Available online at www.sciencedirect.com
ScienceDirect
Transactions of A. Razmadze Mathematical Institute 170 (2016) 352–362
Transactions of A. Razmadze Mathematical Institute
www.elsevier.com/locate/trmi

Original article

A Novel similarity measure based on eigenvalue distribution

Xu Huang [a], Mansi Ghodsi [b], Hossein Hassani [b,*]

[a] *Business School, Bournemouth University, UK*
[b] *Institute for International Energy Studies, Tehran, 1967743 711, Iran*

Received 8 May 2016; received in revised form 2 August 2016; accepted 17 August 2016
Available online 19 September 2016

Abstract

Due to the rapidly increasing interests of effective and efficient data processing, the developments of similarity measure have been significantly expanded. This paper defines the eigenvalue distribution as a criterion of measuring similarity in a multivariate system. The primary evaluations are conducted by simulations with the assistances and comparisons of several empirical statistical tests. Furthermore, the proposed measure is conducted in simultaneous real case scenario by adopting the bootstrap re-sampling technique. It also overcomes the difficulty of different series lengths in the multivariate system. Moreover, it does not have pre-assumptions on distributions, and it can be easily employed and efficiently computed.

Keywords: Similarity measure; Eigenvalue distribution; Singular value decomposition; Multivariate

1. Introduction

The studies of similarity have been overwhelmingly explored and applied in various disciplines on many different formats, for example, numerical values [1,2], images [3,4], genes [5–7], chemical subjects [8–10], words [11,12] and so on. According to [13], the similarity measure is the most essential core element of time series classification and clustering. Therefore, the development of better similarity measure can significantly assist the improvement of data analysis efficiency. According to [14], the similarity measure is closely related to the distance measure, as the distance is defined as a quantitative degree of how far apart two objects are. Consequently, studies of distance and similarity are significantly connected and crucial in terms of solving many pattern recognition related problems, such as clustering technique [15,16], Taxonomy [17,18], image registration [19,20], etc.

As one of the crucial difficulties in similarity measure is that the different types of features are not comparable, this paper proposes the novel similarity measure based on the eigenvalue distribution, which is inspired by the dynamical approach and embedding theorem where a one dimensional time series will be transferred to multidimensional time series in a Hankel matrix. Hankel matrix has many features as a square matrix, where gives a sequence of the

* Corresponding author.
E-mail address: hassani.stat@gmail.com (H. Hassani).
Peer review under responsibility of Journal Transactions of A. Razmadze Mathematical Institute.

http://dx.doi.org/10.1016/j.trmi.2016.08.001

one dimensional time series, also defines the dynamical state-space. This paper is the initial attempt of adopting eigenvalue distribution into formulating a similarity measure in the multivariate system. Time series under evaluation are embedded into multidimensional matrices and combined either vertically or horizontally to be transformed into a Hankel matrix, where the eigenvalues can be extracted by Singular Value Decomposition (SVD) technique accordingly. As Aristotle claimed in [21], the Formal Cause is "the account of what-it-is-to-be", or "what makes a thing one thing rather than many things". Based on the "formal cause" claimed by Aristotle, here in this paper, we define the corresponding distribution of extracted eigenvalues as the "formal" criterion for developing a novel similarity measure. The successful implementation of this novel similarity measure can overcome the limitations of nonlinear dynamic, complex fluctuations and the possibility of distinguishing similarity for particular or selected features.

In order to evaluate the reliability of eigenvalue distribution as the similarity measure, three empirical statistical tests together with the real case scenario are overwhelmingly considered. Possible circumstances during the formulation process of the new measure are comprehensively evaluated with brief introductions and comparisons in following sections.

In general, this paper is structured as follows: Section 2 briefly introduce the techniques for obtaining the corresponding eigenvalue distribution. The review of some empirical methods and the formulation of proposed novel similarity measure are listed in Section 3. Section 4 provides the empirical results and evaluations by simulations, whilst the real case scenario results are stated in Section 5. Finally, the discussion and conclusion are summarized in Sections 6 and 7 respectively.

