Applied Soft Computing 13 (2013) 361–371

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

Large margin mixture of AR models for time series classification

B. Venkataramana Kini*, C. Chandra Sekhar

Dept. of Computer Science and Engg., Indian Institute of Technology Madras, Chennai 600036, India

ARTICLE INFO

Article history:
Received 18 October 2011
Received in revised form 10 June 2012
Accepted 7 August 2012
Available online 14 September 2012

Keywords:
Large margin autoregressive model
Large margin mixture autoregressive model
Time series classification
Outlier detection
Rejection option
Generative and discriminative hybrid models

ABSTRACT

In this paper, we propose the large margin autoregressive (LMAR) model for classification of time series patterns. The parameters of the generative AR models for different classes are estimated using the margin of the boundaries of AR models as the optimization criterion. Models that use a mixture of AR (MAR) models are considered for representing the data that cannot be adequately represented using a single AR model for a class. Based on a mixture model representing each class, we propose the large margin mixture of AR (LMMAR) models. The proposed methods are applied on the simulated time series data, electrocardiogram data, speech data for E-set in English alphabet and electroencephalogram time series data. Performance of the proposed methods is compared with that of support vector machine (SVM) based classifier that uses AR coefficients based features. The proposed methods give a better classification performance compared to the SVM based classifier. Being generative models, the LMAR and LMMAR models provide a generative interpretation that enables utilization of the rejection option in the high risk classification tasks. The proposed methods can also be used for detection of novel time series data.

1. Introduction

Time series form an important class of data objects in tasks such as speech recognition and medical signal analysis. An important property shared by time series data is that the neighboring values in time series are similar (temporal correlation). Therefore, the current value of the time series can be expressed as a finite, *linear aggregate of previous values of the series* and noise (autoregressive model).

The autoregressive (AR) model is a generative model in the sense that the current value of time series is generated as a linear combination of previous values and a probability density function can be defined on the AR model [1]. The AR model is widely used in the tasks such as time series prediction and parametric spectrum estimation. In the current work, our objective is to utilize AR models for time series classification.

Time series classification is an important problem in pattern recognition. In time series classification, a set of time series with class labels is given. This data is used to build a model for each class (training phase). When a new time series is given, an appropriate label is assigned to it (testing phase).

Pattern recognition methods for classification can be broadly divided into generative and discriminative methods. Generative models learn a joint probability distribution $p(\mathbf{x}, y)$, of the input $\mathbf{x}$ and the label y, and perform classification using Bayes rule, $y = \underset{i}{argmax} p(y_i|\mathbf{x})$. Discriminative models learn the mapping from the input $\mathbf{x}$ to the label y. The generative model provides a conducive framework for imposing structure and prior knowledge on a given problem. Due to the generative nature of these models, the outliers (data points that do not belonging to any of the known classes) can be detected easily and uncertain classifications can be referred to domain experts. Discriminative methods give a superior performance by focusing on the given task of classification [2]. Discriminative models have been used in diverse time series classification tasks [3–6].

The black-box nature of discriminative models makes incorporation of the structure of the problem difficult. For example, in time series classification using the discriminative models, it is not possible to make use of the temporal correlation present in a time series directly. Discriminative models have to depend on the temporal methods such as AR modeling for feature extraction. This motivates combining these two methods synergistically, leading to a hybrid method that retains the richness of generative models, at the same time providing a superior classification performance. There is an evidence that the models constructed in this manner can capture the subtleties of the time series data being analyzed [7]. In the hybrid method, the parameters of the generative model are estimated to maximize the classification performance rather than maximizing the likelihood of the training data [7–9].

Recently, there has been a significant interest in the discriminative training of generative models for classification tasks, particularly in the area of speech recognition (for a review see [7–11]). Most of these works use hidden Markov models or Gaussian mixture models as the generative models. The current work is on discriminatively training AR models for time series classification.

* Corresponding author.
E-mail addresses: venkataramana.kini@gmail.com (B.V. Kini), chandra@cse.iitm.ac.in (C.C. Sekhar).

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.08.027

Also, the probabilistic kernel functions can be used to capture temporal characteristics of time series [12,13]. These methods develop temporal models such as hidden Markov model and linear dynamic model for each time series data and then compute Kullback–Leibler (KL) divergence between these models. These methods cannot be considered to perform discriminative training of generative models. These methods define kernel functions in the space of generative model parameters.

In [14,15] the AR coefficients are extracted from the time series data and the discriminative classifiers are used for classification of time series using AR coefficients based features.

We propose an approach to discriminatively train the generative AR models using the large margin method [17], that retains the rich interpretation of AR models. This interpretation enables the model to utilize the rejection option in cases of high risk classification tasks and the detection of outliers when all classes are not covered in the training data set. The model uses the large margin concepts, that are utilized in support vector machines (SVM). However, unlike SVMs, large margin AR (LMAR) models retain generative interpretation. Similar to other discriminative methods, temporal dependencies cannot be directly captured in the SVMs. Therefore, AR modeling or other temporal methods are used for feature extraction.

In certain time series classification problems, some classes may not be adequately represented using a single AR (MAR) model. In such cases, it is useful to consider a mixture of AR models. We propose to build a mixture of AR models [24] for each class and further train these MAR models using the large margin method (LMMAR model).

The mixture of AR models presented in our paper is related to a mixture of ARMA models [24], which is used for time series clustering. It should be noted that in [25] a single time series is modeled using a mixture of AR models for segmenting the time series into homogeneous parts. However, in our case a set of multiple time series is modeled using a mixture of AR models.

In the next section, we present the AR model based methods for time series classification. We first present the method that uses a single AR model for each class and the large margin method for estimation of parameters of AR models. Then we present the method that uses a mixture of AR models for each class trained using the large margin method. In Section 3, we present our studies using the proposed methods for time series classification on different data sets. We compare the performance of the proposed method with that of the SVM based classifiers that use AR coefficients based features.

