown. Viruses such as hepatitis B and hepatitis C have been shown to increase the risk of HCC. The hepatitis temporal database was collected between 1982-2001 at the Chiba university hospital. This database, which is the biggest one in Japan in the hepatitis domain, consists of the results of 983 laboratory tests involving 771 patients. It is a large un-cleansed 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 equipment, the database also contains inconsistent measurements, many missing values, and a large number of non-unified notations [11]. The hepatitis database was given as discovery challenge in 2002, 2003 of PKDDs(http://www.cs.helsinki.fi/events/eclpkdd/challenge.html). The doctors posed six problems that are expected to be solved by data mining techniques:

1. Discover the differences in temporal patterns between hepatitis B and C.
2. Evaluate whether laboratory tests can be used to estimate the stage of liver fibrosis.
3. Evaluate whether interferon therapy is effective or not.
4. Discover the relations between the stage of liver fibrosis and the onset of hepatocarcinoma.

5. Discover the relations between hematological status and time to the onset of hepatocarcinoma.
6. Validate if GOT and GPT can be used to measure inflammation speed.

One of the main approaches to mining medical temporal data is temporal abstraction (TA). The key idea of TA methods is to transform time-stamp points by abstraction into an interval-based representation of data. The common tasks in temporal abstraction are detecting trends and states of some variables (medical tests) from temporal sequences. Typical works on temporal abstraction are those presented in [1], [2], [8], [10], [7], [4], [13], [14]. These works focus on processing short and regular temporal data (data observed at regular time-stamped points). However, there are many real-life applications where temporal data are irregularly collected. The problem we face with hepatitis data is to find trends and states of tests in long and irregular time-stamp sequences. Different from separately finding "states" and "trends" as done in related work, we introduce the notion of "changes of state" to simultaneously characterize trends and states in long-term changed tests and the notions of "base state" and "peaks" to characterize short-term changed tests, as well as algorithms to detect them. This approach, as well a large part of the obtained results, have been evaluated as new and interesting by medical doctors.

The paper starts by a a brief introduction to temporal abstraction research and the preprocessing of the hepatitis data in section 2. Section 3 presents our methods to abstract short-term and long-term changed tests. Section 4 presents discovered results based on abstracted data.

2 Temporal abstraction and related work

Temporal abstraction (TA) is one approach to deal with time-related data in medical research. The key idea is to transform time-stamp points by abstraction into an interval-based representation of data.

The TA task can be defined as follows: The *input* is a set of time-stamped data points (events) and abstraction goals; the *output* is a set of interval-based, context-specific unified values or patterns (usually qualitative) at a higher level of abstraction.

TA task can be generally considered in two phases: (1) *basic temporal abstraction* that concerns with abstracting time-stamped data within episodes (which are significant intervals for the investigation purpose). Basic temporal abstractions typically extract states (e.g., low, normal, high), and/or trends (e.g., increase, stable, decrease) from a uni-dimensional temporal sequences, and (2) *complex temporal abstraction* that concerns with temporal relationships between findings from a basic temporal abstraction or from other complex temporal abstractions.

Our work on temporal abstraction related to the work of several research groups, typically Sharhar et al. [13], [14]; Haimowitz et al. [4], Miksch et al. [10], [7]; Bellazzi et al. [8], [1], [2].

In [13], [14], the authors developed a knowledge-based framework for the creation of abstract, interval concepts from time-stamped clinical data. The principles underlying this framework are generality and reusability where the use of knowledge is emphasized. This proposal has been realized in the system RESUME. A significant novelty of this approach is the dynamic derivation of interpretation contexts. Interpretation contexts are induced by events, such as therapeutic actions. Abstractions are generated on the basis of interpretation contexts, thus the interpretation of the patient data is context sensitive.

The other works, unlike the approach in [13], [14], the aim is at a specific type of application. In [4], the authors focus on medical trend diagnosis. Generic trends are defined through the notion of a trend template that gives great power of expression. This is both the strength and the limitation of this approach(see [9]). In [10], [7] the authors developed methods for context-sensitive and expectation-guided TA of high-frequency data. The interpretation contexts are not dynamically derived, but they are defined through schemata with thresholds that can be dynamically tailored to the patient under examination. In [8], [1], [2], the authors focus on using and combining statistical and probability techniques in/with temporal abstraction.

