Biomedical Signal Processing and Control 15 (2015) 60–66

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Analysis of similarity measure in the longitudinal study using improved longest common subsequence method for lung cancer

LiYing Fang[a], Min Wan[a,*], MingWei Yu[b], JianZhuo Yan[a], Zheng Liu[a], Pu Wang[a]

[a] College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China
[b] Hospital of Traditional Chinese Medicine, CPUMS, Beijing 100010, China

ARTICLE INFO

Article history:
Received 4 November 2013
Received in revised form
18 September 2014
Accepted 23 September 2014

Keywords:
Longitudinal data
Time-series
Similarity measure
Longest common subsequence
Hierarchical clustering

ABSTRACT

Background: In clinical practice, longitudinal data can be used to find trend patterns of pathema progress, such as tumour progress, along a time axis. This kind of data can be treated as time-series data. The maximum common sub-sequence is the most common method for calculating similarity of time-series data; and each point is normally treated as having the same weight. However, not all points of data within the time series should be given the same importance. According to clinical experience, the later period sub-sequence (closer to death) has a more significant effect than earlier periods in a trend analysis.
Results: A weighted-similarity measure based on LCSS with Constraint Window (W-LCSS-CW) Method is proposed. The results obtained from the time-series data using different weighting factors are discussed. In a study of non-small cell lung cancer using time-series data, the relative evaluation method and external evaluation method were adopted to calculate cluster effect. The results show that the proposed method, W-LCSS-CW, can improve clustering performance significantly. Clustering performance of various methods was performed using a comparison of (C_{index}/M_{index}). The proposed W-LCSS-CW Method was evaluated to 1.55 which was 37.02%, 48.01%, 49.64% higher than other common methods (Euclidean, DTW, STS) respectively.
Conclusions: The proposed W-LCSS-CW Method is recommended for monitoring time-series data of tumour patients because the incorporated weighting factor provides more convincing cluster results for medical assist support.

1. Background

Longitudinal data, unifying the merits of cross-sectional data and time-series data, are prevalent in the biological and social sciences, in which repeated measurements have been taken on a cohort of subjects at a sequence of time points or other conditions. When focusing on a single cross-sectional characteristic, the target is to excavate embedded trend patterns, and to assist in classification or prediction indirectly. Thus, the longitudinal data are treated as time-series data for which similarity can be calculated. Similarity search, first described by Agrawal [1] of IBM (1993) and Faloutsos et al. [2], has become an important branch of Time Series Data Mining (TSDM). Due to the increasing need for trend analysis, TSDM combined with query, match, classification and clustering can be applied to finding DNA patterns [3], analyzing and retrieving object trajectories in two or three dimensional space [4], and explaining some phenomena of earthquakes [5], for example.

During the clinical cancer treatment period, a patient's follow-up records include massive symptom-measurement data (such as routine blood tests or tumour size), intervention (such as medicine or radiotherapy), and periodic evaluation (such as FACT-L score or progress report). A patient's follow-up record forms a group of time-series data, which implies the trend of the tumour's progress. By utilizing longitudinal data, especially the embedded trend information in time-series aspect, the unknown future disease status can be predicted, and the dynamic symptom progress characteristics can be inferred from growth rates of tumour size over different time periods, rather than analysing the baseline data only. From the perspective of statistical analysis, a similarity measurement method can provide assist support and reveal progress rules for doctors, and ameliorate the defect that clinicians can only predict the disease progression by subjective experiences.

The clinical follow-up records, treated as a special case of time series data, have the following five features:

* Corresponding author. Tel.: +86 15810127825.
E-mail address: wanmin625@139.com (M. Wan).

http://dx.doi.org/10.1016/j.bspc.2014.09.010
1746-8094/

(1) *Short time-series*. Different from financial and signal data, the length of follow-up records of advanced stage cancer patients cannot be very long, which limited to the patients' disease progression and the privacy protection. In this experiment, the time-series have no more than 30 sampling points, generally within a 2 year observation period.
(2) *Unequal sequence length*. Because of individual differences in the rate of death or expulsion from the study, patients' follow-up records are of unequal length. The average length is 15–18 sampling points in this experiment.
(3) *Shape characteristic*. Based on specific clinical research background, the shape characteristic of time-series sequence can indicate the embedded information. For the FACT-L Score as an example, its shape characteristic is more significant than the accurate value.
(4) *Sequence with different starting times*. As clinical observation does not begin at exactly at the same stage for each patient; a patent's whole time-series data may start at different time on time axis.
(5) *Incomplete sampling point matching*. Patients with similar pathologies may exhibit dissimilar progress trends. Also, the clinical time-series data is short and of unequal length, explained in feature 1 and 2. Furthermore, all of the patients physical status are not exactly the same own to its personality. Thus, as long as the sequence is long enough to show the trend pattern, the sampling points with several outlier are accepted.

