Evaluating an Integrated Time-Series Data Mining Environment
- A Case Study on a Chronic Hepatitis Data Mining -

Hidenao Abe, Miho Ohsaki, Hideto Yokoi, and Takahira Yamaguchi

Department of Medical Informatics, Shimane University, School of Medicine
abe@med.shimane-u.ac.jp,
Faculty of Engineering, Doshisa University
mohsaki@mail.doshisha.ac.jp,
Department of Medical Informatics, Kagawa University Hospital
yokoi@med.kagawa-u.ac.jp,
Faculty of Science and Technology, Keio University
yamaguti@ae.keio.ac.jp

In this paper, we present the implementation of an integrated time-series data mining environment. Time-series data mining is one of key issues to get useful knowledge from databases. However, users often face difficulties during such time-series data mining process for data pre-processing method selection/construction, mining algorithm selection, and post-processing to refine the data mining process as shown in other data mining processes. It is needed to develop a time-series data mining environment based on systematic analysis of the process. To get more valuable rules for domain experts from a time-series data mining process, we have designed an environment which integrates time-series pattern extraction methods, rule induction methods and rule evaluation methods with active human-system interaction. After implementing this environment, we have done a case study to mine time-series rules from blood/urine biochemical test database on chronic hepatitis patients. Then a physician has evaluated and refined his hypothesis on this environment. We discuss the availability of how much support to mine interesting knowledge for an expert.

1 Introduction

In recent years, KDD (Knowledge Discovery in Databases)[3] has been known as a process to extract useful knowledge from databases. In the research field of KDD, 'Time-Series Data Mining' is one of important issues to mine useful knowledge such as patterns, rules, and structured descriptions for a domain expert. Although time-series data mining can find out useful knowledge in given data, it is difficult to find out such knowledge without cooperation among data miner, system developers, and domain experts.

Besides, EBM (Evidence Based Medicine) has been widely recognized as a new medical topic to care each patient with the conscientious, explicit and judicious use of

current best evidence, integrating individual clinical expertise with the best available external clinical evidence from systematic research and the patient's unique values and circumstances. These evidences have been often found out in clinical test databases, which are stored on HIS (Hospital Information Systems). Looking at such findings, time-series rules are one kind of important medical evidences related to clinical courses of patients. However, it is difficult to find out such evidences systematically. Medical researchers need some systematic method to find out these evidences faster.

To above problems, we have developed a time-series data mining environment, which can apply medical data mining to find out medical evidences systematically, considering cooperation among data miners, system developers, and medical experts. Through a case study with a chronic hepatitis database, we have identified the procedures, which have been needed to execute time-series data mining cooperatively with active human-system interaction. With this analysis, we have developed a time-series data mining environment, integrating time-series pattern extraction, rule induction, and rule evaluation through visualization/evaluation/operation interfaces.

In this paper, we present an implementation of the integrated time-series data mining environment. Then we discuss about the process of development of system, evaluating with a case study of time-series rule mining on chronic hepatitis dataset.

2 Related Work

Many efforts have been done to analyze time-series data at the field of pattern recognitions. Statistical methods such as autoregressive model and ARIMA (AutoRegressive Integrated Moving Average) have been developed to analyze time-series data, which have linearity, periodicity, and equalized sampling rate. As signal processing methods, Fourier transform, Wavelet, and fractal analysis method have been also developed to analyze such well formed time-series data. These methods based on mathematic models restrict input data, which are well sampled. However, time-series data include ill-formed data such as clinical test data of chronic disease patients, purchase data of identified customers, and financial data based on social events. To analyze these ill-formed time-series data, we take another time-series data analysis method such as DTW (Dynamic Time Wrapping)[1], time-series clustering with multiscale matching[5], and finding Motif based on PAA (Piecewise Approximation Aggregation)[6].

As the one of the methods to find out useful knowledge depended on time-series, time-series/temporal rule induction methods such as Das's framework[2] have been developed. We can extract time-series rules in which representative patterns are expressed as closes of their antecedent and consequent with this method.

