

Development and Evaluation of an Integrated Time-Series KDD Environment — A Case Study of Medical KDD on Hepatitis —

MIHO OHSAKI,^{†1} HIDENAO ABE,^{†2} SHINYA KITAGUCHI,^{†3}
SHUNJI KUME,^{†3} HIDETO YOKOI^{†4} and TAKAHIRA YAMAGUCHI ^{†5}

We have been conducting case studies for clinical dataset on hepatitis, a common chronic disease, to realize a comprehensive KDD environment for medical time series data. We here propose a new pattern extraction algorithm and embed the algorithm in our latest mining system. This paper reports the rules generated by the mining system and their evaluation results by a medical expert. It also reports the KDD methodology obtained through the case studies and how to extend the mining system to a comprehensive KDD environment.

1. Introduction

Much attention has been given to time series data mining, and a lot of algorithms have been proposed to cluster, classify, and estimate time series data¹⁾. It has however not been studied enough to realize an environment to comprehensively support Knowledge Discovery from Databases (KDD) especially for time series data through human-system interaction.

On the other hand, medical data mining has been also remarkable due to its scientific and social contribution²⁾. The research topics in medical data mining are gradually moving toward the discovery of deeper knowledge that explains complex phenomena such as chronic disease symptom and the support for medical experts to generate new knowledge by polishing up mined rules³⁾.

With these backgrounds, we have been conducting case studies to realize a KDD environment for medical time series data that integrates and works preprocessing, mining, post-processing, user interface, and a human user, cooperatively. What we have done until our last study are: the development and improvement of a mining system consisting of pattern extraction and pattern combination rule generation, the discovery of rules expressing symptom change by applying the system to a clinical dataset on hepatitis⁴⁾, and the rule evaluation

by a medical expert.

This paper reports the latest mining system improved in our current study, summarizes the know-how and methodology of KDD for medical time series data, and discusses how to extend the mining system to an integrated time-series KDD environment.

2. Previous Mining System

In our previous study⁵⁾, we developed and improved a mining system shown in **Fig. 1** based on the typical framework of time series data mining that extracts patterns by clustering and generates pattern combination rules by decision tree learning⁶⁾. We applied the system to a clinical dataset on hepatitis (the results of sequential medical tests on hepatitis that were clinically collected), obtained prognosis prediction rules, and visualized them as graphs. We designed the KDD process to repeat a set of the rule generation by a mining system and the rule evaluation by a medical expert twice for polishing up the obtained rules.

In the first mining, we data miners generated and presented many rules to a medical expert under the several conditions of the start point and observation term of patterns. The medical expert evaluated all rules and was interested in a few rules generated under the condition of the first time point of medical tests and 36 months as the start point and observation term of patterns, respectively. These rules inspired the medical expert to make a hypothesis, namely a seed of new medical knowledge: Contradict to medical common sense, GPT, which is an important medical test result to grasp hepati-

†1 Doshisha University

†2 Shimane University

†3 Shizuoka University

†4 Chiba University Hospital

†5 Keio University

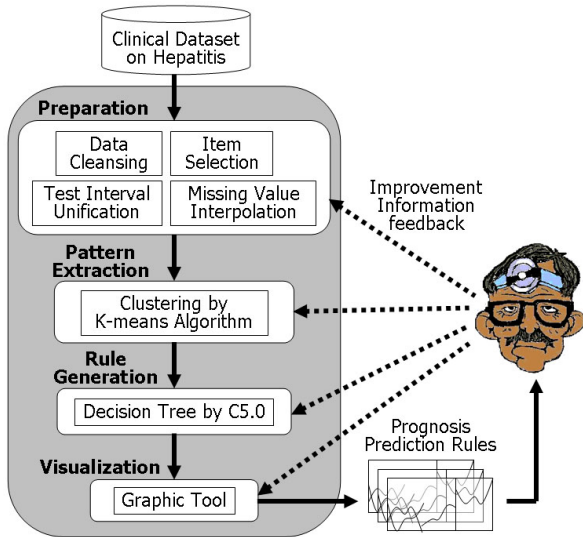


Fig. 1 Our previous mining system.

tis symptom, may change with a cycle of three years (See the upper side in Fig. 2).

