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

Hidenao Abe (Department of Medical Informatics, Shimane University, School of Medicine)
Miho Ohsaki (Faculty of Engineering, Doshisha University)
Hideto Yokoi (Department of Medical Informatics, Kagawa University Hospital)
Takahira Yamaguchi (Faculty of Science and Technology, Keio University)
Contents

- Background
- The Integrated Time-Series Data Mining Environment
- Case Study on Chronic Hepatitis Data Mining
- Conclusion
Background

- KDD (Knowledge Discovery in Databases) has been widely known as a powerful process to extract useful knowledge.
- Collaboration of data miners, domain experts and system developers is important to success a data mining process.
- Knowledge depended on time stream is useful to predict some risk in future.
Issues and Our approach

- Many DM tools only supply DM methods
- There are no systematic support to carry out time-series data mining processes.

- Systematic support with preparing data mining methods from systematic analysis
- Human-system interaction
## Map of our research

<table>
<thead>
<tr>
<th></th>
<th>Input time-series data</th>
<th>Mining approach</th>
<th>Post-processing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our approach</strong></td>
<td>ill-formed/well formed</td>
<td>Rule induction based on time-series patterns</td>
<td>Visualizing patterns as graphs, Active interaction</td>
</tr>
<tr>
<td><strong>Pattern extraction methods</strong></td>
<td>ill-formed/well formed</td>
<td>Particular pattern extraction algorithm</td>
<td>Visualizing patterns as graphs</td>
</tr>
<tr>
<td><strong>Statistical methods, Signal processing methods</strong></td>
<td>well formed</td>
<td>Particular time-series analysis method</td>
<td>(Visualizing as graphs)</td>
</tr>
</tbody>
</table>
Contents

- Background
- The Integrated Time-Series Data Mining Environment
- Case Study on Chronic Hepatitis Data Mining
- Conclusion
The Integrated Time-Series Data Mining Environment

- **Input**
  - ill-formed/well-formed time-series data

- **Output**
  - IF-THEN rule based on time-series patterns
Procedures to Mine Time-Series Rules

- Data pre-processing
  - Pre-processing for data construction
  - Time-series pattern extraction
  - Attribute selection

- Mining
  - Rule induction

- Post-processing of mined results
  - Visualizing mined rules
  - Rule selection
  - Rule evaluation support

- Other database procedures
  - Selection with conditions
  - Join
System Overview

Domain Experts

Data Miners

Time-series data

IF:
Courses of observation period
THEN:
A course of prediction period

Test item A
Test item C
Test item B
Test item A

Human-system interfaces

Input

Output

System developers

Development/Improvement of each method

Time-series data pre-processing

Procedures to construct a dataset

Pattern extraction

Attribute selection

Rule induction
(selection of a proper mining algorithm)

Post-processing of mined results

Visualization

Evaluation support

SystemOverview

Input

Output

Time-series data

Data Miners

Domain Experts

Human-system interfaces

Input

Output

Time-series data

Data Miners

Domain Experts

Human-system interfaces

Input

Output

Time-series data

Data Miners

Domain Experts

Human-system interfaces

Input

Output

Time-series data

Data Miners

Domain Experts

Human-system interfaces

Input

Output

Time-series data

Data Miners

Domain Experts

Human-system interfaces

Input

Output

Time-series data

Data Miners

Domain Experts

Human-system interfaces

Input

Output

Time-series data

Data Miners

Domain Experts

Human-system interfaces

Input

Output
System Flow

Pre-Processing

- Cleaning
- Integration
- Equalizing
- Sampling cycle
- Interpolation
- Procedures for structuring a dataset

Pattern Extraction

- Extracting sub-sequences
- Clustering
- Attribute Selection

Mining

- Rule Induction

Post-processing

- Sorting with an index
- Prediction of interest
- Visualizing as graphs

Commands for users

- Time-series data pre-processing
- Selecting Mining algorithms
- Rule & Data visualization
- Rule evaluation
- Database operations

Validation on raw data

Setting up the mining target and observation period.
Evaluating patterns

Setting up an index for sorting.
Input rule evaluations

Visualized Rules
Procedures to Mine Time-Series Rules

- Data pre-processing
  - Pre-processing for data construction
  - Time-series pattern extraction
  - Attribute selection

