Environmental Modelling & Software xxx (2017) 1–9

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Imbalanced classification techniques for monsoon forecasting based on a new climatic time series

A. Troncoso [a, *], P. Ribera [b], G. Asencio-Cortés [a], I. Vega [b], D. Gallego [b]

[a] Division of Computer Science, Universidad Pablo de Olavide, ES-41013 Seville, Spain
[b] Division of Environmental Sciences, Universidad Pablo de Olavide, ES-41013 Seville, Spain

ARTICLE INFO

Article history:
Received 17 February 2017
Received in revised form
29 September 2017
Accepted 9 November 2017
Available online xxx

Keywords:
Climatic time series
Monsoon forecasting
Imbalanced classification

ABSTRACT

Monsoons have been widely studied in the literature due to their climatic impact related to precipitation and temperature over different regions around the world. In this work, data mining techniques, namely imbalanced classification techniques, are proposed in order to check the capability of climate indices to capture and forecast the evolution of the Western North Pacific Summer Monsoon. Thus, the main goal is to predict if the monsoon will be an extreme monsoon for a temporal horizon of a month. Firstly, a new monthly index of the monsoon related to its intensity has been generated. Later, the problem of forecasting has been transformed into a binary imbalanced classification problem and a set of representative techniques, such as models based on trees, models based on rules, black box models and ensemble techniques, are applied to obtain the forecasts. From the results obtained, it can be concluded that the methodology proposed here reports promising results according to the quality measures evaluated and predicts extreme monsoons for a temporal horizon of a month with a high accuracy.

1. Introduction

The Asian Summer Monsoon is one of the atmospheric phenomena with a highest socio-economic impact in the World. The length and intensity of the monsoon season determines total precipitation in a wide and very densely populated area, extending through the southern and eastern coasts of Asia, including part of the continental areas and most of the islands in that zone (Wang et al., 2001; Weng et al., 2011). But even when it is usual to read about the Asian Summer Monsoon, there is not one but three summer monsoons in Asia. Probably, the best known and the one with most publications about its characteristics, predictability and variability is the Indian Summer Monsoon (ISM), which controls the precipitation amount and the duration of the rainy season over southern Asia, in a broad area centered over the Indian subcontinent (Ordóñez et al., 2016). Then, the East Asian Summer Monsoon (EASM) is a monsoon system affecting the continental eastern Asia, including southeastern China, Korea and Japan. Finally, the Western North Pacific Summer Monsoon (WNPSM) is often considered as an oceanic monsoon, and its area of influence includes part of the South China Sea, the Philippine Sea, their continental coasts, including the Indochina Peninsula, and the Islands in the area, mainly the Philippines (Murakami and Matsumoto, 1994).

This third monsoon, the WNPSM, was shown to have a great impact both over the global climate and over the total precipitation of very densely populated areas (Wang et al., 2001; Lee et al., 2014; Hsu et al., 2014). Precisely due to the oceanic character of this monsoon and the lack of continuous instrumental observations over the oceans, it has been difficult to obtain long lasting series representing its evolution and characteristics. Publications about this monsoon analyze a period of approximately 50 years starting in the mid 20th century. Those analyses are mostly based on reanalyzed data (i.e. relying on climate assimilation models) and, thus, are sensible to the possible lack of observed data over some of the areas analyzed. Recently, taking advantage of the meteorological data available through the ICOADS Project (International Comprehensive Ocean-Atmosphere Data Set [http://icoads.noaa.gov]) and using a method similar to the one proposed in (Gallego et al., 2015), a new, completely instrumental, long lasting series representing the WNPSM evolution and extending its length from about 50 years to more than 100 years has been developed. In particular, this new index consists on a monthly time series quantifying the persistence of a particular wind direction in two

* Corresponding author.
E-mail addresses: atrolor@upo.es (A. Troncoso), pribrod@upo.es (P. Ribera), guasecor@upo.es (G. Asencio-Cortés), ivegmar@upo.es (I. Vega), dgalpuy@upo.es (D. Gallego).

https://doi.org/10.1016/j.envsoft.2017.11.024
1364-8152/

sectors of the Western North Pacific (WNP) monsoon area, and the value of the index determines the intensity of the monsoon. The new index is perfectly comparable with the previous ones during the second half of the 20th century and was named Western North Pacific Directional Index (WNPDI) (Vega et al., 2017).

In recent years, different authors have proposed that the variability of the WNPSM can be associated to many different well-known -and usually better predictable-climatic indices such as El Niño, El Niño Modoki or the Indian Ocean Dipole (Weng et al., 2011; Feng and Chen, 2014; Lu and Lu, 2015). It has been widely accepted the idea proposed in (Wang et al., 2001), where it was shown that the phase and intensity of El Niño-Southern Oscillation (ENSO) events, represented by the Niño3.4 index, affected the intensity of the WNPSM. In this way, a weak WNPSM would be expected to occur the summer after an intense El Niño event; and a strong WNPSM would develop the year after an intense La Niña. More recently (Feng and Chen, 2014), proposed that El Niño Modoki, a climatic pattern somehow related to El Niño but with some important differences, was a better predictor for the intensity of the WNPSM than other ENSO indices. In their paper, it is suggested that WNPSM tends to be stronger during El Niño Modoki intense positive events. Nevertheless (Feng and Chen, 2014), suggested that another climatic oscillation, the Indian Ocean Dipole (IOD), modulated the impact of El Niño Modoki over the WNPSM intensity. They proposed that the relationship between WNPSM and El Niño Modoki becomes weaker when IOD is in its positive phase. Additionally, Zhang et al. in (Zhang et al., 2015) pointed out that zonal position of El Niño Modoki, rather than its intensity, is related to the IOD intensity. On the other hand, the reverse linear relationship between Indian Ocean sea surface temperature and precipitation over WNP area is modulated by the ENSO (Lu and Lu, 2015). Finally, the Pacific Decadal Oscillation (PDO) is a slow varying climatic oscillation which affects to the characteristics of most of the previous climatic patterns and has been proposed to affect to the characteristics of the precipitation of the WNP area (Chan and Zhou, 2005; Yoon and Yeh, 2016).

