Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 16, No. 1, March 2016, pp. 1-12
http://dx.doi.org/10.5391/IJFIS.2016.16.1.1

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

Constructing Efficient Regional Hazardous Weather Prediction Models through Big Data Analysis

Jaedong Lee and Jee-Hyong Lee
Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, Korea

Abstract

In this paper, we propose an approach that efficiently builds regional hazardous weather prediction models based on past weather data. Doing so requires finding the proper weather attributes that strongly affect hazardous weather for each region, and that requires a large number of experiments to build and test models with different attribute combinations for each kind of hazardous weather in each region. Using our proposed method, we reduce the number of experiments needed to find the correct weather attributes. Compared to the traditional method, our method decreases the number of experiments by about 45%, and the average prediction accuracy for all hazardous weather conditions and regions is 79.61%, which can help forecasters predict hazardous weather. The Korea Meteorological Administration currently uses the prediction models given in this paper.

Keywords: Attribute selection, Big data, Hazardous weather, Regional prediction, Support vector machine

1. Introduction

The accurate analysis and prediction of hazardous weather are closely related to real life and can be used in various areas. Therefore, the creation of hazardous weather forecasting systems and the related technologies have always been in demand [1-5]. However, it is difficult to create an accurate hazardous weather forecasting system because the occurrence of hazardous weather is influenced by regional characteristics, so similar meteorological conditions can produce dramatically different weather on the ground in different places. Thus, to create an accurate hazardous weather prediction system, separate prediction models need to be made for each region.

Building regional hazardous weather prediction models requires selection of the weather attributes that strongly affect hazardous weather for each region. Many researchers have used data mining techniques to create hazardous weather prediction models based on past weather data [6-19]. Most researchers simply used all available weather attributes without attribute selection or attributes selected by experts [1, 3, 10, 11].

But using all available weather attributes has several disadvantages, most notably computational cost and system performance [6]. For example, the meteorological data of the European Centre for Medium-Range Weather Forecast (ECMWF) contain 254 weather attributes. Using all the available weather attributes would require a large computational cost to build a hazardous weather model for even one region [20]. Predicting 7 types of hazardous weather in

Received: Feb. 27, 2016
Revised : Mar. 23, 2016
Accepted: Mar. 24, 2016

Correspondence to: Jee-Hyong Lee
(john@skku.edu)

16 regions would require 112 models (7×16), which represents a huge computational cost. Also, using all available weather attributes to build prediction models is ineffective. Because the interactions among weather attributes are complex, the prediction performance with all available attributes might not be better than the performance with only some of the available attributes. Considering weather attributes unrelated to hazardous weather might not improve or could even deteriorate the performance of the prediction models [21-23].

On the other hand, using weather attributes chosen by experts to make regional hazardous weather prediction models will probably show the best performance and require less computational cost. However, using experts to select the correct weather attributes for 112 models is still a problem because it requires a tremendous amount of expert knowledge.

Therefore, in this paper, we make regional hazardous weather prediction models for each hazardous weather condition in each region while minimizing the intervention of the experts. Experts choose only 5 weather attributes and 3 isobaric surfaces that can affect hazardous weather conditions, which results in 15 attributes. Using those 15 attributes, we find the optimal combination of attributes to build regional hazardous weather prediction models.

It remains difficult to select the most effective attributes from 15 candidates. If we limit our model to 3 attributes to minimize the computational cost, we still have 575 candidate models to consider to find the optimal attribute combination for each hazardous weather in each region: the number of single attributes (15) plus the number of 2-weather-attribute combinations (105) plus the number of 3-weather-attribute combinations (455). To make prediction models for 7 types of hazardous weather conditions (heavy rainfall, heat wave, strong winds, wind waves, heavy snowfall, cold wave, and lightning) for 16 regions (Seoul, Incheon, Gangneung, Chuncheon, Chungju, Daejeon, Seosan, Daegu, Andong, Busan, Ulsan, Jeonju, Gwangju, Yeosu, Mokpo, and Jeju). Thus, we would need to conduct 64,400 (16 regions × 575 candidates × 7 hazardous weather conditions) experiments, and that is inefficient. Therefore, to find the best weather attributes for each region and each type of hazardous weather, we adopt a modified top-down attribute selection method that allows us to reduce the number of experiments.

