New Generation Computing, 25(2007)213-222
Ohmsha, Ltd. and Springer

Numerical Time-Series Pattern Extraction Based on Irregular Piecewise Aggregate Approximation and Gradient Specification

Miho OHSAKI
Doshisha University
1-3 Tataramiyakodani, Kyotanabe-shi, Kyoto 610-0321 JAPAN
mohsaki@mail.doshisha.ac.jp
Hidenao ABE
Shimane University
89-1 Enya-cho, Izumo-shi, Shimane 693-8501 JAPAN
abe@med.shimane-u.ac.jp
Takahira YAMAGUCHI
Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, Kanagawa 223-8522 JAPAN
yamaguti@ae.keio.ac.jp

Received 20 January 2006
Revised manuscript received 8 April 2007

Abstract This paper proposes and evaluates a method for extracting interesting patterns from numerical time-series data which takes account of user subjectivity. The proposed method conducts irregular sampling on the data preserving the subjectively noteworthy features using a user specified gradient. It also conducts irregular quantization, preserving the intrinsically objective characteristics of the data using statistical distributions. It then extracts representative patterns from the discretized data using group average clustering. Experimental results using benchmark datasets indicate that the proposed method does not destroy the intrinsically objective features, since it has the same performance as the basic subsequence clustering using K-Means algorithm. Results using a dataset from a clinical hepatitis study indicate that it extracts interesting patterns for a medical expert.

Keywords: Data Mining, Knowledge Discovery in Databases, Numerical Time-Series, Pattern Extraction, Piecewise Aggregate Approximation.

§1 Introduction

Pattern extraction from numerical time-series data is an important research field with many applications, for example, 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.

This paper proposes and evaluates 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 2 introduces related work and discusses the motivation for our study and its problem solving approach. Section 3 explains the pattern extraction processes of our proposed method. In Section 4 we examine the performance of the proposed method in Experiment I which used several benchmark datasets, and Experiment II which used a clinical hepatitis dataset. For experiment II, the results are compared to the pattern evaluation results by a medical expert. Section 5 concludes the paper and discusses future work.

§2 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]

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.

§3 Proposal

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 Knowledge Discovery in Databases (KDD) for chronic diseases.[7][8] 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. Figure 1 shows the framework of our pattern extraction method. It stems from our past studies on

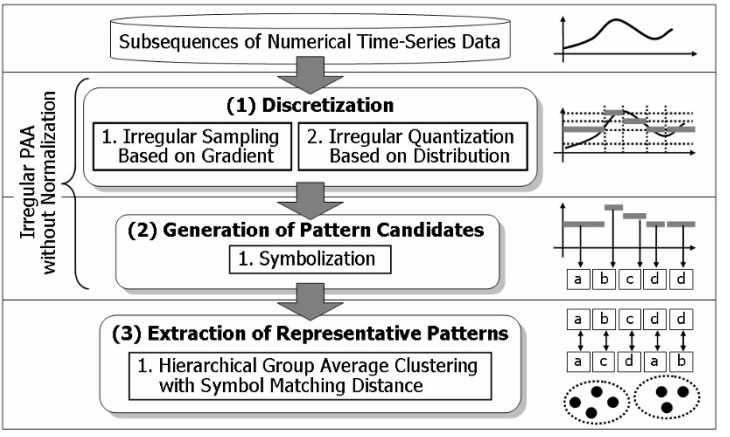

Fig. 1 Framework of our Pattern Extraction Method.

medical KDD[12] 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.

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 2, we note that function (a) itself is a highly challenging problem.[9] 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.[12]