To succeed in a KDD process, human-system interaction especially need at data pre-processing and post-processing of mined result. Current time-series data mining frameworks focus on not human-system interaction but automatic processing. We introduce human-system interaction for data pre-processing and post-processing of mined result with visualization and evaluation interfaces for patterns and if-then rules based on time-series patterns.

3 An integrated time-series data mining environment

Our time-series data mining environment needs time-series data as input. Output rules are if-then rules, which have time-series patterns or/and ordinal closes, which are allowed to represent $A=x$, $A<=y$, and $A>z$ as their antecedent and consequent, depending on selected rule induction algorithm by a user. Fig. 1 illustrates a typical output it-then rule visualized with our time-series data mining environment.

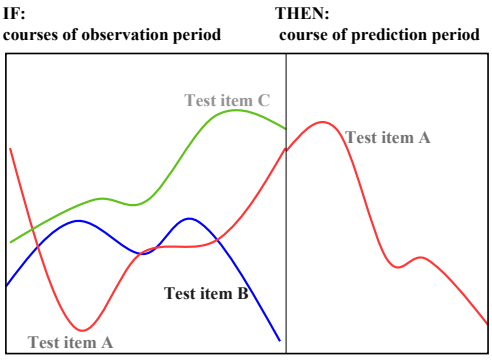

Fig. 1. Typical output if-then rule, which consists of patterns both its antecedent and its consequent.

Our integrated time-series data mining environment combines the following measure functional components: time-series data pre-processing, mining, post-processing for mined results, and other database operators to validate data and results of every phase. The component structure of this environment illustrates as Fig. 2.

With this environment, we aim the following efforts for each agent:

1. Developing and improving time-series data mining procedures for system developers
2. Collaborative data processing and rule induction for data miners
3. Active evaluation and interaction for domain experts

Since we have standardized input/output data formats, data miners and domain experts can execute different algorithms/methods in each procedure seamlessly. They can execute these procedures on graphical human-system interfaces, discussing each other. Beside, system developers can connect new or improved method for a procedure separately. Only following input/output data formats, system developers can also connect a complex sub-system, which selects a proper algorithm/method to the procedure before executing it. If a algorithm/method lacks for a procedure, they are only needed to develop its wrapper to connect the procedure, because each procedure assumes plug-in modules in this environment.

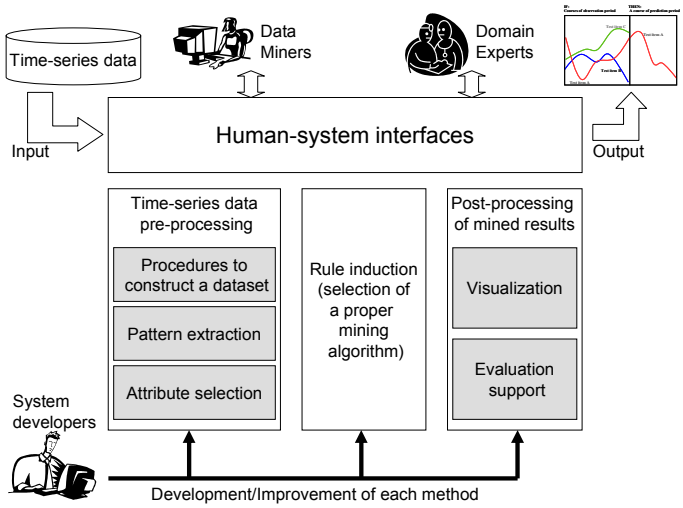

Fig. 2. The component structure of the integrated time-series data mining environment.

To implement the environment, we have analyzed time-series data mining frameworks. Then we have identified procedures for pattern extraction as data pre-processing, rule induction as mining, and evaluation of rules with visualized rule as post-processing of mined result. The system provides these procedures as commands for users. At the same time, we have designed graphical interfaces, which include data processing, validation for patterns on elemental sequences, and rule visualization as graphs. Fig. 3 shows us a typical system flow of this time-series data mining environment.

