ISSN: 2278-4632
Vol-13, Issue-08, August 2023

Scalable and Efficient Prediction of Weather Patterns Using Machine Learning

Rama Krishna Yellapragada[1], Dr. K. Krishna Murthy[2]
[1]Research Scholar, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India
[2]Retd. P.G. Director & Head of the Department of Electronics, P.G. Center, P.B. Siddhartha College of Arts & Science, Vijayawada, Andhra Pradesh, India

Abstract

Weather forecasting has numerous impacts in our daily life from cultivation to event planning. Previous weather forecasting models used the complicated blend of mathematical instruments which was insufficient in order to get higher classification rate. In contrast, simple analytical models are wellsuited for weather forecasting tasks. In this work, we focus on the weather forecasting by means of classifying different weather events such as normal, rain, and fog by applying comprehensible C4.5 learning algorithm on weather and climate features. The C4.5 classifier classifies weather events by building the decision tree using information entropy from the set of training samples. We conducted experiments on LA weather history dataset; from evaluation results, it is revealed that C4.5 classifier classifies weather events with f-score of around 96.1%. This model also indicates that climate features such as rainfall, visibility, temperature, humidity, and wind speed are highly discriminative toward events classification. We study specifically the power of making predictions via a hybrid approach that combines discriminatively trained predictive models with a deep neural network that models the joint statistics of a set of weather-related variables. We show how the base model can be enhanced with spatial interpolation that uses learned long-range spatial dependencies.

Keywords: Machine Learning, Graphical Models, Weather Forecasting, Weather Events

I.INTRODUCTION

The situation of weather plays a crucial role in almost every aspects of human life. Note that intelligent weather analysis techniques can help us to make efficient decisions that can lead us to save valuable lives, properties, and time. As a consequence, researchers focus on the automated analysis of weather and climate data such as forecasting rainfall, predicting air temperature to understand and to extract useful information. As modernization continued, prediction of weather events draws more attention. From the very beginning of civilization, people want to know the pattern of weather change. Discovering the weather pattern and forecasting weather has been a field of interest from the exploration of science and technology. Weather forecasting involves foreseeing how the current situation with the air will change in which present climate conditions are taken by ground perceptions such as from boats, airplane, radiosondes, Doppler radar, and satellites. The collected data is then sent to meteorological focuses in which the information are gathered, analyzed, and made into an assortment of outlines, maps, and charts. Algorithms exchange a huge number of perceptions onto surface and upper air maps and draw the lines on the maps with assistance from meteorologists. Algorithms draw

ISSN: 2278-4632
Vol-13, Issue-08, August 2023

the maps as well as anticipate how the maps will look at some point later on.

Making inferences and predictions about weather has been an omnipresent challenge throughout human history. Challenges with accurate meteorological modeling brings to the fore difficulties with reasoning about the complex dynamics of Earth's atmospheric system. everal challenges must be addressed in taking a datacentric approach to weather prediction. First, we note that the set of weather variables under consideration are tightly coupled. For example, pressure and temperature follow natural gas laws (i.e., the well-known formula, P V = nRT). Similarly, there is a tight relationship between relative humidity and temperature. Consequently, any model that jointly aims to predict the set of weather variables should leverage knowledge of the tight statistical couplings that are based in physics. Secondly, dependencies among the variables may have long-range influences across space and time. For instance, wind vectors across large geographic distances may follow isobaric contours. As another consideration, the weather phenomena may be affected by local geography and associated natural processes (e.g. isolated thunderstorms), as well as shifts in the large-scale structure of atmospheric phenomena (e.g. shifting of jet streams).

We aim to tackle these challenges via a representation that jointly predicts winds, temperature, pressure, and dew point across space and time. The proposed architecture combines a bottom-up predictor for each individual variable with a top-down deep belief network that models the joint statistical relationships. Another key component in the framework is a data-driven kernel, based on a similarity function that is learned automatically from the data. The kernel is used to impose long-range dependencies across space and to ensure that the inferences respect natural laws.

