Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Climate Prediction with Tree Structure Based on Random Forest

Zhouyu Ding*

Wuyi Hongda, Beijing, China

* Corresponding Author Email: kjwgood@tzc.edu.cn

Abstract. Climate prediction refers to the use of scientific methods and techniques to predict and analyze changes in the natural environment at a certain point in the future. This paper aims to improve climate prediction by analyzing its parameters and impacts using Random Forest (RF) techniques. Specifically, firstly, data frames (DF) are manipulated to explore data features, perform visualization, and perform preprocessing. Second, the paper introduces RF as a base model as a backbone network to accurately estimate climate change. RF is robust to outliers in the data using a tree-based approach. By integrating multiple decision trees and introducing stochasticity (e.g., randomly selecting features) during the training process, RFs are effective in reducing the risk of overfitting and improving the model's generalization ability. Third, mean absolute error loss (MAPE loss) is used to compare the errors that are evaluated on the data. The experimental outcomes demonstrate the efficacy of the proposed model in climate prediction, with the accuracy increased to 94%. Applying the proposed model in climate prediction in this paper provides valuable insights for climate prediction.

Keywords: Climate Prediction; Data Frames; Random Forest; Mean Absolute Error Loss.

1. Introduction

Climate prediction refers to using scientific methods and technologies to predict and analyze the changes in the natural environment at a certain time in the future. Such predictions can involve changes in climate and weather. Climate prediction is the process of estimating and predicting the long-term meteorological conditions of the Earth's atmosphere, oceans, and land systems over some time in the future. Typically, climate forecasts have time horizons ranging from months to years, which differs from the short-term projections of weather forecasts. Climate change will alter the diversity and distribution of species, including those related to human health [1]. It can predict weather conditions in advance and reduce property damage caused by extreme weather.

In the last ten years, an increasing number of research efforts have utilized paleoclimatic data on temperature and Carbon dioxide (CO2) levels to estimate Equilibrium Climate Sensitivity (ECS) across various historical climate states, indicating that ECS tends to increase as CO2 concentrations rise [2]. Leo Breiman's Random Forest (RF) algorithm has evolved into a fundamental tool for data analysis in the field of informatics [3]. RF can efficiently handle many large training datasets and a large number of classes of object detection [4]. Previous generations could only predict the climate roughly, but now the use of RFs has made climate predictions more accurate. Every tree relies on the value of a randomly sampled vector, and this distribution remains consistent across all trees within the forest. Trees obtained by traditional methods often fail to grow to arbitrary complexity and may lose some data accuracy [5]. RFs consist of multiple tree predictors combined. As the forest's tree count grows, its generalization error gradually stabilizes. The forest's generalization error hinges on both the individual trees' efficacy and their correlation within the forest [6]. RFs are being used more and more frequently, they can handle "small n big p" (refers to the case of a dataset with a relatively small sample size (small n) but many features or variables (big p)) problems, and complex interactions, and can predict highly correlated variable [7].

The purpose of this paper is to improve the level of climate prediction by introducing RF and analyzing its parameters and effects. Specifically, first, the Data Frame (DF) is manipulated and analyzed to explore the characteristics of the data, perform visualization, perform preprocessing, or

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

for the training and evaluation of machine learning models. Instead of directly applying a specific machine learning technique, the article performed a general step of data preprocessing, separating features and labels for easy training of machine learning models. Second, the article introduces the RF as the base model and uses it as the backbone network to estimate the percentage of climate change more clearly and accurately. RF uses a tree-based approach that is robust to outliers in the data without being overly sensitive to them. RFs can also be used for a variety of machine learning tasks such as classification and regression, performing well on many different types of data sets. By integrating multiple decision trees and introducing randomness (such as randomly selected features) into the training process, RFs can effectively reduce the risk of overfitting and improve the generalization ability of the model. RFs generally have high accuracy and can produce excellent predictions on many problems. Third, the Mean Absolute Percentage Error loss (MAPE loss) comparison error is an evaluation of the data. This experiment demonstrated the accuracy of climate prediction, increasing the accuracy to 94%, indicating that RFs can effectively make climate predictions. This paper puts forward the application of the model to climate prediction and makes sufficient analysis and comparison, which provides valuable tools and insights for climate prediction.

2. Organization of the Text

2.1. Dataset Description and Preprocessing

The dataset used in this study is climate change, which is obtained from the kaggle [8]. There are data dates and values for specific climate changes. The data can then be processed to become a tool for judging the accuracy of climate predictions. The DF model is used to store and manipulate data. Many methods and functions are provided to process data, such as reading and writing files, data cleaning, missing value processing, etc. Climate change is a climate data set that judge climate predictions based on five related parameters.

