A Study on Rainfall Prediction based on Meteorological Time Series

KangWoon Hong
Smart ICT Convergence Research Department
Electronics and Telecommunications Research Institute
Daejeon 34129, Korea
gwhong@etri.re.kr

Taegyu Kang
Smart ICT Convergence Research Department
Electronics and Telecommunications Research Institute
Daejeon 34129, Korea
tgkang@etri.re.kr

***Abstract*—This study aims to present the results of the research and development project on the urban inundation prediction technology during the heavy rain period. In this study, the results of rainfall prediction using heterogeneous weather data and machine learning are presented. In the predictive analysis of univariate time series data, it was confirmed that the CNN-LSTM model showed the best performance among several deep neural network models. In the predictive analysis of multivariate time series data, it was confirmed that the ConvLSTM model showed the best performance among several deep neural network models.**

Keywords—rainfall prediction; deep neural network; meteorological data

I. INTRODUCTION

Rainfall information can be used to analyze damage caused by rain and shows different aspects of rainfall depending on the region, so it is important to provide high-resolution rainfall information for accurate damage analysis. In the existing meteorological field, high-performance computing resources are used to analyze and predict meteorological phenomena, but recently, studies to understand meteorological phenomena through data analysis based on new insights using meteorological big data and machine learning technology have been conducted. This study aims to present the results of research and development carried out as part of research and development project on the urban inundation prediction technology during heavy rain period.

II. RELATED WORKS

A. *Existing research*

Traditionally, in the meteorological field, we predict rainfall using a numerical model or a radar-based model. As machine learning was gradually applied to the meteorological field, a regression model was utilized that enables prediction with an empirical model, easy to understand and interpret, and has an advantage in explainability of prediction. In recent years, there is a trend to apply deep learning technology, which is used in various fields, to the field of weather prediction. This trend can also be seen as an evolution from structured data-based analysis to unstructured data-based analysis.

Regression analysis on time-series data is a representative machine learning technique, and predicts using the relationship between variables derived through regression analysis. Depending on the characteristics of learning speed and large-scale data processing, XGBoost and LightGBM are also highly utilized. Rainfall prediction technology based on deep learning predicts rainfall by learning radar images with deep neural network models such as FC-LSTM, ConvLSTM, Trajectory GRU, and RainNet. More specifically, CNN is used to search and classify radar images, and RNN is used to analyze time series data.

B. *Proposal of prediction models*

We propose models that predict rainfall at locations installed in the region using heterogeneous meteorological data.

In this study, as learning data, meteorological data that can be obtained from the open MET data portal in real time and meteorological data from measuring stations installed by local governments are collected and used for rainfall prediction. Rainfall information can be used to analyze inundation damage. It shows different aspects of rainfall phenomena depending on the region, and requires rainfall information close to the amount of inflow into the drainage system and high-resolution rainfall information. However, even in urban areas, the weather measurement infrastructure operated by public institutions is insufficient to obtain high-resolution rainfall information. Although radar-based measurement stations are operated, it is not accurate to predict the amount of inflow into the drainage system since the radar-based precipitation data is calculated data. Therefore, we supplemented the accuracy of the rainfall prediction method by using the measured rainfall collected from not only local governments, but also nearby their measurement stations.

In this study, multivariate time-series prediction models are used based on the results of reviewing the existing rainfall prediction studies, the correlation matrix based on the structured data. Since it is known as a relatively superior technique compared to other techniques, the above models are utilized. However, the collected data is composed of various variables, and separate variables are newly added to identify seasonal characteristics. Therefore, the model is constructed by determining the input variables of the model through the process of additionally considering significant correlation.

III. DATA ANALYSIS

A. Dataset

The public meteorological data, that can be obtained in real time from the open MET data portal and the local meteorological data from the measuring stations installed by local governments are corresponding to the selected region and period. Public meteorological data collected at public measuring stations (17 locations) include temperature, precipitation, wind speed, wind direction, humidity, vapor pressure, dew point temperature, local air pressure, sea level air pressure, sunlight, solar radiation, total cloud volume, and ground temperature, latitude and longitude information, and local meteorological data consist of temperature, precipitation, wind speed, wind direction, humidity, and latitude and longitude information collected by local measuring stations (39 locations).

