Atmospheric Research 139 (2014) 128–136

Contents lists available at ScienceDirect

Atmospheric Research

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

Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data

E.G. Ortiz-García [a], S. Salcedo-Sanz [b,*], C. Casanova-Mateo [c]

[a] Department of Signal Processing and Communications, Universidad Carlos III de Madrid, Spain
[b] Department of Signal Processing and Communications, Universidad de Alcalá, Spain
[c] Department of Applied Physics, Universidad de Valladolid, Valladolid, Spain

ARTICLE INFO

Article history:
Received 13 March 2013
Received in revised form 17 January 2014
Accepted 17 January 2014

Keywords:
Precipitation prediction
Classification problem
Support Vector Machines
Predictive variables
Observational precipitation data

ABSTRACT

This paper presents a study on the Support Vector Machine (SVM) performance in a problem of daily precipitation prediction. Several novelties are included in the proposed analysis: first, a large set of novel predictive variables is considered, including upper air sounding data, variables derived from a numerical weather prediction model as well as the synoptic pattern of the atmosphere (by means of the Hess–Brezowsky classification). The importance of several of these predictive variables in the SVM performance is analyzed in the paper. In addition, two types of observational rain data are used in the experiments: first data from rain gauges (pluviometers) are considered, in order to establish the precipitation prediction, and then observational data from airports (METAR and SPECI reports) are used to carry out a similar study. The excellent performance of the SVM approach is shown by comparison with several alternative neural computation-based approaches (multi-layer perceptron, Extreme Learning Machine) and with classical algorithms such as decision trees and K-nearest neighbor classifier. Finally, the results of the persistence model are used as reference to certify the good performance of the proposed technique.

1. Introduction

Accurate precipitation (rainfall) forecast models are key in many real-world applications such as agriculture (Wei et al., 2005), water resources management (Yu et al., 2006; Kisi and Cimen, 2011) or facilities maintenance and control (such as airports' management (Benedetto, 2002)), among others (Tapiador et al., 2012). Currently, numerical weather prediction models have improved their performance, but they are stil unable to provide quantitative precipitation forecasts at enough spacial and time resolution as required in some real applications (Kuligowski and Barros, 1997). Several previous works have applied Soft-Computing approaches to overcome this difficulty, mainly based on neural computation approaches. These approaches have several advantages over global numerical models: they are much more simple and fast to train, can be applied to data from a specific point of measurement (a specific area in a river basin, for example), and their performance is really competitive compared to global techniques.

Neural computation models for precipitation prediction started to be applied about twenty years ago (Kuligowski and Barros, 1997; Navone and Ceccatto, 1994; Hall et al., 1999). Some of these pioneering works applied multi-layer perceptrons to a set of predictive variables, carefully chosen to be related to rainfall, and with data from precipitation gauges (pluviometers) to obtain rainfall quantity (Hall et al., 1999; Kuligowski and Barros, 1998; Hung et al., 2009). The majority of these approaches considered short-term

* Corresponding author at: Department of Signal Theory and Communications, Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain. Tel.: +34 91 885 6731; fax: +34 91 885 6699.
E-mail address: sancho.salcedo@uah.es (S. Salcedo-Sanz).

0169-8095/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.atmosres.2014.01.012

precipitation prediction, from 6 h to 24 h time-horizons, obtaining good results in the prediction (Yuval Hsieh, 2003). There are other approaches focused on long-term rainfall prediction and precipitation trends in a given zone, such as (Philip and Joseph, 2003), where the rainfall trend in the southern part of Indian Peninsula is analyzed by using an Adaptive Basis Function Neural Network with a back-propagation training algorithm. In (Silverman and Dracup, 2000) a multi-layer perceptron is applied to a problem of long-term precipitation prediction in California. More recently, in Nastos et al. (2013) an artificial neural network has been applied to model and forecast precipitation in Athens, Greece. In Shukla et al. (2011), a neural network is applied to forecast precipitation during the summer Monsoon station in India, using El Niño South Oscillation (ENSO) indices. In Manzato (2007), a neural computation approach is applied to the short-term forecasting of thunderstorm rainfall.

