社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－ＩＣＳ－133 (14)
2003／9／15

Mining Hepatitis Data with Temporal Abstraction

Tu Bao HO, Trong Dung NGUYEN, Saori KAWASAKI, Si Quang LE, Dung Duc NGUYEN,
Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa, 923-1292 Japan
{bao, nguyen, skawasa, quang, dungduc}@jaist.ac.jp

Hideto YOKOI, Katsuhiko TAKABAYASHI
Chiba University Hospital
Inohana, Chuo-ku, Chiba, 260-8677 Japan
{yokoi,takaba}@ho.chiba-u.ac.jp

ABSTRACT

The hepatitis temporal database collected at Chiba university hospital during 1982-2001 was recently given to challenge the KDD research. The database is large where each patient corresponds to 983 tests represented as sequences of irregular time-stamp points with different lengths. This paper presents a temporal abstraction approach to mining knowledge from this hepatitis database. Exploiting hepatitis background knowledge and data analysis, we introduce new notions and methods for abstracting short-term changed and long-term changed tests. The abstracted data allow us to apply different machine learning methods for finding knowledge that was evaluated with statistical significance and by medical doctors.

Keywords

Hepatitis data, medical data mining, temporal abstraction.

1. INTRODUCTION

The hepatitis temporal database collected during 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. The hepatitis database was given as discovery challenge in 2002, 2003 of PKDD (http://www.cs.helsinki.fi/events/ecmlpkdd/challenge.html). Among problems posed by the doctors we are interested in the following:

P1: Discover the differences in temporal patterns between hepatitis B and C.

P2: Evaluate whether laboratory tests can be used to estimate the stage of liver fibrosis.

P3: Evaluate if 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.

Typical works on temporal abstraction are those in [1], [4], [6]. Temporal abstraction can be generally considered in two phases: basic temporal abstraction that concerns with abstracting time-stamped data within episodes, and complex temporal abstraction that concerns with temporal relationships between findings from a basic temporal abstraction or from other complex temporal abstractions.

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. PREPROCESSING

2.1 General preprocessing

The data cleaning requires eliminate noisy data. The main task is to remove non unified symbols or characters occurred during the data collection.

We also carried out several transformations of data. For example, the test such as CHE was measured before and after the mid-80s by different measurements (with normal regions are [6, 12] and [180, 430], respectively).

We converted the old test values accordingly to the new ones obtained by new measurements.

Another problem is feature selection. By the guide of medical doctors and the statistics on frequencies of tests, from 983 tests we selected the 41 most significant ones. The dataset for investigating each problem will be selected from these tests plus some special tests recommended by the medical doctors.

2.2 Extracting data for P1, P2, and P3

The data extraction aims to create an appropriate dataset for solving each problem by temporal abstraction techniques. According to the medical background knowledge, we focus on exploiting the 15 most frequent tests. It is important to recall that the quality of temporal abstraction also strongly depends on how episodes on which data are abstracted were taken. In this research we adopted a simple technique for determining episodes. Based on suggestions of medical experts, we first determine a pilot point (e.g., the starting day, the last day, the biopsy day of the sequence, etc.), and take episodes (subsequences) from the whole sequence in backward, forward, or to both sides of the pilot point.

In fact, for the problem P1 episodes are forwardly taken from the starting day of the sequence. For the problems P2 and P3 episodes are backwardly taken from the day of doing biopsy or the last day before the treatment with interferon, respectively. For the problem P3 on the effectiveness of interferon, we have to separate the patients into four groups by response to interferon (IFN) therapy based on the domain knowledge of doctors: Response, Partial response, Aggravation, Non-response.

3. BASIC TEMPORAL ABSTRACTION

We started by a separation of two groups of 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.

(1) *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.