Numerical or computational models for weather forecasting are the dynamic representations of the systems is being used in present days. These models discretize regions or bodies in a few measurements by separately utilizing estimated capacities to portray the behavior of the climatic variables of interest [2]. Nowadays, numerical or computational models are irreplaceable for atmosphere estimation. For instance, Bayesian networks [3] with time differed scaling features can be used to review whether there are factually noteworthy patterns in the climate information. In addition, Tae-wong [4] demonstrated a space-time model that displays the short time and geographical conditions of the day by day rain event.

Although there are several techniques available for weather forecasting, weather forecasting is actually a challenging task due to the complicated physics behind weather which depends on numerous features, and which is also boisterous and deterministically confusing natural event. Moreover, people produce numerous disasters, and change of climate or characteristics of climate such as air temperature, rainfall, dew point temperature, visibility, and humidity displays a strong role on the weather. Notice that several automated techniques including Artificial Neural Network (ANN), Support Vector Regressor (SVR), Genetic Algorithm (GA) were applied to forecast or model weather [5] [6] [7] [8], where ANN was the most commonly used technique that can forecast weather with decent performance rate.

On the other hand, public services usually use the data from fixed sensors. The sensors have high quality and potential. However, they can be costly to install and maintain[9]. So the location data collected for weather can also be used to real-time traffic updates as well. In this paper, a system has been designed that will mainly act as the source

ISSN: 2278-4632
Vol-13, Issue-08, August 2023

of weather data along with real-time data for traffic condition with the help of mobile applications and device integrated with Arduino, different sensors and Bluetooth module. The data is quantitative continuous data which is analyzed to show perfect correlation for accurate prediction.

II. RELATED WORK

Over the last few decades, researchers conducted a number of automated analyses on weather and climate data such as dew point temperature prediction using several Artificial Intelligence (AI) techniques from different sub-domains of AI including model output statistics, fuzzy logic, expert system, machine learning, and data mining. For instance, Chevalier et al. [11] trained SVR on small, and minimally pre-processed meteorological dataset to predict air temperature. Devi et al. [12] developed ANN based temperature forecasting model using real-time quantitative data about the current state of the atmosphere. Olaiya and Adeyemo [13] also investigated the performance of ANN and decision trees during the classification of maximum, minimum, and mean temperature, rainfall, evaporation, and wind speed on meteorological data gathered from Nigeria. Lin and Chen [8] designed typhoon rainfall forecasting model using ANN feeding eight typhoon characteristics and spatial rainfall information, where they found that excessive spatial rainfall information may not increase the generalization of the forecasting model. Mohammadi et al. [14] predicted the dew point temperature on the daily scale on different climate conditions applying extreme learning machine algorithm on five common climate-related features such as mean air temperature, relative humidity, atmospheric pressure, vapor pressure and horizontal global solar radiation.

As per our knowledge based on literature review, Awan and Awais's research [15] is the only similar study available in the literature that also attempted to predict weather events. In their research, they aimed to predict weather events based on fuzzy RBS method for Lahore, Pakistan. They used two different datasets of 365 examples with only 4 features, and 2500 examples with 17 features e.g. temperature, dew point, humidity, sea level, visibility, wind speed, respectively, for experimentation. They mentioned in their finding that fuzzy RBS method was sensitive to random sampling with replacement technique that was applied to produce training and test dataset. In contrary, we applied comprehensible tree-based machine learning algorithm for events classification for Los Angeles, California, the USA in which we used 5325 examples with 19 features extending the feature set to include rainfall information to build the weather events prediction model.

