社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－ＩＣＳ－133 (26)
2003／9／15

Empirical Comparison of Clustering Methods for Long Time-Series Databases

Shoji HIRANO† and Shusaku TSUMOTO†

† Department of Medical Informatics, Shimane Medical University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
E-mail: †hirano@ieee.org, ††tsumoto@computer.org

Abstract This paper presents a comparative study of methods for clustering long-term temporal data. We split a clustering procedure into two processes: similarity computation and grouping. As similarity computation methods, we employed dynamic time warping (DTW) and multiscale matching. As grouping methods, we employed conventional agglomerative hierarchical clustering (AHC) and rough sets-based clustering (RC). Using various combinations of these methods, we performed clustering experiments of the hepatitis data set and evaluated validity of the results. The results suggested that (1) complete-linkage (CL) criterion outperformed average-linkage (AL) criterion in terms of the interpret-ability of a dendrogram and clustering results, (2) combination of DTW and CL-AHC constantly produced interpretable results, (3) combination of DTW and RC would be used to find the core sequences of the clusters, (4) multiscale matching may suffer from the treatment of 'no-match' pairs, however, the problem may be eluded by using RC as a subsequent grouping method.

Key words temporal data mining, similarity measure, clustering

長期時系列データ類型化法の比較

平野 章二† 津本 周作†

† 島根医科大学医学部医療情報学講座 〒693–8501 島根県出雲市塩冶町 89–1
E-mail: †hirano@ieee.org, ††tsumoto@computer.org

あらまし 本稿では，慢性ウイルス性肝炎データを対象として長期時系列の類型化法を比較した結果について報告する。

キーワード 時系列データマイニング，類似度，クラスタリング

1. Introduction

Clustering of time-series data [1] has been receiving considerable interests as a promising method for discovering interesting features shared commonly by a set of sequences. One of the most important issue in time-series clustering is determination of (dis-)similarity between the sequences. Basically, the similarity of two sequences is calculated by accumulating distances of two data points that are located at the same time position, because such a distance-based similarity has preferable mathematical properties that extend the choice of grouping algorithms. However instead, this method requires that the lengths of all sequences be the same. Additionally, it cannot compare structural similarity of the sequences; for example, if two sequences contain the same number of peaks, but at slightly different phases, their 'difference' is emphasized rather than their structural similarity [2].

These drawbacks are serious in the analysis of time-series data collected over long time. The long time-series data have the following features. First, the lengths and sampling intervals of the data are not uniform. Starting point of data acquisition would be several years ago or even a few decades ago. Arrengement of the data should be performed, however, shortening a time-series may cause the loss of precious information. Second, long-time series contains both long-term and short-term events, and their lengths and phases are not the same. Additionally, the sampling interval of the data would be variant due to the change of acquisition strategy

over long time.

Some methods are considered to be applicable for clustering long time series. For example, dynamic time warping (DTW) [3] can be used to compare the two sequences of different lengths since it seeks the closest pairs of points allowing one-to-many point matching. This feature also enable us to capture similar events that have time shifts. Another approach, multiscale structure matching [6] [5], can also be used to do this work, since it compares two sequences according to the similarity of partial segments derived based on the inflection points of the original sequences. However, there are few studies that empirically evaluate usefulness of these methods on real-world long time-series data sets.

This paper reports the results of empirical comparison of similarity measures and grouping methods on the hepatitis data set [7]. The hepatitis dataset is the unique, long time-series medical dataset that involves the following features: irregular sequence length, irregular sampling interval and co-existence of clinically interesting events that have various length (for example acute events and chronic events). We split a clustering procedure into two processes: similarity computation and grouping. For similarity computation, we employed DTW and multiscale matching. For grouping, we employed conventional agglomerative hierarchical clustering [8] and rough sets-based clustering [9], focusing that these methods can be used as un-supervised methods and are suitable for handling relative similarity induced by multiscale matching. For every combination of the similarity computation methods and grouping methods, we performed clustering experiments and evaluated validity of the results.

