2004－ICS－138　(9)
2004／12／4

Multi-aspect Hepatitis Data Analysis

Yasuo Fujita,† Muneaki Ohshima,† Ning Zhong †
and Hideto Yokoi††

When IFN (interferon) is used for chronic hepatitis patients, various conceptual knowledge/rules will benefit for giving a treatment. In this paper, we describe an ongoing work on using various data mining agents including the GDT-RS inductive learning system for discovering classification rules, and the LOI (learning with ordered information) for discovering important features, in a multi-phase process for multi-aspect analysis of the hepatitis data. Our methodology and experimental results show that the perspective of doctors will be changed from a single type of experimental data analysis towards a holistic view, by using our *multi-aspect mining* approach.

1. Introduction

Multi-aspect mining in a multi-phase KDD process is an important methodology for knowledge discovery from real-world data[1),4)]. There are two main reasons why a multi-aspect mining approach needs to be used for the hepatitis data analysis. The first reason is that we cannot expect to develop a single data mining algorithm for analyzing all aspects of the hepatitis data towards a holistic view since complexity of the real-world data. Hence, various data mining agents need to be cooperatively used in the multi-phase data mining process for performing multi-aspect analysis as well as multi-level conceptual abstraction and learning. The other reason is that when performing multi-aspect analysis for complex problems, a data mining task needs to be decomposed into sub-tasks. Thus these sub-tasks can be solved by agents distributed over different computers.

More specifically, when IFN is used for chronic hepatitis type C patients, various conceptual knowledge/rules will benefit for giving a treatment. The knowledge/rules, for instance, include (1) when the IFN should be used for a patient so that he/she will be able to be cured, (2) what kinds of inspections are important for a diagnosis, and (3) whether some peculiar data/patterns exist or not. In this paper, we describe an ongoing work on using various data mining agents including the GDT-RS[5)] inductive learning system for discovering classification rules[5)], and the LOI for discovering important features[2),6)]. So that such rules mentioned above can be discovered automatically.

In our experiments, the inspection data of patients in one year before using IFN, is first pre-processed.After that, the pre-processed data are used for each data mining agent, respectively. By using the GDT-RS, the rules of whether a medical treatment is effective or not, can be found. And, by using the LOI, what attributes affect the medical treatment of hepatitis C greatly can be investigated.

2. Rule Discovery by GDT-RS

GDT-RS is a soft hybrid induction system for discovering classification rules from databases with uncertain and incomplete data[5)]. The system is based on a hybridization of the *Generalization Distribution Table (GDT)* and the *Rough Set* methodology.

2.1 Pre-processing

Before using GDT-RS, the condition attributes and the determination class must be defined. The following 11 attributes are selected as condition attributes:

T-CHO, CHE, ALB, TP, T-BIL, D-BIL,
I-BIL, PLT, WBC, HGB, GPT

And the decision attribute named class (see Table 1) is defined by the effect of IFN, which is decided by whether a hepatitis virus exists or not.

Furthermore, only the patients who satisfy the following conditions are extracted:

- Patient who has been medicated with IFN;
- Patient with the data of whether the hepatitis virus exists or not;

† Department of Information Engineering, Maebashi Institute of Technology
†† School of Medicine, Chiba University

Table 1 The decision attribute (class)

Class	The effect of IFN	No. Patients
R	Virus disappearance	58
N	Virus existence	86
?	Data lack	53

- Patient with inspection data collected in one year before IFN is used.

Thus, 197 patients are used in our data mining.

As the condition attributes are continuation values, they are quantized as follows:

(1) As shown in Fig. 1, all the inspection values of a patient within one year before IFN is used are divided into two groups: the first half and the second half,

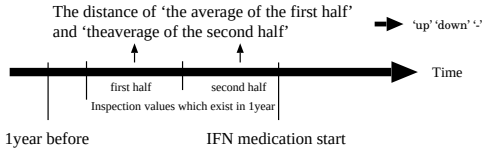

Fig. 1 The evaluation method of condition attributes

(2) Let D be the difference between average values of the first half and the second half,

(3) The attribute value is estimated as "down" if $D >$ threshold, "up" if $D <$ -threshold, "–" (i.e. no change) if the absolute of $D <$ threshold, and "?" if there is no inspection data or only one (i.e. a patient is examined only once).

