Intelligent Weather Forecasting Using Hybrid CNN–VGG Feature Learning with Ensemble Models (Random Forest and AdaBoost) for Enhanced Spatiotemporal Prediction

Dr. Amanullah M[1], Dr. Basant Kumar [2]
[1] Postdoctoral scholar, Lincoln University College, Malaysia
[2] Modern College of Business and Science, Oman;
Email ID: [1]amanhaniya12@gmail.com, [2]basant@mcbs.edu.om

Abstract

The numerous weather forecast components in climatology dynamics and the quickly evolving technology of intelligent weather data processing complicate the situation of Weather prediction. To overcome the problem an advanced hybrid Machine Learning framework which is based on Deep Feature extraction technique with CNN-VGG and Ensemble learning technique is proposed in this paper to enhance the accuracy of the weather prediction system by adopting Random Forest (RF) and AdaBoost algorithms. Preprocessing techniques such as k-Nearest Neighbor (k-NN) imputation, Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) are also incorporated in the framework to further enhance the quality of the data and reduce the dimensionality of the data.

Experimental evaluation shows that our hybrid CNN–VGG feature extractor along with AdaBoost and Random Forest approach is very effective as compared to the baseline machine learning models. The performance of ensemble strategy involved in the generalization performance is improved, prediction error (RMSE, MAE) is decreased, robustness with the extreme weather variability is boosted. In conclusion, the results of the experimental evaluation confirm the effectiveness and intelligence of deep Spatial feature learning and the Adaptive Ensemble methods in predicting weather in a context of intelligent and scalable weather forecasting system.

Keywords: Weather Forecasting, CNN–VGG, Random Forest, AdaBoost, Ensemble Learning, Spatiotemporal Prediction

1. Introduction

Weather prediction is a basic science and at the same time it has got a variety of applications in agriculture, disaster management, assessment of climate risk etc., With increasingly frequent extreme weather events in climate, the quantity and expectations of localized, real time, ultra-accurate forecasting systems have significantly increased [1] [2]. The description of the fine-grained fluctuations of the weather with a deterministic model is extremely difficult due to the chaotic behavior of the atmosphere [3] [4].

A deep layered network (VGG) with small size of convolutional filter has proven itself to the problem of image-based learning tasks. With the heterogeneous climatic data the deep learning could face the issue of overfitting and instability, as well [5].

To further increase the robust and robust its generalization ability, an ensemble learning framework can be used to train a deep learning feature extractor including multi-models such as Random Forest & AdaBoost. Random Forest appears to be more stable and try to create multiple decision trees, AdaBoost tries to be 'adaptive' in learning by giving high weightages to the hard elements. To improve the prediction capability, a highly hybrid approach CNN–VGG + ensemble models are used [6] [7].

In the present research, an intelligent weather forecasting system is presented by using CNN-VGG based spatial feature extraction in addition to the ensemble learning models such as Random Forest Classifier, AdaBoost Classifier. Enhancement of the framework is carried out with the help of advanced pre-processing approaches like k-NN imputations, PCA based dimensionality reduction and RFE feature selection [8] [9]. The ultimate aim is to develop a high level accurate, scalable, and computation efficient forecasting model that would be able to handle complex spatial and temporal data in the meteorological world.

2. Methodology

The proposed system is developed in a multi-stage pipeline so as to overcome above limitations – processing of raw meteorological data, meaningful spatial representation, and using intelligent ensemble based prediction techniques. Five major phases are involved in mapping the whole space, namely: data collection, data preprocessing, feature extraction using CNN-VGG, ensemble learning and performance analysis. Clearly all these steps are important to develop an accurate, stable and transferable climate forecasting model.

2.1 Data Collection

Ground-based weather stations and IoT sensors, which measure weather temperature, humidity, wind speed, rainfall intensity and atmospheric pressure, on the other hand, provide very accurate data at their particular location.

2.2 Data Preprocessing

The first preprocessing takes account of imputing the missing values by K-nearest neighbor (k-NN). The K-NN imputation method will retain the local data structure, also have the realistic trend which would benefit data integrity in meteorology, compared to other simple statistical imputation method such as mean substitution and median substitutions.