We also need to consider the ratio of hazardous weather conditions and non-hazardous weather conditions when constructing training and test data for the prediction models. Non-hazardous weather conditions naturally outnumber hazardous weather conditions. However, if we use training data that reflect the true ratio of non-hazardous to hazardous weather conditions, the prediction models will be over-fitted to non-hazardous weather conditions, which means they will deem all weather conditions "non-hazardous" [24]. Consequently, we maintain an equal ratio of hazardous and non-hazardous weather conditions when constructing the training and test data sets.

Finally, we can make efficient regional hazardous weather prediction models that minimize the intervention of experts by using 10-year accumulated weather data. Our models are currently used in the Korea Meteorological Administration to aid forecasters in making decisions about potentially hazardous weather.

The rest of this paper is organized as follows. Section 2 briefly describes previous research about weather prediction using machine learning methods and its weaknesses compared with our proposed method and provides a brief explanation of the support vector machine (SVM) technique we used to generate the prediction models in this paper. Section 3 describes the weather data, hazardous weather, and regions we used in this paper. Section 4 describes the details of our proposed method, hazardous weather prediction using SVM. Section 5 shows our experimental results. Section 6 summarizes the paper and offers suggestions for future work.

2. Related Work

2.1 Previous Research

The use of machine learning methods to predict the weather has been studied in various ways. Romani et al. [12] used time-series weather data to extract a pattern and detected abnormal weather. In that study, weather data were generated and observed every week in terabytes. Because the authors used every weather attribute to generate the prediction model, it had a high computational cost that is inefficient. Efficiently generating a regional hazardous weather prediction system requires studies on the selection of weather attributes that well represent specific types of hazardous weather for specific regions.

Olaiya and Adeyemo [8] used a decision tree and artificial neural network method to predict daily maximum and minimum temperature, rainfall, evaporation, and wind speed. He conducted experiments that predicted the weather of a certain region and compared his data mining method with the weather forecasting numerical models that are widely used in the meteorological centers of many countries. Because the data mining model is generated using all observed weather attributes, its

calculation costs are high, and its performance is not guaranteed. If a prediction model is generated using efficient weather attributes for the region, the computational cost can be reduced, and the prediction accuracy can be increased.

Radhika and Shashi [10] conducted a study that used SVM to predict the time series atmospheric temperature. They compared predictions of the maximum temperature for the following day from the SVM and artificial neural network methods. The SVM method showed better results than the artificial neural network, but they built their prediction model using only the daily maximum temperature as input, which might not reflect the optimal weather attributes for even a temperature prediction model, much less a hazardous weather prediction model.

Nayak et al. [25] used an enhanced approach to the artificial neural network method to predict the daily maximum temperature. Through a comparison of results from other machine learning methods, they proved that their method offered higher performance. They used 8 weather attributes selected by experts, including temperature, wind speed, and relative humidity. However, they did not analyze how each weather attribute affected the prediction of daily temperature. If they had analyzed each weather attribute and used those results as the input for their prediction model instead of just using all 8 weather attributes, their prediction model would be more efficient in forecasting the daily maximum temperature.

Nikam and Meshram [7] used data mining techniques for modeling rainfall prediction. Out of 36 weather attributes they used 7 attributes as input of model with the decision that the other weather attributes are less relevant. They also did not analyze the information amount of each weather attribute to identify regional characteristics.

As just described, data mining methods have been used in different ways to conduct studies on climate forecasting, but few studies have been associated with regional climate forecasting. In particular, practically no studies have sought the weather attributes needed to make a hazardous weather prediction system that considers regional characteristics. Hazardous weather can affect different regions differently even if they share similar weather conditions. Therefore, a consideration of regional characteristics is important to select the right weather attributes when making a regional hazardous weather prediction model. For this paper, we asked experts to delineate several regions according to the importance of the region and frequency of each type of hazardous weather. We also conducted experiments to determine whether a certain climate affected a particular type of hazardous weather in a region. We used SVM, described in the following sub-section, to build the prediction model.

