Numerical Time-Series Pattern Extraction
Reflecting User’s Interest

Miho Ohsaki  
Faculty of Engineering  
Doshisha University  
1-3 Tataramiyakodani, Kyotanabe-shi,  
Kyoto 610-0321, JAPAN  
Email: mohsaki@mail.doshisha.ac.jp

Hidenao Abe  
Department of Medical Informatics  
Shimane University  
89-1 Enya-cho, Izumo-shi,  
Shimane 693-8501 JAPAN  
Email: abe@med.shimane-u.ac.jp

Takahira Yamaguchi  
Faculty of Science and Technology  
Keio University  
3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi,  
Kanagawa 223-8522 JAPAN  
Email: yamaguti@ae.keio.ac.jp

Abstract—In this paper, we propose a method for extracting interesting patterns from numerical time-series data which takes account of user subjectivity. The proposed method irregularly samples the data preserving the subjectively noteworthy features using a user specified gradient. It also irregularly quantizes the data preserving the intrinsically objective characteristics of the data using statistical distributions. It then extracts representative patterns from the discretized data using group average clustering. We conducted the two experiments to evaluate the extraction performance of the proposed method. It is indicated that the proposed method does not destroy the intrinsically objective features by the experimental results using benchmark datasets. It is also indicated that the proposed method has the possibility to extract interesting patterns for a medical expert by the experimental results using a real clinical dataset on hepatitis.

I. INTRODUCTION

One of important research areas in Data Mining (DM) and Knowledge Discovery in Databases (KDD) is pattern extraction from numerical time-series data: It has many applications such as disease detection and diagnosis, stock price prediction, and traffic flow analysis. (From here on we simply use the term “data” to refer to such numerical time-series data). Typically, conventional methods analyze data by making a mathematical model and extracting patterns based on this model. However, to obtain interesting patterns for real applications, it is necessary to incorporate user subjectivity, i.e., the user’s domain knowledge and point of view, into the extraction process, while remaining faithful to the mathematical data structure. It is also necessary to recognize that typical users who are experts in an application domain but not in pattern extraction frequently prefer intuitively understandable methods rather than black-box approaches even if mathematical strictness of patterns is sacrificed to a certain degree.

We then propose and evaluate a simple method for extracting numerical time-series patterns which uses mathematical data structures and also takes account of user subjectivity. The organization of the paper is as follows: Section II discusses related work and the motivation behind this work. Section III describes our problem solving approach and proposed method, which is based on piecewise aggregate approximation and subsequence clustering. Section IV shows two evaluation experiments, and discusses the extraction performance of the proposed method based on the experimental results. Section V concludes the paper and discusses future work.

II. RELATED WORK

Numerical time-series data mining tries to discover useful patterns and pattern combinations in large real datasets based on statistics, signal processing, and machine learning techniques. It consists of all or some of the following: data feature analysis, pattern extraction, pattern association analysis, and pattern classification/prediction. We focus on pattern extraction. Pattern extraction methods depend on the features of the data. If the data has mathematically clear characteristics such as linearity, periodicity, low noise, and few missing values, precise mathematical modeling approaches utilizing statistics and signal processing are useful for pattern extraction. Methods which use autoregression analysis, ARIMA, Fourier transforms, wavelet transforms, chaos and fractal analysis will be appropriate tools to consider.

However, in the case of data with blurred features, noise and missing values, more flexible approaches are required. The basic functions required in such an approach are: (a) to find appropriate subsequences, (b) to formalize the subsequences and generate pattern candidates, and (c) to select noteworthy patterns. Some such pattern extraction methods implementing these functions are listed below. A method that cuts out subsequences from data, conducts clustering on them, and regards the centers of clusters as extracted patterns [1]. One that prepares prospective patterns, conducts pattern matching on data using Dynamic Time Warping, and regards matched waveforms as extracted patterns [2]. One that conducts pattern matching on data using Multiscale Matching controlling analysis granularity [3]. One that discretizes data using Piecewise Aggregate Approximation [4] and conducts pattern matching on the discretized subsequences removing trivial patterns [5]. One that adjusts discretization intervals using the Minimum Description Length Principle [6]. One that applies a genetic algorithm to the parameter setting of pattern matching using the values of $t$-statistic and false positive for the fitness function [7].
Although these methods are well suited to data with blurred features, noise, and missing values, many of them exclude user subjectivity as much as possible in order to ensure objective reliability. Too much user subjectivity in a pattern extraction method may yield unreliable results, since the method will generate patterns that the user likes. On the other hand, too little user subjectivity may yield uninteresting results for the user, since the method generates patterns taking no account of domain knowledge and the user’s point of view. We should carefully consider this kind of trade-off between objectivity and subjectivity when we design and use a pattern extraction method.

