Pattern Recognition Letters 41 (2014) 23–33

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A novel associative model for time series data mining ☆

Itzamá López-Yáñez [b,*], Leonid Sheremetov [a], Cornelio Yáñez-Márquez [c]

[a] Mexican Petroleum Institute (IMP), Av. Eje Central Lázaro Cárdenas Norte 152, Mexico City, Mexico
[b] Instituto Politécnico Nacional, Centro de Innovación y Desarrollo Tecnológico en Cómputo (CIDETEC – IPN), Av. Instituto Politécnico Nacional 2580, Mexico City, Mexico
[c] Instituto Politécnico Nacional, Centro de Investigación en Computación (CIC – IPN), Av. Juan de Dios Bátiz s/n Edificio CIC, Mexico City, Mexico

ARTICLE INFO

Article history:
Available online 23 November 2013

Keywords:
Time series data mining
Supervised classification
Associative models
Mackey–Glass benchmark
CATS benchmark
Oil production time series

ABSTRACT

The paper describes a novel associative model for time series data mining. The model is based on the Gamma classifier, which is inspired on the Alpha–Beta associative memories, which are both supervised pattern recognition models. The objective is to mine known patterns in the time series in order to forecast unknown values, with the distinctive characteristic that said unknown values may be towards the future or the past of known samples. The proposed model performance is tested both on time series forecasting benchmarks and a data set of oil monthly production. Some features of interest in the experimental data sets are spikes, abrupt changes and frequent discontinuities, which considerably decrease the precision of traditional forecasting methods. As experimental results show, this classifier-based predictor exhibits competitive performance. The advantages and limitations of the model, as well as lines of improvement, are discussed.

1. Introduction

Time series (TS) analysis has become a relevant tool for understanding the behavior of different processes, both naturally ocurring and human caused (Pollock, 1999; Schelter et al., 2006). One of the latter kind of processes is the study of oil production through time, more specifically in fractured oil reservoirs, given their non-homogenous nature (van Golf-Racht, 1982). One of the tasks involved in such TS analysis is the prediction of future values (also known as TS forecasting), which is of particular interest in the context of industrial processes. For instance, accurate long-term oil production forecasting plays an essential role for the design of an oilfield. Unfortunately, it is difficult to forecast the production accurately over a planning period of several years. This fact is due to the complexity of the natural phenomena resulting in the uncertainty of the forecasting process. This difficulty increases for heterogeneous oilfields where different wells can exhibit diverse behaviors.

Computational Intelligence and Machine Learning have become a standard tool for the modeling and prediction of industrial processes in recent years, contributing models related mainly to artificial neural networks (ANN) (Palit and Popovic, 2005; Sheremetov et al., 2005). On the other hand, more classical approaches, such as Box–Jenkins Auto-Regressive Integrated Moving Average (ARIMA) models, are still widely in use (Pollock, 1999; Schelter et al., 2006). However, time series which exhibit non-linear, complex behavior tend to pose difficulties to such methods.

Time series data mining (TSDM) techniques permit exploring large amounts of TS data in search of consistent patterns and/or interesting relationships between variables. The goal of data mining is the analysis of large observational data sets to find unknown relationships and to summarize the data in novel ways that are both understandable and useful for decision making (Hand et al., 2001). The following list contains some traditional TSDM tasks (Agrawal et al., 1995; Cohen and Adams, 2001; Das et al., 1998; Sripada et al., 2002; Sheremetov et al., 2012; Mitra et al., 2002):

- Segmentation: Split a time series into a number of "meaningful" segments.
- Clustering: Find natural groupings of time series or time series patterns.
- Classification: Assign given time series or time series patterns to one of several predefined classes.
- Indexing: Realize efficient execution of queries.
- Summarization: Give a short description of a time series which retains its essential features in considered problem.
- Anomaly Detection: Find surprising, unexpected patterns or behavior.
- Motif Discovery: Find frequently occurring patterns.
- Forecasting: Forecast time series values based on time series history or human expertise.

☆ This paper has been recommended for acceptance by Jesús Ariel Carrasco Ochoa.

* Corresponding author. Tel.: +52 (55) 57296000x52535.
E-mail addresses: ilopezy@ipn.mx (I. López-Yáñez), sher@imp.mx (L. Sheremetov), cyanez@cic.ipn.mx (C. Yáñez-Márquez).
URL: http://www.alfabeta.org.mx (C. Yáñez-Márquez).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.11.008

- Discovery of association rules: Find rules relating patterns in time series.

These tasks are mutually related, for example: segmentation may be used for indexing, clustering, or summarization. Pattern Recognition techniques (supervised, such as associative models, and unsupervised, such as clustering methods) have shown to be very useful for solving some of these tasks, such as segmentation, clustering, classification, and indexing. Nevertheless, they are seldom used for forecasting. Thus, despite such advantages of associative models as low complexity, high capacity for non-linearity, or high performance, their feasibility as a non-linear tool for univariate time series modeling and prediction has not been fully explored yet.

