Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 19 (2006) 196–207
Neural Networks
www.elsevier.com/locate/neunet

2006 Special issue

Time dependent neural network models for detecting changes of state in complex processes: Applications in earth sciences and astronomy

Julio J. Valdés [a,*], Graeme Bonham-Carter [b]

[a] National Research Council, Institute for Information Technology, M50, 1200 Montreal Road, Ottawa, Ont., Canada K1A 0R6
[b] Geological Survey of Canada, 601 Booth Street, Ottawa, Ont., Canada K1A 0E8

Abstract

A computational intelligence approach is used to explore the problem of detecting internal state changes in time dependent processes; described by heterogeneous, multivariate time series with imprecise data and missing values. Such processes are approximated by collections of time dependent non-linear autoregressive models represented by a special kind of neuro-fuzzy neural network. Grid and high throughput computing model mining procedures based on neuro-fuzzy networks and genetic algorithms, generate: (i) collections of models composed of sets of time lag terms from the time series, and (ii) prediction functions represented by neuro-fuzzy networks. The composition of the models and their prediction capabilities, allows the identification of changes in the internal structure of the process. These changes are associated with the alternation of steady and transient states, zones with abnormal behavior, instability, and other situations. This approach is general, and its sensitivity for detecting subtle changes of state is revealed by simulation experiments. Its potential in the study of complex processes in earth sciences and astrophysics is illustrated with applications using paleoclimate and solar data.

Keywords: Heterogeneous neuron models; Time dependent neuro-fuzzy-genetic architectures; Multivariate time series; Heterogeneous data; Forecasting and prediction; Changes of state

1. Introduction

Time series data in the geosciences present difficult problems for a variety of reasons. The data are often incomplete (missing observations), imprecise, and heterogeneous (mixed scales of measurement for the variables involved), making conventional time series approaches difficult, or unsatisfactory. Neural network models using similarity-based heterogeneous neurons and systematic analysis of complex lags offer an alternative approach that is robust and sensitive. The ability to predict such changes of state (long- or short-term) has many important applications to natural systems. In complex or poorly known processes, knowledge discovery designed to uncover the underlying structure of the physical process is crucial, especially for revealing patterns and time dependencies, detecting abnormal behavior, instabilities, changes of state, deriving prediction criteria, and constructing forecasting procedures.

There is a considerable amount of scientific literature devoted to this topic, which clearly indicates its importance. The traditional approach has been from the domain of statistics, with the classical work of Box and Jenkins (1976).

Univariate, stationary series without missing values with a parametric approach were considered in the first research studies; which later progressed towards more complex cases like multivariate, non-stationary series with missing values and outliers, incorporating non-parametric approaches (Bustos & Yohai, 1986; Chang, Tiao, & Chen, 1988; Chen & Liu, 1993; Li & Tsay, 1998; Maddala & Yin, 1997; Martin & Yohai, 1985; Vaage, 2000). Within the computational intelligence field, there are also a wide variety of approaches (Cellier & Nebot, 2004; Cellier, Nebot, Múgica, & Albornoz, 1992; Cherkassky & Mulier, 1998; Giles, Lawrence, & Tsoi, 2001; Kasabov, 2004; Kehagias & Petridis, 1997; Mukherjee, Osuna, & Girosi, 1997; Principe & Kuo, 1995; Saad, Prokhorov, & Wunsch, 1998).

The use of a computational intelligence approach for model discovery and model-change detection in multivariate time processes with heterogeneity, different kinds of variables, missing data and uncertainty is discussed. It is a hybrid approach to time dependent model discovery, based on a

* Corresponding author. Tel.: +1 613 993 0887; fax: +1 613 952 0215.
E-mail addresses: julio.valdes@nrc-cnrc.gc.ca (J.J. Valdés), gbonhamc@nrcan.gc.ca (G. Bonham-Carter).

0893-6080/$ - see front matter
doi:10.1016/j.neunet.2006.01.006

combination of neural networks and evolutionary algorithms (in particular, genetic algorithms). A model of a time dependent target variable is understood as composed of two elements: (i) a prediction function of the value of the variable at a given time (generally non-linear); and (ii) the set of function arguments which are past values of some or all of the variables involved in the multivariate process. This is called the dependency pattern. The mathematical description used here is that of a non-linear multivariate autoregressive (AR) model (i.e. when both the arguments and the prediction function are also functions of time). In non-linear dynamics and chaos theory, there are AR-types of dependencies, and important lag terms are also found using genetic algorithms. However, in the presented approach the nature of the time series is more general (composed of heterogeneous data), with the further stipulation that chaos is not assumed.