The common points of the above works are their basic temporal abstraction methods Were developed for in short periods and/or regular time-stamp points. The work in [8], [1]related to temporal data of an individual measured on consecutive days in a short period; the work in [10] on insulinormal and dependent diabetes related to temporal data measured on consecutive days within two weeks; and the work in the field of artificial ventilation of newborn infants in [7] related to temporal data regularly measured every minute. Generally, doing detection of trends and characterization of states for such sequences is different (easier) from doing these tasks for irregular time-stamp sequences.

3 Data Preprocessing and Basic Temporal Abstraction

3.1 Preprocessing and mining the hepatitis data

Data cleaning involves elimination of noisy data. The main task is to remove non unified symbols or characters that were included in the data collection. For example, we removed characters such as "H" or "L" or other unexpected numeric values, because they are redundant and not suitable for further processing.

Table 1 shows a part of the integrated data table that contains about one thousand columns and fifty thousands rows. The numbers of tests for each patient are different, and for each test (column) different patients can have sequences of different lengths.

When we look at the complete integrated table of temporal data, we see, for example, that the patient MID 1 has undergone tests of GOT, GPT, ALB, etc. 189 times (sequences of length 189) during 1981-2001, while the patient MID 2 has undergone only 88 during 1991-2001. As mentioned, the cause of the most difficulty in processing is the tests were irregularly done. Some early investigation

Table 1. Part of integrated table of temporal data

MID	Data	Sex	IFN	GOT	GPT	ALB	...
1	19810218	M	n	55	65	5.4	...
1	19810316	M	n	51	87	5.2	...
1	19810513	M	n	47	64	4.8	...
...	...	...	...	...	...	...	...
1	20010108	M	y	68	100	5.5	...
1	20010210	M	y	57	93	5.1	...
2	19911021	F	n	54	82	4.5	...
2	19911118	F	n	77	114	4.4	...
...	...	...	...	...	...	...	...

on the histogram of the number of test items in sampling intervals show that most consecutive tests were done within the interval of 28 and 56 days. This observation is adopted as a basis for our further investigation.

We also carried out several data transformations. For example, tests such as CHE was taken before and after the mid-80s by different measurements (with normal regions of [6, 12] and [180, 430], respectively). We accordingly converted the old test values to the new ones obtained by the new measurements.

Another problem is feature selection. With the guidance of medical doctors and statistics on the frequencies of tests, we selected the 41 most significant tests from a total of 983. The dataset for investigating each problem will be selected from these tests plus some special tests recommended by the medical doctors. These tests can be divided into four groups:

1. The most frequent tests: GPT, GOT, LDH, ALP, TP, T-BIL, ALB, D-BIL, I-BIL, UA, UN, CRE, LAP, G-GTP, CHE, ZTT, TTT, T-CHO, oudan, nyuubi, youketsu.
2. The high frequent tests: NA, CL, K
3. The frequent tests: F-ALB, F-A2.GL, G.GL, F-A/G, F-B.GL, F-A1.G
4. The low frequent but significant tests: F-CHO, U-PH, U-GLU, U-RBC, U-PRO, U-BIL, U-SG, U-KET, TG, U-UBG, AMY, and CRP.

3.2 Methods for basic temporal abstraction

We started by separating test into two groups, one with values that can change in the short-term and the other with values that can change in the long-term when hepatitis B or C occur.

1. Tests with values that can change in the short term: 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 are destroyed by inflammation.

Table 2. The temporal abstraction primitives

$< pattern > ::= < state\ primitive >$

$< pattern > ::= < state\ primitive >< relation >< state\ primitive >$

$< pattern > ::= < state\ primitive >< relation >< peak >$

$< pattern > ::= < state\ primitive >< relation >< state\ primitive >< relation >$
$< state\ primitive >$

2. Tests with values that can change in the long term: The tests in the second group can slowly change (within months or years). Liver has a reserve capacity so that some products of liver (T-CHO, CHE, ALB, and TP) do not have low values until the 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:
   - Tests with a "going down" trend: T-CHO, CHE, ALB, TP, PLT, WBC, and HGB.
   - Test with "going up" trend: D-BIL, I-BIL, T-BIL, and ICG-15.

Temporal abstraction primitives Based on visual analysis of various sequences, we determined the following temporal abstraction primitives:

1. *State primitives*: N (normal), L (low), VL (very low), XL (extreme low), H (high), VH (very high), and XH (extreme high).
2. *Trend primitives*: S (stable), I (increasing), FI (fast increasing), D (decreasing), and FD (fast decreasing).
3. *Peak primitives*: P (peaks occurred).

We also determined the following *relations* between the primitives: > ("change state to"), & ("and"), – ("and then"), / ("majority/minority", X/Y means that the majority of points are in state X and the minority of points are in state Y).