2. Materials

We employed the chronic hepatitis dataset [7], which were provided as a common dataset for ECML/PKDD Discovery Challenge 2002 and 2003. The dataset contained long time-series data on laboratory examinations, which were collected at Chiba University Hospital in Japan. The subjects were 771 patients of hepatitis B and C who took examinations between 1982 and 2001. We manually removed sequences for 268 patients because biopsy information was not provided for them and thus their virus types were not clearly specified. According to the biopsy information, the expected constitution of the remaining 503 patients were, B / C-noIFN / C-IFN = 206 / 100 / 197. However, due to existence of missing examinations, the numbers of available sequences could be less than 503.

The dataset contained the total of 983 laboratory examinations. However, in order to simplify our experiments, we selected 13 items from blood tests relevant to the liver function: ALB, ALP, G-GL, G-GTP, GOT, GPT, HGB, LDH, PLT, RBC, T-BIL, T-CHO and TTT. Details of each examination are available at the URL [7].

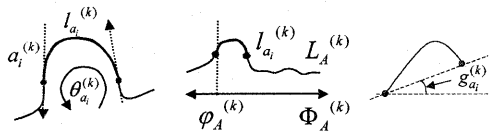

図 1 Segment difference.

Each sequence originally had different sampling intervals from one day to one year. From preliminary analysis we found that the most frequently appeared interval was one week; this means that most of the patients took examinations on a fixed day of a week. According to this observation, we determined resampling interval to seven days. A simple summary showing the number of data points after resampling is as follows (item=ALB, $n = 499$) : mean=456.87, sd=300, maximum=1080, minimum=7. Note that one point equals to one week; therefore, 456.87 points equals to 456.87 weeks, namely, about 8.8 years.

3. Methods

We have implemented algorithms of symmetrical time warping describe briefly in [2] and one-dimensional multiscale matching described in [4]. We modified segment difference in multiscale matching as follows.

$$d(a_i^{(k)}, b_j^{(h)}) = \max(\theta, l, \phi, g), \tag{1}$$

where θ, l, ϕ, g respectively represent differences on rotation angle, length, phase and gradient of segments $a_i^{(k)}$ and $b_j^{(h)}$ at scales k and h. These differences are defined as follows:

$$\theta(a_i^{(k)}, b_j^{(h)}) = \mid \theta_{a_i}^{(k)} - \theta_{b_j}^{(h)} \mid /2\pi, \tag{2}$$

$$l(a_i^{(k)}, b_j^{(h)}) = \left| \frac{l_{a_i}^{(k)}}{L_A^{(k)}} - \frac{l_{b_j}^{(h)}}{L_B^{(h)}} \right|, \tag{3}$$

$$\phi(a_i^{(k)}, b_j^{(h)}) = \left| \frac{\phi_{a_i}^{(k)}}{\Phi_A^{(k)}} - \frac{\phi_{b_j}^{(h)}}{\Phi_B^{(h)}} \right|, \tag{4}$$

$$g(a_i^{(k)}, b_j^{(h)}) = \mid g_{a_i}^{(k)} - g_{b_j}^{(h)} \mid . \tag{5}$$

Figure 1 provides an illustrative explanation of these terms. Multiscale matching usually suffers from the shrinkage of curves at high scales caused by excessive smoothing with a Gaussian kernel. On one-dimensional time-series data, shrinkage makes all sequences flat at high scales. In order to elude this problem, we applied shrinkage correction proposed by Lowe [10].

We also implemented two clustering algorithms, agglomerative hierarchical clustering (AHC) in [8] and rough sets-

based clustering (RC) in [9]. For AHC we employed two linkage criteria, average-linkage AHC (CL-AHC) and complete-linkage AHC (AL-AHC).

In the experiments, we investigated the usefulness of various combinations of similarity calculation methods and grouping methods in terms of the interpretability of the clustering results. Procedures of data preparation were as follows. First, we selected one examination, for example ALB, and split the corresponding sequences into three subsets, B, C-noIFN and C-IFN, according to the virus type and administration of interferon therapy. Next, for each of the three subgroups, we computed dissimilarity of each pair of sequences by using DTW. After repeating the same process with multiscale matching, we obtained 2×3 sets of dissimilarities: one obtained by DTW, and another obtained by multiscale matching.

Then we applied grouping methods AL-AHC, CL-AHC and RC to each of the three dissimilarity sets obtained by DTW. This yielded $3 \times 3 = 9$ sets of clusters. After applying the same process to the sets obtained by multiscale-matching, we obtain the total of 18 sets of clusters.