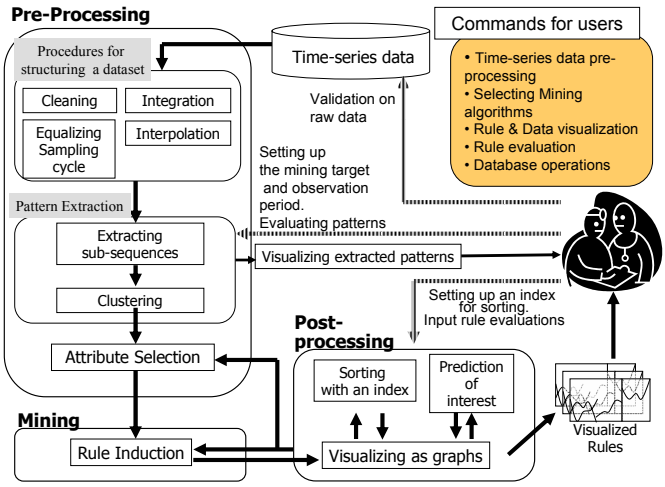

Fig. 3. A system flow view of the integrated time-series data mining environment.

3.2 Details of procedures to mine time-series rules

We have identified procedures for time-series data mining as follows:

Data pre-processing
- pre-processing for data construction
- time-series pattern extraction
- attribute selection

Mining
- rule induction

Post-processing of mined results
- visualizing mined rule
- rule selection
- supporting rule evaluation

Other database procedures
- selection with conditions
- join

As data pre-processing procedures, pre-processing for data construction procedures include data cleaning, equalizing sampling rate, interpolation, and filtering irrelevant data. Since these procedures are almost manual procedures, they strongly depend on given time-series data and a purpose of the mining process. Time-series pattern extraction procedures include determining the period of sub-sequences and finding representative sequences with a clustering algorithm. We have taken our original pattern extraction method[11] for the case study described in Section IV. Attribute selection procedures are done by selecting relevant attributes manually or using attribute selection algorithms[7].

At mining phase, we should choose a proper rule induction algorithm with some criterion. There are so many rule induction algorithms such as Version Space[9], AQ15[8], C4.5 rules[12], and any other algorithm. To support this choice, we have developed a tool to construct a proper mining application based on constructive meta-learning called CAMLET. However, we have taken PART[4] implemented in Weka[14] in the case study to evaluate improvement of our pattern extraction algorithm.

To validate mined rules correctly, users need readability and ease for understand about mined results. We have taken 39 objective rule evaluation indexes to select mined rules[10], visualizing and sorting them depended on users' interest. Although these two procedures are passive support from a viewpoint of the system, we have also identified active system reaction with prediction of user evaluation based on objective rule evaluation indexes and human evaluations.

Other database procedures are used to make target data for a data mining process.

Since the environment has been designed based on open architecture, these procedures have been able to develop separately. To connect each procedure, we have only defined input/output data format.

3.2 Designing interfaces to team up domain experts and data miners

We have designed user interfaces as shown in Fig. 4.