3.1 Temporal abstraction primitives

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

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

We define four structures of abstraction patterns:

<pattern> ::= <state primitive>
<pattern> ::= <state primitive> <relation> | <state primitive> | <trend primitive>
<pattern> ::= <state primitive> <relation> <peak>
<pattern> ::= <state primitive> <relation> <state primitive> <relation> | <state primitive>|<trend primitive>

Examples of abstracted patterns in a given episode are:

- "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 sequennces. Figure 2 shows typical possible patterns (8 and undetermined)

for short-term changed tests, and Figure 3 shows typical possible patterns (21 and undetermined) for long-term changed tests. Suppose that S is a sequence to be considered. The following notations will be used to describe 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.
- $First_{\sigma}(S)$: State of the first σ points in S
- $Last_{\sigma}(S)$: State of the last σ points in S
- State(S): State of S (one of the state primitives)
- Trend(S): Trend of S (one of trend primitives).

3.2 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 that of a "stable" state, called *base state* (BS), and the position and value of *peaks*, where the tests suddenly go up. Based on this remark, we develop the following algorithm to find the base state and peaks of a short term changed test.

3.3 Abstraction of long term changed tests

The 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 are can be at one of the three states "N", "H", or "L". It will happen that either the sequence changes from one state to another state, smoothly or variably (at boundaries), or the sequence remains in its state without changing. As changes can generally happen in long-term, it is possible to consider the trend of a sequence after changing of the state.

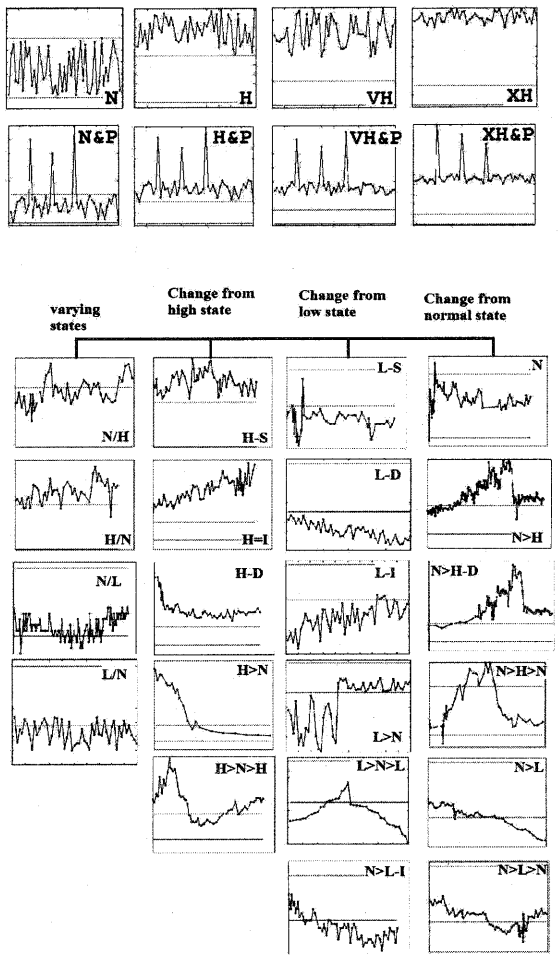

Figure 1: Typical abstraction patterns

Algorithm (for long-term changed tests)

Input: A sequence of patient's values of a test with length N denoted as $S_{00} = \{s_1, s_2, \ldots, s_N\}$ in a given episode.

Output: An abstracted pattern of the sequence derived from the sequence.

Parameters: α, δ, ε, σ (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** $Cross(S_{00}) > \alpha \wedge In(S_{00}) > Out(S_{00}) \wedge High(S_{00}) > Low(S_{00})$ **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 with many crosses

