Feature Discovery in Temporal Data

Ryutaro ICHISE† and Masayuki NUMAO††

In mining time series data, the graph similarity of the data can be used as an effective tool. However, when the time series has missing data, the utility of graph similarity in analyzing time series data is limited. The present study investigates experimentally the impact of missing values in time series data on dynamic time warping, a method that is commonly used in determining graph similarity. Based on the results of the investigation, we propose a new method by which to treat time series data having missing values. The proposed method uses point similarity rather than graph similarity. Experiments were conducted in order to evaluate the performance of the proposed method, and the results indicate that the proposed method is effective for finding features in time series data.

1. Introduction

Advances in computer technology have made time series data, such as stock market prices, sensor data collected in space and medical test data, readily available. There are several data mining approaches that can be applied to analyze these types of time series, including the window sliding approach[2)] and dynamic time warping[5)]. These methods can be used to identify similar graphs. However, in these methods the time series data are assumed to have been collected continuously. This assumption is not valid for some domains, such as medical test data, which involves numerous missing values. In the present study, we analyze experimentally the impact of missing values in time series data on dynamic time warping, a commonly used method involving graph similarity. We then present a new method that is suitable for time series data having missing values.

This paper is organized as follows. Section 2 discusses time series with missing data. Section 3 describes a typical time series data mining method, which uses the dynamic time warping method and the clustering method. In addition, we show the performance of this typical method for time series data having missing values. In Section 4, a new approach, called the point similarity method, is proposed, and the performance of this new method is investigated experimentally. In Section 5, we discuss related studies. Finally, conclusions are presented in Section 6.

† Intelligent Systems Research Division,
National Institute of Informatics
Hitotsubashi 2–1–2, Chiyoda-ku, Tokyo, 101–8430 Japan
E-Mail: ichise@nii.ac.jp
†† The Institute of Scientific and Industrial Research, Osaka University
Mihogaoka 8–1, Ibaraki, Osaka, 567–0047 Japan
E-Mail: numao@sanken.osaka-u.ac.jp

2. Time Series with Missing Data

Recently, time series data has come to be readily available. For example, web log data is available from web servers and stock market data is easily obtained from securities companies. The extraction of useful knowledge from such data requires the application of a mining method. One of the most common methods of data mining is clustering, which reveals common trends among the data based on the similarity of graphs. However, when there are missing data in the time series, or when the time series data can not be obtained continuously, can methods based on the similarity of graphs be assumed to work well? Let us consider the data listed in Table 1. Each row represents one set of time series data, and each set consists of four data, collected at four different points in time. First, we will examine the problem in which conventional mining methods, which are based on graph similarity, can not find a feature among the data. Assume that we have time series ID1 and ID2, given in Table 1. As shown in the upper left-hand graph in Figure 1, the two example time series look very different. However, if we exclude the datum at time 2 for ID1 and that at time 3 for ID2, we obtain the upper right-hand graph in Figure 1. The shapes of the graphs are now identical. Thus, the graph

Table 1 Example time series.

Data	1	2	3	4
ID1	50	100	50	50
ID2	50	50	0	50
ID1'	-	100	50	50
ID1"	50	-	50	50

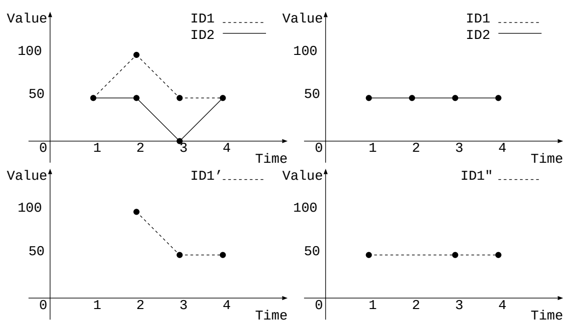

Fig. 1 Graphs of the example time series of Table 1.

similarity approach can not find a feature common to ID1 and ID2. This type of problem may also occur under different circumstances. Looking at the ID1 time series, we first suppose the datum at time 1 is missing. We will call this time series ID1'. Now suppose that for the same original time series (ID1), the datum at time 2 is missing. Let us call this time series ID1". These time series are given in Table 1 and are graphed in the lower left-hand and lower right-hand graphs, respectively, in Figure 1. Even when generated from the same actual time series, the graph shape may differ greatly when data are missing.