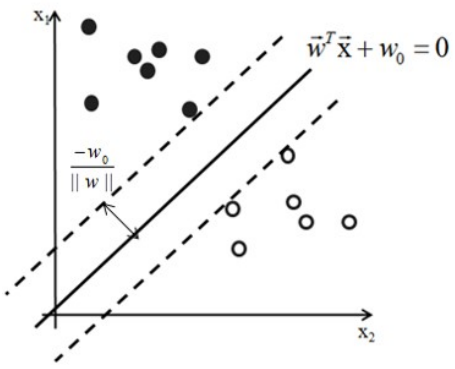


Figure 1. How support vector machine works.

2.2 Support Vector Machine

SVM is known to outperform other classification techniques. SVM sets a hyperplane that fully classifies the training set containing two classes with different values.

Figure 1 shows the classification by drawing in a hyperplane between two data with different properties. Black dots and white dots represent the data with two different properties, and a hyperplane is set between the different data. Here in these two data sets, the nearest point from the hyperplane is called the support vector, and the distance between the support vector and the hyperplane is called the margin. It is best to maximize the distance between the hyperplane and support vector for the best classification.

It is almost impossible to separate the data linearly in most cases, but those problems can be solved using a kernel. A kernel maps the low-dimensional input data into a high dimensional space to solve the nonlinearity problem. SVM seeks a linear separating hyperplane with the maximal margin in this higher dimensional space. The kernel function is defined as Eq. (1) shown below.

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j). \tag{1}$$

The so-called *kernel method* solves the nonlinearity problem by linearizing the data through high dimensional mapping, and that solves the problem of increasing computational complexity. In this paper, we configure the SVM with each attribute of each isobaric surface. The SVM predicts weather data as, for example, heavy rain or not heavy rain, and judges how much effect each attribute will have in predicting heavy rain or not

heavy rain. We implemented the binary classification in every experiment using SVM through the SVM Light tool, and we used the radial basis function kernel.

3. Weather Data, Hazardous Weather, and Region Description

In this section, we describe the characteristics of the weather data, criteria of hazardous weather, and regions we used, along with the weather attributes, isobaric surfaces, and ranges of weather data. We use the same criteria of hazardous weather that are used for a special weather statement from the Korea Meteorological Administration. An expert selected several regions where hazardous weather predictions are especially important. In each region, hazardous weather not only occurs frequently but also has a social and economic influence.

3.1 Weather Data

We use UM N512 meteorological data generated using ECMWF 1.125 degree data. The data consist of 254 weather attributes and 7 isobaric surfaces: 200, 300, 500, 700, 850, 925, and 1000 hPa. UM N512 data consist of 228×257 grids representing the Eastern Asia, for a total of 410,172 (228×257×7) grids. Each grid includes 254 weather attribute values measured in the corresponding isobaric surface and spot. The total number of different values in a weather map is 102,812,850 (228×257×7×254). The data are produced every 6 hours (00:00, 06:00, 12:00, 18:00 UTC).

A prediction model can be built based on accumulated values from the past. Because the UM N512 data set is huge with a large number of attributes, it is inefficient to use all the attributes, and most attributes do not affect meteorological analysis anyway. Therefore, in this study, we use five attributes chosen by experts as empirically known to be effective in the prediction of hazardous weather. The five attributes are *Height* (Z), *Humidity* (R), *Temperature* (T), *Uwind* (U), and *Vwind* (V). The meaning of each attribute is shown in Table 1.

In addition, using all isobaric surfaces to generate a prediction model creates an issue of low accuracy. We use only the isobaric surfaces of 500, 700, and 850 hPa and exclude those of 200, 300, 925, and 1000 hPa. The isobaric surfaces of 1000 and 925 hPa are too close to the ground and can produce unstable data. The isobaric surfaces of 200 and 300 hPa are too far from the ground, so they show little effect on weather prediction.

