Pattern Recognition 73 (2018) 202–215

Contents lists available at ScienceDirect

Pattern Recognition

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

Autoregressive forests for multivariate time series modeling

Kerem Sinan Tuncel*, Mustafa Gokce Baydogan

Department of Industrial Engineering, Boğaziçi University, Bebek/İstanbul 34342, Turkey

ARTICLE INFO

Article history:
Received 13 March 2017
Revised 4 August 2017
Accepted 15 August 2017
Available online 24 August 2017

MSC:
00-01
99-00

Keywords:
Multivariate time series
Vector autoregression
Time series representation
Ensemble learning
Classification

ABSTRACT

Multivariate Time Series (MTS) modeling has received significant attention in the last decade because of the complex nature of the data. Efficient representations are required to deal with the high dimensionality due to the increase in the number of variables and duration of the time series in different applications. For example, model-based approaches such as Hidden Markov Models (HMM) or autoregressive (AR) models focus on finding a model to represent the series with the model parameters to handle this problem. Both HMM and AR models are known to be very successful in the representation of the time series however most of the HMM approaches assume independence and traditional AR models consider linear dependence between the variables of MTS. As most of the real systems exhibit nonlinear relations, traditional approaches fail to represent the time series. To handle these problems, we propose an autoregressive tree-based ensemble approach that can model the nonlinear behavior embedded in the time series with the help of tree-based learning. Multivariate autoregressive forest, namely mv-ARF, is a nonparametric vector autoregression approach which provides an easy and efficient representation that scales well with large datasets. An error-based representation based on the learned models is the basis of the proposed approach. This is very similar to time series kernels used for multivariate time series classification problems. We test mv-ARF on MTS classification problems and show that mv-ARF provides fast and competitive results on benchmark datasets from several domains. Furthermore, mv-ARF provides a research direction for vector autoregressive models that breaks from the linear dependency models to potentially foster other promising nonlinear approaches.

1. Introduction

Research on the efficient representation of the high-dimensional time series is ever-increasing due to the ubiquitous nature of the multivariate and temporal data. Many researchers in different fields such as statistics, machine learning and signal processing focus on modeling multivariate time series (MTS) data. For example, in motion recognition, electromyography (EMG) signals from multiple muscles are recorded for four separate upper-arm movements and the aim is to classify upper-arm movements [17]. Also, a public sector use is shown, where a novel MTS clustering technique is discussed for clustering of crime locations, such as states and districts, with similar crime trends [9]. Another example from hydrology is provided by Raman and Sunilkumar [22], in which MTS modeling of water resources is considered in water management and planning.

Most of the existing studies focus on obtaining a rectangular representation for MTS data. In other words, each MTS is mapped to a feature vector to be used as an input to machine learning approaches. If one can learn a model characterizing the dynamics of the system generating the MTS data, model parameters can be used as a feature vector. For example, transition and emission probabilities of a Hidden Markov Model (HMM) trained on MTS can be used to represent it. On the other hand, some of the time series data mining approaches make use of the similarity information between the MTS. Traditional similarity-based approaches consider the distance between each univariate series of MTS and aggregates this information to obtain the similarity between MTS. Model-based approaches (i.e. HMM) either compare the model parameters for MTS (i.e. feature vector) or the likelihood of observing the MTS given a model. Suppose we are interested in the classification of an MTS as in the arm-movement example [17]. Each training time series is represented by an HMM and the parameters of an HMM trained on the test series can be compared to the ones for training series. The other option is to compute the likelihood of observing the test series using the HMMs of the training series. Similar to nearest-neighbor classification, based on the likelihood over all HMMs (i.e. one HMM for each training series), one can determine the training series that is similar to test series (i.e. the one with the largest likelihood).

* Corresponding author.
E-mail addresses: keremsinantuncel@gmail.com, kerem.tuncel@boun.edu.tr (K.S. Tuncel).

http://dx.doi.org/10.1016/j.patcog.2017.08.016
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

Benefiting from the ideas in model-based approaches, this paper provides an autoregressive (AR) modeling strategy for MTS. As a common approach, we consider classification as an application to evaluate our modeling strategy. We assume that the time series can be modeled as an autoregressive process and time series from different classes has a distinct model. Using these models, each time series can be represented by their goodness of fit to the models of each class. There are numerous algorithms proposed to solve the classification problem, the methods mostly fall into two main categories; feature-based methods and distance-based methods. In this sense, model-based approaches can go into both categories depending on how the models are utilized.

Due to the complexity of MTS data, there is exhaustive research on finding representations and extracting meaningful patterns from MTS data sets. The most popular and simple approach is using rectangularization approaches. For example, a two dimensional Singular Value Decomposition (SVD) approach is proposed in [28]. Consequently, Symbolic Aggregate Approximation (SAX) [20] is another representation technique designed for univariate time series (UTS). Similarly, Baydogan and Runger [4] propose a symbolic representation for MTS which generates a codebook from the terminal nodes of decision trees. Moreover, for longer time series, alternative transformations such as wavelets and Fourier transforms are discussed [8]. HMMs to represent MTS is also considered by several studies in the literature. For example, an HMM-based methodology with a principal component analysis (PCA) representation for motion recognition is provided by Bashir et al. [2].