The threshold values used above are defined as follows:

- **The threshold values for each attribute except GPT**
  The threshold values are set up to 10% of the normal range of each inspection data.
- **The threshold value for GPT**
  As the change of a hepatitis patient's GPT value will exceed the normal range greatly, the threshold value for the GPT needs to be calculated in a more complex method.
  At first, as shown in Fig. 2, the standard deviation of the difference of the adjacent inspection values of each hepatitis patient's GPT is calculated respectively. And then the standard deviation of such standard deviation is used as a threshold value.
  Let M be the number of patients, $t_m (1 \leq m \leq M)$ the times of inspection of patient m, $d_{mi} (1 \leq i \leq t_m - 1)$ the difference of adjacent inspection values. Thus, the threshold value of GPT can be calculated in Eq. 1.

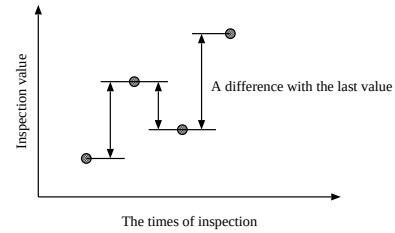

Fig. 2 Standard deviation of the difference of adjacent values

$$\text{Threshold}_{\text{GDT}} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(s_m - \overline{s})^2} \quad (1)$$

Where $s_m (1 \leq m \leq M)$ is the standard deviation of the difference d_{mi} of the inspection value that is calculated for each patient, respectively, and $\overline{s}$ is the average value of s_m.

The threshold values used for evaluating each condition attribute are shown in Table 2.

Table 2 The threshold values for evaluating condition attributes

T-CHO = 9.5	CHE = 25	ALB = 0.12
TP = 0.17	T-BIL = 0.1	D-BIL = 0.03
I-BIL = 0.07	PLT = 20	WBC = 0.5
HGB = 0.6	GPT = 54.56	

2.2 Results of Post-processing

In the experimental results at the accuracy 60%, only the rules with which the number of condition attributes is less than or equal to three are extracted. This is because it will become unclear if the number of condition attributes increases. Tables 3 shows such rules that are divided into classes R and N, respectively.

In our post-processing, each discovered rule is checked and the result is shown in Tables 4 and 5, where the $Pos.$ (or $Neg.$) ID means that the patient is covered by a rule as a *positive* (or *negative*) instance. From these tables, we can see clearly what patients are covered by each rule. It is useful for finding the main features of a patient group.

2.3 Analyses and Evaluations

The rules derived by the GDT-RS and the results acquired from post-processing have been evaluated by a doctor. The evaluations of the

Table 3 Rules discovered by GDT-RS

ID	Rules of class R	Accuracy
001	GPT(up)	(10/16)=62%
002	T-CHO(down) ∧ PLT(down)	(6/9)=66%
003	T-BIL(up) ∧ GPT(down)	(3/4)=75%
004	TP(down) ∧ GPT(down)	(3/4)=75%
...	...	...

ID	Rules of class N	Accuracy
101	D-BIL(down)	(26/43)=60%
102	T-CHO(down) ∧ I-BIL(down)	(7/11)=63%
103	I-BIL(down) ∧ WBC(down)	(7/8)=87%
104	D-BIL(up) ∧ PLT(down)	(4/6)=66%
105	TP(up) ∧ I-BIL(down)	(5/6)=83%
106	TP(up) ∧ T-BIL(down)	(4/6)=66%
107	TP(up) ∧ PLT(down)	(4/5)=80%
108	CHE(up) ∧ T-BIL(down)	(2/4)=50%
...	...	...

Table 4 Patients covered by rules of class R

Rule-ID	Pos. patient ID				Neg. patient ID		
001	158	351	534	547	35	188	273
	778	801	909	923	452	623	712
	940	942					
002	91	351	650	703	169	712	952
	732	913					
003	431	592	700		122		
004	37	71	730		122		
...	...				...		