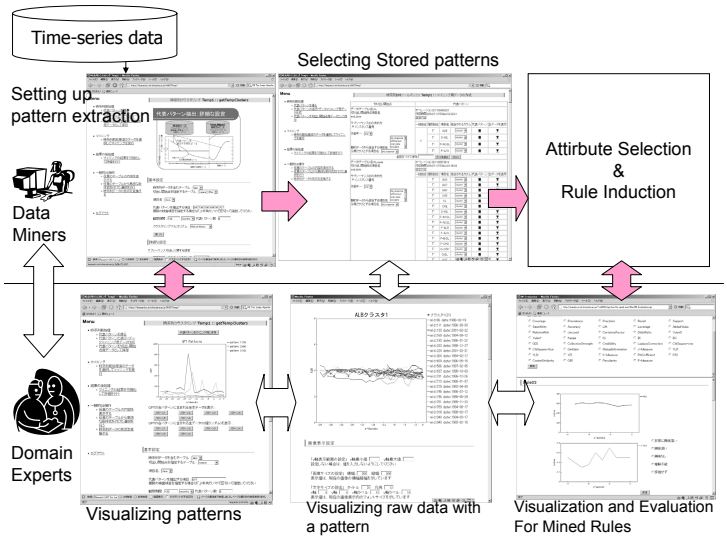

Fig. 4. System-human interfaces of the integrated time-series data mining environment.

For data miners, setting pattern extraction and selecting stored patterns to construct a dataset have been assigned. Pattern extraction setting interfaces are separated two phases. In first phase, a data miner set up basic settings such as an input time-series data set, a period for patterns, a target dataset for the purpose of the data mining process, and a clustering algorithm for pattern extraction. Then other detailed parameters such as the period for equalizing sampling rate and maximum interpolation period are set up at detailed setting phase. A data miner also set up detailed parameters of selected clustering algorithm at this phase. Besides, visualization of patterns, visualization of given raw data included selected pattern, and visualization of mined rules have been assigned for domain experts. Each visualization interface can be switched each other depended on their interest.

4 A case study with a chronic hepatitis database

To evaluate the implementation of integrated time-series data mining environment described in Section 3, we have done a case study to mine interesting rules for domain expert with this system.

In this medical KDD, we have taken a clinical test dataset from Chiba University Hospital[13], which includes clinical blood and urine test data on chronic hepatitis B and C. Their test intervals are not only daily, weekly, or monthly but also randomized intervals depended on patients' statements. This dataset has approximately 1.6 million

records. Each record consists of MID, test date, test item name, and test result. This dataset includes 771 patients, who have up to 20 years as his/her treatment period. We have taken IFN (interferon) treatment results of 195 patients as the target of rules, which represent as if-then classification rules. Each instance of this target data consists of MID, start date of IFN treatment, end date of it, his/her treatment result decided with GPT (ALT) values, and his/her treatment result based on virus markers. We have set up observation periods before finishing IFN treatment.

4.1 Phase1: Rousing new hypothesis in mind

We presented extracted patterns about some observation periods to a physician. He noticed distinguishing patterns on ALB as shown 'pattern 1' and 'pattern 2' in Fig. 5.

He validated this pattern with each sequence of patients. He said that 'pattern 0' indicates typical course while in IFN treatment period. However, the other patterns show remarkable courses based on his knowledge. Since he thought 'pattern 2' shows what patients included in this pattern had adverse reactions at the end of IFN treatment period, he interested in their treatment results. On the other hand, 'pattern 1' indicates less reaction to IFN treatment at the end of the period. He roused a hypothesis that patients, who indicate good reaction after IFN treatment, have moderate adverse reaction during IFN treatment.

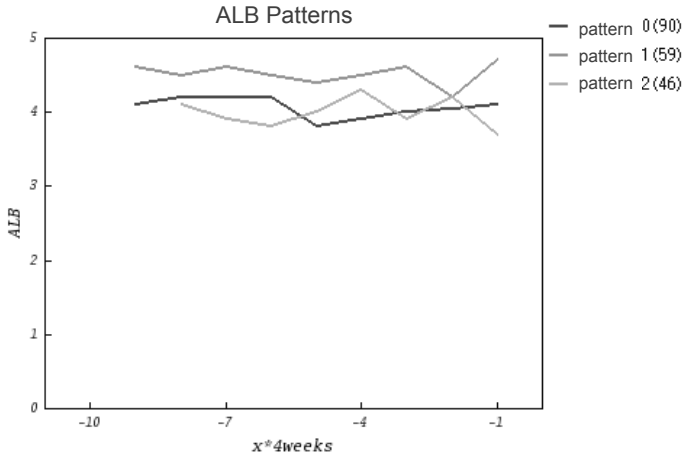

Fig. 5. Representative patterns on ALB while IFN treatment. '-1' on X axis means finishing date of IFN treatment for each patient.

