Matching State-Based Sequences with Rich Temporal Aspects

Aihua Zheng[1,2] *Jixin Ma[2] Jin Tang[1] Bin Luo[1]

[1]Anhui University, Hefei 230039, China, (00)86 0551 5108445

[2]University 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 cluding temporal order, temporal duration and temporal gap, is proposed for state sequence matching. It is believed to be versatile enough to subsume representative existing meas urements 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:

(1) Non-temporal matching between the two sets of states that appear in a given pair of state-sequences, regardless of any temporal issues.
(2) 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.

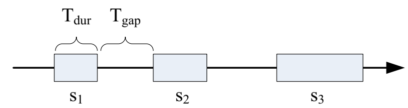

Figure 1. Temporal Gap and Temporal Duration

General Similarity Measurement for Formal State-sequence Matching

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

GTS1) $T_n = [t_1, \ldots, t_n] = [<p_1, q_1>, \ldots, <p_n, q_n>]$
GTS2) $Meets(t_i, t_{i+1}) \vee Before(t_i, t_{i+1})$, for all $i = 1, \ldots, n-1$
GTS3) $T_{dur}(t_i) = q_i - p_i = d_i$, for some i where $1 \le i \le n$
GTS4) $T_{gap}(t_{i-1}, t_i) = p_i - q_{i-1} = g_i$ for $i = 2, \ldots, n$ and $g_1 = 0$

and in turn, the corresponding schema for state-sequence is given as:

GSS1) $S_n = [s_1, \ldots, s_n]$
GSS2) $Holds(s_i, t_i)$, 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:

$S \subset R^d$: *d*-dimensional domain of non-temporal values well-ordered in consequential temporal order;

$D, G \subset R$: the domains of temporal duration and temporal gap respectively.

A given pair of state-sequences can be expressed as $A_m = [a_1, \ldots, a_m]$, $B_n = [b_1, \ldots, b_n] \in U$ where for $i = 1, \ldots, m$, $j = 1, \ldots, n$: $a_i =< s_i^{'}, d_i^{'}, g_i^{'} >$, $b_j =< s_i^{"}, d_i^{"}, g_i^{"} > \in S \times D \times G$.

The general similarity measurement is formulated as:

$$GSM(A_m, B_n) = w_{ntem} Dis_{ntem}(A_m, B_n) + w_{tem} Dis_{tem}(A_m, B_n) \quad (1)$$

where $Dis_{ntem}(A_m, B_n)$and $Dis_{tem}(A_m, B_n)$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 ${}^{m}Pr_n = m!n!/(m-n)!$ ways of pairing A_m and B_n (assuming $m \ge n$). The non-temporal distance is thus defined as below:

*Corresponding author. This research is supported in part by UK RAE funding for Computer Science Group in the School of Computing and Mathematical Sciences, University of Greenwich (9414/0661/76602), National Nature Science Foundation of China (No. 61073116) and 2012 NSFC RSE joint funded project.

$$Dis_{ntem}(A_m, B_n) = min_{pr\in P_r} dis_{ntem}(pr, B_n) \quad (2)$$

Where $pr = [pr_1,\ldots, pr_n]$ and

$$dis_{ntem}(pr, B_n) = \sqrt{\sum_{j=1}^{n} w_{ipr} dis_{Lp}(pr_j, s''_j)^2} \Big/ \sqrt{\sum_{i=1}^{n} w_{ipr}}$$

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:

$$Dis_{tem}(A_m, B_n) = \min \begin{cases} Dis_{tem}(A_{m-1}, B_n) + W_{del} C(a_m \to \phi) \\ Dis_{tem}(A_m, B_{n-1}) + W_{ins} C(\phi \to b_n) \\ Dis_{tem}(A_{m-1}, B_{n-1}) + W_{sub} C(a_m \to b_n) \end{cases} \quad (3)$$

where $C(a_m \to \phi)$, $C(\phi \to b_n)$ and $C(a_m \to b_n)$ denote the cost function for edit operations *deletion*, *insertion* and *substitution*, respectively with m, n≥1.

$$C(x \to y) = \begin{cases} \sum_i w_i \cdot C_i(x \to y) & if\ C_i(x \to y) \le \delta \\ k & else \end{cases} \quad (4)$$

Where $(x \to y) \in \{(a_m \to \phi), (\phi \to b_n), (a_m \to b_n)\}$ and k is a constant usually set either as 0 (to filter out the noise), or as the current maximum cost (to release the influence of the noise).