Despite the success of machine learning in a variety of tasks, applications to the problem of weather forecasting has been limited. Exceptions include the use of Bayesian Networks for precipitation forecasts [3] and temporal modeling via Restricted Boltzmann Machines (RBM) [15]. A separate thread of research has also focused on efficient representation of relational spatiotemporal data in Random Forests for prediction of severe surface-level weather processes, such as droughts and tornadoes. More recently, large-scale wind prediction has been presented [9] using a Bayesian framework with Gaussian Processes.

To date, uses of machine learning for weather prediction have been limited in several ways. First, almost all methods consider only one variable at a time and do not explore the joint spatiotemporal statistic of multiple weather phenomena. Also, to our knowledg

Algorithm 1 C4.5 for weather events classification

1: Check for base cases
2: **for** each feature f **do**
3: Compute normalized I_{Gain} ratio from splitting on f
4: **end for**
5: Let f_{best} be the feature with maximum normalized I_{Gain}
6: Build decision node based on the splitting on f_{best}
7: Recurse on the sublists achieved via splits on f_{best}, and finally, include those node as children of node
8: **return** decision trees

ISSN: 2278-4632
Vol-13, Issue-08, August 2023

e, long-range spatiotemporal dependencies have not been modeled explicitly. We introduce methods that address these limitations, via introduction of a hybrid representation. With a hybrid representation, individual predictors are discriminatively trained from historic data and local inferences from these models are combined with a deep neural network that overlays statistical constraints among key weather variables. We additionally apply a spatial interpolation scheme that respects constraints of long-range statistical dependencies. The methodology employs covariance matrix for Gaussian Process regression constructed from a large dataset. Here, the covariance matrix, also referred to as the kernel, allows us to enforce smoothness constraints over the weather variables. By ensuring that the kernel captures the dynamics of the system as informed by the training data, we are able to align estimates according to spatial constraints imposed by natural laws.

III. C4.5 FOR WEATHER EVENTS CLASSIFICATION

C4.5 is a statistical classifier used to build a decision tree for classification. The key idea beneath C4.5 algorithm is that C4.5 creates decision trees using information entropy H from set of training samples e.g. S = s1 , s2 , . . . , sn of pre-classified samples, where each sample si comprises of Ndimensional vector x1,i , . . . , xN ,i in which xj denotes feature of the sample and class in which si falls. At each node of the tree, C4.5 selects the feature that most effectively splits its set of samples into subsets using normalized information gain as splitting gain criteria. C4.5 makes a decision using the feature with highest information gain where information gain IGain is measured as follows.

$$I_{Gain}(Event, Feature) = H(Event) - H(Event|Feature)$$

Where Event denotes weather event class, and Feature denotes available weather and climate features used in weather prediction model. Figure 1 shows a decision tree produced from C4.5 classifier on LA weather history dataset. The technical description of the C4.5 algorithm is given in Algorithm 1.

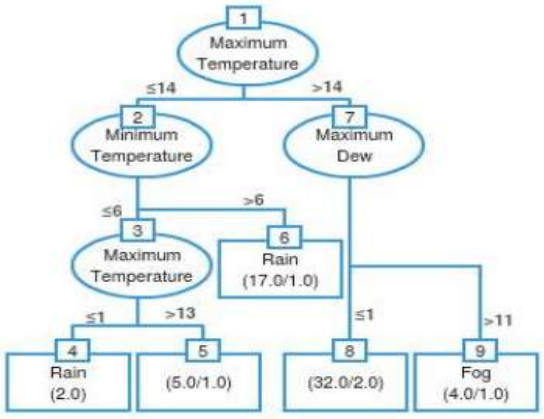

Fig. 1. An example decision tree generated by C4.5 classifier

IV. EXPERIMENTS AND RESULTS

The proposed weather events classification model was evaluated on LA weather history dataset using J48, an open source implementation of C4.5 on data mining and machine learning tool, applying cross-fold validation e.g. 5 − fold, 10 − fold, 20 − fold, and random splitting e.g. 50% − 50%, 60% − 40%, 70% − 30% strategies. We benchmarked C4.5 classifier against classic learning algorithm naive Bayes, and displayed the comparison in Table I and II. Each of the experiments was run for 20 times with different random seeds, and the results were obtained by averaging over 20 different experimental runs. We produced accuracy, precision, recall, and f-score for each weather event class to demonstrate the performance of the models. The larger values of the performance metric accuracy, precision, recall, and f-score indicate the higher weather events

