Atmospheric Research 169 (2016) 424–433

Contents lists available at ScienceDirect

Atmospheric Research

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

Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals

Hanna Meyer *, Meike Kühnlein, Tim Appelhans, Thomas Nauss

Environmental Informatics, Faculty of Geography, Philipps-University Marburg, Deutschhausstr. 12, 35037 Marburg, Germany

ARTICLE INFO

Article history:
Received 10 December 2014
Received in revised form 19 September 2015
Accepted 22 September 2015
Available online 8 October 2015

Keywords:
Machine learning
Rainfall retrieval
Rainfall rate
Rainfall area
MSG SEVIRI

ABSTRACT

Machine learning (ML) algorithms have successfully been demonstrated to be valuable tools in satellite-based rainfall retrievals which show the practicability of using ML algorithms when faced with high dimensional and complex data. Moreover, recent developments in parallel computing with ML present new possibilities for training and prediction speed and therefore make their usage in real-time systems feasible.

This study compares four ML algorithms — random forests (RF), neural networks (NNET), averaged neural networks (AVNNET) and support vector machines (SVM) — for rainfall area detection and rainfall rate assignment using MSG SEVIRI data over Germany. Satellite-based proxies for cloud top height, cloud top temperature, cloud phase and cloud water path serve as predictor variables.

The results indicate an overestimation of rainfall area delineation regardless of the ML algorithm (averaged bias = 1.8) but a high probability of detection ranging from 81% (SVM) to 85% (NNET). On a 24-hour basis, the performance of the rainfall rate assignment yielded R^2 values between 0.39 (SVM) and 0.44 (AVNNET). Though the differences in the algorithms' performance were rather small, NNET and AVNNET were identified as the most suitable algorithms. On average, they demonstrated the best performance in rainfall area delineation as well as in rainfall rate assignment. NNET's computational speed is an additional advantage in work with large datasets such as in remote sensing based rainfall retrievals.

However, since no single algorithm performed considerably better than the others we conclude that further research in providing suitable predictors for rainfall is of greater necessity than an optimization through the choice of the ML algorithm.

1. Introduction

Spatially explicit, continuous and high-resolution monitoring of precipitation is important for a variety of fields in the environmental sciences as well as for the economy and society as a whole. Satellite-based methods are currently the only way to fulfill the requirement of area-wide information. Among the variety of available satellite systems, optical sensors on-board geostationary satellites offer high spatial and temporal resolutions, which are important when considering local and short-term rainfall events (Thies and Bendix, 2011). Furthermore, the latest systems feature adequate spectral resolutions for detecting cloud-top properties such as cloud top height, cloud top temperature, cloud phase and cloud water path (Thies et al., 2008a).

Over the last several decades, many optical satellite-based rainfall retrieval techniques for the detection of precipitating clouds and assignment of rainfall rates have been developed (see valuable overviews by Kidd and Levizzani, 2011; Prigent, 2010; Thies and Bendix, 2011; Kidd and Huffman, 2011; Levizzani et al., 2002). These retrievals are generally based on parametric relations between spectral properties as proxies for cloud-top properties, rainfall areas and rainfall rates. Rainfall areas are commonly delineated from non-raining clouds using thresholds in selected satellite channels and/or derived information (Ba and Gruber, 2001; Feidas and Giannakos, 2012; Roebeling and Holleman, 2009; Thies et al., 2008a,2008b). Rainfall rates are then assigned by relating the spectral information to measured or modelled rainfall rates (Adler and Negri, 1988; Kühnlein et al., 2010; Roebeling and Holleman, 2009; Vicente et al., 1998).

The parametric techniques used within rainfall retrievals have the advantage that they directly map the conceptual knowledge of rainfall processes to their retrieval using remotely sensed proxies. However, machine learning (ML) approaches have generally been shown to be superior when the prediction, and not the understanding of underlying processes, is the focus (Kuhn and Johnson, 2013). Moreover, parametric approaches usually consider only a limited number of predictor variables while ML algorithms can handle the full set of available information.

Precipitation processes leading to different rainfall intensities are very complex. In this context ML algorithms have been deemed

* Corresponding author at: Philipps-University Marburg, Deutschhausstr. 12, 35037 Marburg, Germany. Tel.: +49 6421 2825954.
E-mail address: hanna.meyer@geo.uni-marburg.de (H. Meyer).

http://dx.doi.org/10.1016/j.atmosres.2015.09.021
0169-8095/

valuable tools for dealing with complexity, non-linearity and highly correlated predictor variables. Neural network algorithms are most frequently used in rainfall retrieval techniques to link the input information to rainfall estimates (Behrangi et al., 2009; Capacci and Conway, 2005; Giannakos and Feidas, 2013; Grimes et al., 2003; Hong et al., 2004; Hsu et al., 1997; Rivolta et al., 2006; Tapiador et al., 2004). Random forests is an ensemble technique commonly applied in remote sensing especially for land cover classifications (Gislason et al., 2006; Pal, 2005; Rodriguez-Galiano et al., 2012; Steele, 2000), and its application in rainfall retrievals is very new. Recently, Islam et al. (2014) used random forests to classify rainfall areas from satellite-borne passive microwave radiometers. At the same time, Kühnlein et al. (2014a) and Kühnlein et al. (2014b) investigated the potential of random forests as a tool within satellite-based rainfall retrievals using Meteosat second generation (MSG) spinning enhanced visible and infrared imager (SEVIRI) data. Both obtained promising results for the use of random forests in rainfall retrievals. Support vector machines are less frequently used in remote sensing (Mountrakis et al., 2011) and have yet to be employed in optical rainfall retrievals. However, their potential has been shown in satellite-based land cover classifications (Kavzoglu and Colkesen, 2009; Pal, 2005) and in estimating biophysical parameters like chlorophyll concentration (Bruzzone and Melgani, 2005).

