Discovery of trends and states in irregular medical temporal data

Trong Dung Nguyen, Saori Kawasaki and Tu Bao Ho

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

Abstract. 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

The hepatitis temporal database collected between 1982-2001 at the Chiba university hospital 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 [7]. The hepatitis database was given as discovery challenge in 2002, 2003 of PKDDs(http://www.cs.helsinki.fi/events/eclpkdd/challenge.html). Among problems posed by the doctors we are interested in the following:

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.

One of the main approaches to mining medical temporal data is temporal abstraction (TA). The key idea temporal abstraction 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.

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. Typical works on temporal abstraction are those presented in [1], [4], [2], [8]. 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 [5], [1], [4] related to temporal data of an individual measured on consecutive days in a short period; on normal and dependent diabetes related to temporal data measured on consecutive days within two weeks; and 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.

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.

2 Basic Temporal Abstraction

Each patient is described by 983 temporal sequences corresponding to the 983 hospital tests. As the complexity of learning generally increases with the number of tests under investigation, a small number of selected tests is expected. After selecting 41 tests from 983 tests by statistical frequency check and medical background knowledge.

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.

We firstly focus on the 15 most typical tests as suggested by medical experts. These tests can then be divided into two groups depending whether their values can change in a short term or long term.

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.
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.

Table 1. The temporal abstraction primitives

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

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 1.

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).

We developed and used the following procedure to identify typical abstraction patterns. 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.

1. Consider the patterns structures as formulas and the $< state\ primitive >$, $< trend\ primitive >$ and $< relation >$ as their variables. Create all possible candidate abstraction patterns by replacing the $< state\ primitive >$, $< trend\ primitive >$ and $< relation >$ with their possible values.
2. Randomly take a large number of sequences from the datasets, visualize and manually match them with the candidate abstraction patterns to see each of them matches which candidate abstraction pattern.
3. Eliminate the candidate abstraction patterns that have no or a small number of matched sequences.

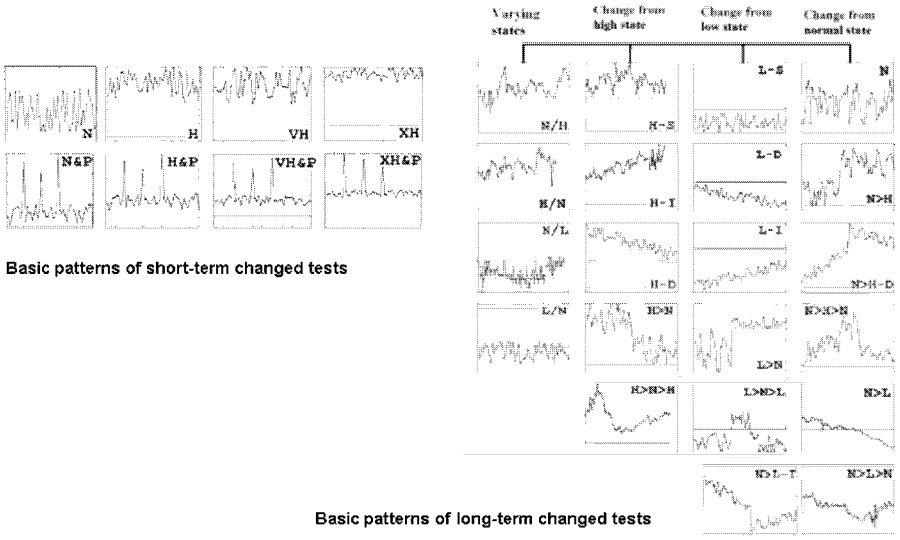

Fig. 1. Abstraction patterns of short-term and long-term changed tests

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, α (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) $BS = N$ *and* $V(M_i) > V(VH)$
   (2) $BS = H$ *and* $V(M_i) > V(XH)$
   (3) $BS = VH$ *and* $V(M_i) > 2 * V(XHU)$
   (4) $BS = XH$ *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 α, δ, σ, ϵ (integer), β (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}) > α *wedge* In(S_{00}) > Out(S_{00}) *wedge* High(S_{00}) > Low(S_{00}) then N/H
2. If Cross(S_{00}) > α *wedge* In(S_{00}) > Out(S_{00}) *wedge* High(S_{00}) ¡ Low(S_{00}) then N/L
3. If Cross(S_{00}) > α *wedge* In(S_{00}) < Out(S_{00}) *wedge* High(S_{00}) > Low(S_{00}) then H/N
4. If Cross(S_{00}) > α *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}) > β 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}) < ϵ 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}) < ϵ then N>H–D
14. If First (S_{00}) = N $\wedge$ Cross(S_{00}) < α High(S_{00}) > $\delta \wedge$ Last (S_{22}) = N
$\wedge$ Trend(S_{22}) = D $\wedge$ Low(S_{00}) < ϵ 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}) < ϵ 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}) < ϵ then N>L–I
17. If First (S_{00}) = N $\wedge$ Cross(S_{00}) < α Low(S_{00}) > $\delta \wedge$ Last (S_{22}) = N
$\wedge$ Trend(S_{22}) = I $\wedge$ High(S_{00}) < ϵ then N>L>N

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

E. Identification of patterns with changes from the low region

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

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

3 Mining abstracted hepatitis data

This step can be considered as complex temporal abstraction with the used of our visual data mining system D2MS [3] or the commercial data mining system Clementine to find useful patterns/models from abstracted data obtained from basic TA.

3.1 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. From the 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%, we 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).
- 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).

3.2 Patterns describing the fibrosis stages

For the problem P2 we found a number of significant rules by D2MS. We can draw interesting patterns:

- 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".

3.3 Patterns describing the effectiveness of interferon therapy

For the problem P3, we found rules 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.

4 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 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., 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.
2. Haimowitz, I.J. and Kohane, I.S. (1996). Managing temporal worlds for medical trend diagnosis. Artificial Intelligence in Medicine 8(3), 299–321.
3. 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.
4. 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.
5. 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.
6. Lavrac, N., Keravnou, E., and Zupan, B. (1997). *Intelligent Data Analysis in Medicine and Pharmacology* (Eds.), Kluwer.
7. Motoda, H. (2002). *Active Mining: New directions of data mining* (Ed.), IOS Press.
8. Shahar, Y. and Musen, M.A. (1997). Knowledge-based temporal abstraction in clinical domains, *Artificial Intelligence in Medicine, 8*, 267–298.