From the above discussion, methods based on graph similarity may not be applicable to time series having missing data. Therefore, in the following, we analyze such time series by Dynamic Time Warping (DTW) and clarify the characteristics of clustering using the similarity of graphs.

3. Impact of Missing Data for Graph Similarity Approach

In the previous section, we discussed some of the characteristics of the graph similarity approach. In this section, we introduce a commonly used method called Dynamic Time Warping (DTW)[5] to determine the graph similarity and a simple clustering method that can be used in conjunction with DTW.

3.1 Dynamic Time Warping

DTW is used to determine the distance between two time series. Suppose that we have two time series, D_1 and D_2, consisting of $d_{1,1}, d_{1,2}, \ldots, d_{1,m}$ and $d_{2,1}, d_{2,2}, \ldots, d_{2,n}$, respectively. In calculating the distance between D_1 and D_2, we must determine how to align the two sequences. The DTW distance is defined as the distance for which the best alignment is achieved, and in given by the following equation:

$$\begin{cases} D(m,n) = min(D(m-1,n-1), \\ \qquad\qquad\qquad D(m-1,n), \\ \qquad\qquad\qquad D(m,n-1)) \\ \qquad\qquad\qquad +d(m,n) \\ D(1,1) = d(1,1) \end{cases}$$

$d(i,j)$ denotes distance between $d_{1,i}$ and $d_{2,j}$. Although this equation appears complicated, $D(m,n)$ is easily determined using Dynamic Programming[5].

3.2 Clustering using DTW distance

In this section, a simple clustering method is used to examine the characteristics of the DTW distance. First, we consider each element, which represents one time series data as cluster. Next, based on DTW distance, we select and merge the most similar clusters. We then calculate the distance of the merged cluster. The average linkage between the merged group is used for the calculation. For the average linkage between the merged group, the distance between two clusters is the average distance of the elements among the clusters. Let us now explain the clustering method using a number of example time series. For three example time series e_1, e_2, and e_3, the DTW distances e_1-e_2, e_2-e_3 and e_1-e_3 are assumed to be 5, 8 and 10, respectively. Each example time series is first considered as a cluster. Then, the clustering method merges the e_1 and e_2 clusters because the DTW distance is shortest. A new cluster consisting of e_1 and e_2 is created, and the distance between the new cluster and the e_3 cluster is then calculated to be 9, based on the average DTW distances e_2-e_3 and e_1-e_3. This clustering method continues until user defined number of clusters is obtained.

3.3 Experiments

In order to analyze the performance of the graph similarity approach with regard to time series having missing data, we conducted experiments to determine clusters in synthetic control chart data[1]. The data set examined in this experiment contains 600 time series comprising six different classes of control charts: normal, cyclic, increasing trend, decreasing trend, upward shift, and downward shift. Each time series contains 60 different points in time. Since the original data set does not have missing data, the missing data were generated artificially. The rate of missing data was selected and points in time were deleted from the time series according to the rate of missing data. If all of the data in the time series were deleted, we reapplied the method until a series was obtained that contained at least one point in time.

We tested the performance of the clustering method using the DTW distance for each possible pair of classes. Thus, data from only two classes, containing 100 data each, were used to form two clusters (one for each class). We varied the missing data rate from 0.1 to 0.9 for each pair and generated ten data sets for each missing data rate for each pair. Each data set is different due to the randomness of the point at which the data was deleted.

3.4 Number of Successfully Generated Clusters

The experimental results for the number of successfully generated clusters are listed in Table 2. The experimental domains of normal, cyclic, increasing trend, decreasing trend, upward shift, and downward shift are denoted as NO, CY, IN, DE, UP and DO, respectively. The table indicates, for example, that for the ten data sets for the normal/cyclic pair domain at the missing data rate of 0.4, 60% were successfully classified using DTW.

The above results help to clarify a number of characteristics of the DTW method. First, the simple clustering method using DTW distance is good for data with few missing values. This method has almost perfect classification for missing data rates lower than 0.3, indicating the effectiveness of DTW for time series having few missing data. Next, the performance of DTW distance is low when the missing data rate is high. With the exception of domains having opposite characteristics, the method fails to classify time series with missing data rates above 0.7. This indicates that care must be taken when applying the DTW method to time series having high missing data rates, such as medical data, as described in[3].