Data analytics has turned into an emerging research field due to the increasing amount of data being created and stored. In particular, data mining techniques try to infer knowledge from data with the purpose of automatically predicting trends and behaviors or describing models that simulate a system. In the last few years, data mining techniques have been successfully applied to forecast time series in different areas such as energy, seismology or environment (Martínez-Álvarez et al., 2015).

In this work, we have applied data mining techniques, namely imbalanced classification techniques, in order to check the capability of traditional climate indices such as El Niño, El Niño Modoki, the IOD or the PDO to capture and forecast the evolution of a monsoonal system as the one represented by the WNPSM. The main goal is to predict the occurrence of extreme monsoons for a temporal horizon of a month. Given that the number of extreme monsoons is much lower than the number of non-extreme monsoons, the resultant classification problem is highly imbalanced, where the class representing the extreme monsoons is a minority class, but the class of interest. Therefore, the main contributions of this work can be summarized as the novel formulation of the forecasting problem based on imbalanced classification techniques, which has not been to the best of authors' knowledge exploited to forecast the occurrence of monsoons so far, and on the other hand, to deal with the WNPSM, since most of the published papers deal with the ISM forecasting due to the WNPSM is predominantly an oceanic monsoon and historical data about its evolution are much scarcer.

The rest of the paper is structured as follows: next section makes a revision of the main data mining works published in the literature about the monsoon forecasting. Section 3 describes the methodology and the imbalanced classification techniques evaluated in this paper. Section 4 presents the experimental part of the paper, where we carry out a comparative evaluation of the results when predicting the intensity of the WNPSM. Section 5 closes the paper giving some final conclusions.

2. Related work

Data mining techniques have been used in the literature to deal with the problem of predicting if a monsoon is extreme. Nevertheless, the intensity of a monsoon is related to the amount of rainfalls, and therefore, the main approaches have been focused on two kind of problems: to forecast the monsoon rainfalls and to find the best attributes to be predictors to carry out the prediction. The majority of the current papers published deal with the forecasting for the ISM due to the WNPSM is predominantly oceanic and fewer historical data are recorded.

An autoencoder neural network has been proposed to obtain variables to be predictors to forecast the rainfalls for ISM (Saha et al., 2016). Namely, climatic indices such as air temperature, sea surface temperature (SST) and sea level pressure are the inputs of the autoencoder.

A graphical analysis is carried out to determine predictors for forecasting the rainfalls for the ISM in (Cannon and Mckendry, 2002). Namely, the relationships between the predictions obtained by a multiple linear regression and neural networks and their inputs (sea level pressure and geopotential height) are illustrated graphically.

Recently, clustering techniques have been applied as a previous step to improve the forecasting. In (Saha and Mitra, 2015) the authors show that the same set of climatic predictors do not have to be good for all the years. Thus, a clustering technique is applied in two dimensions to obtain groups of years and predictors simultaneously. Unsupervised neural networks (Chattopadhyay and Chattopadhyay, 2016) have also been used to obtain different clusters composed of years depending on the variability of the annual precipitations. Later, a support vector machine (SVM) is applied for each cluster to provide the forecasting of annual rainfalls for ISM. In this work, climatic predictors were not used, but the input for each SVM was the monthly rainfalls for March, April and May for the years belong to the corresponding cluster.

The anomalies of the SST are climatic indices, which are related to the rainfall variability. It is well-known that the tropical Pacific SST anomalies are related to ENSO. Therefore, many works have proposed the SST and El Niño indices as predictors, which have been successfully applied in the last years as inputs of neural networks. In (Shukla et al., 2011) the authors analyzed the correlations between the rainfalls for ISM and different indices of El Niño for lags from 1 to 8 seasons. The five indices with the highest correlations along with the principal component were used as inputs for seven linear regression models and for a multi-layer feed-forward neural network. The neural network was trained with the back-propagation algorithm and it provided the best results showing that the relationships between the rainfalls and these climatic indices are mainly non linear. The most highly correlated average of the SST anomalies over different areas with precipitations in June were chosen as predictors for a neural network in (Acharya et al., 2012). An ensemble of five neural networks with a different number of neurons in the hidden layer was proposed to forecast the annual rainfalls and the rainfalls for June, July, August and September separately for the ISM in (Singh and Borah, 2013). In this case, just historical data of the precipitations were used as inputs to the neural networks.

Other unsupervised learning techniques as association rules

Please cite this article in press as: Troncoso, A., et al., Imbalanced classification techniques for monsoon forecasting based on a new climatic time series, Environmental Modelling & Software (2017), https://doi.org/10.1016/j.envsoft.2017.11.024