2.2. Proposed Approach

The study explores more accurate predictions about the climate. Climate prediction provides an important information base for various fields, enabling people to respond more effectively to the challenges posed by climate change. DF is processed and analyzed to explore the characteristics of the data and perform pre-processing [9]. The RF is then referenced as the base model and used as the backbone network. RFs generally have strong accuracy and are not overly sensitive to outliers [3]. The map loss comparison error is then used to evaluate the data. The model cited by MAPE loss is suitable for evaluating prediction errors at different scales and units [10]. The MAPE is averaged by adding the percentage error between the predicted value and the actual value for each sample. Finally, the data are integrated and processed to get the research conclusion. The idea of the experiment and the model used are shown in Figure 1. This allows for precise filtering of the data, helping to predict a more accurate percentage.

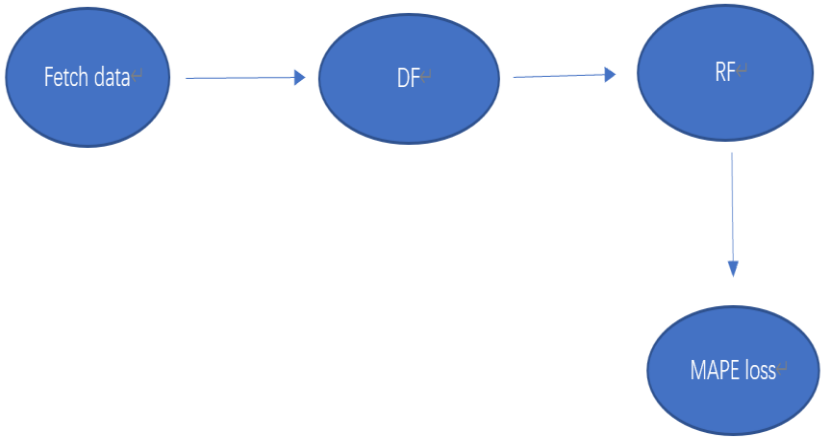

Figure 1. Ideas and models

2.2.1. Data Frame (DF)

DF is a pivotal data structure entrenched within programming environments, serving as a cornerstone for data processing and analysis endeavors. DF operates as a tabular representation of data, aligning with rows and columns akin to a spreadsheet. This construct proves indispensable in a myriad of applications, with its versatility particularly pronounced in the domain of climate change analysis. DF assumes a pivotal role in scrutinizing and sifting through climate change data [11]. Its multifaceted capabilities extend far beyond mere storage, facilitating intricate analytical maneuvers essential for deriving meaningful insights. DFs boast a comprehensive suite of functionalities, enabling seamless manipulation, filtering, sorting, grouping, and merging of data points. This amalgamation of features positions DFs as a quintessential tool in the arsenal of data analysts and researchers alike. Facilitating the seamless transformation and organization of data, DFs empower analysts to discern intricate patterns and trends lurking within voluminous datasets. Leveraging its intuitive interface, users can effortlessly execute a gamut of operations, ranging from basic arithmetic computations to sophisticated statistical analyses. Moreover, DFs foster interoperability with a plethora of auxiliary libraries and tools, augmenting their utility across diverse analytical paradigms.

DF emerges as an indispensable asset in the contemporary landscape of data analysis, epitomizing the convergence of computational prowess and analytical acumen. Its utility extends far beyond the confines of conventional programming paradigms, permeating diverse domains and catalyzing groundbreaking discoveries in the realm of climate change research and beyond.

2.2.2. Random Forest (RF)

The RF method is an advanced ensemble learning approach that leverages multiple decision trees to improve predictive precision and robustness [3,4]. Comprising an ensemble of decision trees, RF operates by randomly sampling the data and selecting subsets of features for each tree. This randomization process injects diversity into the model, mitigating the risk of overfitting and bolstering its generalizability. The predictive prowess of RF stems from its ability to aggregate the predictions of numerous decision trees, thereby leveraging the collective intelligence of the ensemble. By amalgamating the outputs through voting or averaging, RF can furnish predictions that are more robust and reliable than those of individual trees alone. This ensemble approach imbues RF with predictive stability that is invaluable in domains like climate prediction, where accurate forecasts hinge on discerning subtle patterns amidst complex data. Furthermore, RF facilitates a nuanced understanding of feature importance within the dataset. By evaluating the frequency of feature usage across decision trees or gauging the improvement in node purity upon feature division, RF elucidates the relative influence of each feature on the prediction outcomes. This feature importance analysis

not only enhances interpretability but also guides feature selection and refinement processes, augmenting the efficacy of the predictive model.

A notable advantage of RF lies in its inherent parallelizability. Each decision tree within the ensemble can be constructed independently, allowing for concurrent training across multiple processors or computing nodes. This parallelization capability expedites the model training process, affording researchers and practitioners greater efficiency in handling large-scale datasets and accelerating the pace of scientific inquiry. In essence, RF epitomizes the synergy between computational ingenuity and predictive finesse, offering a robust framework for tackling complex prediction tasks in climate research and beyond. Its ensemble nature not only fortifies predictive accuracy but also furnishes insights into the intricate interplay of features within the data, thereby empowering researchers to unravel the mysteries of climate dynamics with unprecedented clarity and precision.