In the last few years, alternative classification and regression algorithms have been applied to precipitation prediction and modeling, mainly Support Vector Machines (SVMs). SVMs are a Machine Learning method for classification and regression problems, that reduces the problem of over-fitting by means of adopting the theory of structural risk minimization. The original idea is due to V. Vapnik (1998), and since then, this methodology has been applied to a huge number of forecasting problems. In fact, SVMs have been quite used recently in precipitation prediction and other related applications, due to its excellent characteristics of robustness and generalization performance. In Ingsrisawang et al. (2008), a comparison of machine learning algorithms (decision trees (DT), neural networks (ANN) and SVMs)) has been carried out for a short-term precipitation prediction problem in Thailand. In Wu et al. (2010), a hybrid SVM for regression with particle swarm optimization was applied to a problem of rainfall prediction. In Lu and Wang (2011), a SVM approach with different kernel functions is presented to predict monthly rainfall in a region of China. In Kisi and Cimen (2012), a novel wavelet-SVM approach was applied to precipitation forecasting from past data. SVMs have also been recently applied to precipitation related studies, such as precipitation downscaling (Tomassetti et al., 2009; Chen et al., 2010) or streamflow prediction (Kisi and Cimen, 2011). In spite of this huge work on precipitation prediction using SVM algorithms, there are still some open research lines to investigate, in order to improve the performance of the systems.

The study described in this paper is focused on the SVM performance in a problem of daily precipitation prediction, including several novel points that have not been taken into account in previous works in the literature: first, new predictive variables are considered. Last studies with SVMs did not include some important meteorological variables coming from atmospheric soundings. To our knowledge, the synoptic configuration of the atmosphere (synoptic situation using Hess–Brezowsky classification) has not been either considered in precipitation prediction studies with machine learning techniques, in spite of its significance to establish precipitation regimes in mid-latitude regions (Trigo and DaCamara, 2000). The importance of other predictive variables such as humidity and Equivalent Potential Temperature (both measured in vertical soundings) is also evaluated, and it is studied as well the influence of groups of these variables in the SVM performance. In addition, a variation on the rainfall data used in this paper has been included: the majority of previous studies consider precipitation gauge data as real rainfall measurement. In this paper it is proposed that the use of rain gauge data and observational data (METAR and SPECI reports) from airports as well, and the algorithms' performance using both rain measures has been evaluated. Finally, the performance of the SVM approach is compared with that of different machine learning techniques: a multi-layer perceptron, an Extreme Learning Machine and a Decision Tree, and also with a reference prediction quite used in meteorology: the persistence. It is shown that the SVM approach is able to obtain the best results among all compared approaches.

The rest of this paper is structured as follows: next section presents the modeling of the precipitation prediction problem as a classification problem. Section 3 reviews the main concepts of Support Vector Machines for classification, and Section 4 presents the experimental part of the paper. Finally, some concluding remarks for closing the paper are given in Section 5.

2. Problem statement, predictive variables and rainfall measurements used

Let $S \triangleq \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^n, y_i \in \{-1, 1\}, i = 1, \ldots, N\}$ be a data set that can be divided into a training set $\aleph_T \triangleq \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^n, y_i \in \{-1, 1\}, i = 1, \ldots, N_T\}$ and an independent test set $\aleph_E \triangleq \{(\mathbf{x}_i, y_i | \mathbf{x}_i \in \mathbb{R}^n, t_i \in \{-1, 1\}, i = 1, \ldots, N_E\}$, where x_i stands for the input (predictive) variables at 00:00 h of day i and y_i stands for the dependent (output or *result*) variable, i.e. $y_i = 1$ if it rains during day i and -1 otherwise, i.e., a 24 hour horizon is considered in this problem, and $N = N_T + N_E$. The problem consists of constructing a machine $f(\mathbf{x})$ such that, after a training process involving the set $\aleph_T$, it is able to provide the best possible generalization (in terms of probability of error $P_e(f(\mathbf{x}))$ or accuracy, obtained by comparing the outcome of $f(\mathbf{x})$ with the corresponding label y) in the set $\aleph_E$. Thus, it is ensured that, if a new sample $\mathbf{x}_k$ (representing a set of predictive variables for a new day k) is considered, the already trained machine $f(\mathbf{x})$ will estimate the optimal classification for sample k, i.e. the classification with the highest reliability that the used training set allows.

Observe that the definition above leaves open the process of constructing the classification machine $f(\cdot)$ for the rainfall prediction (many methods or algorithms can be used to this end), and it also leaves open the number and type of predictive variables used in the problem. Regarding this point, rainfall requires the existence of adequate clouds to produce precipitation. Therefore in order for precipitation to occur, three basic factors should be combined in an adequate way: condensation nuclei, enough water vapor (moist) and vertical movements (updrafts and downdrafts as well as the atmospheric stability). As a consequence, data selection should cover all these three elements so as to obtain a robust group of predictive meteorological variables related to the physical processes involved in the production of precipitation. Fortunately, an adequate number of condensation nuclei (such as smoke from industrial, particles of salt, etc.) on which water vapor undergoes condensation to form water droplets or deposition to form ice crystals are almost always present in the atmosphere. Then, it is only necessary to