The above process is repeated with the remaining 12 examination items. Consequently, we constructed 12×18 clustering results. Note that in this experiments we did not perform cross-examination comparison, for example comparison of an ALB sequence with a GPT sequence.

We used the following parameters for rough clustering: $\sigma = 5.0$, $T_h = 0.3$. In AHC, cluster linkage was terminated when increase of dissimilarity firstly exceeded mean+SD of the set of all increase values.

4. Results

Table 1 provides the numbers of generated clusters for each combination. Let us explain the table using the raw whose first column is marked ALB. The second column "Number of Instances" represents the number of patients who took the ALB examination. Its value 204/99/196 represents that 204 patients of Hepatitis B, 99 patients of Hepatitis C (who did not take IFN therapy) and 196 patients of Hepatitis C (who took IFN therapy) took this examination. Since one patient has one time-series examination result, the number of patients corresponds to the number of sequences. The third column shows the number of generated clusters. Using DTW and AL-AHC, 204 hepatitis B sequences were grouped into 8 clusters. 99 C-noIFN sequences were grouped into 3 clusters, as well as 196 C-IFN sequences.

4.1 DTW and AHCs

Let us first investigate the case of DTW-AHC. Comparison of DTW-AL-AHC and DTW-CL-AHC implies that the results can be different if we use different linkage criterion.

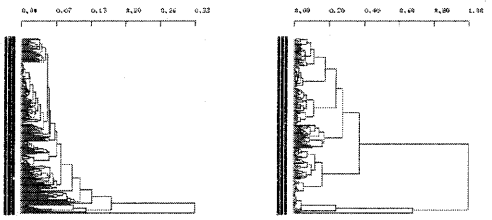

図 2 Dendrograms for DTW-AHC-B. Left: AHC-AL. Right: AHC-CL.

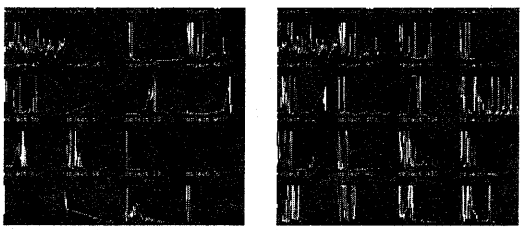

図 3 Examples of the clusters. Left: AHC-AL. Right: AHC-CL.

Figure 2 left image shows a dendrogram generated from the GTP sequences of type B hepatitis patients using DTW-AL-AHC. It can be observed that the dendrogram of AL-AHC has an ill-formed structure like 'chaining', which is usually observed with single-linkage AHC. For such an ill-formed structure, it is difficult to find a good point to terminate merging of the clusters. In this case, the method produced three clusters containing 193, 9 and 1 sequences respectively. Figure 3 left image shows a part of the sequences grouped into the largest cluster. Almost all types of sequences were included in this cluster and thus no interesting information was obtained.

On the contrary, the dendrogram of CL-AHC shown in the right of Figure 2 demonstrates a well formed hierarchies of the sequences. With this dendrogram the method produced 7 clusters containing 27, 21, 52, 57, 43, 2, and 1 sequences. Figure 3 right image and Figure 4 show examples of the sequences grouped into the first three clusters respectively. One can observe interesting features for each cluster. The first cluster contains sequences that involve continuous vibration of the GPT values. These patterns may imply that the virus continues to attack the patient's body periodically. The second cluster contains very short, meaningless sequences, which may represent the cases that patients stop or cancel receiving the treatment quickly. The third cluster contains another interesting pattern: vibrations followed by the flat, low values. This case may represent the cases that the patients were cured by some treatments, or naturally.

表 1 Comparison of the number of generated clusters. Each item represents clusters for Hepatitis B / C-noIFN / C-IFN cases.