Table 2 Number of successfully generated clusters.

	Rate of missing data								
Domain	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
NO-CY	0	0	0	0	0	6	10	9	10
NO-IN	1	5	9	10	10	9	10	10	10
NO-DE	1	8	9	10	10	10	10	10	10
NO-UP	0	1	6	7	10	10	10	10	10
NO-DO	0	1	4	9	9	10	10	10	10
CY-IN	0	3	9	9	10	10	10	10	10
CY-DE	0	3	7	10	10	10	10	10	10
CY-UP	0	0	7	9	9	9	10	10	10
CY-DO	0	0	1	8	10	10	10	10	10
IN-DE	5	10	10	10	10	10	10	10	10
IN-UP	0	0	0	0	0	0	0	0	0
IN-DO	1	10	10	10	10	10	10	10	10
DE-UP	2	10	10	10	10	10	10	10	10
DE-DO	0	0	0	0	0	0	0	0	0
UP-DO	0	7	10	10	10	10	10	10	10

Finally, the DTW method is ineffective for the cyclic domain for time series with missing data. DTW distance fails to classify time series even when the missing data rate is as low as 0.5. This is an example of the problem discussed in Section 2 and indicates that care must be taken when applying the graph similarity method to time series with missing data.

4. Point Similarity Approach

In this section, we propose a new approach for finding effective features by which to distinguish time series data with missing values. As explained in Section 2, one of the problems associated with graph similarity approaches, such as DTW, is the treatment of graph shape. The shape of a graph is not stable for time series data having a large number of missing values. Therefore, the graph shape should be ignored and only the observed data should be used to find features. This means not adding false data when actual data cannot be obtained☆. Then, the problem of finding effective features for distinguishing time series data is how to find similar features in the observed data.

We propose the use of the point similarity approach which is dependent only on the observed data. When different time series have similar

☆ In the graph similarity approach, the time point, which is not obtained as observed data, is created from the nearest data, even when this is not the intent, because the graph is drawn by lines connecting the nearest data.

observation data, we can consider these data as an important feature, even when there is no information regarding the remainder of the data. Let us now return to the previous example, in Table 1. If ID1 and ID2 are in the same class, we can easily find a common feature for those data for which the value of time period 1 is 50 and the value of time period 4 is 50. We propose the use of this type of data point similarity. In other words, if the some of the observed data points are similar, we can treat this similarity as a feature of the data. However, the sparseness of time series data with missing values reduces the number of data obtained for the nearby time point. This phenomena can also occur for the value of the data. A method by which to handle such a problem is needed.

4.1 A Method for Finding Features

In order to determine the similar points in the same class, we introduce the following function:

$$f(t,v) = \frac{g(e)}{p_t p_v} \times exp\left(-\frac{1}{2}\left(\left(\frac{t-t'}{p_t}\right)^2 + \left(\frac{v-v'}{p_v}\right)^2\right)\right)$$

$$g(e) = \begin{cases} +1 & \text{(e is a positive example)} \\ -1 & \text{(x is a negative example)} \end{cases}$$

The function $f(t,v)$ is used for calculating the important feature in the time series data. Here, t' and v' are observed data, namely, the time point and the value in the time series data, respectively. p_t and p_v are domain specific parameters for the weight of importance of different time points and values, and e denotes an example from the entire set of examples E. The above function represents the impact of the observed data for time t and value v.

Let us next present a simplified example to help explain this function. For a positive example observed for time period 3 and value 20, the nearby point of (3, 20) may be a feature of positive examples. The possibility of being a feature is reduced as the distance from the point increases. This possibility is indicated by the function $f(t,v)$. Figure 2 shows a graph having $v = 20$. Note that the function is maximized in time period 3.

Next, we define the following value:

$$V = \sum_{e \in E} f(t,v)$$

We can now obtain features among the data using V. The region having a high value of V is the convergence area of positive examples, and the regions having low values of V are convergence areas of negative examples. Such areas are candidates by which to distinguish positive and negative examples. Note that the regions that are not useful for distinguishing the positive and negative examples are the regions of mid-range values. The proposed method is not affected by missing values. In addition, the method can be applied even when a large number of values are missing.