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: i = 0;   j = 0;
02: Let the user specify the upper threshold of gradient T_G; // 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:     i++; }
```

Fig. 2 Pseudo Code for Irregular Sampling.

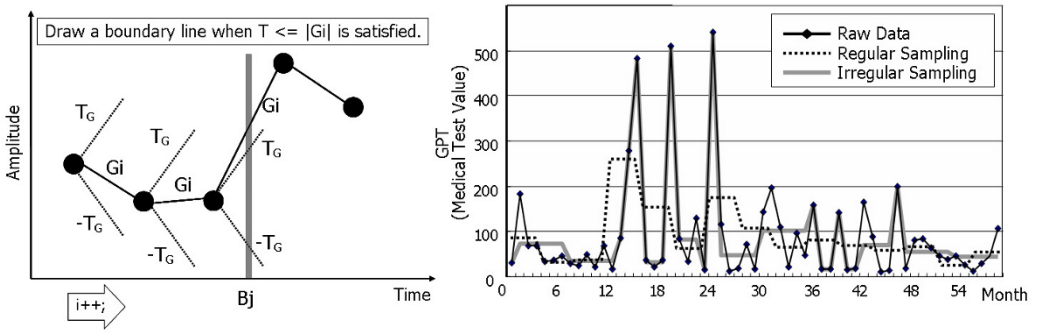

Fig. 3 Conceptual Scheme (left) and Result Example (right) of Irregular Sampling.

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

01: $p = 0$;
02: Set the lower threshold of kurtosis T_{Ku} at the default value of 3; // User-Settable
03: Calculate the median Me of all points;
04: while ((p) percent <= 100 percent) {
05: Take (p) percent points toward both sides of Me in the frequency distribution;
06: Calculate the standard deviation Sd and kurtosis Ku of the selected points;
07: if ($Ku <= T_{Ku}$)
08: break;
09: p++; }
10: $i = Me$; $j = 0$;
11: while (true) {
12: Set the weight of quantization interval W_j at the default value of 1; // User-Settable
13: Calculate (j)-th quantization interval QI_j; $i = i + QI_j$;
14: if ((i)-th point does not exist in the upper part of Me)
15: break;
16: Register i as (j)-th upper boundary UB_j;
17: j++; }
18: Do the same procedure from 10 to 17 with $i = i$ - QI_j for (j)-th lower boundary LB_j;

Fig. 4 Pseudo Code for Irregular Quantization.

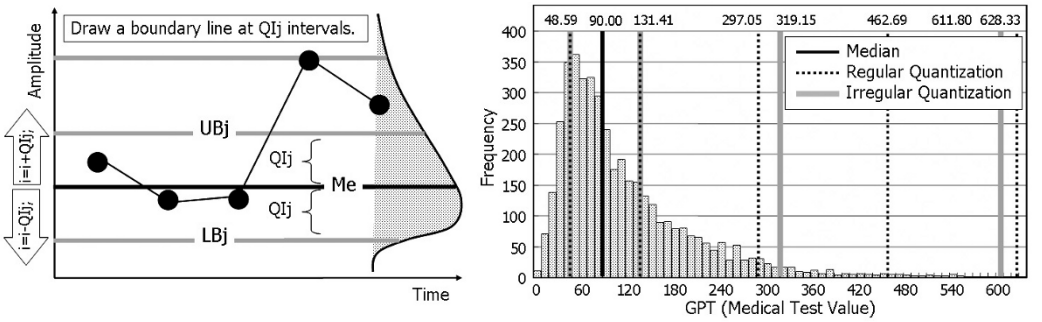

Fig. 5 Conceptual Scheme (left) and Result Example (right) of Irregular Quantization.

definition of the j-th quantization interval QI_j. It includes a correction factor $jSd\sqrt[3]{Sk}$ 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.

$$QI_j = \frac{1}{W_j}\{(2j+1)\frac{Sd}{2} + jSd\sqrt[3]{Sk}\} \quad (1)$$

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. Finally, the process (3) extracts representative patterns by the clustering of pattern candidates. We adopted group average clustering,[11] which utilizes hierarchical clustering and is applicable to both concentric and chain-like distributions. We empirically examine some distance metrics for patterns in Section 4, because they have a significant influence on clustering performance.

§4 Performance Evaluation

4.1 Experiment I: Evaluation Using Benchmark Data

First, we conducted Experiment 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.[10] 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 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 1 shows the results of Experiment 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 1 Results of Experiment I.

	Proposed Method								K-Means	
Distance	(1-1)		(1-2)		(2-1)		(2-2)		Euclidean	
Criteria	CR	FV	CR	FV	CR	FV	CR	FV	CR	FV
GunX	61.50	141.74	61.50	135.98	64.50	144.18	65.50	123.84	50.00	152.05
ECG_znorm205	100.00	61.84	100.00	61.84	100.00	61.06	100.00	67.43	100.00	57.72
Tracedata	52.50	119.18	55.50	197.12	52.50	113.64	52.50	113.64	53.00	265.89
Leaf_all	32.81	48.10	33.26	46.36	32.81	44.63	32.81	45.91	32.13	18.38
cbf	62.36	160.03	62.34	158.83	67.90	169.67	67.46	165.49	71.92	129.14
cbf-tr	44.62	143.70	44.60	144.00	45.78	142.39	42.68	151.93	43.38	133.22
two-pat	35.46	84.78	35.52	84.48	35.54	84.98	35.38	84.75	32.38	37.31
Average	55.61	108.48	56.10	118.37	57.00	108.65	56.62	107.57	54.69	113.39

4.2 Experiment II: Evaluation Using Clinical Data

Next, we conducted Experiment II to examine whether the proposed method extracts really interesting patterns for a user. The input was a clinical hepatitis dataset in ECML/PKDD2002 Discovery Challenge.[13] The distance metric used was the mean absolute error of step-like sequences, which performed best in Experiment I. The gradient was specified by a medical expert. The object of comparison with the proposed method was a method having the same processes as the proposed one except for irregular PAA, namely SC with regular PAA using group average clustering. Note that we excluded normalization from

this regular PAA to make it suitable for medical KDD.

