Pattern Recognition 45 (2012) 570–577

Contents lists available at ScienceDirect

Pattern Recognition

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

The copula echo state network

Sotirios P. Chatzis *, Yiannis Demiris

Department of Electrical and Electronic Engineering, Imperial College London, United Kingdom

ARTICLE INFO

Article history:
Received 31 January 2011
Received in revised form
24 March 2011
Accepted 24 June 2011
Available online 19 July 2011

Keywords:
Copula
Echo state network
Sequential data modeling
Maximum-likelihood

ABSTRACT

Echo state networks (ESNs) constitute a novel approach to recurrent neural network (RNN) training, with an RNN (the reservoir) being generated randomly, and only a readout being trained using a simple, computationally efficient algorithm. ESNs have greatly facilitated the practical application of RNNs, outperforming classical approaches on a number of benchmark tasks. This paper studies the formulation of a class of copula-based semiparametric models for sequential data modeling, characterized by nonparametric marginal distributions modeled by postulating suitable echo state networks, and parametric copula functions that help capture all the scale-free temporal dependence of the modeled processes. We provide a simple algorithm for the data-driven estimation of the marginal distribution and the copula parameters of our model under the maximum-likelihood framework. We exhibit the merits of our approach by considering a number of applications; as we show, our method offers a significant enhancement in the dynamical data modeling capabilities of ESNs, without significant compromises in the algorithm's computational efficiency.

1. Introduction

Recurrent neural networks (RNNs) constitute a significant nonlinear approach for modeling dynamical systems as they entail recurrent connections between neurons, thus allowing for direct processing of temporal dependencies. This way, they are capable of modeling a rich set of dynamical behaviors [1–5]. Among the numerous successful applications of RNNs, one might mention grammatical inference studies, recognition and generation of strings from finite state machines, speech recognition, data classification, and interval and rate invariance studies [6,7].

Nevertheless, although much effort has been spent on the development of effective model parameter estimation schemes for RNNs [8,9], most optimization methods lead to only mediocre results compared to alternative methods for sequential data modeling [10,11]. One reason for this is the ill-posed nature of the problem, i.e., parameter (synapse weight) estimation involves inversion of a nonlinear dynamical system from finite and noisy data which typically is ill-posed [9,4]. Regularization has been considered in the past as a method to ameliorate these issues. Regularization in neural networks is usually achieved through the addition of a penalty term in the cost function [4] which favors simpler models over complex mappings. Penalization shrinks nonsignificant weights, decreases the model variability, and improves predictions. A principled approach to the estimation of the regularization parameter(s) has been proposed in a Bayesian setting in [12]. This probabilistic setting facilitates inference of the regularization hyperparameters which are viewed as beliefs in the uncertainties of the model parameters. However, this procedure entails offline estimation of the covariance matrix (Hessian) that might be computationally inappropriate, since it is often the case that eigenvalues of the Hessian matrix turn out to decay to zero causing numerical instabilities (i.e., Hessian singularities).

A groundbreaking and surprisingly efficient network structure for RNNs, resolving all the aforementioned issues, was invented independently in the seminal works of Jaeger [13], who called these RNNs echo state networks (ESN), and Maass et al. [14], who developed a similar approach for spiking neural networks and called the derived model the liquid state machine (LSM). These two innovative methodologies have given rise to the novel paradigm of reservoir computing (RC) [15], under which both the ESN and LSM network structures are usually subsumed. The RC paradigm avoids the shortcomings of typical, gradient-descent-based RNN training by setting up the network structure in the following way [16]:

- A recurrent neural network is randomly created and remains unchanged during training. This RNN is called the *reservoir*. It is passively excited by the input signal and maintains in its state a nonlinear transformation of the input history.
- The desired output signal is generated by a linear *readout* layer attached to the reservoir, which computes a linear combination of the neuron outputs from the input-excited reservoir (*reservoir states*).

* Corresponding author.
E-mail address: soteri0s@me.com (S.P. Chatzis).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.06.022

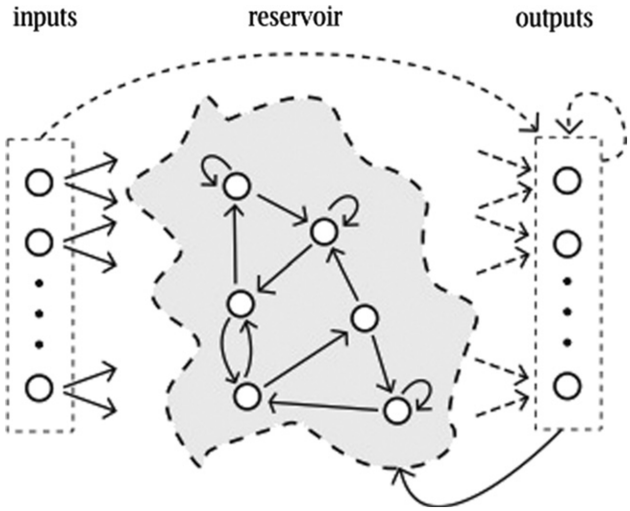

Fig. 1. Schematic overview of the reservoir computing approach [30].