Table 5 Patients covered by rules of class N

Rule-ID	Pos. patient ID				Neg. patient ID		
101	2	104	125	182	37	71	133
	184	191	203	208	169	180	206
	239	290	546	439	248	276	413
	493	498	529	578	593	610	683
	585	634	652	653	702	713	732
	669	715	719	743	771	948	
	750	756					
102	2	239	563	634	169	413	650
	652	653	952	732			
...	...				...		

rules are based on acceptability and novelty, each of which are divided into five levels from 1 (lowest) to 5 (highest), as shown in Table 6.

2.3.1 Evaluation of Rules

From the viewpoint of the rules with a higher support (i.e. *rule-001* and *rule-101*), we observed that:

- It will heal up in many cases if a patient is medicated with IFN at the time when GPT is going up (hepatitis is getting worse);
- It does not heal up in many cases even if a patient is medicated with IFN at the time when D-BIL is descending.

As we have already known that the effect of IFN is (1) relevant to different types of hepatitis viruses, and (2) hard to be effective when there are large amounts of hepatitis virus, we can see that *rule-001* and *rule-101* do not conflict with the existing medicine knowledge.

Table 6 Evaluation of rules

Class	Rule-ID	Acceptability	Novelty
R	001	4	5
	002	3	5
	003	4	5
	004	4	5
N	101	4	5
	102	2	3
	103	2	3
	104	1	1
	105	3	4
	106	3	4
	107	2	3
	108	3	4

From these two rules, the hypothesis: "IFN is more effective when the inflammation of hepatitis is stronger" can be formed. Based on this hypothesis, we can evaluate the rules discovered as follows.

- In class R, the rules with respect to aggravation of liver function have a good acceptability.
- In class N, the rules with respect to recovery of liver function have good acceptability.

Hence, the evaluations shown in Table 6 can be obtained. In class N, we can see that the acceptability of some rules is 2. This is because both the recovery and aggravation of liver function are included in the premise of the rules.

2.3.2 Evaluation of Post-processing

From the discovered rules in class N, we found that there are some relevances among the patients supported by bilirubin (T-BIL, D-BIL, I-BIL). From T-BIL = D-BIL + I-BIL, the well known background knowledge, it is clear that the rules with respect to bilirubin are relevant.

In order to examine that the condition in a rule is not contradictory to a patient's condition, the discovered rules are categorized, based on liver function, into three categories: recovery, aggravation, and mixture of recovery and aggravation, as shown in Table 7.

From Table 7, we observed that there are many rules with the same conditions in the rule group supported by a patient group, and it may conflict with unknown medical background that is not represented in the conditions of the rules. However, it does not mean that the rules are incorrect. The reason may be that the rules cannot be simply categorized by recovery and aggravation.

Table 7 Category of discovered rules

Class	Recovery	Aggravation	Rec. & Agg.
R	rule 007	rule 001	rule 003
	rule 008	rule 002	rule 004
	rule 009		rule 005
	rule 011		rule 006
			rule 010
N	rule 101	rule 104	rule 102
	rule 105	rule 109	rule 103
	rule 106		rule 107
	rule 108		rule 111
	rule 110		

For example, although it can show liver function aggravation, the lower values of WBC and ALB may not be the real reason of liver function aggravation. On the other hand, since WBC and PLT are the same blood cell ingredient, and T-CHO and ALB are relevant to protein that makes liver, they may be relevant from this point of view. However, T-CHO and ALB do not only provide for liver, but also, for example, T-CHO is related to eating, and ALB is related to the kidney, respectively. Hence it cannot declare there is such correlation.

In summary, there is correlation if we are mentioning about mathematical relevance like BIL. However, it is difficult to find out correlation for others. We need the following methods to solve the issue.

- Finding out what rules are significant from the statistical point of view, based on rough categorizing such as recovery and aggravation.
- Showing whether such rough categorizing is sufficient or not.

3. Rule Discovery by LOI

The LOI uses background knowledge called *order relation* for discovering *ordering rules* and important attributes for an ordered decision class[2),6)]. For example, since the larger the value of T-CHO is, the better, the order relation can be set to (VH $\succ$ H $\succ$ N $\succ$ L $\succ$ VL), where "$\succ$" denotes a weak order. Furthermore, if a decision attribute has two classes: R (response) and N (no response), the order relation can be set to $R \succ N$.