Medical doctors give thresholds for distinguishing the state primitives of tests, 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. We define four *structures of abstraction patterns* as shown in Table 2.

Examples of abstracted patterns in a given episode are as follows: "ALB = N" (ALB is in the normal region), "CHE = H–I" (CHE is in the high region and then increasing), "GPT = XH&P" (GPT is extremely high and with peaks), "I-BIL = N>L>N" (I-BIL is in the normal region, then changed to the low region, and finally changed to the normal region).

Also, based on a careful investigation of various sequences from the hepatitis database, we found and defined possible patterns of sequences. Figure 1 shows 8 typical possible patterns for short-term changed tests (left) and 21 typical possible patterns for long-term changed tests (right). Several notations will be used to describe algorithms for detecting short-term and long-term changed tests.

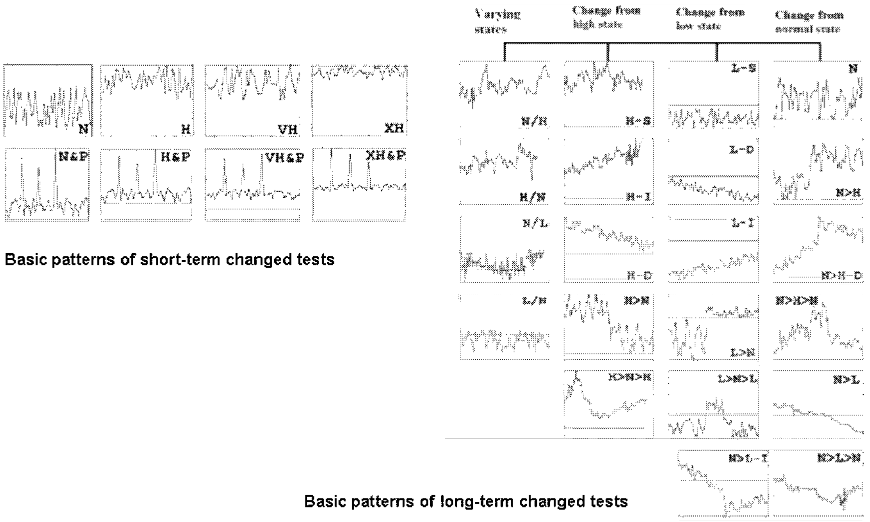


Fig. 1. Basic patterns of short-term and long-term changed tests

Abstraction of short-term changed tests Our observation and analysis showed that the short-term changed tests, especially GPT and GOT, can go up in some very short period of time and then go back to some "stable" state. We found that the two most representative characteristics of these tests are a "stable" state, called *base state* (BS), and the position and value of *peaks*, where the tests suddenly go up. Based on this observation, we have developed an algorithm to find the base state and peaks of a short-term changed test, as shown in Figure 2.

Abstraction of long-term changed tests Our key idea is to use the "change of state" as the main feature to characterize sequences of the long-term changed tests. The "change of state" contains information of both state and trend, and can compactly characterize the sequence.

At the beginning of a sequence, the first data points can be at one of the three states, "N", "H", or "L". Either the sequence changes from one state to another state, smoothly or variably (at boundaries), or it remains in its state without changing. Because changes can generally happen in the long-term, it is possible to consider the trend of a sequence after changing of the state. We have developed an algorithm to find the base state and peaks of a short-term changed test, as shown in Figure 3.

Notations used in temporal abstraction algorithms: High(S): # points of S in the high region; VeryHigh(S): # points of S in the very high region; ExtremeHigh(S): # points of S in the extreme high region; Low(S): # points of S in the low region; VeryLow(S): # points of S in the very low region; Normal(S): # points of S in the normal region; Total(S) = High(S) + VeryHigh(S) + ExtremeHigh(S) + Normal(S) + Low(S) + VeryLow(S); In(S) = Normal(S)/Total(S); Out(S) = (Total(S - In(S)) /Total(S); Cross(S): # times S crosses the upper and lower boundaries of the normal region; First (S): State of the first points in S; Last (S): State of the last points in S; State(S): State of S (one of the state primitives); Trend(S): Trend of S (one of trend primitives).