4.2 Phase2: Ensuring expert's hypothesis

To endure the hypothesis, we have extracted patterns of 0.8 years (40 weeks) as observation period for 40 test items. After joining these patterns as attributes of the dataset, we have induced if-then rules with the dataset.

Then the patterns and rules are evaluated by the physician, visualizing patterns, rules and patient data on demand. He interested in the rule shown in Fig. 6. The right hand end of each pattern means the end date of IFN treatment for the patients. The class values were labeled with GPT values after IFN treatment by the physician. "Response" means what the treatment succeeded. Besides, "Non-Response" means what the treatment failed. These rules shows one of the reasons why IFN treatment success or not.

Although the patients included in the right rule failed with of accuracy, the patients included in the left rule succeeded in IFN treatment with of accuracy. The difference is one of the patterns of RBC(Red Blood-cell Count). This pattern shows trend of anemia, which is one of typical adverse reactions of IFN treatment. Validating row data with the graphical interface shown in Fig.1, these patients included in this pattern would be anemia, he said.

After evaluating all of mined if-then rules, he noticed that some combination of patterns show good result after IFN treatment but other combinations of patterns show no good result. He has ensured his hypothesis with this mining result.

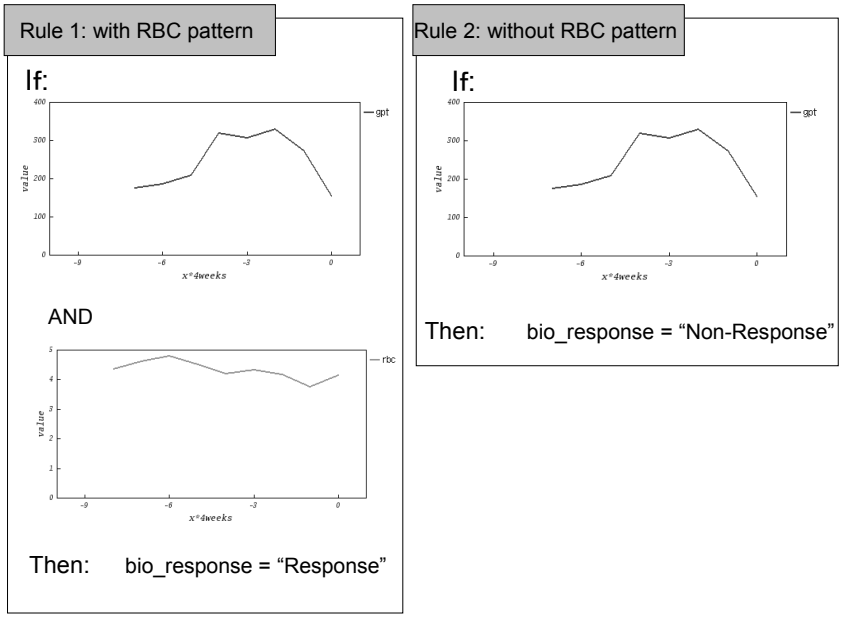

Fig. 6. Typical pair of interesting rules, having opposite class values. Only one pattern is deleted from the antecedent of the left rule.

At the same time, we have also developed pattern extraction method based on irregular sampling and quantization. Table 1 shows us numbers of expert's

evaluations. The number of 'very interesting' increases after improvement. This result has been caused by one of the reasons why the improvement of pattern extraction algorithm extracts patterns, which can be suitable to his knowledge about patients' course.

Table 1. Numbers of evaluated rules from before and after implovement of the pattern extraction algorithm.