In this experiment, we used the following 12 atrributes as condition attributes :

T-CHO , CHE , ALB , TP, T-BIL , D-BIL , I-BIL , PLT, WBC , HGB , GPT , GOT

and use the same determination class as that used in GDT-RS (see Table 1).

After the patients who have no data of whether the hepatitis virus exists or not are deleted, the data of 142 patients was used in this experiment.

3.1 Rule Discovery by LOI Method 1

According to the background knowledge:

GOT: $N \leq 40 < H \leq 100 < VH \leq 200 < UH$
T-BIL: $L \leq 0.6 < M \leq 0.8 < H \leq 1.1 < VH \leq 1.5 < UH$
PLT: $UL \leq 50 < VL \leq 100 < L \leq 150 < N \leq 350 < H$
WBC: $UL \leq 2 < VL \leq 3 < L \leq 4 < N \leq 9 < H$

we have the order relation as follows:

GOT: $N \succ H \succ VH \succ UH$
GPT: $N \succ H \succ VH \succ UH$
T-CHO: $VH \succ H \succ N \succ L \succ VL$
CHE: $VH \succ H \succ N \succ L \succ VL$
Class: $R \succ N$

An ordered information table can be made by changing the attribute values to symbols and comparing the order of each patient data (see Fig. 3). An ordered information table may be viewed as information tables with added semantics (background knowledge).

1. Change the attribute values to symbols:

	ALB	CHE	..	GOT	CLASS
p1	VH	VH	..	UH	R
p2	N	H	..	N	R
p3	L	L	..	VH	N
⋮			⋮		

2. Create the ordered information table by comparing each patient with others.

Object	ALB	CHE	..	GOT	CLASS
(p1,p2)	N$\succ$	H$\succ$	..	N$\prec$	=
(p1,p3)	L$\succ$	L$\succ$	..	VH$\prec$	N$\succ$
			⋮		
(p2,p1)	VH$\prec$	VH$\prec$	..	UH$\succ$	=

Fig. 3 The creation of an ordered information table

After this transformation, the ordered information table can be used in GDT-RS rule mining system, and the ordering rules can be discoverd.

3.1.1 Results and Evaluation

The rules discovered by LOI are shown in Table 8, where the condition attributes in the rules denote the inspection value "go better" or "go worse".

The evaluation of acceptability and novelty for the rules heavily depends on the correctness of the ordered information. By investigating the values of each attribute in the ordered information table by using Eqs. (2) and (3), the correction rate of the background knowledge with respect to "go better" (or "go worse") can be

Table 8 Rules of class N

Rules(SupportNUM > 10)	Support
PLT(N$\prec$) $\wedge$ DBIL(H$\succ$) $\wedge$ GPT(VH$\succ$)	27/30=90%
WBC(L$\succ$) $\wedge$ GOT(VH$\succ$) $\wedge$ GPT(VH$\succ$)	20/22=90%
PLT(VL$\succ$) $\wedge$ GOT(VH$\succ$) $\wedge$ GPT(VH$\succ$)	16/18=88%
PLT(L$\succ$) $\wedge$ TP(H$\prec$) $\wedge$ GOT(VH$\prec$)	15/16=93%
DBIL(H$\succ$) $\wedge$ GPT(VH$\succ$)	14/17=82%
HGB(N$\prec$) $\wedge$ WBC(L$\succ$) $\wedge$ GOT(VH$\succ$)	16/16=100%
DBIL(H$\succ$) $\wedge$ GOT(H$\succ$)	14/16=87%
DBIL(H$\succ$) $\wedge$ GOT(H$\succ$) $\wedge$ GPT(VH$\succ$)	12/15=80%
ALB(L$\prec$) $\wedge$ WBC(L$\succ$) $\wedge$ GPT(VH$\succ$)	10/12=83%
...	...

obtained as shown in Table 9.