We then extended the observation term to 66 months and conducted the second mining based on the comments by the medical expert to improve the system and polish up the rules. A few rules in the second mining supported him to confirm the hypothesis and enhanced its re-

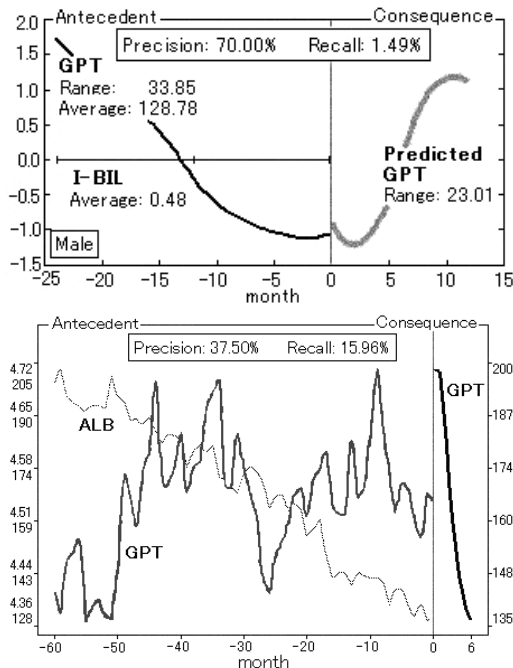


Fig. 2 Examples of highly valued rules in first (upper) and second mining (lower).

liability (See the lower side in Fig. 2). The raw data supporting the rules added information that GPT has cyclic peaks rather than sinusoidal change to the hypothesis. The medical expert commented that the GPT peaks with a cycle of three years will be an important medical knowledge if it is verified by a well-controlled medical experiment. He also commented that too much summarizing of patterns by the mining system rather disturbs the medical insight into symptom by medical experts.

As a consequence of the previous study, our mining system for medical time series data was able to discover new knowledge with a certain amount of real medical worth. However, the following issues are left: the algorithm improvement of pattern extraction and rule generation, the refinement of knowledge on cyclic GPT peaks, and the systemization and semi-automation of KDD process by extending the mining system to an integrated time-series KDD environment. We try to solve the issues in our current study; the trials for the first, second, and third issues are shown in 3, 4, and 5, respectively.

3. Current Mining System

We propose an algorithm of pattern extraction and redesign the mining system here. Although we used EM algorithm for pattern extraction in the first mining, it turned out not to be easily understandable and adjustable for medical experts. We then used K-means algorithm without such disadvantages in the second mining. However, the organized clusters sometimes do not reflect the medically important features of patterns. The main reasons why K-means algorithm did not work well are that it was applied directly to raw data including noises and that it used a too simple distance measure, Euclidean distance.

We then propose a pattern extraction algorithm referring the idea of finding motifs⁷⁾. This algorithm aims to moderately and irregularly descritize raw data for the suppression of noise effects and the preservation of distinctive pattern features and to efficiently cluster the motifs found through the descritization. It consists of descritization (irregular resampling and conversion into motifs) and clustering (cluster generation and cluster merging).

3.1 Descritization of Our Pattern Extraction Algorithm

```

i = 1 ;
while i < Number_of_points_in_observation_term
    Calculate the gradient, G, between (i)-th and (i+1)-th points ;
    if Threshold_of_gradient <= the absolute value of G
        Regard ((i)+(i+1))/2 as the end of current window ;
        i++ ;
    end_while
Output the set of windows ;

```

Fig. 3 Irregular resampling.

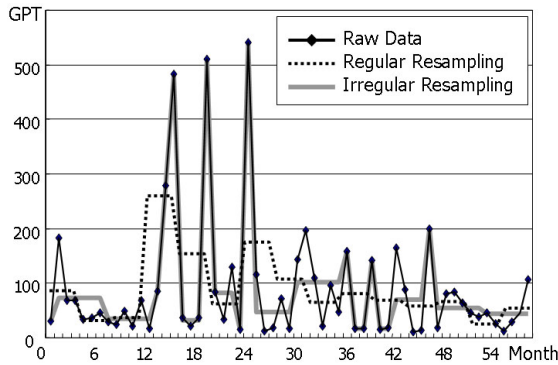


Fig. 4 One of the results of irregular resampling.

```

# There is NO amplitude normalization
# differently from Lin's original motif finding algorithm.

for i=1 to Number_of_windows
    Calculate the amplitude average, AA, of the points in i-th window ;
    for j = 1 to (i-1)
        if AA in i-th window == AA in j-th window
            Assign the same symbol in j-th window
            to the points in i-th window ;
            Break ;
        if no symbol was assigned
            Assign a new symbol to the points in i-th window ;
        end_for
    end_for
Output the symbol array as a motif ;

```

Fig. 5 Conversion into motifs.