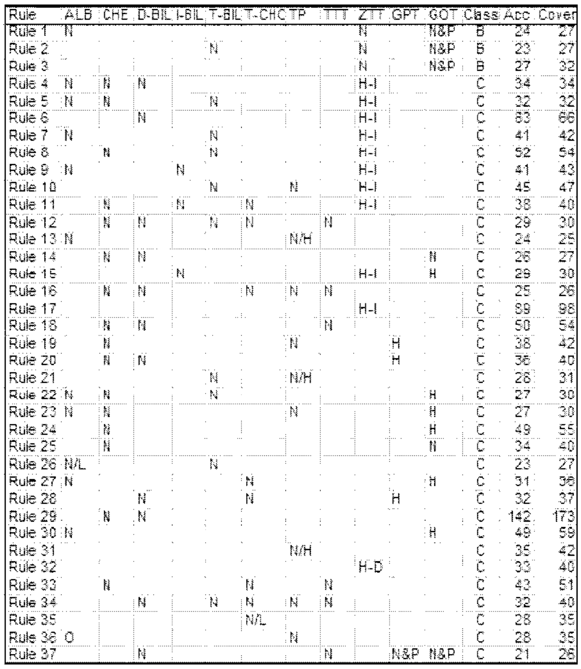
```
Input       Sequence of values of a test $S_{00} = \{s_1, s_2, ..., s_N\}$ in a given episode.
Result      A base state and peaks, a set of peaks PEi, and an abstracted pattern.
Parameters  NU, HU, VHU, XHU: upper thresholds of normal, high, very high,
            extreme high regions of a test, $\alpha$ (real).

A. Searching for base state

1. Based on NU, HU, VHU, and XHU, calculate the quantities Normal(S), High(S),
   VeryHigh(S), and ExtremeHigh(S)
2. Take MV = max {Normal(S), High(S), VeryHigh(S), Extreme-High(S)}.
   If MV/Total(S) > $\alpha$ then BS := MS.
3. Else BS := NULL

B. Searching for peaks

1. For every element $s_i$ of S, if $s_i > s_{i-1}$ and $s_i > s_{i+1}$ then $s_i$ is a local maximum of S.
2. For every element $M_i$ of the set of local maximum points, $P_j = M_i$ will be a peak,
   if one of the following conditions is true, where $V(x)$ is the value of $x$:

    (1) $BS = N \text{ and } V(M_i) > V(VH)$
    (2) $BS = H \text{ and } V(M_i) > V(XH)$
    (3) $BS = VH \text{ and } V(M_i) > 2 * V(XHU)$
    (4) $BS = XH \text{ and } V(M_i) > 4 * V(XHU)$

C. Output the basic temporal abstraction pattern

1. If BS = N there is no peak, then N
2. If BS = N there is at least a peak, then N&P
3. If BS = H there is no peak, then H
4. If BS = H there is at least a peak, then H&P
5. If BS = VH there is no peak, then VH
6. If BS = VH there is at least a peak, then VH&P
7. If BS = XH there is no peak, then XH
8. If BS = XH there is at least a peak, then XH&P
9. If BS = NULL then Undetermined.
```