B. Data collection

As the input to the prediction model, we used 1-year (from 1 a.m. on January 1, 2020 to 12 a.m. on January 1, 2021) measurement data collected from the selected local government's own measurement station and the public measurement stations (3 ASOS/AWS stations of the Korea Meteorological Administration) adjacent to the local measurement station on the basis of the haversine distance.

C. Preprocessing

The preprocessing is a step of processing data to be suitable for rainfall prediction in the machine learning, and processing outliers or missing values, feature selection, and data preparation are performed.

In order to avoid the case that the meteorological data contains outlier or missing data at an excessive rate (more than 10%), we perform data search on public and local meteorological data, then select public measuring stations adjacent to the local measurement station.

When predicting multivariate time-series data, correlation analysis between variables is performed after adding new variables such as monthly precipitation and seasonal precipitation to identify seasonal characteristics. As a variable of significant correlation, the precipitation variable at each measurement station is selected and used to construct various models. According to the result of correlation analysis that there is no variable having a high correlation except for rainfall, we select only precipitation variables as features among meteorological data from a local measurement station and its adjacent public measurement stations in multivariate time series prediction.

In this study, we perform both univariate data prediction and multivariate data prediction. Therefore, it is necessary to prepare data for training and prediction in the data preparation stage. In univariate data prediction, only time-series rainfall data is used as both the independent variable and the dependent variable, whereas in multivariate data prediction, the value of the dependent variable, the amount of precipitation, is predicted by using various types of variables constituting meteorological data including precipitation. In addition, data preparation should consider not only the type of data prediction, but also the type of model used for prediction. In other words, in order to apply to the CNN, LSTM and models derived from these models, the input data is specifically transformed from 1D to 2D and used for learning and prediction.

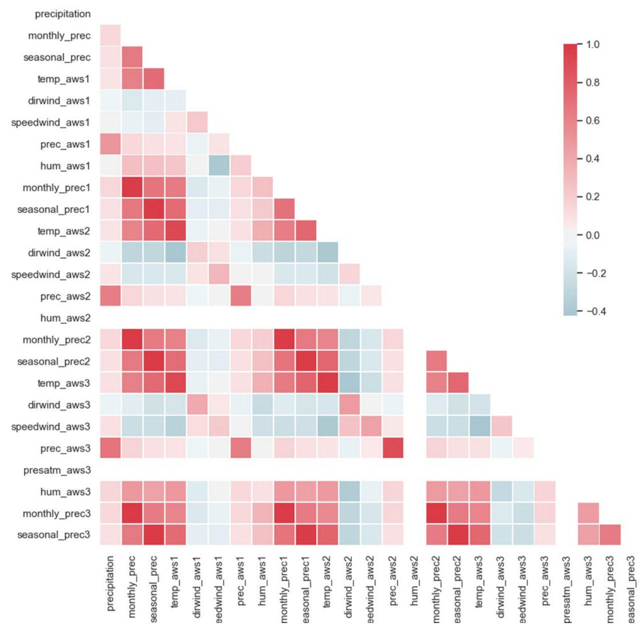

Fig. 1. *Correlation Matrix*

IV. PREDICTION MODEL

A. Univariate Time Series Prediction Model

For the prediction of univariate time series data, we used precipitation data among the weather data collected at the selected local measurement station of the local government. As detailed models, MLP, CNN, LSTM, and other derivative network models are used, and the configuration of these models is as follows.

B. Multivariate Time Series Prediction Model

For the prediction of multivariate time series data, we used meteorological data collected from the selected local government's measurement stations and public weather data collected from adjacent public measurement stations at the selected local measurement station of the local government.

C. Model configuration

As detailed models, MLP, CNN, LSTM, and other derivative network models are used, and the configuration of these models is as follows.

```
model = Sequential()
model.add(Dense(500, activation='relu', input_dim=24))
model.add(Dense(1))
model.compile(loss='mse', optimizer='adam')
model.fit(x, y, epochs=100, batch_size=100, verbose=0)
```