2. Eigenvalue distribution

To overcome the difficulty of existing diverse and incomparable features, the novel similarity measure extracts the corresponding eigenvalue distributions as the formal criterion by considering the elements of time series as a whole without removing any nonlinear or complex features. Note that as the structures of constructing Hankel matrix containing multiple variables differ, including both horizontal and vertical forms.

Consider M time series with different series length N_i $Y_{N_i}^{(i)} = (y_1^{(i)}, \ldots, y_{N_i}^{(i)})(i = 1, \ldots, M)$. In this case, the standard univariate form can be acquired by setting $M = 1$. Firstly, we transfer a one-dimensional time series $Y_{N_i}^{(i)}$ in to a multidimensional matrix $[X_1^{(i)}, \ldots, X_{K_i}^{(i)}]$ with vectors $X_j^{(i)}$ that equals to $(y_j^{(i)}, \ldots, y_{j+L_i-1}^{(i)})^T \in \mathbf{R}^{L_i}$, where $L_i (2 \leq L_i \leq N_i/2)$ is the window length for each series with length N_i and $K_i = N_i - L_i + 1$. We can then get the trajectory matrix $\mathbf{X}^{(i)} = [X_1^{(i)}, \ldots, X_{K_i}^{(i)}] = (x_{mn})_{m,n=1}^{L_i,K_i}$ after this step. The above procedure for each series separately provides M different $L_i \times K_i$ trajectory matrices $\mathbf{X}^{(i)}(i = 1, \ldots, M)$.

To construct a block Hankel matrix in the vertical form we need to have $K_1 = \cdots = K_M = K$. Accordingly, this version enables us to have various window length L_i and different series length N_i, but similar K_i for all series. The result of this step is the following block Hankel trajectory matrix:

$$\mathbf{X}_V = \begin{bmatrix} \mathbf{X}^{(1)} \\ \vdots \\ \mathbf{X}^{(M)} \end{bmatrix}.$$

Note that $\mathbf{X}_V$ indicates that the output of the first step is a block Hankel trajectory matrix formed in a vertical form.

Then, the SVD of $\mathbf{X}_V$ is performed in the following step. Note that the SVD technique is closely related to the Singular Spectrum Analysis technique and its multivariate extension, which have been widely applied in a range of different fields and a multitude of fairly precise results proved it as a powerful and applicable technique [22,29,23–28, 30–35]. Denote $\lambda_{V_1}, \ldots, \lambda_{V_{L_{sum}}}$ as the eigenvalues of $\mathbf{X}_V\mathbf{X}_V^T$, arranged in decreasing order ($\lambda_{V_1} \geq \cdots \lambda_{V_{L_{sum}}} \geq 0$) and $U_{V_1}, \ldots, U_{V_{L_{sum}}}$, the corresponding eigenvectors, where $L_{sum} = \sum_{i=1}^{M} L_i$. Note also that the structure of the matrix $\mathbf{X}_V\mathbf{X}_V^T$ is as follows:

$$\mathbf{X}_V\mathbf{X}_V^T = \begin{bmatrix} \mathbf{X}^{(1)}\mathbf{X}^{(1)T} & \mathbf{X}^{(1)}\mathbf{X}^{(2)T} & \cdots & \mathbf{X}^{(1)}\mathbf{X}^{(M)T} \\ \mathbf{X}^{(2)}\mathbf{X}^{(1)T} & \mathbf{X}^{(2)}\mathbf{X}^{(2)T} & \cdots & \mathbf{X}^{(2)}\mathbf{X}^{(M)T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(M)}\mathbf{X}^{(1)T} & \mathbf{X}^{(M)}\mathbf{X}^{(2)T} & \cdots & \mathbf{X}^{(M)}\mathbf{X}^{(M)T} \end{bmatrix}.$$