Evaluation Labels	Before improvement	After improvement
Very Interesting	4	15
Interesting	7	5
Fair	15	11
Difficult to understand	5	1
TOTAL	31	32

5 Conclusion

We have designed and implemented a time-series data mining environment, which integrates time-series pattern extraction, rule induction, and rule evaluation with active interaction on graphical interfaces. As the result of the case study on a chronic hepatitis dataset, we have succeeded in rousing and ensuring hypothesis about IFN treatment in the mind of a medical expert. We have also developed the pattern extraction method, cooperating with the data mining process. This case study also shows the availability of this environment as a development infrastructure to develop both time-series data mining method and specific data mining process.

As the result of the case study on chronic hepatitis, this environment mines valuable time-series rules based on both data and knowledge of domain experts. With these models, we will be able to predict some risks such as fails of a treatment in medical domain in easier, faster and more properly.

Although we have not tried to select proper algorithms for the attribute selection procedure and the mining procedure, it is also able to connect subsystems for selecting a proper attribute selection algorithm and a proper rule induction algorithm to this environment. Then we will construct time-series data mining application based on active user reactions with rule evaluation interface.

We also plan to apply this environment for other time-series data mining on other domain such as treading, customer purchasing, and so forth.

References

1. D. J. Berndt and J. Clifford, "Using dynamic time wrapping to find patterns in time series", *in Proc. of AAAI Workshop on Knowledge Discovery in Databases*, 1994, pp.359-370.

2. G. Das, L. King-Ip, M. Heikki, G. Renganathan, and P. Smyth, "Rule Discovery from Time Series", *in Proc. of International Conference on Knowledge Discovery and Data Mining*, 1998, pp.16-22.
3. U. M. Fayyad, G. Piatcktsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview", *Advances in Knowledge Discovery and Data Mining,* AAAI Press / The MIT Press, CA, 1996, pp.1-34.
4. E. Frank, I. H. Witten, "Generating accurate rule sets without global optimization", *in Proc. of the Fifteenth International Conference on Machine Learning*, 1998, pp.144-151
5. S. Hirano and S. Tsumoto, "Mining Similar Temporal Patterns in Long Time-Series Data and Its Application to Medicine", *in Proc. of the 2002 IEEE International Conference on Data Mining*, 2002, pp.219-226.
6. J. Lin, E. Keogh, S. Lonardi, and P. Patel, "Finding Motifs in Time Series", *in Proc. of Workshop on Temporal Data Mining*, 2002, pp.53-68.
7. H. Liu and H. Motoda, "Feature selection for knowledge discovery and data mining", Kluwer Academic Publishers, 1998.
8. R. Michalski, I. Mozetic, J. Hong, and N. Lavrac., "The AQ15 Inductive Leaning System: An Overview and Experiments", *Reports of Machine Leaning and Inference Laboratory*, MLI-86-6, George Maseon University, 1986.
9. T. M. Mitchell, "Generalization as Search", *Artificial Intelligence*, 18(2), 1982, pp.203-226.
10. M. Ohsaki, S. Kitaguchi, K. Okamoto, H. Yokoi, and T. Yamaguchi, "Evaluation of Rule Interestingness Measures with a Clinical Dataset on Hepatitis", *in Proc. of ECML/PKDD 2004*, LNAI3202, 2004, pp.362-373.
11. M.Ohsaki, H. Abe, S. Kitaguchi, S. Kume, H. Yokoi, and T. Yamaguchi, "Development and Evaluation of an Integrated Time-Series KDD Environment - A Case Study of Medical KDD on Hepatitis-", *Joint Workshop of Vietnamese Society of Artificial Intelligence, SIGKBS-JSAI, ICS-IPSJ and IEICE-SIGAI on Active Mining*, 2004, No.23.
12. J. R. Quinlan, "Programs for Machine Learning", Morgan Kaufmann, 1992.
13. S. Tsumoto,
Hepatitis Dataset for Discovery Challenge, http://lisp.vse.cz/challenge/ecmlpkdd2002/index.html, 2002.
14. I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", Morgan Kaufmann, San Francisco, 2000.