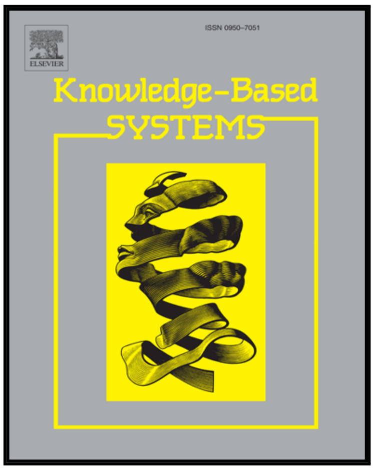

Univariate and Multivariate Time Series Manifold Learning

Colin O'Reilly, Klaus Moessner, Michele Nati

PII: S0950-7051(17)30260-5
DOI: 10.1016/j.knosys.2017.05.026
Reference: KNOSYS 3924

To appear in: *Knowledge-Based Systems*

Received date: 20 September 2016
Revised date: 25 May 2017
Accepted date: 27 May 2017

Univariate and Multivariate Time Series Manifold Learning

Colin O'Reilly[a,*], Klaus Moessner[a], Michele Nati[b]

[a]Institute for Communication Systems (ICS), Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, GU2 7XH

[b]Digital Catapult, 101 Euston Rd, London NW1 2RA, United Kingdom

Abstract

Time series analysis aims to extract meaningful information from data that has been generated in sequence by a dynamic process. The modelling of the non-linear dynamics of a signal is often performed using a linear space with a similarity metric which is either linear or attempts to model the non-linearity of the data in the linear space. In this research, a different approach is taken where the non-linear dynamics of the time series are represented using a phase space. Training data is used to construct the phase space in which the data lies on or close to a lower-dimensional manifold. The basis of the non-linear manifold is derived using the kernel principal components derived using kernel principal component analysis where fewer components are retained in order to identify the lower-dimensional manifold. Data instances are projected onto the manifold, and those with a large distance between the original point and the projection are considered to be derived from a different underlying process. The proposed algorithm is able to perform time series classification on univariate and multivariate data. Evaluations on a large number of real-world data sets demonstrate the accuracy of the new algorithm and how it exceeds state-of-the-art performance.

Keywords: time series, univariate, multivariate, one-class classification, kernel principal component analysis

1. Introduction

Time series data are derived from the measurement of an underlying phenomena and are represented as sequential instances of values that may occur individually (univariate), or concurrently (multivariate). They are generated in a broad range of domains such as aviation [1, 2], financial [3, 4], meteorological [5] and industrial monitoring [6]. Extraction of knowledge from a time series using machine learning or data mining methods can enable classification of current data instances, or the prediction of future instances. Several excellent survey articles have been published in this area, for example those by Fu [7] and Längkvist *et al.* [8].

A task that is often performed is the classification of cyclostationary time series data. Given a sequence of time series data which represents one cycle, the aim is to classify the sequence into the correct category of cycle. There are two approaches that are used to perform the task. One method uses a lazy-learning approach where testing data are compared to training data to determine similarity. No model is constructed during the training phase, however, during the testing phase a window of test data is compared to windows of the training data. A similarity metric, for example Euclidean distance, is used to determine the level of similarity. An alternative approach uses the time series as a representation of a dynamical system to construct a *phase space*. A model is created of the system from the training data set, and then a similarity metric is used to determine if test data instances were generated from the same underlying process.

This research takes the second approach where the aim is to construct a model of a dynamical system using a phase space. The model is then used to identify test data instances that were generated by the same process. State of the art is extended in the following way:

- A univariate time series is represented as a dynamical system using a phase space. In the phase space, the lower-dimensional manifold on which the data lies on or close to is determined using kernel PCA. The similarity of a test data instance to the dynamical system is determined by projecting the data instance onto the manifold and determining the error in its reconstruction.

- The algorithm is extended to operate on multivariate time series using horizontal form SSA. The algorithm extracts the most important information from the multivariate streams by identifying the lower-dimensional manifold that the data lies on or close to.

- A detailed evaluation of the algorithm on both univariate and multivariate real-world time series is pro-

*Corresponding author

Email addresses: c.oreilly@surrey.ac.uk (Colin O'Reilly), k.moessner@surrey.ac.uk (Klaus Moessner), michele.nati@digicatapult.org.uk (Michele Nati)