Fig. 3 shows the procedure of irregular resampling that adaptively determines the start and end points of each window based on the gradient of points included in the window. We designed the calculation of gradient as simple as possible to preserve the local but distinctive change of amplitude and shorten the calculation time. The preset parameter of irregular

resampling is the threshold of gradient and was determined by trial and error. We applied our irregular resampling to the clinical dataset on hepatitis and compared the wave shape of raw data, the result of regular resampling, and that of irregular resampling. As shown in **Fig. 4**, it is obvious that irregular resampling preserves the wave shape of raw data better than regular resampling does.

```

# Initialize Clusters
Calculate the distances among the all motifs
in the set of motifs that do not belong to any clusters, L ;
Extract the set of motifs with the most neighborhood ones, M, from L ;
while M is not empty
    Generate a new cluster ;
    Extract a motif from M ;
    Put it on the root node of the new cluster
    and regard it as the representative one ;
end_while

# Grow Clusters
Calculate the distances between the all motifs in L and those in M ;
while L is not empty && the depth of cluster trees < Threshold_of_depth
    Extract the motif with the smallest distance from L ;
    Put it on the lower node of the nearest cluster ;
end_while
Output the clusters ;

```

Fig. 6 Cluster generation.

```

for the number of clusters = 1 to Threshold_number_of_clusters
    Search the smallest cluster, A ;
    Calculate the distance of A from the other clusters and
    find the nearest cluster, B ;
    Add A to B ;
    Reorganize the tree structure of B following
    the same strategy of "Grow Clusters in Cluster Generation" ;
end_for
Output the final clusters ;

```

Fig. 7 Cluster merging.

Fig. 5 shows the procedure of conversion into motifs that replaces the numeric values of amplitude with symbols. We designed the basic framework of our conversion into motifs as the same in Lin's algorithm⁷⁾. However, we dare not to include amplitude normalization that removes the effect of absolute amplitude level, since medical experts frequently place importance on absolute amplitude level. The conversion into motifs has no preset parameter. Note that the procedures in **Fig. 3** and **Fig. 5** are for one time series and repeatedly executed for all time series.

3.2 Clustering of Our Pattern Extraction Algorithm

Fig. 6 shows the procedure of cluster generation that organizes tree-structured clusters including motifs as their nodes and extracts the representative motifs of the clusters as patterns. The key points of our cluster generation are the same of many tree-structured clustering methods; the definition of distance among motifs and the setup of the threshold of tree depth. As a first step, we tentatively used the concordance rate of symbols as the distance and the depth for which at least two clusters are generated as the threshold of tree depth.

Fig. 7 shows the procedure of cluster merging that searches clusters with a small number of motifs and add them into other nearest large clusters. The preset parameter of cluster merging is the threshold number of clusters and able to be roughly determined since the actual number of clusters ends up being several depending on the distance among motifs and the threshold of tree depth.

3.3 Redesign of Mining System

We redesigned our mining system as shown in Fig. 8 based on the requirements by the medical expert and us data miners. The differences between the previous and current mining systems are the algorithm of pattern extraction and the presence of rule generation (Compare Fig. 1 and Fig. 8).

We thought that it is needed to examine whether generating prognosis prediction rules by decision tree learning is really desirable for medical experts to enhance their hypothesization or not. We also thought that it is simple but possibly useful to visualize and present the members in a cluster of a certain medical test and those in a cooccurring cluster of the other medical test as a graph on the demand of medical experts. This process is like pseudo association rule mining that places priority on the insight of each medical expert rather than rule accuracy.

4. Evaluation Experiment

We conducted an experiment to iteratively execute mining with our current mining system in front of the medical expert. The purposes of experiment were the performance evaluation of the proposed pattern extraction algorithm

