Matching State-Based Sequences with Rich Temporal Aspects

Aihua Zheng1,2 *Jixin Ma2 Jin Tang1 Bin Luo1
1Anhui University, Hefei 230039, China, (00)86 0551 5108445
2University of Greenwich, London SE10 9LS, United Kingdom, (00)44 020 8331 8539
ahzheng214@gmail.com; j.ma@gre.ac.uk; ahhftang@gmail.com; luobin@ahu.edu.cn

Abstract
A General Similarity Measurement (GSM), which takes into account of both non temporal and rich temporal aspects in elucidating temporal order, temporal duration and temporal gap, is proposed for state sequence matching. It is believed to be versatile enough to subsume representative existing measurements as its special cases.

Various similarity measurements have been developed over the past half century for state-sequence matching. However, most existing similarity measurements characterize temporal distance only in terms of the temporal order over state-sequences, where other important temporal features such as temporal duration of each state itself, temporal gap between two adjacent states, etc., have been neglected. The only noted exception is TWED [Marteau 2008] which addresses temporal gap difference in term of the temporal index of states while temporal duration of states is not dealt with at all. In addition, in most existing systems, time-series and state-sequences are simply expressed as lists (time-stamps) in the form of \( t_1, t_2, \ldots, t_n \) (or \( s_1, s_2, \ldots, s_n \)), where the fundamental time theories based on which time-series and sequences are formed up are usually not explicitly specified. Based on a formal characterization of time-series and state-sequence, the objective of this paper is to propose a general similarity measurement (GSM) which accommodates two folds of state-sequence matching:

i. Non-temporal matching between the two sets of states that appear in a given pair of state-sequences, regardless of any temporal issues.

ii. Temporal matching between the given two state-sequences, which deals general temporal aspects, including:

   i. Temporal Order: the order relation over the states to be matched in the two given state-sequences. E.g., state \( s_1 \) is “before” state \( s_2 \). As shown in Figure 1.

   ii. Temporal Duration: the duration of each state. E.g., \( T_{dur} \) as shown in Figure 1.

   iii. Temporal Gap: the possible time delay between two adjacent states. E.g., \( T_{gap} \) as shown in Figure 1.

Figure 1. Temporal Gap and Temporal Duration

General Similarity Measurement for Formal State-sequence Matching

A general time-series is formally defined in terms of the following schema:

GTS1) \( T_n= [t_1, \ldots, t_n] = [\langle p_1, q_1 \rangle, \ldots, \langle p_n, q_n \rangle] \)

GTS2) \( \text{Meets}(t_i, t_{i+1}) \) \( \text{Before}(t_i, t_{i+1}) \) for all \( i = 1, \ldots, n-1 \)

GTS3) \( T_{dur}(t_i) = q_i - p_i = d_i \) for some \( i \) where \( 1 \leq i \leq n \)

GTS4) \( T_{gap}(t_i, t_{i+1}) = p_{i+1} - q_i = g_i \) for \( i = 2, \ldots, n \) and \( g_1 = 0 \) and in turn, the corresponding schema for state-sequence is defined as:

GSS1) \( S_n = [s_1, \ldots, s_n] \)

GSS2) \( \text{Holds}(s_i, t_i) \) for all \( i = 1, \ldots, n \)

where \( [t_1, \ldots, t_n] = T_n \) is a time-series.

Based on above formalization, the triple domain \( U = S \times D \times G \) is defined for state-sequences, where:

- \( S \subseteq \mathbb{R}^d \): \( d \)-dimensional domain of non-temporal values well-ordered in consequential temporal order;
- \( D, G \subseteq \mathbb{R} \): the domains of temporal duration and temporal gap respectively.

A given pair of state-sequences can be expressed as \( A_m = [a_1, \ldots, a_m] \), \( B_n = [b_1, \ldots, b_n] \) \( U \) where for \( i = 1, \ldots, m, j = 1, \ldots, n \):

\[ a_i \prec s_j, d_i, g_i >, b_j \prec s'_j, d'_j, g'_j > \in S \times D \times G. \]

The general similarity measurement is formulated as:

\[
GSM(A_m, B_n) = w_{ntem} \text{Dis}_{ntem}(A_m, B_n) + w_{tem} \text{Dis}_{tem}(A_m, B_n) \quad (1)
\]

where \( \text{Dis}_{ntem}(A_m, B_n) \) and \( \text{Dis}_{tem}(A_m, B_n) \) denote the non-temporal distance and temporal distance, respectively with the corresponding weight \( w_{ntem} \) and \( w_{tem} \).

Non-temporal distance

Regardless of temporal order, there are \( \# \text{Pr}_m = m!n!/(m-n)! \) ways of pairing \( A_m \) and \( B_n \) (assuming \( m \geq n \)). The non-temporal distance is thus defined as below:
The temporal distance between two given state-sequences $A_m$ and $B_n$ with respect to the 3 temporal aspects is recursively defined as below:

$$\text{Dis}_{\text{tem}}(A_m, B_n) = \min_{pr} \text{dis}_{\text{tem}}(pr, B_n)$$

Where $pr = [pr_1, ..., pr_n]$ and

$$\text{dis}_{\text{tem}}(pr, B_n) = \sqrt{\sum_{j=1}^{n} w_{pr} \text{dis}_p(pr_j, s_j^r)^2 / \sum_{j=1}^{n} w_{pr}}$$

### Temporal distance

The temporal distance between two given state-sequences $A_m$ and $B_n$ with respect to the 3 temporal aspects is recursively defined as below:

$$\text{Dis}_{\text{tem}}(A_m, B_n) = \min_{pr} \text{dis}_{\text{tem}}(pr, B_n)$$

$$\text{dis}_{\text{tem}}(pr, B_n) = \sqrt{\sum_{j=1}^{n} w_{pr} \text{dis}_p(pr_j, s_j^r)^2 / \sum_{j=1}^{n} w_{pr}}$$

The initialization is set as below:

$$\text{Dis}_{\text{tem}}(A_0, B_0) = 0$$

$$\text{Dis}_{\text{tem}}(A_0, B_i) = \infty, \text{ for } i \geq 1$$

$$\text{Dis}_{\text{tem}}(A_i, B_0) = \infty, \text{ for } i \geq 1$$

The cost functions of GSM are defined as below:

$$C_{\text{ord}}(a_i \rightarrow b_j) = \begin{cases} \text{dist}_p(0, s_j^r) & \text{if } s_j^r = \phi \\ \text{dist}_p(s_j^r, 0) & \text{if } s_j^r = \phi \\ \text{dist}_p(s_j^r, s_j^r) & \text{else} \end{cases}$$

$$C_{\text{dur}}(a_i \rightarrow b_j) = \begin{cases} \text{dist}_p(0, d_j^r) & \text{if } d_j^r = \phi \\ \text{dist}_p(d_j^r, 0) & \text{if } d_j^r = \phi \\ \text{dist}_p(d_j^r, d_j^r) & \text{else} \end{cases}$$

$$C_{\text{gap}}(a_i \rightarrow b_j) = \begin{cases} \text{dist}_p(0, g_j^r) & \text{if } g_j^r = \phi \\ \text{dist}_p(g_j^r, 0) & \text{if } g_j^r = \phi \\ \text{dist}_p(g_j^r, g_j^r) & \text{else} \end{cases}$$

### Experimental Results

The GSM has been tested on 6 benchmark datasets. Table 1 shows the clustering accuracy of K-means on each of these dataset. Generally speaking, GSM has the highest accuracy which means it outperforms all the other Binary-value Measurements.

### Reference