Pattern distance of time series

D. WangDa & RongGang

National Key Laboratory of Industrial Control Technology,
Zhejiang University, Hangzhou, P. R. China, 310027

Abstract

The pattern model representation (PMR) of time series is proposed in this paper. PMR is based on a piecewise linear representation (PLR) and is effective at describing the tendency of time series. Then, the pattern distance can be calculated to measure the similarity of tendency. This method overcomes the problem of time series mismatch based on point distance. According to the numbers of series’ segmentations, pattern distance has a multi-scale feature and can reflect different similarities with various bandwidths. Because normalization is unnecessary, the calculation consumption of pattern distance is low.

1 Introduction

Knowledge discovery of time series plays an important role in data mining. In particular, dynamic information can be found by analysing the tendency change of time series. Commonly, matching time series is based on the point distance in current methods, see [1, 2]. Normalization is needed because the comparison results are different according to different scales. In addition, the distance value between two time series depends on various normalization methods. Distance based on point is insensitive to the tendency change of time series. It is a statistic value. In Figure 1, the tendency change of series 1 and series 3 are similar, series 2 is relatively not similar to the others. However, a cluster algorithm based on point distance will recognize series 1 and series 2 as one class. On the other hand, time series similarity based on point distance cannot reflect the difference between different analysis frequencies. For example, the point distances, calculated based on 7-day-period and 30-day-period, are approximately equal to each other. We call this a multi-scale feature.

The conception of a pattern model is proposed in this paper. It is based on PLR (piecewise linear representation) series. According to the tendency change,
the pattern model divides time series into some subsets. Each subset represents one pattern model. Then, the distance based on the time series pattern model can be calculated. The experiments show that this distance is useful. It overcomes the shortcoming of point distance and can effectively support multi-scale measurement.

![Figure 1: Example of matching series.](image)

The rest of this paper is organized as follows: In section 2, PLR and PMR of time series are introduced. In section 3, series pattern distance is defined. Its calculating algorithm and performance are discussed. Section 4 provides experiments of time series matching based on series pattern distance and section 5 offers conclusions.

2 Pattern model representation

In this section, PLR time series is introduced. Then PMR is proposed based on PLR and the pseudocode of an algorithm to generate PMR is given.

2.1 PLR&PMR of time series

Piecewise linear representation refers to the approximation of a time series with sequences of straight lines, and has innumerable advantages as a representation, as in Figure 2.

![Figure 2: PLR of time series.](image)
A PLR is represented as
\[ S = \{(y_{iL}, y_{iR}, t_i), \cdots, (y_{K_L}, y_{K_R}, t_K)\} \]  
where \( y_{iL}, y_{iR} \) \( (i = 1, 2, \cdots, K) \) respectively denote the i-th segment line starting value and ending value. \( t_i \) denotes i-th segment ending time and \( K \) is the number of segments. The open problem of PLR, how many segments one time series should be segmented into, is solved by a Bottom Up algorithm, which was proposed by Keogh [3].

A pattern represents one segment monotones tendency. The pattern is a triple, \{Up, Hold, Down\}. To facilitate calculation, the pattern is represented as \( P = \{1, 0, -1\} \). \( p \in P \) is an element of the pattern set. The time series PMR is an order set of the (pattern value, time stamp) pair,
\[ S = \{(p_1, t_1), \cdots, (p_N, t_N)\} \]  
where \( p_i \in P, t_i (i = 1, 2, \cdots, N) \) is the pattern ending time. \( N \) is the number of pattern segments. Formulation (2) denotes the following: pattern \( p_1 \) keeps from zero to \( t_1 \), and pattern \( p_i (i = 2, \cdots, N) \) keeps from \( t_{i-1} \) to \( t_i \).

### 2.2 Algorithm of PLR2PMR

This algorithm is used to convert the piecewise linear representation series to the pattern model representation series.

**Algorithm PLR2PMR**

**Input:**
1. \( S_{PLR} \) series, just as formulation (1);
2. threshold of slope to determine segment’s pattern

**Output:**
\( S_{PMR} \) series of time series, just as formulation (2).

\[
S_{PMR} = [ ];
\]
**For i-th segment in PLR(i = 1, 2, \cdots, K)**

\[
\{
\begin{align*}
p &= \frac{y_{iL} - y_{iR}}{t_i - t_{i-1}}; \\
\text{If} & \ (p \leq -\text{Slope}) \\
& p = -1; \\
\text{Elseif} & \ (p \geq \text{Slope}) \\
& p = 1; \\
\text{Else} & \ p = 0; \\
\text{If} & \ (\text{LastMod}(S_{PMR}) \neq p) \\
& S_{PMR} = [S_{PMR}, (p, t_i)]; \\
\text{Else} & \ LastTime(S_{PMR}) = t_i
\end{align*}
\]
where $K$ is the numbers of time series’ pattern, $LastMod(S)$ is a function to get the last pattern of series PMR $S$, $LastTime(S)$ is a function to get the sample time of series PMR $S$. This algorithm is efficient because it scan the searching space of PLR only once. Let $N$ represent the numbers of time series’ pattern. It’s easy to see that $N < K$. Therefore, the matching space of PMR is smaller than the PLR’s.