Proposing a method for similarity measurement for this special kind of time series data is the core target of this paper. In general, series similarity measurement algorithms include methods based on correlation coefficient, longest common subsequence (LCSS), Euclidean Distance, Dynamic Time Wrapping (DTW), Short Time Series Distance (STS), and frequency range based on Discrete Fourier Transformation (DFT) or Discrete Wavelet Transform (DWT) [6]. These methods have different features, and each has its own limitations in analysis of the data described type of data sets. Agrawal et al. [1] showed that the methods based on DFT or DWT could get better performance as time-series length increases. Refs. [4,7] concluded that, on the one hand, the method based on Euclidean Distance is not suitable for series of unequal length; while on the other hand, incomplete matching points obviously affect the performance of Euclidean and DTW methods. Both STS and LCSS can deal with shape characteristics, while STS is not suitable for unequal series [8]. Both DTW and LCSS can handle unequal series with wrapping; however, DTW needs to match each point, making it more time consuming [9,10]. Therefore, methods based on LCSS are more suitable for clinical follow-up data. Moreover, the above methods generally treat each sample point as having equal importance. However, according to clinical experience, for patients having similar symptoms, later-period sub-sequences (closer to death) have more significant effect than earlier sub-sequences in analysis of trend patterns.

LCSS was proposed by Wagner and Fisher in the 1970s [11], and widely applied in the field of DNA analysis, pattern identification, text match and automation production [12–14]. The proliferation of LCSS enabled many researchers to focus on improving LCSS: such as, CLCSS—adding constraint to the LCSS [15–20], LAPCS—adding nested arc annotations in RNA structure [10,21], ASM—discussing an approximate string matching problem [22–24] and so on. Nowadays, methods based on LCSS have been widely adopted [12–14]. This paper describes an improvement of the similarity measure method based on LCSS within the constraint described in the Similarity Measure Method section. Further, the paper addresses the deficiency in using the weighting factor of time-sequence by describing the use of the discussed follow-up time series data.

2. Methods

To describe the progress trend patterns in clinical follow-up data and to propose a more accurate approach for the time-series analysis, the definitions are described in the Question description part.

2.1. Question description

Definition 1 *(A follow-up event e).* During the observation period, a cancer patient's routine inspection and treatment can be described as a follow-up event, denoted by e. The e can be expressed by $n+1$-tuple $\langle s_1, s_2, \ldots, s_n, t\rangle$, in which $s_1, s_2, \ldots, s_n$ are sample points at time t. For example, s_1 is the value of tumour marker, s_2 is the FACT-L score, etc. In order to reach the target and simplify the issue, data dimension can be reduced from n to 1. In this case, the e can be transformed to 2-tuple $\langle s, t\rangle$. However, according to clinician experiences, a patient's follow-up interval is changed over the treatment period, ranging from 2 weeks, to 1 month or 2 months. So, sampling-interval-changing event is generalized to the number of follow-up times t, $t \in N^+$. Furthermore, s represents different meanings, such as FACT-L score, FACT-L score gradient, and FACT-L score variation tendency. All of these representations will be used in experiments.

Definition 2 *(A follow-up sequence S^m).* A follow-up sequence S^m, S for short, is made up of a patient's m follow-up events according to t time sequence, denoted by $S^m = e_1, e_2, \ldots, e_m$ ($m \in N^+$), in which $e_k (k \in [1, m])$ is the kth follow-up event. $|S|$ represents the length of S, assumed to be the number of follow-up events, $|S| = m$.

Definition 3 *(A limited set U of follow-up sequences).* All of a patient's follow-up sequences form a limited set, denoted by $U = \{S_n^{m_n} | n, m_n \in N^+\}$, in which $S_p^{m_p} (p \in [1, n])$ is the pth patient follow-up sequence with length of m_p. Because a patient's follow-up intervals could be unequal, m_p, m_q may be unequal when $p, q \in [1, n]$ and $p \neq q$.

Definition 4 *(Similarity of follow-up sequences $sim(S_p, S_q)$).* If S_p and S_q are two random sequences of U, $sim(S_p, S_q)$ represents their similarity, $sim(S_p, S_q) \in [0, 1]$, and $sim(S_p, S_q) = sim(S_q, S_q)$. If $S_p = S_q$, $sim(S_p, S_q) = 1$. If the difference between S_p and S_q is larger than a given threshold σ, $sim(S_p, S_q) = 0$.

Definition 5 *(A cluster of follow-up sequences).* The limited set U can be divided into K clusters according to their similarity, i.e. $U = \{C_i | i \in [1, K]\}$; meanwhile $\bigcup_{i=1}^{K} C_i = U$ and $C_i \cap C_j = \Phi$, so that the objects in the same cluster are more similar (in some sense) to each other than to those in other clusters.

2.2. Similarity measurement

From the background analyses, the measurement based on LCSS was chosen to calculate similarity of follow-up sequences. However, its measurement performance was unsatisfactory because it couldn't accurately distinguish different categories of samples, especially if ignoring the weighting factor of different sample points. Therefore, considering the data features and relative medical knowledge, this paper describes an improved method for the traditional similarity measure based on LCSS and proposes a weighting similarity measure based on LCSS with Constraint Window (W-LCSS-CW) for categorical variables.

2.2.1. The weighted similarity measure based on LCSS with Constraint Window (W-LCSS-CW)

For any S_p and S_q, their LCSS is the maximum common subsequence, denoted by $LCS(S_p, S_q)$. The length of the $LCS(S_p, S_q)$