$$att_{\text{pos}} = \frac{\#ATT_{\succ,\succ}}{\#ATT_{\succ,\succ} + \#ATT_{\succ,\prec}} \qquad (2)$$

$$att_{\text{neg}} = \frac{\#ATT_{\prec,\prec}}{\#ATT_{\prec,\succ} + \#ATT_{\prec,\prec}} \qquad (3)$$

where $\#ATT$ is the number of different attribute values of attribute ATT in the ordered information table; $\succ,\succ$ denotes that the attribute value is "go better" and the patient is cured; $\succ,\prec$ denotes that the attribute value is "go better" but the patient is not cured; $\prec,\prec$ denotes that the attribute value is "go worse" and the patient is not cured; and $\prec,\succ$ denotes that the attribute value is "go worse" but the patient is cured.

Table 9 The correction rate of the background knowledge

Attribute	att_{pos}	att_{neg}
ALB	58.9%	67.2%
CHE	54.7%	70.6%
D-BIL	26.8%	35.3%
GOT	34.2%	47.3%
GPT	36.7%	47.8%
HGB	78.0%	81.4%
I-BIL	32.5%	53.1%
PLT	36.6%	49.5%
T-BIL	44.0%	47.1%
T-CHO	30.6%	48.3%
TP	62.7%	78.3%
WBC	9.5%	30.5%

The higher correction rate of the background knowledge (i.e. TP and HGB) can be explained that the background knowledge is consistent with the specific characteristics of the real collected data. On the contrary, the lower correction rate (i.e. WBC) may mean that the order relation given by an expert may not suitable for the specific data analysis. In this case, the order relation as common background knowledge needs to be adjusted according to specific characteristics of the real data such as the distribution and clusters of the real data. How to adjust the order relation is an important ongoing work.

3.2 Rule Discovery by LOI Method 2

Generally, for each inspection data, the normal range exists. However, if the condition attribute value is devided into normal and unusual, it is difficult to discover useful rules for our hepatitis patient data. Therefore, the inspection data are divided into sections by the following method, and the condition attributes are symbolized by these sections.

(1) Use the average value of the inspection data within one year before IFN medication of each patient as the attribute value;
(2) Divide each attribute into 20 sections, and then, merge the adjacent sections if the class distributions are similar;
(3) Symbolize each section.

The result is shown as follows.

Att	1	2	3	4	5	6	7
ALB	104	136	151	159	182	198	260
CHE	3	200	250	299	373	496	
HGB	3.2	3.8	4.1	4.4	4.8		
PLT	5.7	7.3	7.5	8.0	9		
T-BIL	0.4	0.63	0.86	1.09	2.7		
I-BIL	0.03	0.14	0.31	1.15			
D-BIL	0.31	0.50	0.63	1.6			
T-CHO	56.3	120	167	183	215	374	
TP	9.5	13.2	14.0	14.3	14.7	15.8	16.9
WBC	3.1	4.0	4.8	5.4	6.3	6.9	8.9
GOT	30	88	174	231	289	605	
GPT	24	55	101	192	330		

3.2.1 Background Knowledge

Generally speaking, aggravation of hepatitis makes some inspection values going up or falling down. However, there is no background knowledge about the relation of the effect of IFN and the inspection values. So that, as the order relation, the ascent order or descent order of attributes are used.

Condition attribute: $\infty \succ 0$
Class: $R \succ N$

3.2.2 Results

The ordered information table is created by comparing every records. Therefore, the table becomes very large, and the number of the rules discovered also increases very much. Therefore, analysis and integration of rules are needed in the stage of post-processing. Table 10 shows the results.

3.2.3 Post-processing and Evaluation

In LOI method 2, the ranges of attribute value are included in the discovered rules. It is possible to look for a tendency by comparing the range of the attribute. For example, for the patients with ALB value in (198, 260], the possibility of healing up is low; and for the patients with ALB in (136, 151] the possibility of

Table 10 Discovered rules

ID	Rules of class R	Accuracy
001	D-BIL(0.31, 0.5]∧ T-CHO(183, 215]	72 / 90=80%
002	PLT(8, 9]∧ D-BIL(0.31, 0.5]	72 / 72=100%
003	ALB(136, 151]∧ I-BIL(0.03.0.14]	56 / 56=100%
...	...	...