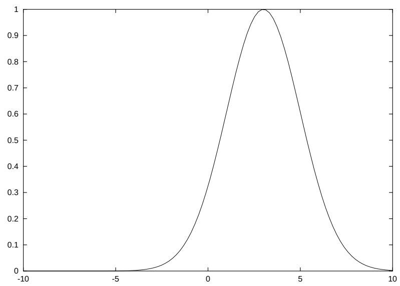

Fig. 2 Example graph of $f(t, 20)$.

4.2 Experiments

In order to evaluate the proposed method, we conducted two experiments.

The first experiment examines time series data without missing values. The synthetic data introduced in Section 3.3 is used. The domain of the data is *increase* and *up*, which is difficult to distinguish using the DTW method. The original data is shown in Figure 3. A visualization method was used in order to simplify the results. Regions having high values of V are presented in lighter shades of grey, and regions having low values of V are presented in darker shades of grey. The visualized image indicates areas that are suitable for features of positive examples (*increase* data) as lighter areas, and areas that are suitable for features of negative examples (*up* data) as darker areas. The visualized image is shown in Figure 4.

The second experiment is conducted using time series data with missing values. The rate of missing values is 0.9, and the data is generated using the method described in Section 3.3. As in the previous experiment, the domain of the data is *increase* and *up*. The experimental data with missing values is shown in Figure 5, and the results of applying the proposed method and visualization technique are shown

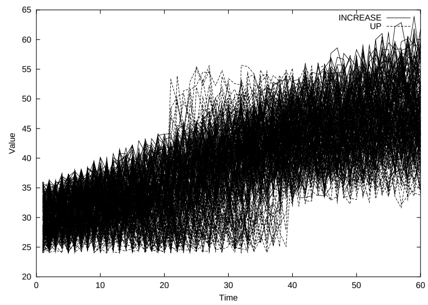

Fig. 3 Example time series of *increase* and *up* without missing values.

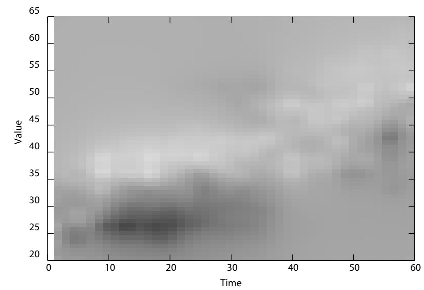

Fig. 6 Obtained image for *increase* and *up* with missing values.

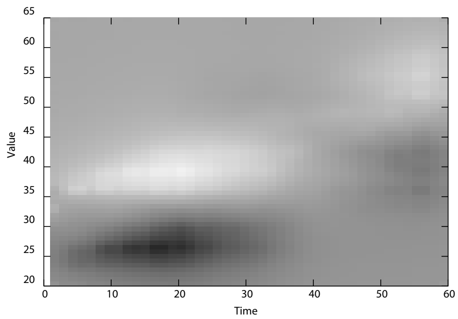

Fig. 4 Obtained image for *increase* and *up* without missing values.

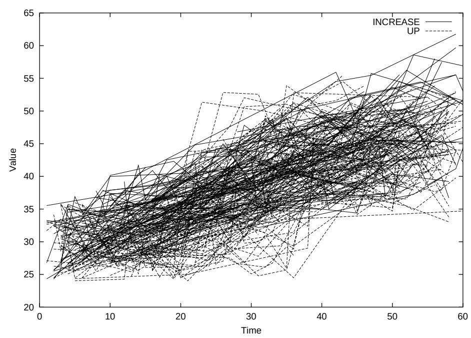

Fig. 5 Example time series of *increase* and *up* with missing values.

in Figure 6.

4.3 Discussion

First, we discuss data without missing values. The proposed method performed well with respect to the graph shown in Figure 3. Upon observing the graph, we select a feature around time period 20 as an important feature. The proposed method successfully derives the feature from the complicated graph. One reason for this is that the function defined in Section 4.1 represents the possibility of occurrence for each example. As such, the values in the regions having several positive examples become high, as illustrated by lighter shades of grey (high luminosity) in Figure 2, and the values in the regions having several negative examples become low, as illustrated by darker shades of grey (low luminosity). On the other hand, the data from time period 40 to time period 50 contains mixed positive and negative examples. Therefore, this region is not useful for distinguishing the examples. This property can also be seen in Figure 4. The luminosity of the region from time period 40 to time period 50 is neither high nor low in the figure.