2.2.3. Mean Absolute Percentage Error Loss (MAPE loss)

MAPE loss is a measure used to evaluate the performance of a predictive model, rather than a loss function. It plays a role in predicting the percentage of climate change in the experiment. MAPE calculates the average of the model's percentage error concerning observations. The role of MAPE is to provide insight into the relative error of the model in the prediction [10]. The lower the average absolute value of the percentage error, the more accurate the model's prediction, as:

$$MAPE = \frac{1}{n}\sum\left|A_i - \frac{F_i}{A_i}\right| \times 100 \quad (1)$$

where Ai stands for the real observed value for the ith sample, Fi represents the predicted value for the sample i. n is the number of samples. The initial step involves calculating the percentage error between the actual and predicted values for each sample, then taking its absolute value, averaging the percentage error for all samples, and finally multiplying by 100 to get the average absolute percentage error in the form of a percentage.

3. Results and Discussion

First, the DF model is used to process the data, filter, sort, and group the data. Make the data clearer [12]. Then the RF model is used to collect the results of the decision tree for prediction, which improves the stability and accuracy of prediction [4]. Often used for tasks such as classification, regression, and feature importance assessment, RF performs well when dealing with large data sets and high-dimensional feature Spaces, reducing the potential for error. Finally, MAPE loss was used for a clearer percentage calculation. The lower the average absolute value of the percentage error, the more accurate the model's prediction. Make the whole experiment perfect.

DF is used to filter the data, and five groups of data were selected in Table 1, which clearly expresses the selected data in the library. DF performs data manipulation, filtering, sorting, grouping, merging, and more. DF stores different types of data and provides rich capabilities for data manipulation and analysis. The data clearly lists the year, month, date, average and actual data. DF was used to screen out five data, which clearly reflected the data.

Table 1. DF Indicates the Filtered Data

Year	Month	Day	Week	Temp2	Temp1	Average	Actual
2019	1	1	Fri	45	45	45.6	45
2019	1	2	Sat	44	45	45.7	44
2019	1	3	Sun	45	44	45.8	41
2019	1	4	Mon	44	41	45.9	40
2019	1	5	Tues	41	40	46	44

Scikit-Learn library will create and train a random forest regressor model and train it with training data for subsequent use to predict labels or target values for unknown data. Make a tree graph of data that is initially irrelevant and make statistics on the data.

4. Conclusion

In this paper, the stochastic forest model is used to predict the climate. In this paper, the extracted data is analyzed in detail, and DF is used to filter and merge the data. Subsequently, the data is thoroughly examined using the random forest model. Finally, the MAPE loss model was used to conduct a detailed audit of the results. The climate prediction was 94% accurate. Going forward, the hope is to make climate predictions 100 percent accurate, which could benefit agriculture and keep people safe in harsh climates. Data is cleaned, missing values are filled, outliers are removed, and other pre-processing operations are carried out to ensure the quality and consistency of the input data. The model is updated in real-time to introduce the latest meteorological data and observations promptly to ensure the model's accurate prediction of the current climate state.

References

[1] K. Michael, et al. Integrating Biophysical Models and Evolutionary Theory to Predict Climatic Impacts on Species' Ranges: The Dengue MosquitoAedes Aegyptiin Australia. Functional Ecology, 23(3), 2009, pp.528–538.

[2] Tierney, E. Jessica, et al. Past Climates Inform Our Future. Science, 370(6517), 2020.

[3] Boulesteix, A. Laure, et al. Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 2012, pp. 493–507.

[4] Gall, Juergen, et al. An Introduction to Random Forests for Multi-Class Object Detection. Lecture Notes in Computer Science, 2012, pp. 243–263.

[5] L. Bingguo, et al. Scalable Random Forests for Massive Data. Lecture Notes in Computer Science, 2012, pp. 135–146.

[6] Breiman, Leo. Random Forests. Machine Learning, 45(1), 2001, pp. 5–32, pp.1471-2105.

[7] S. Carolin, et al. Conditional Variable Importance for Random Forests. BMC Bioinformatics, 9(11), 2008.

[8] Information on: www.kaggle.com/code/anandhuh/climate-prediction-random-forest-94-accuracy/notebook.

[9] Petersohn, Devin, et al. Towards Scalable Dataframe Systems. ArXiv.org, 2020.

[10] V. Eliana, et al. A Systematic Review of Statistical and Machine Learning Methods for Electrical Power Forecasting with Reported MAPE Score. Entropy, 22(12), 2020, p. 1412.

[11] Information on: www.databricks.com/glossary/what-are-dataframes.

[12] Information on: www.aporia.com/learn/a-comprehensive-guide-to-mean-absolute-percentage-error-mape.