Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2020-0006

Vol. **28**, No. 4, pp. 1305-1320
October 2020

A Data-Driven Random Subfeature Ensemble Learning Algorithm for Weather Forecasting

Chen Yu[1], Haochen Li[2], Jiangjiang Xia[3], Hanqiuzi Wen[1,4] and Pingwen Zhang[1,4,*]

[1] *School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China.*
[2] *School of Science, Beijing University of Posts and Telecommunications, Beijing 100876, P.R. China.*
[3] *Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, P.R. China.*
[4] *National Engineering Laboratory for Big Data Analysis and Application, Peking University, Beijing 100871, P.R. China.*

Received 10 January 2020; Accepted (in revised version) 20 April 2020

Abstract. In this paper, the RSEL (Random Subfeature Ensemble Learning) algorithm is proposed to improve the forecast results of weather forecasting. Based on the classical machine learning algorithms, RSEL algorithm integrates random subfeature selection and ensemble learning combination strategy to enhance the diversity of the features and avoid the influence of a small number of unstable outliers generated randomly. Furthermore, the feature engineering schemes are designed for the weather forecast data to make full use of spatial or temporal context. RSEL algorithm is tested by forecasting the wind speed and direction, and it improves the forecast accuracy of traditional methods and has good robustness.

AMS subject classifications: 62P12, 86A10, 93B15, 97M10

Key words: Weather forecasting, ensemble learning, machine learning, feature engineering.

1 Introduction

Weather forecasting is closely related to various fields, including agriculture, transportation, industry, and energy. In recent years, weather forecasting industry has developed rapidly, which mainly relies on the better theory, the updating of numerical weather prediction (NWP), the increase in the number and accuracy of meteorological observatories, and the improved computational power [1]. A variety of weather prediction methods have been developed in the literature, and they are generally classified into physical methods, statistical methods, machine learning methods, and hybrid methods [2].

*Corresponding author. *Email addresses:* pzhang@pku.edu.cn (P. Zhang), yuchen1995@pku.edu.cn (Y. Chen), lihaochen_bjut@sina.com (H. Li), xiajj@tea.ac.cn (J. Xia), qiuzi.wh@pku.edu.cn (H. Wen)

The physical methods based on the NWP model, which simulates the overall trend of atmospheric motion by solving atmospheric physical equations [3]. However, The NWP models have deficiencies, such as the adaptability of physical equations to local alpine areas, not enough spatial and temporal resolution and bad results of nowcasting and short-term forecasting [4]. Global NWP models include the European Centre for Medium-Range Weather Forecasts (ECMWF), the Global Forecast System (GFS), the Integrated Forecast Model (IFS), etc [5–7]. The statistical method utilizes historical observation data to establish a statistical model for training, which is suitable for short-term prediction. Commonly used statistical methods are Model Output Statistics (MOS) [8–11], Analog Ensemble (ANEN) [12, 13], Kalman Filter (KF) [14, 15] and Markov Chain models [16, 17]. Statistical methods are not available for medium and long-term forecasting, and these methods are not suitable for solving the problem of large data volume.

Machine learning methods can deal with big data in meteorological fields, such as meteorological observations and NWP data. There have been many applications of machine learning in meteorological science [18–20]. The features of big data are diverse in machine learning, thus how to extract useful information from the ever-increasing stream of geoscience data and how to obtain effective features from the NWP models are unavoidable problems [21]. But researches on feature engineering in weather forecasting have received little concerns [22, 23]. Li et al. (2019) proposed model output machine learning (MOML) method to process spatiotemporal features and solved the grid temperature forecasting problem [24]. Nevertheless, due to the spatial and temporal complexity of weather forecasting, it is difficult for current methods to give an optimal scheme directly. A new approach is a hybrid model, coupling physical NWP models with the versatility of data-driven machine learning [21]. Most of the existing hybrid models only mix several statistical methods with weighted strategies and do not form an integrated machine learning algorithm [25–27]. Thus, these methods lack the general optimal strategy. Ensemble learning achieves learning tasks by building and combining multiple base learners. It has superior generalization than a single learner. The representative methods of ensemble learning include boosting and bagging [28–30].

In this paper, an innovative random subfeature ensemble learning algorithm (RSEL) is proposed for weather forecasting. RSEL is a data-driven hybrid ensemble learning algorithm, it brings forth new ideas in the feature engineering scheme and the strategy of ensemble learning algorithm, which also couples the NWP model data and the observational data. To test the application in practical problems, we applied the RSEL algorithm to forecast the wind speed and wind direction at two weather stations that are located in the alpine region, and focused on the next 12-240 h forecasting results. We performed experiments to verify the root mean square error and forecast accuracy of these results and compared them with the ECMWF model, the classical multivariate linear MOS algorithm [10], and MOML algorithm, which has certain innovative meanings [24].

The remainder of the paper is organized as follows. In Section 2, the data concerned in this study are described. Feature engineering scheme and random subfeature ensemble learning algorithm are proposed in Section 3 and Section 4 respectively. Section 5 gives