Next, we discuss data having missing values. The data and the experimentally obtained result are shown in Figure 5 and 6, respectively. As shown in Figure 5, since the data with missing values comprises only 10% of the complete data, the shapes of the graphs for *increase* and *UP* are similar. However, the feature of negative examples is induced around time period 20, as indicated by the darker shades of grey in Figure 6. The feature induced from the complete data set is shown in Figure 4. The proposed method successfully induces the important feature needed to distinguish positive and negative examples, even when using time series data having missing values. In addition, we obtained almost the same image in Figure 4 and 6. This means that the proposed method is robust with respect to data having missing values.

5. Related Studies

The proposed method can be considered as both a visualization technique for finding features of time series data having missing values and a feature discovery method for time series data.

Zooming is a useful visualization technique[6),7)] in which row data is summarized and the user interactively zooms in on the data of interest. The proposed method summarizes the row data in the same manner, but focuses on finding a feature among examples with classes. Prototypeline[8)] is another type of visualization technique that clusters the time series data. However, this method does not treat the row data, whereas the proposed method does.

There are several feature discovery methods that use the graph similarity approach, including the application of DTW[5)]. Another graph similarity approach uses the window sliding approach to find features[2)]. However, this approach is not suitable for time series data with missing values, as discussed in the previous section. TimeSleuth[4)] uses C4.5 to classify examples with time series data. However, due to the strict treatment of time points in this approach, it is also unsuitable for time series data with missing values. Although a method that uses point similarity was proposed previously[3)], unlike the present method, this previous method cannot treat row data including numerical data.

6. Conclusion

The present study investigated two important properties of time series data having missing values. We discussed the limitations of previous approaches and proposed a new method for finding useful features in time series data having missing values. The present study investigated the utility of the DTW distance for time series having missing data. Although DTW is an effective tool for clustering time series data, the DTW method is not suitable for application to time series having missing data in some cases. We analyzed the DTW method using various time series having missing values, and the experimental results indicate the utility of the DTW distance for time series having missing data. In addition, we proposed a point similarity approach that focuses on the similarity of the observed data, rather than the similarity of the graph shape. The basic concept behind this method is the possibility of occurrence of data in the same class. Experiments to examine the performance of this method revealed that features were determined successfully even with 90% of the data missing.

In the future, we hope to develop an effective data mining method for time series having missing data. As indicated by the results of the present study, the use of graph similarity is not compatible with such time series. However, the proposed method can determine the features of examples in the same class. We plan to use these features to construct rule-based knowledge. In addition, we hope to compare the performance of the proposed method with the performances of other systems for various examples.

References

1) Alcock, R. J. and Manolopoulos, Y.: Time-Series Similarity Queries Employing a Feature-Based Approach, *Proceedings 7th Panhellenic Conference in Informatics (PCI'99)*, pp. 1–9 (1999).
2) Das, G., Lin, K.-I., Mannila, H., Renganathan, G. and Smyth, P.: Rule Discovery from Time Series, *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, AAAI Press, pp. 16–22 (1998).
3) Ichise, R. and Numao, M.: Discovery of Temporal Relationships using Graph Structures, *Proceedings of the 2nd International Workshop on Active Mining*, pp. 118–129 (2003).
4) Karimi, K. and Hamilton, H. J.: TimeSleuth: A Tool for Discovering Causal and Temporal Rules, *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence* (2002).
5) Keogh, E. J. and Pazzani, M. J.: Scaling up Dynamic Time Warping for Datamining Applications, *In the Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining*, pp. 285–289 (2000).
6) Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D. and Shneiderman, B.: LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records, Technical Report HCIL-98-08, University of Maryland (1998).
7) Spenke, M.: Visualization and interactive analysis of blood parameters with InfoZoom, *Artificial Intelligence in Medicine*, Vol. 22, pp. 159–172 (2001).
8) Watanabe, T., Suzuki, E., Yokoi, H. and Takabayashi, K.: Prototyping Medical Test Results in Chronic Hepatitis Data with the EM Algorithm on Multi-Dice Models, *Proceedings of International Workshop on Active Mining*, pp. 45–51 (2002).