Distance-based methods mainly consists of modifications of univariate time series (UTS) similarity approaches to work in MTS framework. Among the distance measures in the literature, Dynamic Time Warping (DTW) [6] has become the benchmark due to its high accuracy in many of the commonly used datasets. Subsequently, many researchers focus on improvements on various aspects of DTW methodology such as efficiency and accuracy. Also, there is research being conducted to find more robust distance measures than DTW. For example, Yang and Shahabi [29] propose a PCA-based distance measure for MTS called *Eros* (Extended Frobenius Norm). High dimensionality is another problem to be addressed in MTS datasets. For that purpose, Cuturi and Doucet [12] propose an autoregressive kernel-based distance measure for MTS data. Another AR modeling strategy aims at learning MTS representation based on local autopatterns [5], which also introduces a similarity measure called learned pattern similarity (LPS). Distance-based approaches mostly work by taking each attribute of an MTS as an independent UTS. However, this methodology may sometimes cause some important data loss since MTS are not only defined by separate attributes but also by relationships between them.

The difference of mv-ARF from the works in literature is that it provides a framework for modeling the dynamic behavior of MTS. The strategy incorporates the strengths of AR modeling with the capabilities of tree-based ensembles. Use of tree-based learners enables a flexible and simple generative approach with only a few parameters. Moreover, combining AR components with the multi-response learning scheme provides comprehensive information to model the interactions between features. In that sense, the representation learned by mv-ARF is comparable to the AR-based strategies such as LPS and AR Kernel in addition to HMM-based strategies due to its generative and AR approach.

The primary contribution of this work is the introduction of a new autoregressive modeling strategy. Provided approach aims to capture the dynamics of MTS by using a representation based on a vector autoregressive model. Proposed methodology is denoted as "*M*ulti*V*ariate *A*uto*R*egressive *F*orests", mv-ARF, which trains tree-based ensemble learners with autoregressive components in a multitask setting. Fig. 1 illustrates a summary of the mv-ARF methodology. There are some key features of mv-ARF model; the multitask learning approach utilized in the model provides a way to consider all the variables at the same time. The representation is acquired from a supervised learner which uses AR components as input. Therefore, the only form of feature extraction is to generate the AR components with a predefined lag value leading to a very simple feature extraction scheme with only lag parameter. A simple representation is used, which consists of time index and the observation values of each attribute as columns. Most AR models either aim to capture non-linear relations or non-stationary behavior, mv-ARF is a robust AR model which is able to capture both. Tree-based ensembles utilized by the model are able to capture non-linear relations between AR components and incorporation of time order enables mv-ARF to capture non-stationary behavior. Furthermore, the lag (or order) of autoregression serves as an upper-bound, that is, the model is able to detect and use the lag value that gives the best accuracy, within the range of lag values defined by the upper-bound. Lastly, the computational complexity of mv-ARF is almost linear to both problem and algorithm parameters and each step of mv-ARF is embarrassingly parallel.

The capabilities and the effectiveness of the mv-ARF model is tested within MTS classification framework. For a classification task with C class labels, the model ends up with C explanatory models. A novel classifier is introduced which utilizes a representation using the prediction errors acquired from the constituent models as illustrated in Fig. 1. The representation and the tree-based approach allows mv-ARF to efficiently deal with datasets that contain time series with differing lengths as well as different data types (i.e. categorical, ordinal). The final error-based representation can be acquired by calculating the predictions and the prediction errors for each time series from each model. Effectiveness of the proposed model is demonstrated by experiments on a large amount of commonly used datasets in MTS classification literature.

Another important motivation for this work, is to underline the importance of utilizing multivariate learning methods instead of modifying univariate models, in MTS analysis. Multivariate models are able to capture the interactions between attributes, which may be of importance. In order to show the superiority of multivariate approaches, the univariate approach for mv-ARF framework is also considered, that is, univariate forests are built and the final representation is acquired by combining separate predictions, instead of using multivariate learning methods. The results of mv-ARF is compared with the univariate approach, along with several other benchmark approaches used in MTS classification.

The rest of the paper is organized as follows; Section 2 introduces some relevant works in the literature and Section 3 provides background. Section 4 explains the methodology of the method in detail. Section 5 provides the experiments from a list of datasets [10,19,21,24]. Section 6 gives the concluding remarks.

2. Literature review

There are two modeling strategies that are common in the literature. Feature-based strategies aims to find a rectangular representation for MTS by extracting descriptive features from the time series. On the other hand, distance-based strategies aim to provide similarity measures that works well with MTS. Moreover, model-based strategies can be used for both extracting features and to define a similarity depending on how the model is utilized. As mv-ARF is a model-based framework, the main focus is on model-based approaches.

As one of the many feature-based approaches, a recent successful approach by Grabocka et al. [16] extends the shapelet discovery techniques for MTS classification. Shapelet discovery is a methodology to discover subsequences (which are called shapelets) to represent a time series. Generally, shapelets with high