3 Time series pattern distance

Pattern distance essentially is a code distance. However the series pattern distance cannot be computed directly. It needs an ‘equal pattern number’ process. The last part of this section discusses the feature of series pattern distance.

3.1 Pattern distance

Let $S = \{(p_1, t_1), \cdots, (p_N, t_N)\}$ be a PRM series. Let $s_i$ be its $i$-th pattern, $s_i = (p_i, t_i)$, where $i = 1,2, \cdots, N$. The difference between the $k$-th pattern $s_j = (p_j, t_j)$ and the $j$-th pattern $s_k = (p_k, t_k)$ can be measured with pattern distance $D_p$ as

$$D_p(s_j, s_k) = |p_j - p_k|$$

(3)

It can be seen that pattern distance has the following properties:

$$D_p(s_j, s_k) = D_p(s_k, s_j)$$

$$D_p(s_j, s_k) \geq 0$$

$$D_p(s_j, s_k) = 0 \iff p_j = p_k$$

$$D_p \in \{0,1,2\}$$

3.2 Series pattern distance

Series pattern distance denotes the similarity of time series, which are the same length. Let $S_1$ and $S_2$ be PMR series to be matched and defined as

$$S_1 = \{(p_{1_1}, t_{1_1}), \cdots, (p_{1_N}, t_{1_N})\}$$

(4)

$$S_2 = \{(p_{2_1}, t_{2_1}), \cdots, (p_{2_M}, t_{2_M})\}$$

(5)

Let $s_{1i}$ and $s_{2j}$ be respectively the $i$-th and $j$-th patterns of $S_1$ and $S_2$ which are defined as

$$s_{1i} = (m_{1i}, t_{1i}) \quad i = 1,2, \cdots, N$$

(6)

$$s_{2j} = (m_{2j}, t_{2j}) \quad j = 1,2, \cdots, M$$

(7)

Let $t_{1ih}$ and $t_{2jh}$ be respectively the hold times of pattern $s_{1i}$ and $s_{2j}$, which are defined as
Because it is whole series matching, \( t_{1N} \) must be equal to \( t_{2M} \). Assume that
\[
t_{1ih} = t_{2jh} \quad i = j \quad \text{Con. 1}
\]

\( S_1 \) and \( S_2 \) are named as EPN (equal pattern number) series if they satisfy Con.1. Then, series pattern distance can be computed as
\[
D_{sp}(S_1, S_2) = \frac{1}{t_N} \sum_{i=1}^{N} t_{ih} D_p(s_{iu}, s_{iu})
\]
where \( t_N \) denotes the length of time series, \( t_{ih} \) denotes the i-th pattern hold time; \( N \) is the total numbers of patterns. \( t_{ih} \) in the formulation is a weight. The longer the hold time is, the more ratios it has in SPD. From formulation (10), it is known that \( D_{sp}(S_1, S_2) \in [0,2] \). From this value, the following can be archived:

1) The compared series tendency is more similar when the SPD runs to zero;
2) The compared series tendency is more reverse when the SPD runs to 2.

Let \( Sim = 1 - D_{sp}(S_1, S_2) \), because the hold pattern is very few in actual series (see example series in Section 4), \( Sim \) represents the similar extent of \( S_1 \) and \( S_2 \). However, there is almost no pair of series which can satisfy Con.1 and the pattern distance cannot be calculated directly just as formulation (10). So an EPN process is needed.

\[
\text{Figure 3: Non-EPN series.}
\]
After the EPN process, EPN PRM is

\[ S'_1 = \{(1,t_1),(-1,t_2),(-1,t_3),(0,t_4)\} \]

\[ S'_2 = \{(1,t_4),(1,t_2),(-1,t_4),(-1,t_4)\} \]

where \( t_1 = t_{11}, t_2 = t_{21}, t_3 = t_{12}, t_4 = t_{13} = t_{22} \).

In fact, during the EPN processing, SPD is archived (see the following algorithm).