ISSN: 2278-4632
Vol-13, Issue-08, August 2023

classification performance. Note that, we modeled the weather events prediction task as classification problem as we aimed to estimate the probable weather event using weather and climate features. Table I displays the performance of C4.5 and naive Bayes classifiers for multiple cross-fold validations and random splitting strategies, where Table II shows the performance of C4.5 and naive Bayes classifiers during the classification of three weather events e.g. normal, rain, and fog. From Table II, it can be indicated that C4.5 classifier was better than naive Bayes since C4.5 classified all three events with higher f-score than naive Bayes. More precisely, naive Bayes classified fog event with a low precision rate of 34.5%, and f-score rate of 49.9% that was extremely worst than the performance of C4.5 classifier. Another important point to note is that C4.5 basically confused fog and rain events with normal, and normal event with rain event, while naive Bayes widely confused fog event with both normal and rain events, and rain event with the normal event. According to the experimental results from Table II, we can outline that the proposed C4.5 classifier can efficiently classify each of the three weather events. Hence, C4.5 can be extensively utilized for weather event prediction or forecasting. In addition, C4.5 is extremely viable weather event classifier as the C4.5 classifier is comprehensible and interpretable, can deal with the over-fitting issue and may take care of persistent features.

V. HYBRID MODEL

We seek a prediction model that respects spatiotemporal dependencies among weather variables induced by atmospheric physics. We test the framework with data drawn from a continental scale weather corpus composed of data captured via balloons. In particular, we consider the IGRA dataset consisting of balloon observations made at 60 stations across the U.S. These balloons transmit observations about wind speed and direction, temperature, geopotential height, dew point, and other weather variables.

These observations are released in real time by the NOAA and later by the National Climatic Data Center following preprocessing. The data is eventually integrated into the curated IGRA dataset which is updated daily and contains historical weather data spanning decades compiled from eleven source datasets. Any data added to the archive undergoes a cycle of quality assurance to resolve potential inconsistencies among variables [4, 5]. Formally, we consider four weather variables in the model: wind velocity, v; pressure, p; temperature, t and dew point, d. The wind observations are represented as a two-dimensional vector, v = [v x , vy] while all other weather variables are scalars. We represent weather stations (where the balloons are released) as SL = {s1,...,sNs } where Ns is the total number of weather stations. For each of these stations, we have historical weather data logged at a frequency of approximately six hours over several years. Our approach to building the weather model was governed by the following guidelines:

1. **Temporal mining:** Our model should be able to identify and learn from recurring weather patterns over time.

2. **Spatial interpolation:** The dynamic influence of atmospheric laws on weather phenomena need to be accounted for in our predictions.

3. **Inter-variable interactions:** The local interdependencies between weather variables should be captured by our model.

VI. CONCLUSION

We presented a weather forecasting model that makes predictions via considerations of the joint influence of key weather variables. We introduced a

ISSN: 2278-4632
Vol-13, Issue-08, August 2023

data-centric kernel and showed how using GPR with such a kernel can effectively interpolate over space, taking into account weather phenomena such as turbulence. We performed temporal analysis using short- and longer-term features within a gradient-tree based learner. We augmented the system with a deep belief network and tuned the parameters to model the dependencies among weather variables. A set of experiments on real-world data shows that the new methodology can provide better results than NOAA benchmarks, as well as recent research that had demonstrated improvements over the benchmarks.