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** $First_{\sigma}(S_{00}) = N \wedge Cross(S_{00}) < \alpha \wedge Last_{\sigma}(S_{22}) = H \wedge Trend(S_{22}) = I \wedge Low(S_{00}) < \varepsilon$ **then** N>H
13. **If** $First_{\sigma}(S_{00}) = N$ & $Cross(S_{00}) < \alpha$ & $Last_{\sigma}(S_{22}) = H$ & $Trend(S_{22}) = D \wedge Low(S_{00}) < \varepsilon$ **then** N>H–D
14. **If** $First_{\sigma}(S_{00}) = N \wedge Cross(S_{00}) < \alpha \wedge High(S_{00}) > \delta \wedge Last_{\sigma}(S_{22}) = N \wedge Trend(S_{22}) = D \wedge Low(S_{00}) < \varepsilon$ **then** N>H>N
15. **If** $First_{\sigma}(S_{00}) = N \wedge Cross(S_{00}) < \alpha \wedge Last_{\sigma}(S_{22}) = L \wedge Trend(S_{22}) = D \wedge High(S_{00}) < \varepsilon$ **then** N>L
16. **If** $First_{\sigma}(S_{00}) = N \wedge Cross(S_{00}) < \alpha \wedge Last_{\sigma}(S_{22}) = L \wedge Trend(S_{22}) = I \wedge High(S_{00}) < \varepsilon$ **then** N>L–I
17. **If** $First_{\sigma}(S_{00}) = N \wedge Cross(S_{00}) < \alpha \wedge Low(S_{00}) > \delta \wedge Last_{\sigma}(S_{22}) = N \wedge Trend(S_{22}) = I \wedge High(S_{00}) < \varepsilon$ **then** N>L>N

D. Identification of patterns with changes from the high region

18. **If** $First_{\sigma}(S_{00}) = H \wedge Cross(S_{00}) < \alpha \wedge Last_{\sigma}(S_{22}) = N \wedge Low(S_{00}) < \varepsilon$ **then** H>N
19. **If** $First_{\sigma}(S_{00}) = H \wedge Cross(S_{00}) < \alpha \wedge Normal(S_{00}) > \delta \wedge Last_{\sigma}(S_{22}) = H \wedge Trend(S_{22}) = I \wedge Low(S_{00}) < \varepsilon$ **then** H>N>H

E. Identification of patterns with changes from the low region

20. **If** $First_{\sigma}(S_{00}) = L \wedge Cross(S_{00}) < \alpha \wedge Last_{\sigma}(S_{22}) = N \wedge Low(S_{00}) < \varepsilon$ **then** L>N
21. **If** $First_{\sigma}(S_{00}) = L \wedge Cross(S_{00}) < \alpha \wedge Normal(S_{00}) > \delta \wedge Last_{\sigma}(S_{22}) = L \wedge Trend(S_{22}) = D \wedge High(S_{00}) < \varepsilon$ **then** L>N>L
22. **If** NULL **Then** Undetermined.

Figure 2: An algorithm for temporal abstraction

4. COMPLEX TEMPORAL ABSTRACTION

In this section we report applications of different machine learning methods to abstracted data obtained by basic temporal abstraction, including our system D2MS [4], C4.5 and Clementine.

	Minδ no (%)	Min Conf (%)	# rules with statistical significance over # rules found			
			α=0.05	α=0.1	α=0.15	α=0.2
HBV and HCV	1	75	341/4529	508/4529	687/4529	974/4529
	2	75	239/1584	334/1584	493/1584	669/1584
	3	75	197/948	267/948	394/948	506/948
	12	75	7/25	8/25	10/25	17/25
	Average		17.65%	23.1%	31.85%	46.25%
Fibrosis stage	2	70	6/319	8/319	14/319	45/319
	2	75	6/224	6/224	6/224	12/224
	2	80	6/178	6/178	6/178	6/178
	3	75	2/89	2/89	2/89	8/89
	Average		2.5%	2.67%	3.13%	7.95
IFN	5	75	5/1201	19/1201	66/1201	116/1201
	10	80	5/517	19/517	66/517	116/517
	20	75	2/386	8/386	45/386	95/386
	30	75	0/218	4/218	37/218	73/218
	Average		0.53%	2.28%	11.7%	22.53%

Figure 3: Ratio of statistically significant rules by

A critical question is which of them are statistically significant and were not found due to chance? It is worth noting that this important question has not been fully investigated in the KDD community. Figure 3 shows the ratio of statistically significant rules evaluated by the method in [2] that performs three kinds of hypothesis testing:

- the significance of the consequence of the rule;
- the significance of antecedent of the rule; and
- the significance of the accuracy/confidence of the rule.