Table 1. The definition of each weather attribute

Attributes	Definition
Height	Vertical coordinate referenced to earth's mean sea level
Humidity	Amount of water vapor in a mixture of air and water vapor
Temperature	Temperature of the air
Uwind	East-west component of the wind
Vwind	North-south component of the wind

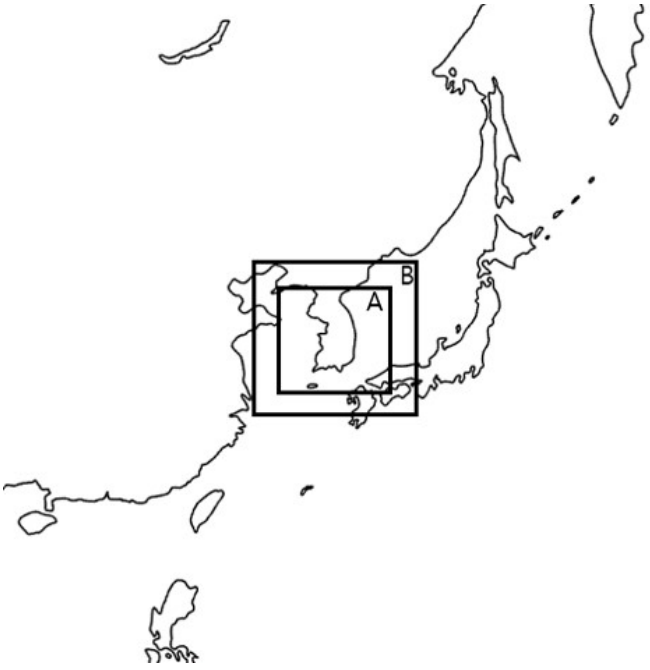


Figure 2. Range of area used in experiments: square A, 3030 grid; square B, 4040 grid.

3.2 Hazardous Weather and Region

In the UM N512 data, the Eastern Asian region is expressed using a 228×257 grid. Forecasting hazardous weather on the Korean Peninsula using all the weather data from the entire region is still cost-prohibitive. Moreover, including unnecessary regional weather data will adversely affect the predictions. For a more efficient experiment to predict hazardous weather on the Korean Peninsula, we limited our experimental data to the areas surrounding the Peninsula. We considered the movement of air to determine the area. We use an area of 30×30, square A in Figure 2, for the 6 hour forecast and an area of 40×40, square B in Figure 2, for the 24 hour forecast.

Prediction of hazardous weather shows regional peculiarities even under similar meteorological conditions, which makes it difficult to accurately predict hazardous weather for all regions using a single model. Therefore, characteristics that affect a given region's hazardous weather must be identified and used to generate a prediction model for each region and each hazardous weather type.

For this paper, we choose several metropolitan areas on the Korean Peninsula where hazardous weather occurs frequently and build hazardous weather prediction models for each region. We choose metropolitan areas that need a hazardous weather prediction model by taking into account the importance of the regions. Regions with many people experience more bad influence from hazardous weather than other regions. We also consider the number of hazardous weather occurrences and their frequency. We select 16 metropolitan areas chosen by experts to predict heavy rainfall, lightning, heat wave, and heavy snowfall: Seoul, Incheon, Gangneung, Chuncheon, Chungju, Daejeon, Seosan, Daegu, Andong, Busan, Ulsan, Jeonju, Gwangju, Yeosu, Mokpo, and Jeju. We select 14 metropolitan areas, excluding Mokpo and Jeju, to predict cold waves; 4 metropolitan areas, Busan, Yeosu, Mokpo, and Jeju, to predict strong wind; and 4 regions, Deokjeokdo, Chilbaldo, Geomundo, and Geojedo for wind waves.

We use the following criteria for hazardous weather in this study. They are the same as the criteria used for special weather statements from the Korea Meteorological Administration.

- Heavy rainfall: 60 mm of accumulated rainfall or more over a 6 hour period
- Heat wave: Daily maximum temperature of 33°C or more
- Strong winds: Wind speed of 14 m/s or more
- Wind waves: Wave height of 3 m or higher
- Heavy snowfall: 5 cm or more of accumulated snow over a 24 hour period
- Cold wave: A drop of 10° or more from the previous day
- Lightning: Occurrence

We use 6 hour prediction systems for heavy rainfall, strong winds, wind waves, heavy snowfall, and lightning. For heat and cold waves, we use a 24 hour prediction system because we need daily information to determine whether the hazardous weather has occurred.