We also outline relevant and influential weather event features e.g. rainfall, visibility, temperature, wind speed, dew point computing their relative importance scores. In addition, feature correlation plot demonstrates that air temperature and dew point, dew point and humidity, humidity, and temperature, temperature, and visibility, humidity and rainfall are highly correlated. In our future work, we will include an extension of the weather event class to more complex events such as rainfog, thunderstorm, rain-thunderstorm, tornado, rain-tornado, rain-thunderstorm-tornado, and fog-rain-thunderstorm. The proposed event prediction model on more complex and unbalanced weather dataset with different climate conditions.

REFERENCES

1] F. Olaiya and A. B. Adeyemo, “Application of data mining techniques in weather prediction and climate change studies,” International Journal of Information Engineering and Electronic Business (IJIEEB), vol. 4.1, p. 51, 2012.

[2] G. M. S.-Z. R. G.-G. De la Torre-Gea, Guillermo and E. Rico-Garca., “Bayesian networks for defining relationships among climate factors,” Int. J. Phys. Sci., vol. 6, no. 1, pp. 4412–4418.

[3] C. H. Lima and U. Lall, “Spatial scaling in a changing climate: A hierarchical bayesian model for non-stationary multi-site annual maximum and monthly streamflow,” Journal of Hydrology, vol. 383, no. 3, pp. 307–318, 2010.

[4] T.-w. Kim, H. Ahn, G. Chung, and C. Yoo, “Stochastic multi-site generation of daily rainfall occurrence in south florida,” Stochastic Environmental Research and Risk Assessment, vol. 22, no. 6, pp. 705– 717, 2008.

[5] J. Wu, L. Huang, and X. Pan, “A novel bayesian additive regression trees ensemble model based on linear regression and nonlinear regression for torrential rain forecasting,” in Computational Science and Optimization (CSO), 2010 Third International Joint Conference on, vol. 2. IEEE, 2010, pp. 466–470.

[6] A. Bautu and E. Bautu, “Meteorological data analysis and prediction by means of genetic programming,” in Proceedings of the 5th Workshop on Mathematical Modeling of Environmental and Life Sciences Problems Constanta, Romania, 2006, pp. 35–42.

7] E. Berndt, A. Molthan, W. Vaughan and K. Fuell, "Transforming Satellite Data into Weather Forecasts", Eos, 2017.

[8] "Weather Analysis and Forecasting", American Meteorological Society (AMS), 2015.

[9] G. Leduc, "Road Traffic Data: Collection Methods and Applications",Institute for Prospective Technological Studies, pp. 5-13, 2008.

[10] V. M. Krasnopolsky and M. S. Fox-Rabinovitz. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. Neural Networks, 19(2):122–134, 2006.

ISSN: 2278-4632
Vol-13, Issue-08, August 2023

[11] R. J. Kuligowski and A. P. Barros. Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. Weather and Forecasting, 13(4):1194–1204, 1998.

[12] G. Marchuk. Numerical methods in weather prediction. Elsevier, 2012.

[13] A. McGovern, D. John Gagne, N. Troutman, R. A. Brown, J. Basara, and J. K. Williams. Using spatiotemporal relational random forests to improve our understanding of severe weather processes. Statistical Analysis and Data Mining: The ASA Data Science Journal, 4(4):407–429, 2011.

[14] A. McGovern, T. Supinie, I. Gagne, M. Collier, R. Brown, J. Basara, and J. Williams. Understanding severe weather processes through spatiotemporal relational random forests. In 2010 NASA conference on intelligent data understanding, 2010.

[15] R. Mittelman, B. Kuipers, S. Savarese, and H. Lee. Structured Recurrent Temporal Restricted Boltzmann Machines. In Proceedings of the 31st International Conference on Machine Learning (ICML), pages 1647–1655, 2014.

[16] Y. Radhika and M. Shashi. Atmospheric temperature prediction using support vector machines. International Journal of Computer Theory and Engineering, 1(1):1793–8201, 2009.