4.1 Mining abstracted hepatitis data with system D2MS

D2MS is a visual data mining system with visualization support for model selection. D2MS

facilitates the trials of various alternatives of algorithm combinations and their settings. The data mining methods in D2MS consists of programs CABRO for tree learning and LUPC for rule learning [4]. CABRO produces decision trees using R-measure and graphically represents them in particular with T2.5D tool (trees 2.5 dimension). LUPC is a separate-and-conquer algorithm that controls the induction process by several parameters that allow obtaining different results. This ability supports the user plays a central role in the mining process.

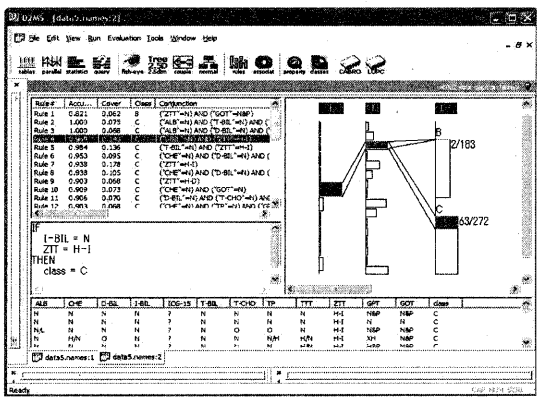

Figure 4: A rule found for hepatitis type C

For the problem P1, different datasets were found by using LUPC with different parameters. Figure 4 presents one of rules describing the type C of hepatitis that is considered interesting by medical doctors. Table 2 summarizes a rule set discovered by LUPC under the constraints that each of them covers at least 20 cases and with accuracy higher than 80%.

From this table some remarks can be drawn:

- Few rules are statistically significant.
- GOT, GPT, CHE, D-BIL, and ZTT often occur in rules for types B and C.
- There are not many rules with large coverage for type B. "if ZTT is in the boundary of the normal and high regions and GOT or GPT are high with peaks then the patient likely has type B".

For the problem P2 we found a number of interesting rules by D2MS. For the problem P3, with the rules found we can be observed that many rules describing the non-response class are with patterns on GPT and/or GOT having values "XH&P", "VH&P", "XH", or "H", while many rules describing the response class are with patterns on GPT or GOT having 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.

Table 1: Statistically significant rules for types B and C

CLASS	ALB	CHE	D-BIL	GOT	GPT	I-BIL	T-BIL	T-CHO	TP	TTT	ZTT	Acc	#correct	#cover
B (60%)				H&P	H&P		N/H					0.75	9	12
B (95%)			H>N	H&P								0.75	6	8
B (80%)			N/H	H&P					N/L			0.8	4	5
B (80%)				H&P			N/H		N/L			0.83	5	6
B (95%)				H&P							N/H	0.9	9	10
B (80%)				N&P						N/H	N/H	0.83	5	6
B (80%)				XH	XH							0.8	4	5
B (90%)					H&P						N/H	0.91	11	12
B (95%)					H&P						N/H	0.92	11	12
B (80%)					N&P	N/H	H/N					0.8	4	5
B (90%)			N/H		VH&P							0.75	3	4
B (80%)											N>H>N	0.7	7	10
B (90%)		H/N										0.7	14	20
B (80%)		N/H										0.74	23	31
B (80%)		H-I										0.75	3	4
B (95%)	N	N/H										0.78	22	28
C (80%)	N			H	H							0.82	51	62
C (95%)					H							0.75	78	103
C (95%)					H							0.78	78	103
C (95%)					H				N/H			0.91	11	12
C (95%)					H		N/H					0.93	14	15
C (95%)				H&P	H&P						H-I	0.95	20	21
C (95%)			N		N&P					N		0.82	28	34
C (90%)		N	N		N&P							0.83	51	61
C (95%)				H	XH							0.91	11	12
C (95%)				H								0.76	89	117
C (95%)	N			H								0.8	84	104
C (95%)			N	N								0.76	38	50
C (80%)		N		N								0.83	41	49
C (95%)				N&P					N/H			0.91	11	12
C (95%)			N					N				0.7669	125	163
C (95%)		N	N									0.82	142	173

4.2 Mining abstracted hepatitis data with Clementine

The complex temporal abstraction can be done by different data mining and machine learning methods depending on the purpose. Together with using D2MS we also use Clementine to investigate the abstracted hepatitis data, in particular the association rule mining and See5 programs in Clementine.