Fig. 2. TA algorithm for short-term changed tests

```
Input       Sequence of values of a test $S_{00} = \{s_1, s_2, ..., s_N\}$ in a given episode.
Result      An abstracted pattern derived from the sequence.
Parameters  $\alpha$, $\delta$, $\sigma$, $\epsilon$ (integer), $\beta$ (real).
Notation: $S_{10} = [s_1, median], S_{20} = [median, s_N], S_{11} = [s_1, 1st\ quartile]$,
    $S_{12} = [1st\ quartile, median], S_{21} = [median, 3rd\ quartile], S_{12} = [3rd\ quartile, s_N]$

A. Identification of patterns with many crosses
1. If Cross($S_{00}$) $> \alpha$ wedge In($S_{00}$) $>$ Out($S_{00}$) wedge High($S_{00}$) $>$ Low($S_{00}$) then N/H
2. If Cross($S_{00}$) $> \alpha$ wedge In($S_{00}$) $>$ Out($S_{00}$) wedge High($S_{00}$) ¡ Low($S_{00}$) then N/L
3. If Cross($S_{00}$) $> \alpha$ wedge In($S_{00}$) $<$ Out($S_{00}$) wedge High($S_{00}$) $>$ Low($S_{00}$) then H/N
4. If Cross($S_{00}$) $> \alpha$ wedge In($S_{00}$) $<$ Out($S_{00}$) wedge High($S_{00}$) ¡ Low($S_{00}$) then L/N

B. Identification of patterns without changes of state
5. If In($S_{00}$) $> \beta$ then N
6. If Out($S_{00}$) $> \beta \wedge$ State($S_{00}$) = H $\wedge$ Trend($S_{00}$) = S then H–S
7. If Out($S_{00}$) $> \beta \wedge$ State($S_{00}$) = H $\wedge$ Trend($S_{00}$) = I then H–I
8. If Out($S_{00}$) $> \beta \wedge$ State($S_{00}$) = H $\wedge$ Trend($S_{00}$) = D $\wedge$ Last($S_{22}$) = H then H-D
9. If Out($S_{00}$) $> \beta \wedge$ State($S_{00}$) = L $\wedge$ Trend($S_{00}$) = S then L–S
10. If Out($S_{00}$) $> \beta \wedge$ State($S_{00}$) = L $\wedge$ Trend($S_{00}$) = D then L–D
11. If Out($S_{00}$) $> \beta \wedge$ State($S_{00}$) = L $\wedge$ Trend($S_{00}$) = I $\wedge$ Last($S_{22}$) = L then L-I

C. Identification of patterns with changes from the normal region
12. If First ($S_{00}$) = N $\wedge$ Cross($S_{00}$) $< \alpha \wedge$ Last ($S_{22}$) = H $\wedge$ Trend($S_{22}$) = I
    $\wedge$ Low($S_{00}$) $< \epsilon$ then N>H
13. If First ($S_{00}$) = N $\wedge$ Cross($S_{00}$) $< \alpha \wedge$ Last ($S_{22}$) = H $\wedge$ Trend($S_{22}$) = D
    $\wedge$ Low($S_{00}$) $< \epsilon$ then N>H–D
14. If First ($S_{00}$) = N $\wedge$ Cross($S_{00}$) $< \alpha$ High($S_{00}$) $> \delta \wedge$ Last ($S_{22}$) = N
    $\wedge$ Trend($S_{22}$) = D $\wedge$ Low($S_{00}$) $< \epsilon$ then N>H>N
15. If First ($S_{00}$) = N $\wedge$ Cross($S_{00}$) $< \alpha \wedge$ Last ($S_{22}$) = L $\wedge$ Trend($S_{22}$) = D
    $\wedge$ High($S_{00}$) $< \epsilon$ then N>L
16. If First ($S_{00}$) = N $\wedge$ Cross($S_{00}$) $< \alpha \wedge$ Last ($S_{22}$) = L $\wedge$ Trend($S_{22}$) = I
    $\wedge$ High($S_{00}$) $< \epsilon$ then N>L–I
17. If First ($S_{00}$) = N $\wedge$ Cross($S_{00}$) $< \alpha$ Low($S_{00}$) $> \delta \wedge$ Last ($S_{22}$) = N
    $\wedge$ Trend($S_{22}$) = I $\wedge$ High($S_{00}$) $< \epsilon$ then N>L>N

D. Identification of patterns with changes from the high region
18. If First ($S_{00}$) = H $\wedge$ Cross($S_{00}$) $< \alpha$ Last ($S_{22}$) = N $\wedge$ Low($S_{00}$) $< \epsilon$ then H>N
19. If First ($S_{00}$) = H $\wedge$ Cross($S_{00}$) $< \alpha$ Normal($S_{00}$) $> \delta$ Last ($S_{22}$) = H
    $\wedge$ Trend($S_{22}$) = I $\wedge$ Low($S_{00}$) $< \epsilon$ then H>N>H

E. Identification of patterns with changes from the low region

20. If First ($S_{00}$) = L $\wedge$ Cross($S_{00}$) $< \alpha$ Last ($S_{22}$) = N $\wedge$ Low($S_{00}$) $< \alpha$ then L>N
21. If First ($S_{00}$) = L $\wedge$ Cross($S_{00}$) $< \alpha$ Normal($S_{00}$) $> \delta \wedge$ Last ($S_{22}$) = L
    $\wedge$ Trend($S_{22}$) = D $\wedge$ High($S_{00}$) $< \epsilon$ then L>N>L
22. If NULL Then Undetermined.
```