Exam	Number of Instances	Number of Generated Clusters					
		DTW			Multiscale Matching		
		AL-AHC	CL-AHC	RC	AL-AHC	CL-AHC	RC
ALB	204 / 99 / 196	8 / 3 / 3	10 / 6 / 5	38 / 22 / 32	19 / 11 / 12	22 / 21 / 27	6 / 14 / 31
ALP	204 / 99 / 196	6 / 4 / 6	7 / 7 / 10	21 / 12 / 29	10 / 18 / 14	32 / 16 / 14	36 / 12 / 46
G-GL	204 / 97 / 195	2 / 2 / 5	2 / 2 / 11	1 / 1 / 21	15 / 16 / 194	16 / 24 / 194	24 / 3 / 49
G-GTP	204 / 99 / 196	2 / 4 / 11	2 / 6 / 7	1 / 17 / 4	38 / 14 / 194	65 / 14 / 19	35 / 8 / 51
GOT	204 / 99 / 196	8 / 10 / 25	8 / 4 / 7	50 / 18 / 60	19 / 12 / 24	35 / 19 / 19	13 / 14 / 15
GPT	204 / 99 / 196	3 / 17 / 7	7 / 4 / 7	55 / 29 / 51	23 / 30 / 8	24 / 16 / 16	11 / 7 / 25
HGB	204 / 99 / 196	3 / 4 / 13	2 / 3 / 9	1 / 16 / 37	43 / 15 / 15	55 / 19 / 22	1 / 12 / 78
LDH	204 / 99 / 196	7 / 7 / 9	15 / 10 / 8	15 / 15 / 15	20 / 25 / 195	24 / 9 / 195	32 / 16 / 18
PLT	203 / 99 / 196	2 / 13 / 9	2 / 7 / 6	1 / 15 / 19	33 / 5 / 12	34 / 15 / 17	1 / 11 / 25
RBC	204 / 99 / 196	3 / 4 / 6	3 / 4 / 7	1 / 14 / 26	32 / 16 / 13	40 / 23 / 17	1 / 6 / 17
T-BIL	204 / 99 / 196	6 / 5 / 5	9 / 5 / 4	203 / 20 / 30	17 / 25 / 6	20 / 30 / 195	11 / 23 / 48
T-CHO	204 / 99 / 196	2 / 2 / 7	5 / 2 / 5	20 / 1 / 27	12 / 13 / 13	17 / 23 / 19	12 / 5 / 23
TTT	204 / 99 / 196	7 / 2 / 5	8 / 2 / 6	25 / 1 / 32	29 / 10 / 6	39 / 16 / 16	25 / 16 / 23

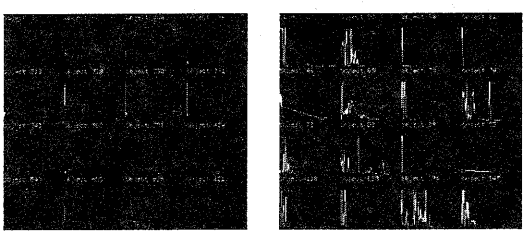

図 4 Other examples of the clusters obtained by AHC-CL. Left: the second cluster containing 21 sequences. Right: the third cluster containing 52 sequences.

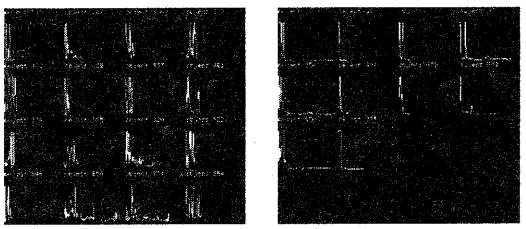

図 5 Examples of the clusters obtained by RC. Left: the second cluster containing 16 sequences. Right: the third cluster containing 10 sequences.

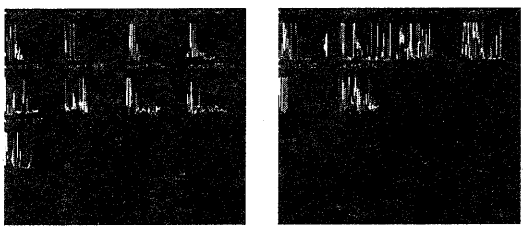

図 6 Other examples of the clusters obtained by RC. Left: the fourth cluster containing 9 sequences. Right: the fifth cluster containing 6 objects.