- Mining
  - Rule induction

- Post-processing of mined results
  - Visualizing mined rules
  - Rule selection
  - Rule evaluation support

- Other database procedures
  - Selection with conditions
  - Join
Data pre-processing for extracting time-series patterns from ill-formed data

- Pre-processing for data construction
  - Data cleaning, Integration of values, Equalizing sampling cycle, Interpolation

- Time-series pattern extraction
  - Extracting sub-sequences, Clustering (K-means, EM, our original pattern extraction algorithm)

- Attribute selection for a mining algorithm

![Diagram showing the flow of data pre-processing stages from ill-formed time-series data to a dataset for a mining algorithm]
Post-processing with active human-system interaction

- Visualize mined rules
  - Visualizing text rules as graphical rules based on patterns

- Rule selection
  - Sorting graphical rules with indexes called objective measurement values

- Rule evaluation support
  - Predicting users’ interest with re-using evaluated results
Graphical interfaces of the integrated time-series data mining environment

Setting up pattern extraction

Selecting stored patterns & generating the dataset

Attribute Selection & Rule Induction

Data Miners

Domain Experts

Visualizing patterns

Visualizing raw data included in a pattern

Visualization and Evaluation For Mined Rules

Graphical interfaces of the integrated time-series data mining environment
Contents

- Background
- The Integrated Time-Series Data Mining Environment
- Case Study on Chronic Hepatitis Data Mining
- Conclusion
Description of the chronic hepatitis data mining

- Blood and urine laboratory test data
  - 1.9 million records
  - 965 test items
  - 771 patients (Hepatitis type B and C)
    - Up to 20 years for each sequence

- To find out risks related to IFN treatment results
  - 195 patients
  - Decided with GPT(ALT) values after finishing his/her IFN treatment
  - The risk means failures of his/her IFN treatment
Phase 1:
Focusing expert’s interest

- Pattern extraction about ALB values during IFN treatment

- Data miners:
  - Setting up observation period, pattern extraction algorithm and its parameters
  - Taking the original pattern extraction algorithm based on irregular sampling to calculate similarities between two sub-sequences

- Physician:
  - Evaluating patterns with visualized patterns and raw data included in interesting pattern as graphs on the interfaces
Pattern 2 indicates adverse reaction of IFN treatment. On the other hand, pattern 1 indicates less reaction than typical pattern (pattern 0). Does the degree of adverse reaction relate to treatment result?
Phase 2:
Ensuring expert’s hypothesis

- Inducing rules to predict IFN treatment results from patterns of treatment periods

- Data miners:
  - Setting up observation period, pattern extraction algorithm and its parameters
  - Taking the original pattern extraction algorithm based on irregular sampling to calculate similarities between two sub-sequences
  - Selecting rule induction algorithm -> PART

- Physician:
  - Evaluating rules with visualized rules, patterns and raw data included in interesting pattern as graphs on the interfaces
Representative rules having opposite IFN treatment results

**Rule 1: with RBC pattern**

**IF:** GPT pattern = 3  
**AND** RBC pattern = 1  
**THEN:** bio_response = “Response”

**Rule 2: without RBC pattern**

**IF:** GPT pattern = 3  
**THEN:** bio_response = “Non-Response”

Trend of anemia while IFN treatment (anemia is a major adverse reaction)
Improvement of pattern extraction algorithm

- System Developers:
  - Improving the algorithm to calculate similarities between two sub-sequences
Evaluation results of the improvement of the pattern extraction algorithm

<table>
<thead>
<tr>
<th>Evaluation Labels</th>
<th>Before Improvement</th>
<th>After Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Interesting</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Interesting</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Fair</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>Difficult to understand</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>TOTAL</td>
<td>31</td>
<td>32</td>
</tr>
</tbody>
</table>
Contents

- Background
- The Integrated Time-Series Data Mining Environment
- Case Study on Chronic Hepatitis Data Mining
- Conclusion
Conclusion

- Implemented a time-series data mining environment, integrating time-series pattern extraction, rule induction, and rule evaluation support with active human-system interaction.
- Succeeded in finding out a new hypothesis related to risks of IFN treatment result.
- Developing active evaluation support re-using evaluations of domain experts.
- Introducing algorithm selection sub-systems for each procedure to support data miners.
- Applying this environment to other domains.