Though some rainfall retrieval techniques use different ML algorithms, to our knowledge, no study has compared different algorithms for rainfall assessment on the same dataset up until now. Hence, this study compares random forests (RF), neural networks (NNET), its extension averaged neural networks (AVNNET) and support vector machines (SVM) for their applicability in rainfall retrieval techniques.

This paper is structured as follows: Section 2 explains the methodology of the comparison study including data preprocessing, model training and the validation strategy. Section 3 presents the results of the comparison study which are then discussed in Section 4.

2. Data and methodology

Following the approach developed by Kühnlein et al. (2014a), rainfall area and rainfall rates were predicted for Germany during summer 2010. Day, twilight and night precipitation events were all treated separately due to differing information content about the cloud properties at different times of day. MSG SEVIRI data were used since they permit a quasi-continuous observation of the rainfall distribution and rainfall rate in near-real time. A radar-based precipitation product from the German Weather Service, RADOLAN RW (Bartels et al., 2004), was used for ground truth data.

The general work flow included preprocessing the data to provide three datasets for model training: a day, twilight and a night dataset. The retrieval process was two-fold and consists of (i) the identification of precipitating cloud areas and (ii) the assignment of rainfall rates. Since the focus of this study is on the comparison of the algorithms, the validation of rainfall rate assignments was based on rainfall areas derived from the radar network rather than the results from step (i). This ensures that the performance of rainfall rate models is comparable without confusion based on errors from the prior rainfall area delineation. Fig. 1 shows the work flow of the model training and comparison: For each dataset one model for rainfall area delineation and one model for rainfall rate assignment was tuned and trained for each of the chosen ML algorithms. The final models were applied to a test dataset and their performance was compared between the ML algorithms.

The following sections describe these steps in detail. All modeling and analysis were completed using the R environment for statistical computing (R Core Team, 2014). Model tuning, training and prediction were performed using the caret package (Kuhn, 2014a) as a wrapper package for a large list of machine learning algorithms implemented in R. Parallel processing was performed on 16 cores using the R package “doParallel” (Revolution Analytics and Weston, 2014).

2.1. Datasets

2.1.1. Satellite data

MSG SEVIRI (Aminou et al., 1997) scans the full disk every 15 min with a spatial resolution of 3 by 3 km at sub-satellite point. Reflected and emitted radiances are measured by 12 channels, three channels at visible and very near infrared wavelengths (between 0.6 and 1.6 μm), eight from near-infrared to thermal infrared wavelengths (between 3.9 and 14 μm), and one high-resolution visible channel.

MSG SEVIRI data were downloaded from the EUMETSAT data center (www.eumetsat.int) and were preprocessed based on a newly designed Meteosat processing scheme implemented in co-operation with the computer science department at Marburg University. The processing chain uses xxl technology and custom raster extensions which were

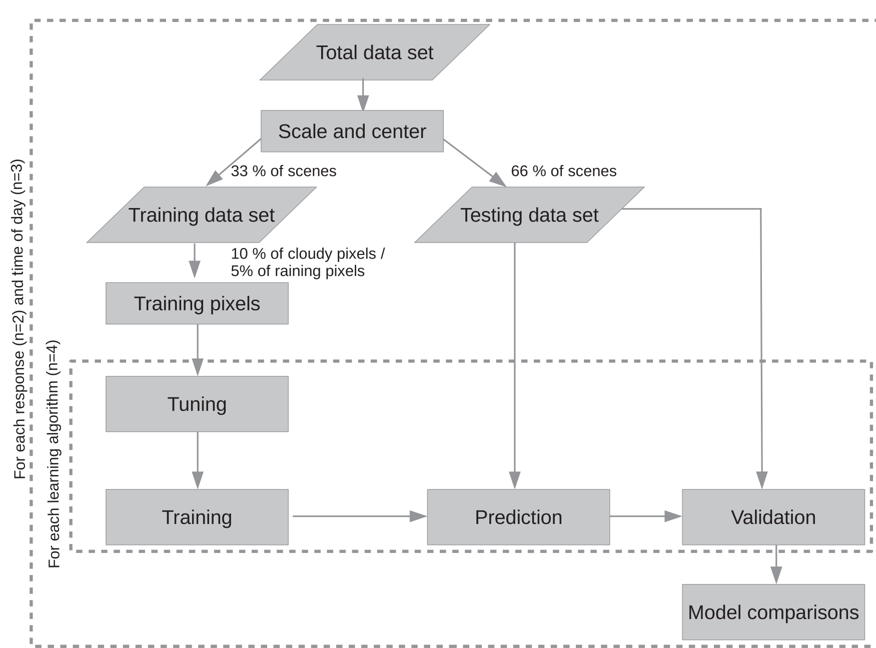

Fig. 1. Flow chart of the main methodology applied in this study.