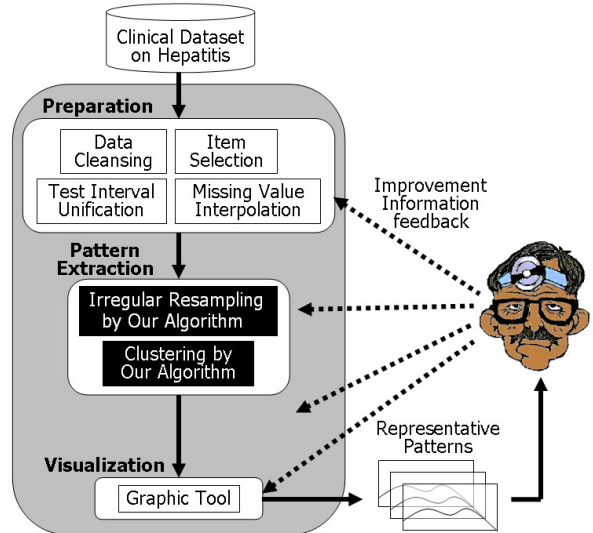


Fig. 8 Our current mining system.

and the hypothesis refinement and generation by the medical expert.

In the hypothesis refinement, the medical expert tried to refine the hypothesis of cyclic GPT peaks obtained in our previous study. He focused on the reason why cyclic GPT peaks occur. Therefore, we extracted distinctive GPT patterns including peaks with a term of 18 months and the subsequences of medical test results with a term of 18 months just before the start point of each distinctive GPT pattern using the mining system. We then presented the sets of a distinctive GPT pattern and a subsequence as graphs to the medical expert on his demand.

In the hypothesis generation, the medical expert tried to generate a new hypothesis on the therapy effect of interferon, which is a remedy of hepatitis. We extracted and presented the subsequences of medical test results with several terms just before and after the interferon therapy.

As experimental results, the medical expert said that the descritization of the pattern extraction algorithm actually preserved the medically important wave shape of raw data and that the clustering however did not work well. It is estimated that the definition of distance among motifs and the setup of the threshold of tree depth were not proper in cluster generation. Now, we are considering to utilize dynamic time warping instead of the cluster gen-

eration based on distance and tree depth.

He also said that the hypothesis refinement on GPT and the hypothesis generation on interferon did not succeed. Although the medical expert examined the subsequences of GPT, ALB, TTT, and ZTT, which are major medical test results, just before distinctive GPT patterns or just before and after interferon, remarkable symptom did not appear. It is difficult to identify the cause of this unsucces, in other words, to determine that the performance of mining system was too low or that the used dataset originally have no remarkable symptom. We should estimate the performance of each module in the mining system using artificial data and determine the cause in our future work.

5. Conclusions and Future Work

In this study, we proposed a new pattern extraction algorithm and improved our mining system for medical time series data by embedding the algorithm. We then applied the system to a clinical dataset on hepatitis and had a medical expert evaluate the obtained rules through an experiment.

Although the experimental results were not desirable, irregular resampling succeeded, and some information for system improvement was obtained. We will improve and confirm the pattern extraction process of mining system and verify the possibility of generalization of our mining system for other kinds of chronic disease.

Finally, we summarize the know-how and methodology of KDD for medical time series data through our previous and current case studies and discuss how to extend the mining system to an integrated time-series KDD environment. **Fig. 9** shows the roles of a medical expert and a mining system and the hypothesis stages. This summary made us notice that it is considerably important to reconstruct mining algorithms according to a hypothesis stage. We expect that the summary will help data miners and domain experts to conduct KDD for medical time series data.

Fig. 10 shows the extension of our mining system to an integrated time-series KDD environment. The drastic differences between them are that the meta learning by CAM-




	 Medical Expert KDD Scenario	 System Tuning Information	 Mining System Algorithms
Phase 1 Hypothesis Generation	Motivation Interest in the temporal change of important medical test results. Discovery GPT may change with a cycle of three years.	Start_point Various Term Various	Clustering by EM algorithm + Decision tree learning by C5.0
Phase 2 Hypothesis Confirmation	Motivation Interest in GPT change with a long term. Discovery GPT produces peaks with a cycle of three years.	Start_point Various Term 66 months	Clustering by K-means algorithm + Decision tree learning by C5.0
Phase 3 Hypothesis Refinement	Motivation Interest in the trigger of GPT cyclic peaks. Discovery Not yet.	Start_point Start and end of a peak Term 18 month before and after a peak	Clustering by our algorithm + Pseudo association rule mining

Fig. 9 Summary of our previous and current case studies.

LET⁸) and the user interface by interestingness measures⁹) (Compare Fig. 8 and Fig. 10). We expect that this environment will be able to semi-automatically reconstruct mining algorithms according to a hypothesis stage based on meta learning scheme.