Fig. 3. TA algorithm for long-term changed tests

Table 3. Discovered rules describing hepatitis B and C



Rule	ALB	CHE	D-BIL	I-BIL	T-BIL	T-CHC	TP	TTT	ZTT	GPT	GOT	Class	Acc	Cover
Rule 1	N								N		N&P	B	24	27
Rule 2					N				N		N&P	B	23	27
Rule 3									N		N&P	B	27	32
Rule 4	N	N	N						H-I			C	34	34
Rule 5	N	N			N				H-I			C	32	32
Rule 6			N						H-I			C	63	66
Rule 7	N				N				H-I			C	41	42
Rule 8		N			N				H-I			C	52	54
Rule 9	N			N					H-I			C	41	43
Rule 10					N		N		H-I			C	45	47
Rule 11		N		N		N			H-I			C	38	40
Rule 12		N	N		N	N		N				C	29	30
Rule 13	N						N/H					C	24	25
Rule 14		N	N								N	C	26	27
Rule 15				N					H-I		H	C	29	30
Rule 16		N	N			N	N	N				C	25	26
Rule 17									H-I			C	89	98
Rule 18		N	N					N				C	50	54
Rule 19		N					N			H		C	38	42
Rule 20		N	N							H		C	36	40
Rule 21					N		N/H					C	26	31
Rule 22	N	N			N						H	C	27	30
Rule 23	N	N					N				H	C	27	30
Rule 24		N									H	C	49	55
Rule 25		N									N	C	34	40
Rule 26	N/L				N							C	23	27
Rule 27	N					N					H	C	31	36
Rule 28			N			N				H		C	32	37
Rule 29		N	N									C	142	173
Rule 30	N										H	C	49	59
Rule 31							N/H					C	35	42
Rule 32									H-D			C	33	40
Rule 33		N				N		N				C	43	51
Rule 34			N		N	N	N	N				C	32	40
Rule 35						N/L						C	28	35
Rule 36	O						N					C	28	35
Rule 37			N					N		N&P	N&P	C	21	26

4 Mining abstracted hepatitis data

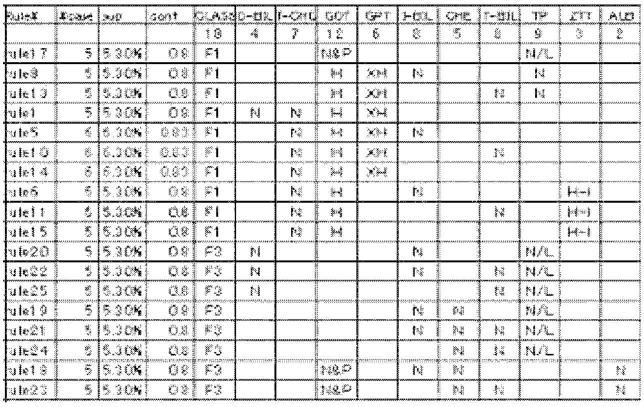
This step can be considered as complex temporal abstraction with the used of our visual data mining system D2MS [5] or the commercial data mining system Clementine to find useful patterns/models from abstracted data obtained from basic TA.

4.1 Patterns describing hepatitis B and C

For the problem P1, different rule sets were found by using program LUPC in system D2MS with different parameters. Table 3 summarizes a rule set discovered by LUPC that describes hepatitis B and C under the constraints that each of them covers at least 20 cases and has accuracy higher than 80%. From this table the medical doctors and us have drawn a number of interesting conclusions.

- The tests ALB, CHE, D-BIL, TP, and ZTT often occur in rules describing types B and C of hepatitis. The test GPT and GOT are not necessarily the key tests to distinguish types B and C of hepatitis (though they are important for solving other problems).

Table 4. Discovered rules describing fibrosis stages



Rule#	#case	sup	conf	CLASS	D-BIL	T-CHO	GOT	GPT	I-BIL	CHE	T-BIL	TP	ZTT	ALB
				18	4	7	12	6	8	5	8	9	3	2
ule17	5	5.30%	0.9	F1			N&P					N/L		
ule8	5	5.30%	0.8	F1			H	XH	N			N		
ule13	5	5.30%	0.9	F1			H	XH			N	N		
ule1	5	5.30%	0.8	F1	N	N	H	XH						
ule5	6	6.30%	0.83	F1		N	H	XH	N					
ule10	6	6.30%	0.83	F1		N	H	XH			N			
ule14	6	6.30%	0.83	F1		N	H	XH						
ule6	5	5.30%	0.8	F1		N	H		N				H-I	
ule11	5	5.30%	0.8	F1		N	H				N		H-I	
ule15	5	5.30%	0.8	F1		N	H						H-I	
ule20	5	5.30%	0.9	F3	N				N			N/L		
ule22	5	5.30%	0.8	F3	N				N		N	N/L		
ule25	5	5.30%	0.8	F3	N						N	N/L		
ule19	5	5.30%	0.9	F3					N	N		N/L		
ule21	5	5.30%	0.8	F3					N	N	N	N/L		
ule24	5	5.30%	0.8	F3						N	N	N/L		
ule18	5	5.30%	0.8	F3			N&P		N	N				N
ule23	5	5.30%	0.8	F3			N&P			N	N			N