For further optimization of the data set, Principal Component Analysis (PCA) is used for dimensionality reduction. PCA transforms the original feature space into a reduced number of orthogonal features which retain a majority of the information about the data set. Recursive Feature Elimination (RFE) is also employed in order to reduce the dimensions while selecting the features.

2.3 CNN–VGG Feature Extraction

The main block of the proposed system is encoder-decoder hybrid Convolutional Neural Network (CNN) that is specific to spatial features of the meteorological data. All the CNN architectures used are designed with the VGG network, which consists of multiple stacks of conv layers and small filters (3×3) are used. Pooling layers are placed between the Convolutional layers to reduce the spatial dimensions, while keeping the relevant features. The down sampling also reduces the number of network parameters, which can further help reduce the computational complexity of the network, and can also avoid overfitting problems.

2.4 Ensemble Learning Framework

To improve the class prediction of the system, ensemble learning framework has been integrated in the system which makes the system confident on prediction.

Random Forest (RF)

Random Forest is a great and powerful Ensemble method, which will create multiple Decision Trees from Bootstrap samples of training data. Random Forest, in particular, is effective in capturing the nonlinear relation between the meteorological variables, in the context of weather forecasting.

AdaBoost

An ensemble method which attempts to enhance the predictive ability is to follow the hard to predict examples — AdaBoost or Adaptive Boosting does that. The mechanism of adaptive weighting is the main advantage of AdaBoost; it finds its way to rectifying the mistakes produced by the former models in order to progressively improve itself.

2.5 Model Training and Evaluation

The entire data set is divided in three sets: Train, Validation and Test. Three widely used regression metrics are used to measure our forecasting accuracy: the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Coefficient of Determination (R^2).

3. Results and Discussion

The ensemble of the proposed hybrid CNN–VGG + structure was tested. The developed system was compared with various baseline systems namely Support Vector Regression (SVR), Decision Tree, Random Forest (RF) alone and adaBoost. Evaluation was done on various important weather parameters such as temperature and rain, based on standard regression statistics including Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

3.1 Model Performance Comparison

It can be obtained from the comparative performance of all models which is summarized in the Table 1 showing the RMSE and MAE values for temperature and rainfall prediction task of all models.

Table 1: Comparative Performance of Forecasting Models

Model	RMSE (Temp)	MAE (Temp)	RMSE (Rain)	MAE (Rain)
SVR	2.90	2.15	4.85	3.75
Decision Tree	2.60	1.95	4.40	3.50
Random Forest	2.30	1.80	4.10	3.20
AdaBoost	2.10	1.65	3.90	3.00
CNN–VGG + RF	1.55	1.18	3.10	2.40
CNN–VGG + AdaBoost (Hybrid Ensemble)	**1.30**	**1.02**	**2.75**	**2.10**

The findings show a clear trend towards decreasing from the value of RMSE and the MAE with increasing the model complexity. But an SVR and Decision Tree model has less accuracy to model the non-linear relations between spatio-temporal variable. Random Forest adds all the success of classifiers and attempts to make prediction using ensemble averaging method whereas AdaBoost corrects the error that any pathway that misclassified or extremely predicted example will undergo.

However, the significant improvement can be noticed in CNN-VGG + ensemble configurations. The hybrid CNN–VGG + RF model shows significant decrease in prediction error as well, while the CNN–VGG + AdaBoost model shows the overall best performance as it gives lowest RMSE and MAE values at both the temperature and rainfall prediction task. This shows that very high space feature extraction and adaptive ensemble learning has a significant contribution in the enhancement of predictive accuracy.

3.2 R^2 Performance Analysis

To look at the effectiveness of the model further several meteorological parameters were calculated in terms of Coefficient of Determination (R^2). The proportion of variance in the observations that is accounted for by the variance of the predicted values, closer to 1 are good.

Table 2: R^2 Scores for Weather Prediction Models (No. 2)

Parameter	Random Forest	AdaBoost	Proposed Hybrid CNN–VGG + Ensemble
Temperature	0.90	0.92	**0.96**
Rainfall	0.84	0.87	**0.91**
Humidity	0.86	0.89	**0.93**
Wind Speed	0.82	0.85	**0.90**

The R2 results indicate that, proposed hybrid system consistently get the best results for all the meteorological data. This implies that the model is quite good with respect to the representation of the complex relation of the atmospheric parameters. For temporal patterns the prediction of temperature is more accurate, because these are stable, whereas rainfall patterns are irregular and very nonlinear.