Acknowledgments This research was partially supported by the Grant-in-Aid for Scientific Research on the Priority Area (B),13131205, by the Ministry of Education, Science, and Culture for Japan.

References

- 1) Keogh, E. and Kasetty S.: On The Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration, *ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pp. 102–111 (2002).
- 2) Cios, K. J. and Moore G. W.: Uniqueness of medical data mining, *Artificial Intelligence in Medicine*, Vol.26, No.1–2, pp. 1–24 (2002).
- 3) Motoda, H. (eds.): Active Mining, *IOS Press*, Amsterdam, Holland (2002).
- 4) Tsumoto, S.: Hepatitis Dataset for Discovery Challenge. *Web Page of Euro. Conf. on Principles and Practice of Knowledge Discovery in Databases*, <http://lisp.vse.cz/challenge/ecmlpkdd2002/index.html> (2002).
- 5) Ohsaki, M., Sato, Y., et. al: A Rule Discovery Support System for Sequential Medical Data, – In the Case Study of a Chronic Hepatitis Dataset –, *Int'l Workshop on Active Mining in IEEE Int'l Conf. on Data Mining*, pp. 97–102 (2002).
- 6) Das, G., King-Ip, L., et. al: Rule Discovery from Time Series. *Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 16–22 (1998).

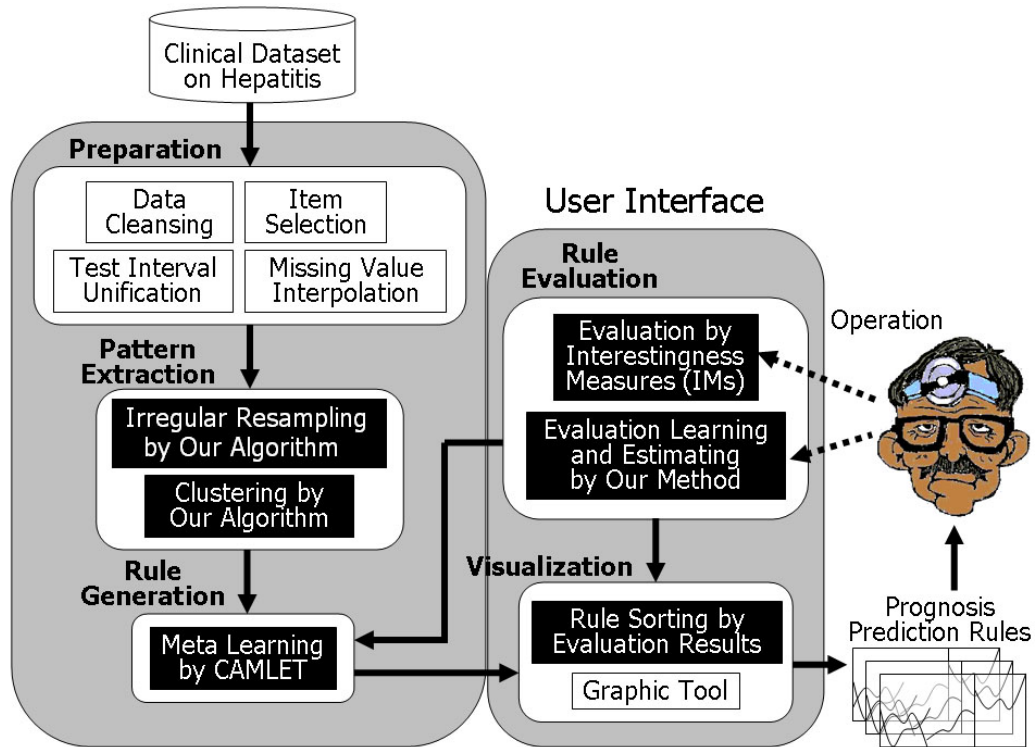


Fig. 10 Integrated time-series KDD environment.

- 7) Lin, J., Keogh, E., et. al: Finding Motifs in Time Series, *Int'l Workshop on Temporal Data Mining in ACM Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 53-68 (2002).
- 8) Abe, H. and Yamaguchi T.: Constructive Meta-Learning with Machine Learning Method Repository, *Lecture Note on Artificial Intelligence*, Vol.3029, pp. 502-511 (2004).
- 9) Ohsaki, M., Sato, Y., et. al: Comparison Between Objective Interestingness Measures And Real Human Interest In Medical Data Mining, *Lecture Note on Artificial Intelligence*, Vol.3029, pp. 1072-1081 (2004).