4.2 DTW and RC

For the same data, rough set-based clustering method produced 55 clusters. Fifty five clusters were too many for 204 objects, however, 41 of 55 clusters contained less than 3 sequences, and furthermore, 31 of them contained only one sequence. This was because of the rough set-based clustering tends to produce independent, small clusters for objects being intermediate of the large clusters. Ignoring small ones, we found 14 clusters containing 53, 16, 10, 9, 6 ... objects. The largest cluster contained short sequences quite similarly to the case of CL-AHC. Figure 5 and 6 show examples of sequences for the 2nd, 3rd, 4th and 5th clusters. Because this method evaluates the indiscernibility degree of objects, each of the generated clusters contains strongly similar sets of sequences. Although populations in the clusters are not so large, one can clearly observe the representative of the interesting patterns described previously at CL-AHC.

4.3 Multiscale Matching and AHCs

Comparison of Multiscale Matching-AHC pairs with DTW-AHC pairs shows that Multiscale Matching's dissimilarities resulted in producing the larger number of clusters than DTW's dissimilarities.

One of the important issues in multiscale matching is treatment of 'no-match' sequences. Theoretically, any pairs of sequences can be matched because a sequence will become single segment at enough high scales. However, this is not a realistic approach because the use of many scales results in the unacceptable increase of computational time. If the upper bound of the scales is too low, the method may possibly fail to find the appropriate pairs of subsequences. For example, suppose we have two sequences, one is a short sequence

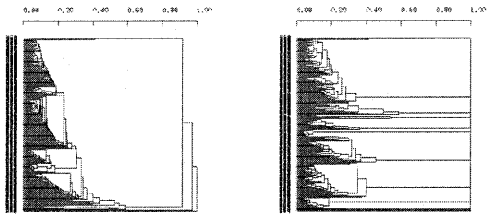

図 7 Dendrograms for MSMmatch-AHC-C-IFN. Left: AHC-AL. Right: AHC-CL.

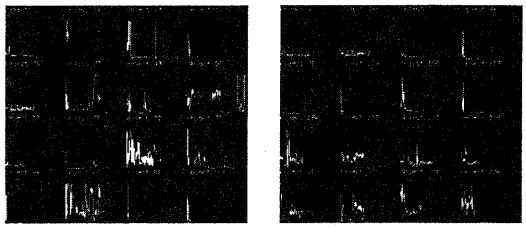

図 8 Examples of the sequences clusters obtained by AHCs. Left: AHC-AL. The first cluster containing 182 sequences. Right: AHC-CL. the first cluster containing 71 sequences.

containing only one segment and another is a long sequence containing hundreds of segments. The segments of the latter sequence will not be integrated into one segment until the scale becomes considerably high. If the range of scales we use does not cover such a high scale, the two sequences will never be matched. In this case, the method should return infinite dissimilarity, or a special number that identifies the failed matching.

This property prevents AHCs from working correctly. CL-AHC will never merge two clusters if any pair of 'no-match' sequences exist between them. AL-AHC fails to calculate average dissimilarity between two clusters. Figure 7 provides dendrograms for GPT sequences of Hepatitis C (with IFN) patients obtained by using multiscale matching and AHCs. In this experiment, we let the dissimilarity of 'no-match' pairs the same as the most dissimilar 'matched' pairs in order to elude computational problems. The dendrogram of AL-AHC is compressed to the small-dissimilarity side because there are several pairs that have excessively large dissimilarities. The dendrogram of CL-AHC demonstrates that the 'no-match' pairs will not be merged until the end of the merging process.

For AL-AHC, the method produced 8 clusters. However, similarly to the previous case, most of the sequences (182/196) were included in the same cluster. As shown in Figure 8 left image, no interesting information was found in the cluster. For CL-AHC, the method produced 16 clusters containing 71, 39, 29, ... sequences. Figure 8 right image and Figure 9 provide examples of the sequences grouped into the three primary clusters, respectively. Similar sequences were found in the clusters, however, obviously dissimilar sequences were also observed in their clusters.

4.4 Multiscale Matching and RC

Rough set-based clustering method produced 25 clusters containing 80, 60, 18, 6 ... sequences. Figures 10 and 11 represent examples of the sequences grouped into the four primary clusters. It can be observed that the sequences were properly clustered into the three major patterns: continuous vibration, flat after vibration, and short. This should result from the ability of the clustering method for handling relative proximity.