2. AR model based methods for time series classification

An AR process models the linear dependency that may exist in a given time series. It models the signal as the output of a linear system driven by white noise of zero mean and unknown variance. Autoregressive moving average (ARMA) model regresses on noise as well. However, there exists an equivalent higher order AR model. Hence, without loss of generality, AR models are considered in this paper.

Let the time series training data be: $\mathbf{X} = \{<\mathbf{x}_1 y_1>, <\mathbf{x}_2 y_2>, \ldots, <\mathbf{x}_n y_n>, \ldots, <\mathbf{x}_N y_N>\}$, $\mathbf{x}_n \in M\Re$ *and* $y_n \in \{1, 2, \ldots, C\}$. Here $\mathbf{x}_n = [x_n(1), x_n(2), \ldots, x_n(M)]^T$ is the nth time series of length M, y_n is the corresponding class label and C is the number of classes.

Using an AR model with order P, the value of time series $\mathbf{x}_n$ at discrete time t can be represented as:

$$x_n(t) = -\sum_{p=1}^{P} a_{np} x_n(t-p) + e_n(t)$$
$$= \hat{x}_n(t) + e_n(t) \quad (1)$$

where $e_n(t) \sim \mathcal{N}(0, \sigma^2)$ is the zero mean white noise with σ^2 as variance, and $\mathbf{a}_n = [a_{n1}, a_{n2}, \ldots, a_{nP}]^T$ are the AR coefficients.

The autocorrelation function (ACF) of $\mathbf{x}_n$ at lag p is estimated using $r_{np} = \sum_t x_n(t) x_n(t+p)$, $p = 1, \ldots, P$ and represented as $\mathbf{r}_n = [r_{n1}, \ldots, r_{nP}]^T$. The temporal characteristic of a time series can be captured using its ACF [1]. The variance of time series, r_{n0}, estimated using $\sum_t x_n(t) x_n(t)$ gives its instantaneous characteristic.

Since $e_n(t) \sim \mathcal{N}(0, \sigma^2)$, the probability density function (pdf) of $\mathbf{x}_n$ can be written as [1,18]:

$$p(\mathbf{x}_n|\mathbf{a}_n, \sigma^2) = (2\pi\sigma^2)^{-M/2} exp(-0.5\sigma^{-2} \sum_{t=1}^{M} e_n^2(t))$$
$$= (2\pi\sigma^2)^{-M/2} exp(-0.5\sigma^{-2} \mathbf{a}_n^T \Sigma_n \mathbf{a}_n) \quad (2)$$

where the autocorrelation matrix, $\mathbf{\Sigma}_n$, is defined as

$$\Sigma_{\rm n} = \begin{pmatrix} 1 & r_1 & r_2 & \ldots & r_{P-1} \\ r_1 & 1 & r_1 & \ldots & r_{P-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{P-1} & r_{P-2} & \ldots & r_1 & 1 \end{pmatrix}_n \quad (3)$$

Using Yule–Walker (Y–W) equations [1], the AR coefficients $\mathbf{a}_n$ can be derived from the autocorrelation function $\mathbf{r}_n$ and the autocorrelation matrix $\mathbf{\Sigma}_n$ as $\mathbf{a}_n = \mathbf{\Sigma}_n{}^{-1} \mathbf{r}_n$.

Therefore, (2) can be written as:

$$p(\mathbf{x}_n|\mathbf{r}_n) \propto \exp\left(-\frac{1}{2}\mathbf{r}_n^T \Sigma_{\rm n}{}^{-1} \mathbf{r}_n\right) \quad (4)$$

The autocorrelation matrix Σ_n represents the temporal structure of one time series $\mathbf{x}_n$. That is, the inherent assumption in Eq. (4) is that AR process is ergodic. We can relax this assumption by using multiple time series from the same class for estimating the ensemble autocorrelation matrix $\mathbf{\Sigma}_c$ for class c. The ensemble average provides a robust estimate of the AR process compared to parameters estimated with ergodic assumption. We propose that each class is an AR process represented by the AR model design matrix $\mathbf{\Sigma}_c$ and the quantity $\mathbf{r}_n^T \mathbf{\Sigma}_c^{-1} \mathbf{r}_n$ is similar to the squared Mahalanobis distance. Using such models for different classes, $c \in \{1, 2, \ldots, C\}$, the classification of a new time series $\mathbf{x}$ with $\mathbf{r}$ as its ACF can be performed using the following decision rule:

$$y = \underset{c}{argmin}\,\{\mathbf{r}^T \mathbf{\Sigma}_c^{-1} \mathbf{r}\} \quad (5)$$

It is to be noted that the autocorrelation matrix $\mathbf{\Sigma}_c$ has the Toeplitz structure and can be characterized only by the ACFs for the examples of the class. In the next subsection, we develop the large margin AR model for classification of time series data.

2.1. Large margin AR model

In the large margin AR (LMAR) model, we propose to maximize the margin (distance of the nearest training example from the decision boundary) by optimizing the parameters involved, i.e., the ACFs representing each class. The ACFs are found not only to classify the training data correctly, but also to place the decision boundaries optimally. We propose to constrain each time series in the training data to be at least one unit distance away from the decision boundary of each of the competing classes (similar to what is done in case of large margin Gaussian mixture model [10]). For a time series $\mathbf{x}_n$ with its class label as y_n, the constraints are as follows:

$$\mathbf{r}_n^T(\mathbf{\Phi}_c - \mathbf{\Phi}_{y_n})\mathbf{r}_n \geq 1, \quad \forall c \neq y_n \quad (6)$$