As is easy to infer from the preceding description, the function of the reservoir in RC networks can be compared to that of the kernel in kernel machine approaches (e.g., support vector machines [17], relevance vector machines [18], and their variants) [19]: input signals drive the nonlinear reservoir and produce a high-dimensional dynamical "echo response," which is used as a non-orthogonal basis to reconstruct the desired outputs. A schematic illustration of the RC approach is provided in Fig. 1.

Among the existing RC implementations, most of the attention of the research community has been concentrated on the design of the network topologies and the selection of the neuron types. In this work, we focus on echo state networks, which usually employ analog neurons, typically linear, sigmoid or leaky-integrator [20] units, and simple sparsely connected graphs as their network topologies. An extensively studied subject in the field of ESN concerns the introduction of appropriate *goodness* measures of the reservoir structure. Indeed, the classical feature that reservoirs should possess is the echo state property. This property essentially states that the effect of a previous reservoir state and a previous input on a future state should vanish gradually as time passes, and not persist or even get amplified. However, for most practical purposes, the echo state property can be easily satisfied by merely ensuring that the *reservoir weight matrix* $\boldsymbol{W}$ is *contractive*, i.e., by scaling the reservoir weight matrix so that its *spectral radius* $\rho(\boldsymbol{W})$ (that is, its largest absolute eigenvalue) is less than one [21].

When modeling sequentially interdependent data, there is always the need of measuring and expressing the nature and degree of dependence by means of an explicit probabilistic model. Such problems are usually addressed in the statistical machine learning literature by postulating conditional models. As we have already discussed, ESNs do not provide such capabilities; instead, to capture the temporal dynamics of the modeled datasets, they rely on the memory capacity of the employed reservoirs, and the quality of the temporal information encapsulated in the generated reservoir outputs (reservoir states). In this work, we seek to provide an explicit expression for the dependence between successive observations modeled by means of an ESN. For this purpose, we postulate an appropriate conditional density model, which is based on a first-order Markov chain-type assumption for the interdependence between the ESN-generated outputs, the formulation of which is facilitated by utilization of the statistical tool of *copula* [22].

The seminal work of Sklar [22] shows how to come up with one form of dependence between random variables with given marginal distributions; the statistical tool developed for this purpose is called copula. The application of copulas in various fields pertaining to data modeling is a rather recent trend; in particular, modeling temporal dependence of sequentially appearing data using copulas has recently gained much attention [23–25]. Since the emergence of the concept of copula, s everal copula families have been constructed, e.g., Gaussian, Clayton, Frank, Gumbel, Joe, etc. [23] that enable capturing of any form of dependence structure. By coupling different marginal distributions with different copula functions, copula-based time series models are able to model a wide variety of marginal behaviors (such as skewness and fat tails), and dependence properties (such as clusters, positive or negative tail dependence) [23].

Copulas are powerful tools in statistical modeling because the copula-based modeling problem can be always split into two stages: the first stage deals with the identification of the marginal distributions, and the second stage involves defining the appropriate copula for adequately modeling the dependence structure. Such a two-stage approach is a convenient and common procedure in copula modeling. Typical application areas of copula-based models include gene prediction and cancer classification based on gene-expression measurements from microarrays [26], analyzing and pricing volatility of investment portfolios, credit risk analysis, as well as reliability analysis of highway bridges [27], and analysis of spike counts in neuroscience [28]. Note though that a shortcoming of the copula approach consists of the fact that it is not always obvious how to identify the copula that adequately represents a needed dependence structure. Nevertheless, selection of the best-fit copula has been a topic of rigorous research efforts during the last years, and motivating results have already been achieved (see, e.g., [25]).

In this paper, we introduce a novel probabilistic regard towards ESNs which utilizes the concept of copulas to yield a conditional predictive distribution for the modeled sequential data. Specifically, we begin by postulating an ESN to model the examined dynamical observations; based on the postulated ESN, a marginal distribution for the modeled data can be straightforwardly obtained, by introducing a simple probabilistic assumption to allow for the case of noise-contaminated observations. Subsequently, we introduce the obtained marginal distributions in the context of copula-based modeling, to eventually yield a first-order Markov chain-like conditional predictive distribution for the modeled data.

Estimation of copula-based multivariate density models is often computationally difficult to perform by means of maximum-likelihood. To overcome optimization problems which can be encountered when using simple algorithms such as Newton–Raphson or the expectation-maximization (EM) algorithm, inference for our model is conducted by employing a two-step estimation method known as inference functions for margins (IFM) [29]. On the first step of the IFM method, the marginal model is maximized with respect to its entailed (marginal) parameters, while, in the second step, the copula model is maximized with respect to the entailed (copula) parameters, using the marginal estimates obtained from the first step. This way, model estimation becomes computationally efficient, while comparison of different copulas can also be conducted in a convenient way, by means of standard methodologies for assumption testing.

The efficacy of the proposed approach is evaluated considering a number of application scenarios, and its performance is compared to conventional echo state network formulations employing ridge regression-based readout training. The remainder of this paper is organized as follows: In Section 2, we provide a brief overview of the basic configuration of ESNs. In Section 3, we concisely review the basic mathematical expressions pertaining