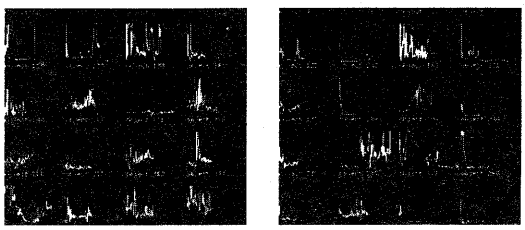

図 9 Other examples of the clusters obtained by AHC-CL. Left: the second cluster containing 39 sequences. Right: the third cluster containing 29 sequences.

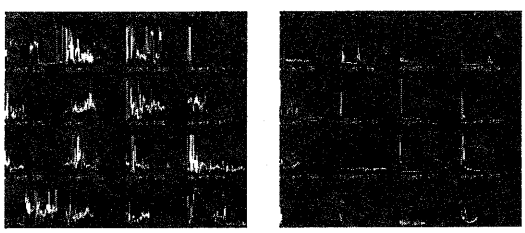

図 10 Examples of the clusters obtained by RC. Left: the second cluster containing 16 sequences. Right: the third cluster containing 10 sequences.

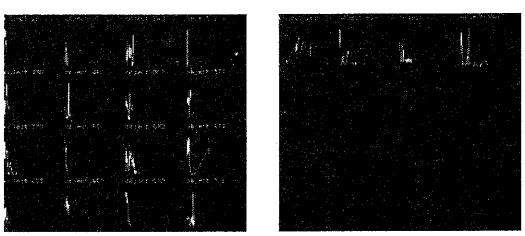

図 11 Other examples of the clusters obtained by RC. Left: the fourth cluster containing 9 sequences. Right: the fifth cluster containing 6 objects.

5. conclusions

In this paper we have reported a comparative study of

clustering methods for long time-series data analysis. Although the subjects for comparison were limited, the results suggested that (1) complete-linkage criterion outperforms average-linkage criterion in terms of the interpret-ability of a dendrogram and clustering results, (2) combination of DTW and CL-AHC constantly produced interpretable results, (3) combination of DTW and RC would be used to find core sequences of the clusters. Multiscale matching may suffer from the problem of 'no-match' pairs, however, the problem may be eluded by using RC as a subsequent grouping method.

Acknowledgments

This work was supported in part by the Grant-in-Aid for Scientific Research on Priority Area (B)(No.759) "Implementation of Active Mining in the Era of Information Flood" by the Ministry of Education, Culture, Science and Technology of Japan.

文　献

[1] E. Keogh (2001): Mining and Indexing Time Series Data. Tutorial at the 2001 IEEE International Conference on Data Mining.

[2] Chu, S., Keogh, E., Hart, D., Pazzani, M. (2002). Iterative Deepening Dynamic Time Warping for Time Series. In proceedings of the second SIAM International Conference on Data Mining.

[3] D. J. Berndt and J. Clifford (1994): Using dynamic time warping to find patterns in time series. Proceedings of AAAI Workshop on Knowledge Discovery in Databases: 359-370.

[4] S. Hirano and S. Tsumoto (2002): Mining Similar Temporal Patterns in Long Time-series Data and Its Application to Medicine. Proceedings of the IEEE 2002 International Conference on Data Mining: pp. 219–226.

[5] N. Ueda and S. Suzuki (1990): A Matching Algorithm of Deformed Planar Curves Using Multiscale Convex/Concave Structures. IEICE Transactions on Information and Systems, J73-D-II(7): 992–1000.

[6] F. Mokhtarian and A. K. Mackworth (1986): Scale-based Description and Recognition of planar Curves and Two Dimensional Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(1): 24-43

[7] URL: http://lisp.vse.cz/challenge/ecmlpkdd2003/

[8] B. S. Everitt, S. Landau, and M. Leese (2001): Cluster Analysis Fourth Edition. Arnold Publishers.

[9] S. Hirano and S. Tsumoto (2003): An Indiscernibility-based Clustering Method with Iterative Refinement of Equivalence Relations. Journal of Advanced Computational Intelligence and Intelligent Informatics, (in press).

[10] Lowe, D.G (1980): Organization of Smooth Image Curves at Multiple Scales. International Journal of Computer Vision, 3:119–130.