Fig. 2. *MLP model*

```
model = Sequential()
model.add(TimeDistributed(Conv1D(filters=64, kernel_size=3,
activation='relu', input_shape=(None,12,1))))
model.add(TimeDistributed(Conv1D(filters=64, kernel_size=3,
activation='relu')))
model.add(TimeDistributed(MaxPooling1D(pool_size=2)))
model.add(TimeDistributed(Flatten()))
model.add(LSTM(100, activation='relu'))
model.add(Dense(100, activation='relu'))
model.add(Dense(1))
model.compile(loss='mse', optimizer='adam')
model.fit(x, y, epochs=200, batch_size=100, verbose=0)
```

Fig. 3. *CNN-LSTM model*

```
model = Sequential()
model.add(ConvLSTM2D(filters=256, kernel_size=(1,3), activation='relu',
input_shape=(3, 1, 12, 1)))
model.add(Flatten())
model.add(Dense(200, activation='relu'))
model.add(Dense(1))
model.compile(loss='mse', optimizer='adam')
model.fit(x, y, epochs=200, batch_size=100, verbose=0)
```

Fig. 4. *ConvLSTM model*

D. Damage Prediction Model

Accurate prediction of rainfall and inundation is also necessary, but efforts to solve and reduce inundation damage by predicting the various types of damage expected due to inundation are also important. Although appropriate weather forecasts are announced in a timely manner, there are cases where they do not always lead to good results. This phenomenon comes from differences in understanding of weather forecasts, the increased vulnerability to disasters due to changes in social factors, and the spread of the effects of meteorological phenomena on the entire industry. Accordingly, the Korea Meteorological Administration provides an impact forecast separately from the weather forecast, and it is necessary to predict the damage caused by direct or indirect effects even in the event of a disaster such as flooding, solve the disaster problem, and utilize it for reduction.

For this, we generate the trend/scenario targeting the risk factors derived as damage types such as floods, floating population, vehicles, traffic jams, etc., and a method of predicting flood damage by linking the trend/scenario with the predicted rainfall is presented.

V. PERFORMANCE

In comparison of performance, we used MLP, CNN, LSTM, CNN-LSTM, and ConvLSTM models to learn weather data collected from public and local measurement stations for every hour of 2020 and to create a predictive model.

Table-1 shows the accuracy evaluation results of the univariate and multivariate time series prediction model. According to the evaluation results, the CNN-LSTM model and the ConvLSTM model show the best performance, respectively.

VI. CONCLUSION AND FUTURE RESEARCH

In this study, the results of rainfall prediction using heterogeneous weather data and machine learning are presented. In the predictive analysis of univariate time series data, the CNN-LSTM model showed the best performance among several deep neural network models. In the predictive analysis of multivariate time series data, the ConvLSTM model showed the best performance among several deep neural network models.

We predict rainfall using structured data. In recent years, unstructured data is used in the field of rainfall prediction, so the prediction accuracy of this study may be poor. However, this study can find results in suggesting whether features included in meteorological data other than rainfall should be considered and which types of deep neural networks should be applied.

In the future, it is expected that the prediction accuracy can be improved by performing additional experiments such as using long-term weather data or using radar images as additional features.

TABLE I. ACCURACY EVALUATION

DNN	Univariate time series		Multivariate time series	
	RMSE	Std dev	RMSE	Std dev
MLP	0.043	0.026	0.442	0.021
CNN	0.029	0.02	0.581	0.053
LSTM	0.398	0.657	0.454	0.284
CNN-LSTM	0.012	0.009	0.452	0.026
ConvLSTM	0.036	0.026	0.386	0.036

ACKNOWLEDGEMENT

This research was conducted with the support of Data Driven Solutions(DDS) Convergence Research Program funded by the National Research Council of Science and Technology "Development of solutions for region issues based on public data using AI technology -Focused on the actual proof research for realizing safe and reliable society-"(1711101951).

REFERENCES

[1] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description”, CoRR, abs/1411.4389, 2014

[2] Shi, X. et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. Adv. Neural Inf. Process. Syst. 28, 2015, pp. 802–810

[3] Shi, X. et al., “Deep learning for precipitation nowcasting: a benchmark and a new model”, Adv. Neural. Inf. Process. Syst. 30, 2017, pp. 5617–5627

[4] Ayzel, G., Scheffer, T., and Heistermann, M., “RainNet v1.0: A convolutional neural network for radar-based precipitation nowcasting”, Geoscientific Model Development, Vol. 13, 2020, pp. 2631-2644

[5] Brownlee, J., “Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python”, Machine Learning Mastery, 2018