The initialization is set as below:

$$Dis_{tem}(A_0, B_0) = 0,$$
$$Dis_{tem}(A_0, B_j) = \infty, \text{ for } j\ge 1 \quad (5)$$
$$Dis_{tem}(A_i, B_0) = \infty, \text{ for } i\ge 1$$

The cost functions of GSM are defined as below:

$$C_{Tord}(a_i \to b_j) = \begin{cases} dist_{Lp}(0, s''_j) & if\ s'_i = \phi \\ dist_{Lp}(s'_i, 0) & if\ s''_j = \phi \\ dist_{Lp}(s'_i, s''_j) & else \end{cases} \quad (7)$$

$$C_{Tdur}(a_i \to b_j) = \begin{cases} dist_{Lp}(0, d''_j) & if\ d'_i = \phi \\ dist_{Lp}(d'_i, 0) & if\ d''_j = \phi \\ dist_{Lp}(d'_i, d''_j) & else \end{cases} \quad (8)$$

$$C_{Tgap}(a_i \to b_j) = \begin{cases} dist_{Lp}(0, g''_j) & if\ g'_i = \phi \\ dist_{Lp}(g'_i, 0) & if\ g''_j = \phi \\ dist_{Lp}(g'_i, g''_j) & else \end{cases} \quad (9)$$

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.

Dataset / Method	AT&T face	USPS	MNIST	COIL20	Isolet1	Bin Alpha
OED	65.39	60.50	54.95	59.84	65.85	68.96
EDR	76.92	66.87	66.31	61.28	70.49	71.32
LCSS	74.57	66.25	52.96	53.74	60.37	56.44
CLCS	60.23	57.64	50.35	51.87	55.24	53.49
ACS	75.84	73.85	55.66	60.55	64.85	60.55
T-WLCS	72.59	70.17	58.23	66.62	66.36	61.21
GSM	**78.36**	**76.41**	**66.35**	**69.20**	**75.58**	**72.66**

Table 1. Clustering accuracy comparison

Table 2 below shows the average mean and standard deviation (STD) of retrieval precision on each noised dataset with Gaussian noise with respect to different means and variances which verifies the effectiveness of GSM.

Dataset / Statistic		AT&T face	USPS	MNIST	COIL20	Isolet1	Bin Alpha
ERP	**Mean**	63.71	65.60	59.48	61.53	74.66	71.25
	STD	0.1249	0.1391	0.1742	0.2519	0.1285	0.1595
DTW	**Mean**	73.37	72.29	65.79	73.11	78.51	74.29
	STD	0.1932	0.1128	0.1890	0.1438	0.0891	0.1032
TWED	**Mean**	79.95	75.30	68.80	72.96	79.38	76.90
	STD	0.0993	0.1025	0.1359	0.1235	0.0940	0.0895
GSM	**Mean**	**85.65**	**80.54**	**74.82**	**78.44**	**84.19**	**82.84**
	STD	**0.0632**	**0.0738**	**0.1022**	**0.0983**	**0.0593**	**0.738**

Table 2. Statistic of the retrieval precision of noised dataset

Table 3 presents the classification precision with different combinations of temporal aspects. It shows that GSM is capable of tackling most matching tasks involving time-series and state-sequence data, especially with different temporal matching requirements.

Dataset / Aspects	AT&T face	USPS	MNIST	COIL20	Isolet1	Binary Alpha
$\boldsymbol{T_{ord}}$	87.50	90.69	85.40	87.08	89.23	86.00
$\boldsymbol{T_{dur}}$	91.00	86.56	82.20	88.75	90.13	87.18
$\boldsymbol{T_{gap}}$	88.50	87.12	83.80	88.47	89.87	87.77
$\boldsymbol{T_{ord}+T_{dur}}$	89.50	89.61	86.80	89.86	92.69	90.73
$\boldsymbol{T_{ord}+T_{gap}}$	90.50	91.44	89.20	89.72	93.21	89.15
$\boldsymbol{T_{dur}+T_{gap}}$	87.50	90.77	86.60	89.86	92.82	90.34
$\boldsymbol{T_{ord}+T_{gap}+T_{dur}}$	**94.00**	**93.53**	**89.80**	**91.81**	**94.23**	**92.90**

Table 3. Classification precision with various combinations

Reference

Marteau P.F. “Time Warp Edit Distances with Stiffness Adjustment for Time Series Matching”. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, In Press(0):1-15, April 2008.