Using the Apriori program we have discovered several interesting properties of hepatitis. Table 1 shows the rules obtained by one of our experiments when investigating the problem P1. These rules cover more than 60% of the database. There are 18 over 20 found rules sharing a lot of common cases and all of them contain the condition "ZTT = H–I". On the other hand, the only one rule on hepatitis type B covering 77 cases says that "if ALB = N and ZTT = N then type B", and another rule covering 173 cases says that "if D-BIL = N and CHE = N then type C" which does not relate

with the condition on ZTT.

Table 2 shows summaries of 10 rules discovered for fibrosis stages F1 and 8 rules for fibrosis stage F3 when investigating the problem P2. In this figure, 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:

- The 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".
- The 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 2: Typical rules describing non-response and response cases to interferon

Rule	D-BIL	T-CHO	GOT	GPT	I-BIL	CHE	T-BIL	TP	ZTT	ALB	Class	Min.Sup.	Conf.
17			N&P					N/L			F1	5.3%	0.8
3			H	XH	N			N			F1	5.3%	0.8
13			H	XH			N	N			F1	5.3%	0.8
1	N	N	H	XH							F1	5.3%	0.8
5		N	H	XH	N						F1	6.3%	0.83
10		N	H	XH			N				F1	6.3%	0.83
14		N	H	XH							F1	6.3%	0.83
6		N	H		N				H-I		F1	5.3%	0.8
11		N	H				N		H-I		F1	5.3%	0.8
15		N	H						H-I		F1	5.3%	0.8
20	N				N			N>L			F3	5.3%	0.8
22	N				N		N	N>L			F3	5.3%	0.8
25	N						N	N>L			F3	5.3%	0.8
19					N	N		N>L			F3	5.3%	0.8
21					N	N	N	N>L			F3	5.3%	0.8
24						N	N	N>L			F3	5.3%	0.8
18			N&P		N	N				N	F3	5.3%	0.8
23			N&P			N	N			N	F3	5.3%	0.8

5. CONCLUSION

We have presented a temporal abstraction approach to mining the temporal hepatitis data. From the results obtained so far, several lessons have been learned and in some issues could be further investigated.

The temporal abstraction approach in our work differs from related temporal abstraction approaches in two points: the irregular data-stamped points and abstraction of multiple variables. Different from related work, the irregularity in measuring the hepatitis data requires a careful statistical analysis basing on and combining with the expert's opinion, in particular in the determination of episodes. The interactive and visual system D2MS provides us a powerful tool for complex temporal abstraction not only in combining obtained abstractions but also in visualizing them in order to give a understanding of relationships between basic temporal abstractions. Not only D2MS, See5, and Clementine but many other machine learning methods can be applied to the abstracted data to find other kinds of new patterns/models in the hepatitis domain.

REFERENCES

[1] Bellazzi, R., Larizza, C., Magni, P., Monntani, S., and Stefanelli, M., "Intelligent Analysis of Clinic Time Series: An Application in the Diabetes Mellitus Domain", *Artificial Intelligence in Medicine 20* (2000), 37-57.

[2] Bruzzese, D. and Davino, C., "Statistical Pruning of Discovered Association Rules", Computational Statistics 16 (3), pp. 387 -398, 2001.

[3] Ho, T.B., Nguyen, T.D., and Nguyen, D.D., "Visualization Support for a User-Centered KDD Process", *ACM International Conference on Knowledge Discovery and Data Mining KDD-02*, Edmonton, 519-524.

[4] Horn, W., Miksch, S., Egghart, G., Popow, C., and 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), 389-409, 1997.

[5] Lavrak, N., "Selected Techniques for Data Mining in Medicine", *Artificial Intelligence in Medicine*,

[6] Shahar, Y., "A Framework for Knowledge-based Temporal Abstraction", *Artificial Intelligence*, 90 (1997), 79-133.