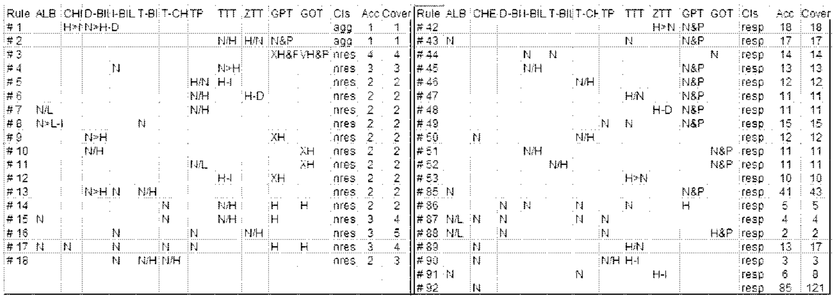
- There are not many rules with large coverage for type B.
- Rule 32 is simple and interesting as it confirms that among four typical short-term changed tests, TTT and ZTT have sensitivity to inflammation but they do not have enough specificity to liver inflammation. The rule says that "if ZTT is high but decreasing we can predict the type C with accuracy 83% (5.1)".
- Rule 29 "IF CHE = N and D-BIL = N THEN Class = C" is also typical for type C as it covers a large population of the class (173/272 or 63.6%) with accuracy 82.08% (3.42).

4.2 Patterns describing the fibrosis stages

For the problem P2 we found a number of significant rules by D2MS. Table 4 shows summaries of 10 rules discovered for fibrosis stages F1 and 8 rules for fibrosis stage F3. In this Table, says, the first rule describing fibrosis stage F1 can be read as "if GOT = N&P and TP = N/L then the class is F1". It is interesting that the rules describing fibrosis stage F1 and F3 are well separated:

- Rules describing the fibrosis stage F1 except the first one are typically related to the combinations of "GOT = H and GPT = XH and (T-CHO = N or TP = N)", or "T-CHO = N and GOT = H and ZTT = H–I".
- Rules describing the fibrosis stage F3 can be distinguished from those of F1 by the combinations "TP = N/L and (D-BIL = N or CHE = N)", or "GOT = N&P and CHE = N".

Table 5. Discovered rules describing the effectiveness of interferon therapy



Rule	ALB	CHI	D-BI	I-BIL	T-BI	T-CH	TP	TTT	ZTT	GPT	GOT	Cls	Acc	Cover	Rule	ALB	CHE	D-BI	I-BIL	T-BIL	T-CH	TP	TTT	ZTT	GPT	GOT	Cls	Acc	Cover
# 1		H>I	N>H-D									agg	1	1	# 42									H>N	N&P		resp	18	18
# 2								N/H	H/N	N&P		agg	1	1	# 43	N							N		N&P		resp	17	17
# 3										XH&P	VH&P	nres	4	4	# 44				N	N						N	resp	14	14
# 4				N				N>H				nres	3	3	# 45				N/H						N&P		resp	13	13
# 5							H/N	H-I				nres	2	2	# 46						N/H				N&P		resp	12	12
# 6							N/H		H-D			nres	2	2	# 47								H/N		N&P		resp	11	11
# 7	N/L						N/H					nres	2	2	# 48									H-D	N&P		resp	11	11
# 8	N>L-I				N							nres	2	2	# 49							N	N		N&P		resp	15	15
# 9			N>H							XH		nres	2	2	# 50		N				N/H						resp	12	12
# 10			N/H								XH	nres	2	2	# 51				N/H							N&P	resp	11	11
# 11							N/L				XH	nres	2	2	# 52					N/H						N&P	resp	11	11
# 12								H-I		XH		nres	2	2	# 53								H>N				resp	10	10
# 13			N>H	N	N/H							nres	2	2	# 85	N									N&P		resp	41	43
# 14						N		N/H		H	H	nres	2	2	# 86			N	N		N		N		H		resp	5	5
# 15	N					N		N/H		H		nres	3	4	# 87	N/L	N	N			N	N					resp	4	4
# 16				N			N		N/H			nres	3	5	# 88	N/L		N				N				H&P	resp	2	2
# 17	N	N		N		N	N			H	H	nres	3	4	# 89		N						H/N				resp	13	17
# 18				N	N/H	N/H						nres	2	3	# 90		N					N/H	H-I				resp	3	3
															# 91	N					N			H-I			resp	6	8
															# 92		N										resp	85	121

4.3 Patterns describing the effectiveness of interferon therapy

For the problem P3, Table 5 shows the rules found for two classes of "non-response" and "response" patients with interferon therapy. It can be observed that many rules for the "non-response" class containing GPT and/or GOT with values "XH&P", "VH&P", "XH", or "H", while many rules for the "response" class containing GPT or GOT with values "N&P" or "H&P". The results allows us to hypothesize that the interferon treatment may have strong effectiveness on peaks (suddenly increasing in a short period) if the base state is normal or high. It can be hypothesized that when the base state is very high or extremely high, the interferon treatment is not clearly effective.