III. PROPOSAL

A. Pattern Extraction Framework

For users who are experts in their application domain but not in pattern extraction, there is a need for pattern extraction methods that accept more user subjectivity and consist of more intuitively understandable processes than do conventional methods. This need is particularly apparent in the area of medical KDD for chronic diseases [8][9]. Therefore we take a different stance from proponents of conventional methods. We emphasize the role of the user and propose a pattern extraction method which allows for user subjectivity. The proposed method aims not to perfectly ensure the objective reliability of extracted patterns but to support a user in obtaining interesting ones. It defers the final judgment of objective reliability to the user, who has sophisticated expertise in a specific domain.

Fig. 1 shows the framework of our pattern extraction method. It stems from our past studies on medical KDD [14] and others on Subsequence Clustering (SC) [1] and Piecewise Aggregate Approximation (PAA) [4][5]. We adopted the basic concepts of SC and PAA because they are readily understood by users. We then extended these concepts and their concretization, considering how to include user subjectivity.

Fig. 1. Framework of our pattern extraction method.

The proposed method consists of (1) discretization, (2) generation of pattern candidates, and (3) extraction of representative patterns. With reference to the three fundamental functions discussed in Section II, we note that function (a) itself is a highly challenging problem [10]. Consequently, we decided to implement functions (b) and (c) on the assumption that proper subsequences are given as a first step. Processes (1) and (2) correspond to the function (b), and the process (3) corresponds to the function (c). The proposed method excludes normalization differently from PAA for motif discovery that emphasizes pattern shape rather than pattern position [5][6], since pattern position is important for the judgment of normal or abnormal symptoms in medical KDD [14].

B. Irregular Discretization

Irregular sampling in the process (1) allows a user to specify a waveform gradient $T_G$ for reflecting a remarkable change (See the line 02 in Fig. 2). It then discretizes the data along the time axis preserving the local waveform features that are subjectively remarkable using $T_G$ (See the lines from 03 to 08 in Fig. 2 and on the left of Fig. 3). As shown on the right of Fig. 3, irregular sampling actually preserves the subjectively remarkable features better than does regular sampling with no gradient specification; it samples roughly in the flat regions and finely in the more volatile ones.

01: \( s = 0 \); \( j = 0 \);
02: Let the user specify the upper threshold of gradient $T_G$; // $T_G$ is User-Initiative
03: while ( (i+1)-th point exists in the time-series ) { 
04: Calculate (i)-th gradient \( G_i \) between (i)-th and (i+1)-th points;
05: if ( \( T_G < \) the absolute value of \( G_i \) ) { 
06: Register ((i+(i+1))/2 as (j)-th boundary \( B_j \); 
07: \( j++ \); 
08: 

Fig. 2. Pseudo code for irregular sampling.

Fig. 3. Conceptual scheme (upper) and result example (lower) of irregular sampling.
The irregular quantization in the process (1) discretizes the data along the amplitude axis preserving objective global waveform features using a statistical data distribution. Real data frequently has a leptokurtic, platykurtic, and/or a skewed distribution. We selected some nonparametric statistics and used them for extreme outlier removal and quantization boundary positioning, taking into account the kurtosis and skewness of the distribution.