The contribution of the present work is twofold. First, we introduce a novel non-linear forecasting technique, which is based on the Gamma classifier associative model (López-Yáñez et al., 2011), which is a supervised learning pattern classifier of recent emergence, belonging to the associative approach to Pattern Recognition. The second contribution of this paper is the results of encompassing tests for long-term time horizons performed both on forecasting benchmarks and a specific data set of oil wells monthly production. Our computations show that for the tested data sets, the Gamma classifier model can be preferable in terms of forecast accuracy compared to the previously reported techniques.

The rest of the paper is organized as follows. Section 2 is dedicated to describing some related works, while Section 3 presents the Gamma classifier, which is the basis for the proposal. The method presented here is further discussed in Section 4, while Section 5 introduces the experiments done, as well as the time series used. The experimental results and their discussion are included in Section 6, and the conclusions and future work are included in the final section.

2. Related work

This section focuses on discussing several previous works related to TSDM in general, and more particularly to the task of TS forecasting. In the second subsection, specific methods that have previously been applied to the benchmark TS are also included.

2.1. Previous works in time series data mining and forecasting

Generally, long-term TS oil production forecasting methods can be classified into two broad categories: parametric methods and Artificial Intelligence (AI) based methods. The parametric methods are based on relating production to its affecting factors by some mathematical model, like material balance and volumetric equations, or estimation of historical data of production TS. The classical way of long time production forecast in petroleum engineering are decline curve analysis (DCA) and nodal analysis, which are a kind of graphical (Fetkovich, 1987) and analytical (Sonrexa et al., 1997) methods for production forecast respectively. DCA can be seen as a method of curve fitting to one of three types: exponential, hyperbolic, and harmonic.

Recently several novel AI-based techniques have been reported for long-term (multi-step-ahead) prediction (Weiss et al., 2002; Mohaghegh, 2005; Khazaeni and Mohaghegh, 2010). Most of them are based on the identification of patterns which are later used for forecasting.

TS shape patterns usually can be defined by signs of first and second derivatives (Baldwin et al., 1998). Triangular episodes representation language was formulated in Cheung and Stephanopoulos (1990) for representation and extraction of temporal patterns. These episodes defined by the signs of the first and second derivatives of time dependent function can be linguistically described as "A: Increasing and Concave; B: Decreasing and Concave; C: Decreasing and Convex; D: Increasing and Convex; E: Linearly Increasing; F: Linearly Decreasing; G: Constant". These episodes are used for coding TS patterns as a sequence of symbols like ABCDAB. Such coded representation of TS is used for dimensionality reduction, indexing and clustering.

Linear trend patterns are frequently used for diagnosis. A Shape Definition Language (SDL) was developed in Agrawal et al. (1995) for retrieving objects based on shapes contained in the histories associated with these objects. These methods are very useful for situations when the user cares more about the overall shape but does not care about specific details.

Methods of perception-based forecasting are usually used when historical data either are not available or are scarce, for example to forecast sales for a new product (Batyrshin and Sheremetov, 2008). For instance, in predicting sales of a new product, the product life cycle is usually thought of as consisting of several stages: "growth", "maturity" and "decline". Each stage is represented by qualitative patterns of sales, e.g. a "Growth" stage can be given as follows (Bowerman and O'Connell, 1979): "Start Slowly, then Increase Rapidly, and then Continue to Increase at a Slower Rate". Such perception-based description is subjectively represented as S-curve, which could then be used to forecast sales during this stage. As our previous studies showed, to predict time intervals for each step of petroleum production can be a very difficult task for heterogeneous reservoirs.

Also, diverse techniques such as Multilayer Perceptron Backpropagation Networks, matrix memories and Support Vector Machines have proved their effectiveness in solving data classification problems.

Yet the associative memory paradigm is the one that best reflects the intention of emulating the associative nature of the brain. For this reason, these memories have been widely studied during the last years and, although other more expressive paradigms have emerged lately, in many important application fields these memories show advantages to other methods.

Particularly, associative memories are stronger than other techniques in those tasks where data is represented as static vectors and fast training and testing times are required. One of such areas is pollution forecasting. This particular application is currently being actively developed (López-Yáñez et al., 2011).

When the TS is noisy and the underlying dynamical system is nonlinear, ANN models have frequently outperformed standard linear techniques, such as the well-known Box–Jenkins models (Box et al., 1994). Additionally, other classification methods have recently been used for forecasting (Piao et al., 2008).

2.2. Previous works applied to the Mackey–Glass and CATS benchmarks

Due to the diversity of forecasting methods reported in the literature, to compare the forecasting accuracy of the proposed associative model a selection of reported models, which have been tested previously on the forecasting benchmarks, is presented here (De Gooijer and Hyndman, 2006). This representative sample includes models such as MLMVN (Aizenberg and Moraga, 2007), NARX (González et al., 2012; Menezes and Barreto, 2008), CNNE (Islam et al., 2003), SuPFuNIS (Paul and Kumar, 2002), GEFREX (Russo, 2000), EPNet (Yao and Liu, 1997), GFPE (Kim and Kim, 1997), Classical BP NN (Aizenberg and Moraga, 2007; Kim and Kim, 1997), ANFIS (Jang, 1993), Ensemble models (Wichard and Ogorzalek, 2004), Fuzzy inductive reasoning (Cellier et al., 1996), and the Kalman smoother (Sarkka et al., 2004).