ID	Rules of class N	Accuracy
004	PLT(5.7, 7.5]∧ GPT(24, 55]	462 / 650=71%
005	ALB(198, 260]	306 / 420=72%
006	D-BIL(0.50, 0.63]∧ GPT(24, 55]	210 / 210=100%
007	D-BIL(0.50, 0.63]∧ GOT(30, 88]	182 / 182=100%
008	WBC(6.9, 8.9]	110 / 156=70%
...	...	...

healing up is high. These are verified by the original data: in class R (healing up), 60% patients' ALB values are in (136, 151] and in class N (not healing up), 85% patients' ALB values are in (198, 260]. Thus, the discovered rule can be said to be right also in the original data.

3.2.4 Speculation

As the hepatitis data, if no background knowledge about the order relation of attribute values can be used, the attribute values can be ordered in ascent or descent at first. After the rule discovery on this ordered information, the order relation can be adjusted more rationally by analyzing the rules. If such a process is repeated, it is sure that the better order relation can be found, and the better rules can be discovered.

4. Conclusions and Remarks

Peculiarity represents a new interpretation of interestingness, an important notion long identified in data mining. Peculiarity, unexpected relationships/rules may be hidden in a relatively small number of data. *Peculiarity rules* are a typical regularity hidden in many scientific, statistical, medical, and transaction databases. They may be difficult to find by applying the standard association rule mining method, due to the requirement of large support.

We presented a multi-aspect mining approach in a multi-phase, multi-aspect hepatitis data analysis process. Both pre-processing and post-processing steps are important before/after using data mining agents. Informed knowledge discovery in real-world hepatitis data needs to use background knowledge obtained from medical docotors to guide the multi-phase discovery process such as pre-processing, rule mining, and post-processing, towards finding interesting and novel rules/features hidden in data.

Our methodology and experimental results show that the perspective of doctors will be changed from a single type of experimental data analysis towards a holistic view, by using our multi-aspect mining approach in which various data mining agents are used in a distributed cooperative mode in the spiral discovery process.

References

1) Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. "From Data Mining to Knowledge Discovery: an Overview", *Advances in Knowledge Discovery and Data Mining*, MIT Press (1996) 1-36.

2) Sai, Y., Yao, Y.Y., and Zhong, N. "Data Analysis and Mining in Ordered Information Tables", *Proc. 2001 IEEE International Conference on Data Mining (ICDM'01)*, IEEE Computer Society Press (2001) 497-504.

3) Yokoi, H., Hirano, S., Takabayashi, K., Tsumoto, S., and Satomura, Y. "Active Mining in Medicine: A Chronic Hepatitis Case – Towards Knowledge Discovery in Hospital Information Systems", *Journal of Japanese Society for Artificial Intelligence*, Vol.17, No.5 (2002) 622-628 (in Japanese).

4) Zhong, N. and Ohsuga, S. "Toward A Multi-Strategy and Cooperative Discovery System", *Proc. First Int. Conf. on Knowledge Discovery and Data Mining (KDD'95)*, AAAI Press (1995) 337-342.

5) Zhong, N., Dong, J.Z., and Ohsuga, S. "Rule Discovery in Medical Data by GDT-RS", Special Issue on Comparison and Evaluation of KDD Methods with Common Medical Datasets, *Journal of Japanese Society for Artificial Intelligence*, Vol.15, No.5 (2000) 774-781 (in Japanese).

6) Zhong, N., Yao, Y.Y., Dong, J.Z., and Ohsuga, S. "Gastric Cancer Data Mining with Ordered Information", *Rough Sets and Current Trends in Computing*, LNAI 2475, Springer (2002) 467-478.

7) Greco, S., Mastarazzo, B., and Slowinski, R. "Dominance-Based Rough Set Approach to Knowledge Discovery (I):General Perspective", *Intelligent Technologies for Information Analysis*, Springer (2004) 513-548.

8) Greco, S., Mastarazzo, B., and Slowinski, R. "Dominance-Based Rough Set Approach to Knowledge Discovery (II):Extensions and Applications", *Intelligent Technologies for Information Analysis*, Springer (2004) 553-607.