Initially, the irregular quantization calculates the kurtosis $Ku$ of the data distribution and removes unnecessary extreme outliers by kurtosis thresholding (See the lines from 04 to 09 in Fig. 4). Although the kurtosis threshold $T_{Ku}$ cannot be altered by a user in the default configuration, it can be set by a user when more user subjectivity is allowed. Next, the irregular quantization calculates intervals based on standard deviation $Sd$ and skewness $Sk$ and determines boundary positions by assigning the intervals relative to the median $Me$ (See the lines from 11 to 18 in Fig. 4 and on the left of Fig. 5). Equation (1) is the definition of the $j$-th quantization interval $QI_j$. It includes a correction factor $jSd\sqrt{Sk}$ to reflect the distribution profile over the interval. The weight of the interval $W_j$ may be set manually when the user is being allowed more control over the granularity of observation.

$$QI_j = \frac{1}{W_j}\{((2j+1)\frac{Sd}{2} + jSd\sqrt{Sk})\}$$

As shown on the right of Fig. 5, the irregular quantization reflects the objective features better than does the regular quantization with no correction factor; the further from the peak of data distribution, the wider the quantization interval. Note that the irregular quantization contains the regular one as a special case, because the median equals the mean and the skewness is zero if the data has a symmetrical distribution.

The process (2) generates pattern candidates by the symbolization of irregularly discretized subsequences. It calculates the mean of the amplitudes in a sampling interval, finds the nearest quantization boundary to the mean, and flattens the amplitude in the sampling interval to the boundary value. It repeats the flattening operation for each sampling interval and consequently transforms a subsequence into a step-like sequence. To remove the effect of slight differences caused by noise among the step-like sequences, it converts them into symbol strings as pattern candidates preserving the magnitude relation of boundaries. The simplest way to carry out this conversion is to alphabetically assign symbols to the numerical values of boundaries in ascending order.

C. Subsequence Clustering

Finally, the process (3) extracts representative patterns by the clustering of pattern candidates. We adopted group average clustering [12], which utilizes hierarchical clustering and is applicable to both concentric and chain-like distributions. The
number of clusters are determined according to target data based on the domain knowledge on the data. We empirically examine some distance metrics for patterns in Section IV, because they have a significant influence on clustering performance.

IV. PERFORMANCE EVALUATION

A. Exp. I: Performance Evaluation Using Benchmark Datasets

First, we conducted Exp. I to examine whether the proposed method extracts patterns without destroying their objective reliability and also to check which distance metric results in best performance. The input datasets were GunX, ECG znorm205, Tracedata, Leaf_all, cbf, cbf-tr, and two-pat in the UCR Time Series Data Mining Archive [11]. The details of the proposed method are as shown in the next paragraph. The proposed method was compared with one of the most basic methods, namely SC using the K-Means algorithm [13] and Euclidean distance. The criteria used to evaluate performance were the values for Correct Rate (CR) and statistic F Value (FV), where CR was the percentage of the patterns assigned to correct clusters in all patterns, and FV was the ratio of the unbiased variance of clusters to that of patterns in the clusters. The larger CR and FV, the more accurate the pattern extraction results. The values of CR and FV between the proposed method and the comparison method were statistically tested with the Wilcoxon matched-pair signed-rank test.

The distance metrics used in the proposed method were (1-1) the mean absolute error of symbol strings, (1-2) the mean square error of the symbol strings, (2-1) the mean absolute error of step-like sequences, and lastly (2-2) the mean square error of step-like sequences. In the calculation of the distance metrics (1-1) and (1-2), the distance between a symbol and a neighbor was set uniformly to 1 because of the intended purpose of symbolization, namely to remove the effect of slight differences. For instance, the distance between ‘a’ and ‘b’ was 1, and that between ‘a’ and ‘c’ was 2. In the calculation of the distance metrics (2-1) and (2-2), step-like sequences before symbolization were used to examine the influence of symbolization. The distance between a step of a step-like sequence and that of another one was the actual numerical value of the difference. We initially tried various thresholds of gradient for each dataset and then adopted the threshold that achieved the highest performance.