The criteria used to evaluate performance were the number of extracted beneficial patterns (NB) and the number of extracted non-beneficial ones (NN). We utilized the patterns of hepatitis symptoms judged especially interesting, interesting, not interesting, or not understandable by the medical expert in our past study.[12] NB was the total number of especially interesting or interesting patterns, and NN was the total number of not interesting or not understandable ones. The differences of NB and NN between the proposed method and the object of comparison were not statistically tested due to the single dataset.

Table 2 shows the results of Experiment II. NB for the proposed method was greater than NB for the object of comparison, and the opposite trend appeared for NN. 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 2 Results of Experiment II.

	Proposed Method	Regular PAA
All Extracted	32	26
Especially Interesting	15	4
Interesting	5	7
NB	20	11
Not Interesting	11	15
Not Understandable	1	5
NN	12	20

§5 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 and comparisons between the proposed method and methods based on other types of irregular PAA.

Acknowledgements

This research was partially supported by the Ministry of Education, Science, and Culture, by a Grant-in-Aid for Scientific Research in a Priority Area (B) 13131205 and Young Scientists (B) 17700162.

References

1) Das, G., King-Ip, L., Heikki, M., Renganathan, G. and Smyth, P., "Rule Discovery from Time Series," in *Proc. of Int. Conf. on Knowledge Discovery and Data Mining*, pp. 16–22, 1998.
2) Berndt, D. J. and Clifford, J., "Using dynamic time warping to find patterns in time series," in *Proc. of AAAI Workshop on Knowledge Discovery in Databases*, pp. 359–370, 1994.
3) Hirano, S. and Tsumoto, S., "Mining Similar Temporal Patterns in Long Time-series," *Data and Its Application to Medicine*, pp. 219–226, 2002.
4) Yi, B-K. and Faloutsos, C., "Fast Time Sequence Indexing for Arbitrary Lp Norms," in *Proc. of Int. Conf. on Very Large Databases*, pp.385–394, 2000.
5) Lin, J., Keogh, E., Lonardi, S. and Patel, P., "Finding Motifs in Time Series," in *Proc. of Workshop on Temporal Data Mining*, pp. 53–68, 2002.
6) Tanaka, Y. and Uehara, K., "Discover Motifs in Multi Dimensional Time-Series Using the Principal Component Analysis and the MDL Principle," in *Proc. of Int. Conf. on Machine Learning and Data Mining in Pattern Recognition*, pp. 252–265, 2003.
7) Motoda, H., *Active Mining*, IOS Press, Amsterdam, 2002.
8) Tsumoto, S., Yamaguchi, T., Numao, M. and Motoda, H., "Active Mining Project: Overview," *Lecture Notes in Artificial Intelligence, 3403*, pp. 1–10, 2005.
9) Keogh, E. and Lin, J., "Clustering of Time-Series Subsequences is Meaningless: Implications for Previous and Future Research," *Knowledge and Information Systems, 8 (2)*, pp.154–177, 2005.
10) Keogh, E., "Time Series Data Mining Archive," http://www.cs.ucr.edu/~eamonn/TSDMA/, 2005.
11) Yang, Y., Carbonell, J., Brwon, R. and Pierce, T., "Learning Approaches for Detecting and Tracking News Events," *IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, 14*, pp.32–43, 1999.
12) Ohsaki, M., Sato, Y., Yokoi, H. and Yamaguchi, T., "A Rule Discovery Support System for Sequential Medical Data, – In the Case Study of a Chronic Hepatitis Dataset –," in *Proc. of Int. Workshop on Active Mining*, pp. 97–102, 2002.
13) Tsumoto, S., "ECML/PKDD2002 Discovery Challenge," http://lisp.vse.cz/challenge/ecmlpkdd2002/, 2002.

Miho Ohsaki, Ph.D.: She is an associate professor in the Department of Information Systems Design at Doshisha University. She received her Ph.D. in Engineering from Kyushu Institute of Design in 1999. She is interested in support for humans in cognitive, intellectual, and creative activities through human-intelligent system interaction.

Hidenao Abe, Ph.D.: He is a Research Associate in Shimane University, School of Medicine, Japan. He received Ph.D. degree in computer science from Shizuoka University in 2004. Dr. Abe is a member of the IEEE Computer Society, the Japanese Society for Artificial Intelligence, and the Japan Association for Medical Informatics.

Takahira Yamaguchi, Ph.D.: He is a Professor in the School of Science and Technology in Keio University, Yokohama. He received Ph.D. degree in telecommunication engineering from Osaka University in 1984. His major work includes semantic web, ontologies, data mining and knowledge management.