The prediction functions are represented by hybrid neural networks using heterogeneous neurons in the first hidden layer, which accept, as input, heterogeneous, fuzzy and missing data. Instead of trying to find a global model for the whole multivariate process, like in conventional approaches, the discovery process proceeds as a continuous exploration along the multivariate series. The method finds sets of non-linear models for a target signal at time-intervals. The overall dependencies between the multivariate heterogeneous time series are characterized as probability distributions and error-cost functions over the sets of time lags of the ensemble of discovered dependency patterns. These distributions are represented as images (spectra), and are combined with the prediction error curves associated with the discovered models. Their joint interpretation allows the segmentation of the multivariate process (e.g. stable and transient states can be recognized).

From the methodological point of view, this approach can be considered as an abstract, conceptual filtering of multivariate time series of a general nature, which transforms the original heterogeneous, imprecise, and incomplete multivariate collection into time series of local data-driven models (with time-varying dependency patterns, and time-varying neural networks).

It should be emphasized that the main goal of this approach is exploratory, in the spirit of data mining. This is typical for the first stages of the investigation of poorly known or unknown processes. In these cases, fast-screening techniques (perhaps less accurate) may provide an initial insight into the nature of the interdependencies present, and the identification of potentially interesting dependency patterns. At subsequent stages, more elaborated techniques (usually computationally more expensive), should be applied, but this time focusing on the prospective interdependencies found, and/or on specific subsets of the data. The strategy is that of having coarse and refinement stages, and this paper deals specifically with the first of them. The detection of internal changes within the structure of the process is sensed through the changes in the composition of sets of coarse models (quickly constructed and evaluated), which approximate fragments of the time evolving process. Prediction accuracy, an always desirable feature, is compromised in favor of the speed with which models can be built and evaluated.

It often happens that stronger or more accurate methods are based on assumptions which are either more restrictive, difficult to verify, or difficult to satisfy. This is where robust (and hopefully fast) methods compatible with the nature of the problem and the data are desirable. In order to illustrate the potential of this approach, two simple data sets are used; sunspot numbers (related with variations of solar activity), and oxygen isotope ratio from ice-cores (paleotemperature indicator). Simple, should be understood only as a reference to the kind of data (not process), consisting of univariate series with homogenous (real-valued), crisp magnitudes without missing data. More complex cases involving multivariate time series with missing values are studied in Valdés and Barton (2003a,b) and Valdés, Barton, and Paul (2002).

1.1. Heterogeneous domains and multivariate time series

A formal approach for describing heterogeneous information in general observational problems was given in Valdés (2002b), and for constructing neuron models in Belanche (2000), Valdés, Belanche, and Alquézar (2000), and Valdés and García (1997). Different information sources are associated with the attributes, relations and functions, and these sources are associated with the nature of what is observed (e.g. point measurements, signals, documents, images, etc.). They are described by mathematical sets of the appropriate kind called source sets (Ψ_i), constructed according to the nature of the information source to represent (e.g. point measurements of continuous variables by subsets of the real in the appropriate ranges, structural information by directed graphs, etc.). Source sets also account for incomplete information. A heterogeneous domain is a Cartesian product of a collection of source sets: $\hat{H} = \Psi_1 \times \cdots \times \Psi_n$, where $n > 0$ is the number of information sources to consider. For example, consider a domain where objects are described by attributes like continuous crisp quantities, discrete features, fuzzy features, time series, images, and graphs (missing values are allowed). Individually, they can be represented as Cartesian products of subsets of real numbers ($\hat{R}$), nominal ($\hat{N}$) or ordinal sets ($\hat{O}$), fuzzy sets ($\hat{F}$), sets of images ($\hat{I}$), sets of time series ($\hat{S}$) and sets of graphs ($\hat{G}$), respectively, all properly extended for accepting missing values. Thus, the heterogeneous, time dependent domain is $\hat{H}^n(t) = \hat{N}^{n_N}(t) \times \hat{O}^{n_O}(t) \times \hat{R}^{n_R}(t) \times \hat{F}^{n_F}(t) \times \hat{I}^{n_I}(t) \times \hat{S}^{n_S}(t) \times \hat{G}^{n_G}(t)$, where n_N is the number of nominal sets, n_O of ordinal sets, n_R of real-valued sets, n_F of fuzzy sets, n_I of image-valued sets, n_S of time series sets, and n_G of graph-valued sets, respectively ($n = n_N + n_0 + n_R + n_F + n_I + n_S + n_G$). A multivariate, heterogeneous time series is shown in Fig. 1.

1.2. Model mining with heterogeneous neurons and hybrid neural networks

A model expresses a functional relationship between values of a given time series (the target), and a subset of the past values of the entire set of series. The purpose of the kind of