**Algorithm EPN&SPD**

**Input:** PMR Series

**Output:** Series Pattern Distance

\[
\begin{align*}
ST &= 0; \\
ET_1 &= 0; \\
ET_2 &= 0; \\
D_{SP} &= 0; \\
Step_1 &= 1; \\
Step_2 &= 1; \\
\end{align*}
\]

**For** (\( Step_1 < N \) And \( Step_2 < M \))

\[
\begin{align*}
ET_1 &= \text{GetPatternEndTime}(S_1, Step_1); \\
ET_2 &= \text{GetPatternEndTime}(S_2, Step_2); \\
P_1 &= \text{GetPattern}(S_1, Step_1); \\
P_2 &= \text{GetPattern}(S_2, Step_2); \\
\end{align*}
\]

**If** \( ET_1 > ET_2 \)

\[
\begin{align*}
D_{SP} &= (ET_2 - ST) \times \text{Abs}(P_1 - P_2); \\
ST &= ET_2; \\
Step_2 &= +++; \\
\end{align*}
\]

**ElseIf** \( ET_1 < ET_2 \)

\[
\begin{align*}
D_{SP} &= (ET_1 - ST) \times \text{Abs}(P_1 - P_2); \\
ST &= ET_1; \\
Step_2 &= +++; \\
\end{align*}
\]

**Else**

\[
\begin{align*}
D_{SP} &= (ET_1 - ST) \times \text{Abs}(P_1 - P_2); \\
ST &= ET_1; \\
Step_2 &= +++; \\
Step_1 &= +++; \\
\end{align*}
\]

\( D_{SP} = DSP / ET_1 \);

### 3.3 Performance analysis

Immune to Sample Time Noise: There is a difference between actual sample time and record time because of the delay time of transmission and so on. Let \( \varepsilon \) be sample time noise. Thus, the pair of (value, time+\( \varepsilon \)) is used to represent sample result. Matching method based on point distance is sensitive to sample
time noise because it accumulates every point distance error. However, SPD has weight and only at the pattern change point will the sample time noise affect the SPD. At the same time, according to Formulation (10), $\varepsilon$ has little effect on SPD because of $\varepsilon << t_N$. Therefore, SPD is immune to sample time noise.

Multi-Scale Feature: Pavlidis and Horowitz [5] pointed out that PLR can compress and filter raw data. So different segments of PLR have different filter bandwidth. Based on this, series pattern distance has a multi-scale feature. The more segments, the higher frequency signal of series is compared.

4 Empirical comparison

In this section, an example of three stock indexes, which started at 1986-6-6 and has 2700 points, is provided to compare two kinds of distance. After normalization, they are just as in Figure 4, where $S_1$ is Hong Kong (Hang Seng), $S_2$ is New York (S&P 500) and $S_3$ is Paris (CAC 40) [4].

![Figure 4: Normalization of three stock indexes.](image)

4.1 Different distance comparison

The series pattern distance and point distance results are shown in Table 1, where the number of segments is 50 and points distance is Euclidean distance,

$$D_{(s_1,s_2)} = \sqrt{\sum (s_{1i} - s_{2i})^2}$$

where SPD is series pattern distance and PD is point distance. It can be found that three values of SPD are approximately equal to each other. This means that
the tendencies of the three series are similar. However, S1 and S3 will be recognized as one class based on point distance. Obviously, this tendency classification is improper.

Table 1: Comparison of distance.

<table>
<thead>
<tr>
<th></th>
<th>S1 &amp; S2</th>
<th>S1 &amp; S3</th>
<th>S2 &amp; S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPD</td>
<td>0.427</td>
<td>0.413</td>
<td>0.465</td>
</tr>
<tr>
<td>PD</td>
<td>21.197</td>
<td>5.443</td>
<td>21.076</td>
</tr>
</tbody>
</table>

4.2 Multi-scale feature

Series pattern distances, which are calculated under different segments, are shown in Table 2.

Table 2: 2 multi-scale feature.

<table>
<thead>
<tr>
<th>Segs</th>
<th>S1 &amp; S2</th>
<th>S1 &amp; S3</th>
<th>S2 &amp; S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.358</td>
<td>0.724</td>
<td>0.628</td>
</tr>
<tr>
<td>20</td>
<td>0.823</td>
<td>0.375</td>
<td>0.942</td>
</tr>
<tr>
<td>50</td>
<td>0.427</td>
<td>0.413</td>
<td>0.465</td>
</tr>
</tbody>
</table>

From Table 2, the following can be found: the tendencies of S1 and S2 are similar when the series is divided into 10 segments, S1 and S3’s tendencies approximately equal to each other when the number of segments is 20. According to different analysis bandwidths, various results occur.

5 Conclusion

Pattern distance can effectively reflect the tendency similarity of time series. Based on the data filter function of PLR, Pattern distance has a multi-scale feature. It can describe the series tendency with different bandwidths. This provides a new attribute for a data mining job.

Acknowledgement

The National Hi-tech R&D Program of China supported this paper.

References