Finally, the Generalization ability of the hybrid CNN-VGG + ensemble model is better and can be reliably applied to real world forecasting application.

3.3 Key Findings

The results of the experimental evaluation show some key insights into the performance of the proposed forecasting system. On the other hand, CNN–VGG has a strong ability to extract spatial features, which can well utilize the spatial structure of the atmospheric space, such as the formation and the cloud structure of storms, the spatial distribution of precipitations, etc. These spatial features have significant contribution to the overall prediction accuracy level.

The second is the adaptive learning of AdaBoost that will give more weight to the samples that are difficult to predict and consequently will improve the sensitivity to extreme weather changes. However, ensemble averaging the variances, Random Forest gives a more stable and robust model which can be used to make prediction from different sets of data.

Third, efficiency of the model can be improved by removing the redundant information by using Preprocessing techniques such as RFE Feature Selector and PCA dimensionality reducing. The advantage of these steps is that the learning will focus only upon the most relevant meteorological features.

Furthermore, the Ensemble based runtime achieves the desired effect of reducing the prediction errors by merging multiple supplementary learning strategies which, while in place, render stable and accurate forecasts. However, rainfall is one of the most challenging prediction to make, for it is so volatile and stochastic. Future inclusion of attention model in deep learning and real-time adaption learning algorithms could perhaps enhance a deep learning model's prediction of extreme weather events.

4. Conclusion

In this research, the idea was to present an intelligent hybrid weather forecasting framework in which deep spatial features were extracted using CNN–VGG and later on trained to the ensemble learning models which includes Random Forest and AdaBoost. The CNN-VGG based architecture has been successful in the extraction of complex spatial patterns for meteorological images and ensemble methods have been demonstrated to be good in generalization and to reduce the prediction variance.

The outcomes of the experiments show a huge improvement in the performance of the proposed hybrid CNN – VGG + AdaBoost model over the base models because the proposed model resulted in the lowest values of RMSE, MAE and R2 for all the weather parameters it is estimated for. The model system demonstrates high quality in capturing the response of the atmosphere, which is highly nonlinear, and improving the stability of the simulated response, in the various phase states of climate.

Future study may focus on integrating attention mechanism using transformer with real time adaptive learning system that can be used for further improvement of extreme weather prediction.

References

1. R. Singh and A. Yadav, "CNN-LSTM hybrid model for rainfall forecasting," *IEEE Access*, vol. 11, pp. 46512–46520, 2023.
2. Y. Zhou, X. Han, and Z. Wu, "Bi-LSTM for wind speed prediction," *IEEE Transactions on Sustainable Energy*, vol. 14, no. 2, pp. 567–575, 2022.
3. D. Patel, K. Mehta, and B. Shah, "Stacked ensemble models for humidity prediction," *IEEE Sensors Journal*, vol. 23, no. 3, pp. 987–994, 2022.
4. M. Rahman and M. Islam, "Hybrid ARIMA-LSTM model for seasonal weather forecasting," *IEEE Internet of Things Journal*, vol. 10, no. 6, pp. 5563–5571, 2023.
5. J. Li and C. Deng, "Feature selection for weather forecasting using RFE," *IEEE Transactions on Big Data*, vol. 9, no. 2, pp. 330–341, 2022.
6. T. Alam, A. Sayed, and M. Raza, "Handling missing weather data using k-NN imputation," *IEEE Access*, vol. 11, pp. 40356–40365, 2023.
7. A. Martins, J. Costa, and L. Silva, "PCA-based data preprocessing for meteorological datasets," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–10, 2023.
8. S. Jain and M. Thakur, "A comparative study of time series models for weather prediction," *IEEE Access*, vol. 11, pp. 11935–11943, 2023.
9. A. Kumar, R. Shukla, and V. Tiwari, "Spatiotemporal precipitation forecasting using ConvLSTM," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2022.