Exploration of Climate Data and Temperature Forecasting using Machine Learning
Eman AlBalawi

https://doi.org/10.1590/SciELOPreprints.9174

Submitted on: 2024-06-29
Posted on: 2024-07-11 (version 1)
(YYYY-MM-DD)
Exploration of Climate Data and Temperature Forecasting using Machine Learning

Eman Khalid Al-Balawi*
Department of Geography, Umm ul Qura University,
Makkah, Saudi Arabia
*Corresponding author
https://orcid.org/0000-0002-8149-449X

Abstract.

In this short communication, a concept has been presented to model geographical data to predict future temperature of Tabuk, region. Machine learning has been applied to the weather station data to develop a prediction model. The preliminary results are promising and encouraging and are envisaging to further this research towards the determination of unknown temperature rise in the region. This is important to mention here, that the problem has been formulated as a Regression problem, NOT as a classification problem. Hence, applying Convolutional neural networks is not possible, due to the non-existence of classes or converting the temperature values to classes does not make any sense. Hence, this is defined as a regression problem which achieved encouraging desirable results.

Keywords: Machine learning, Geographical data, Temperature prediction.

1. Introduction

Machine learning is widely used as a tool to predict or estimate unknown values or conditional attributes for an environment given the historical data. This technique has widely used in different technologies e.g. computer vision, robotics, signal processing, biomedical, aerospace and their associated frameworks [1, 3, 4, 6, 8, 11, 12]. Many researchers have used machine learning techniques to predict the future temperature values [2, 5, 7, 9, 10].

2. Data Collection and Data Patterns

Meteorological data uploaded here covers 31 years of meteorological information obtained from the Tabuk weather radar station (ID 40375), located at 28° 36’ N, 36° 63’E. The data includes Daily Maximum Temperature (Tmax), Minimum Temperature (Tmin), Temperature (mean), Rainfall (RF), Wind Speed (WS), Relative Humidity (RH), Pressure (Press), and Vapor (V), at the station level.

The data was plotted and is shown in the Fig. 1. This can be observed in the data that it follows a sinusoidal wave pattern as shown in Fig. 1. Moreover, the Fast fourier transformation (FFT) was applied on the data features and a very similar pattern can be seen
This manuscript is a non-peer reviewed version of a conference paper. Yet to be submitted for publication in “Environmental Design, Material Science, and Engineering Technologies 22-25 April 2024, Abu Dhabi University, Dubai Campus, UAE”.

as shown in Fig. 2. To this end, a more subtle verification method was applied using Seaborn Python’s builtin in utility Pair-plot as shown in the Fig. 3. Having said that we prepared the data of these four feature sets available from the Tabouk weather station. The idea was to train the temperature values using different machine learning methods, e.g. Decision Trees (DT), K-Nearest neighbours (KNN), Support Vector Machines (SVMs) and other Deep Learning methods like Multi-Linear Perceptron (MLP) or Artificial Neural Network (ANN). The intent of this exercise is to realise that whether the features also vary w.r.t the temperature values and also whether their is a similar pattern which exists among the individual features, too.

Figure 1 – Five hundred days temperature data points. This can be seen that the data follows sinusoidal curve i.e. a sine wave pattern.
This manuscript is a non-peer reviewed version of a conference paper. Yet to be submitted for publication in “Environmental Design, Material Science, and Engineering Technologies 22-25 April 2024, Abu Dhabi University, Dubai Campus, UAE”.

Figure 2 – Fast Fourier transformation (FFT) was applied to individual features of the data.
This manuscript is a non-peer reviewed version of a conference paper. Yet to be submitted for publication in “Environmental Design, Material Science, and Engineering Technologies 22-25 April 2024, Abu Dhabi University, Dubai Campus, UAE”.

Figure 3 – Pair plot of the major features of the temperature data set. The relative kernel density estimation (KDE) was used to find the inter-relation between the features.

In order to verify all these hypothesis we developed different machine learning models, explained in more details in the results section.

3. Results

The results are presented after building various machine learning models using different algorithmic techniques, e.g. DT, KNN, SVM and multi-layer perceptron (MLP) which treated the data as a regression problem. The results show that SVM as a regression technique (SVR) for this dataset did not perform well as can be seen in the Fig. 4(a). The error was chosen as negative mean squared error, hence it can be seen in Fig. 4(b) that together with the stacking algorithm, all the rest of them were within the bounds of -2 to -2.4. Means, the stacking algorithm can be used to predict the future temperature values of Tabuk region.

A need for Climate and Geographical data specific novel machine learning algorithms and human machine interaction systems

These results pose a challenge that there is a need for the development of temperature values specific and their features specific machine learning algorithms. Although, one algorithm on this dataset performed better but what happens when the feature set increases. Classical machine learning algorithms assign weights in a linear way, there is a need for an algorithm which learns the weights and not just simply adds or subtracts them but assign some intelligent weighting system on the features. One of the theories in Psychological sciences has a potential to do this job of assigning weights with reorganizational capabilities, called as Perceptual control theory (PCT) [13, 14, 15, 16, 17, 18, 19]. There is an effort undergoing to develop machine learning systems similar to human control systems and we hope one day better machine learning systems based on perceptual models will be created which will be more accurate and light in memory.

Moreover, with the geographical and climate data being collectable from synthetic aperture radars (SAR) and satellites. There is definitely a need for designing better human machine learnable interfaces. These digital interfaces can help climate and geographical sciences researchers to manipulate and process their data more conveniently [20, 21, 22, 23, 24 25, 26, 27, 28, 29]. With the advent of virtual reality and augmented reality this would be useful along with machine learning interfaces to facilitate climate sciences researchers.
This manuscript is a non-peer reviewed version of a conference paper. Yet to be submitted for publication in “Environmental Design, Material Science, and Engineering Technologies 22-25 April 2024, Abu Dhabi University, Dubai Campus, UAE”.

![Comparative Model Performance](image)

Figure – 4: The results of the dataset. (a) All machine learning algorithms were trained and tested on the real time data. The negative mean squared error was calculated for all the algorithms. (b) A more detailed analysis can be seen, after zooming on the significant boxplots without SVR.

**Acknowledgment**

I am thankful to my mentor and supervisor Associate Prof. Tauseef Gulrez for his help towards the fulfillment of my post Ph.D work, which I used in this paper.

**Conflict of Interest**

There is no conflict of interest in this paper among the authors.

**References**

This manuscript is a non-peer reviewed version of a conference paper. Yet to be submitted for publication in “Environmental Design, Material Science, and Engineering Technologies 22-25 April 2024, Abu Dhabi University, Dubai Campus, UAE”.


This manuscript is a non-peer reviewed version of a conference paper. Yet to be submitted for publication in “Environmental Design, Material Science, and Engineering Technologies 22-25 April 2024, Abu Dhabi University, Dubai Campus, UAE”.


This preprint was submitted under the following conditions:

- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.

- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.

- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.

- The authors declare that the data, applications, and other content underlying the manuscript are referenced.

- The deposited manuscript is in PDF format.

- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.

- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.

- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.

- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.

- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.

- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.

- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.