Table I shows the results of Exp. I. The highest values of CR and FV in each dataset and the average are underlined. With respect to both measures CR and FV, the differences between the two methods were not statistically significant for any of the distance metrics (1-1), (1-2), (2-1), and (2-2). This result indicates that the proposed method has comparable performance to that of SC using the K-Means algorithm and Euclidean distance and that it does not destroy the objective features of the data. Although the differences were not statistically significant, the distance metric (2-1) produced the largest number of highest values.

<table>
<thead>
<tr>
<th>Distance</th>
<th>Proposed Method</th>
<th>Euclidean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criteria</td>
<td>CR</td>
<td>FV</td>
</tr>
<tr>
<td>GunX</td>
<td>61.50</td>
<td>141.74</td>
</tr>
<tr>
<td>ECG znorm205</td>
<td>100.00</td>
<td>61.84</td>
</tr>
<tr>
<td>Tracedata</td>
<td>52.30</td>
<td>119.18</td>
</tr>
<tr>
<td>Leaf_all</td>
<td>52.81</td>
<td>48.10</td>
</tr>
<tr>
<td>cbf</td>
<td>44.62</td>
<td>143.70</td>
</tr>
<tr>
<td>cbf-tr</td>
<td>35.46</td>
<td>84.78</td>
</tr>
<tr>
<td>two-pat</td>
<td>55.61</td>
<td>108.48</td>
</tr>
</tbody>
</table>

Table I: Results of Exp. I.
differences of \(N_B\) and \(N_N\) between the proposed method and the object of comparison were not statistically tested due to the single dataset.

Table II shows the results of Exp. II. \(N_B\) for the proposed method was greater than \(N_B\) for the object of comparison, and the opposite trend appeared for \(N_N\). This result suggests that the proposed method performs better in extracting really interesting patterns than SC with regular PAA using group average clustering and that it reflects user subjectivity in this particular example.

**TABLE II**

<table>
<thead>
<tr>
<th>RESULTS OF EXP. II.</th>
<th>Proposed Method</th>
<th>Regular PAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Extracted</td>
<td>32</td>
<td>26</td>
</tr>
<tr>
<td>Especially Interesting</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>Interesting</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>(N_B)</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>Not Interesting</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>Not Understandable</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>(N_N)</td>
<td>12</td>
<td>20</td>
</tr>
</tbody>
</table>

Fig. 6 and Fig. 7 are examples of rules obtained by our proposed method, to which the medical expert actually showed an interest. In Fig. 6, he focused on the relation between WBC (White Blood Cells) and bio-response (the degree of interferon effect). His comments on this rule are as follows: The reduction of WBC means that the side effect of interferon medication appears, and the value of bio-response, \(N\), means that the positive effect of interferon medication does not appear. Although the side effect appears, the positive effect does not appear. The mechanism of this contradiction is interesting and should be discussed.

In Fig. 6, he focused on the appearance of K (Kalium). His comments on this rule are as follows: It is generally said that kalium has no relation with interferon effect. However, the time-series of kalium values appear with the value of bio-response, \(R\), which means the positive effect of interferon medication. The relation between them is interesting and should be discussed.

Fig. 7. An example of rules to which a medical expert showed an interest. The medical expert focused on the appearance of K.

V. CONCLUSION

We have proposed a method to extract interesting patterns from numerical time-series data based on irregular Piecewise Aggregate Approximation (PAA) and subsequence clustering. It discretizes and creates patterns from the data reflecting both user subjectivity and the mathematical data structure. The experimental results using benchmark datasets indicated that the proposed method has an accuracy comparable to that of the K-means algorithm and the possibility to extract patterns without destroying their reliability. The results using a clinical hepatitis dataset indicated that the proposed method performs better in extracting really interesting patterns for a medical expert than regular PAA, at least in the hepatitis domain.

Our future work will be the implementation of the indexing of subsequence start and end points, the comparisons between the proposed method and methods based on other types of irregular PAA, and the psychological evaluation on whether the proposed method can extract interesting patterns in other real domains.

ACKNOWLEDGMENT

This research was supported by the Ministry of Education, Culture, Sports, Science and Technology Japan (MEXT), the Academic Frontier Research Project on “New Frontier of Biomedical Engineering Research” and the Grant-in-Aid for Young Scientists (B), 17700162.

REFERENCES