5 Conclusion

We have presented a temporal abstraction approach to mining the temporal hepatitis data. The temporal abstraction approach in our work differs from related temporal abstraction works in two points: the irregular data-stamped points and long periods. Different from these applications, the irregularity in measuring the hepatitis data requires a statistical analysis basing on and combining with the expert's opinion, in particular in the determination of episodes. The key ideas of our method are its combination of "states" and "trends" in the notion of "changes of state" for long-term changed tests, and the combination on "base state" and "peaks" to characterize short-term changed tests.

Our temporal abstraction methods can be applied to other domains where we need process similar temporal data. Also, many other machine learning methods can be applied to the abstracted data to find other kinds of new patterns/models in the hepatitis domain. The findings by our temporal abstraction methods are positively evaluated by medical doctors in terms of novelty, acceptability and utility. They have evaluated many found patterns (rules) as new and interesting, and the sets of rules partially answered the problems under consideration (P1, P2, and P3).

The temporal abstraction approach presented in this paper is carried out in the scope of an on going project in collaboration with medical doctors. The issues to be investigated in the next step include refinement of abstracted patterns (for example, positions of peaks or parameters for abstraction), the post-processing and interpretation of obtained complex temporal abstractions.

References

1. Bellazzi, R., Magni, P., Larizza, C., De Nicolao, G., Riva, A., and Stefanelli, M.(1998). "Mining biomedical time series by combining structural analysis and temporal abstractions", *Journal of American Medical Informatics Association*, Vol. 5, 160–164.
2. Bellazzi, R., Larizza, C., Magni, P., Monntani, S., and Stefanelli, M. (2000). Intelligent analysis of clinic time series: An application in the diabetes mellitus domain, Artificial Intelligence in Medicine, 20, 37–57.
3. Brodley, C. and Smyth, P., (1995). The process of applying machine learning algorithms. *Workshop on Applying Machine Learning in Practice, Twelfth International Machine Learning Conference*, 7–13.
4. Haimowitz, I.J. and Kohane, I.S. (1996). Managing temporal worlds for medical trend diagnosis. Artificial Intelligence in Medicine 8(3), 299–321.
5. Ho, T.B., Nguyen, T.D., Nguyen, D.D., and Kawasaki, S. (2001). Visualization Support for User-Centered Model Selection in Knowledge Discovery and Data Mining, *International Journal of Artificial Intelligence Tools*, Vol. 10, No. 4, 691–713.
6. Ho, T.B., Nguyen, D.D., Nguyen, T.D., and Kawasaki, S. (2002). Extracting knowledge from hepatitis data with temporal abstraction. *Workshop on Active Mining, IEEE Conference on Data Mining ICDM'02*, Maebashi, 91–96.
7. Horn, W., Miksch, S., Egghart, G., Popow, C., and 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. Larizza, C., Bellazzi, R., and Riva, A. (1997)."Temporal abstractions for diabetic patients management", *Artificial Intelligence in Medicine*, Keravnou, E. et al. (eds.), Proc.AIME-97, 319–330.
9. [Lavrac 97] Lavrac, N., Keravnou, E., and Zupan, B. (1997). *Intelligent Data Analysis in Medicine and Pharmacology* (Eds.), Kluwer.
10. Miksch S., Horn W., Popow C., and Paky F. (1996). Utilizing temporal data abstraction for data validation and therapy planning for artificially ventilated newborn infants, *Artificial Intelligence in Medicine, 8*, 543–576.
11. Motoda, H. (2002). *Active Mining: New directions of data mining* (Ed.), IOS Press.
12. Nguyen, D.T. and Ho, T.B. (1999). An interactive-graphic system for decision tree induction. *Journal of Japanese Society for Artificial Intelligence, 14:1*, 131–138.
13. Shahar, Y. and Musen, M.A. (1997). Knowledge-based temporal abstraction in clinical domains, *Artificial Intelligence in Medicine, 8*, 267–298.
14. Shahar, Y. (1997). A framework for knowledge-based temporal abstraction, *Artificial Intelligence, 90*, 79–133.