4. Prediction Model Construction with Modified Top-Down Method

In this section, we describe how to select attributes and compose the training data set to efficiently build prediction models using SVM. We modify the top-down attribute selection method to choose proper weather attributes with fewer computational resources than required by the traditional method. To build SVM models, we down-sample non-hazardous weather data to be equal to hazardous weather data in occurrence for the training data sets to prevent the SVM models from being over-fitted. Finally, we build optimal regional hazardous weather prediction models.

4.1 Modified Top-Down Weather Attributes Selection Method

In this paper, we use the five weather attributes and three isobaric surfaces selected by experts for effective hazardous weather prediction. Thus, overall we have 15 attributes (5 weather attributes $\times$ 3 isobaric surfaces) that can be used to generate each prediction model, which is still an overwhelming number of possible models, as explained above. Therefore, we use a modified top-down attribute selection method to choose the best attributes for a prediction model with high prediction accuracy.

For n single attributes and k attributes combined at maximum, the steps to build prediction models for a given type of hazardous weather in a given region with the modified top-down method are shown in Algorithm 1.

In this paper, we combine a maximum of 3 attributes ($k = 3$) to reduce computational cost. To determine the best-performing attributes for a given type of hazardous weather in a specific region, we first make 15 prediction models with 15 attributes (5 attributes $\times$ 3 isobaric surfaces). Then, we select the 3 best single attributes by their prediction performances. Next, we combine those 3 best attributes as follows: the best and second best attributes $(A_1 + A_2)$, the best and third best attributes $(A_1 + A_3)$, the second and third best attributes $(A_2 + A_3)$, and all three attributes $(A_1 + A_2 + A_3)$. Finally, we select the model that has the best performance among the single attribute models and the combined attribute models. Using this modified approach, we use a total of 19 prediction models to find the final prediction model for a specific type of hazardous weather in a specific region. The traditional top-down method for all attributes requires 42 prediction models to choose the 3 best attributes. Not only does our method require fewer experiments

Algorithm 1 Steps to build prediction models

1: For a specific region and a specific type of hazardous weather, make n prediction models using n single attributes
2: Get prediction results for n single attributes and select k best attributes. Call the i-th best attribute "A_i"
3: Combine two attributes in the k attributes. For example, combine A_1 and $A_2(A_1 + A_2)$, A_1 and $A_3(A_1 + A_3)$, ... , and A_{k-1} and $A_k(A_{k-1} + A_k)$ to obtain $k(k-1)/2$ combinations
4: Get prediction results from $k(k-1)/2$, two-attribute models and select k best attribute combinations. Call the i-th best attribute combination "C_i"
5: Generate three-attribute combinations by adding A_i to C_j's
6: Get prediction results from the three-attribute models and select k best combinations. Call the i-th best attribute combination "C_i"
7: Repeat steps 5 and 6 by increasing the number of attributes to be combined until an n-attribute combination has been created
8: Choose the best model among the single attribute models and all the combined attribute models. If more than two models show the same performance, choose the one with the smallest number of attributes

than the traditional top-down method, it also combines attributes A_2 and A_3, which the traditional method does not do. Using our modified top-down attribute selection method, we need only try 2,128 combinations of 64,400 possible combinations to build optimal prediction models for 7 types of hazardous weather in each of 16 regions.

Minimizing the need for intervention by experts, we find hazardous weather prediction models for each region and each hazardous weather condition using the modified top-down selection method. Experts just chose 5 weather attributes and 3 isobaric surfaces that can affect hazardous weather conditions; our modified top-down attribute selection method uses those choices efficiently to make regional hazardous weather prediction models. In the next section, we explain how we evaluate our prediction models with weather data using SVMs.

4.2 SVM Adaptation to Weather Data

We use meteorological data from 2002 to 2011 in our hazardous weather predicting experiments. For 6 hour prediction, we use 6 hours of past data before the current time to predict whether hazardous weather occurs. The number of hazardous weather conditions varies by region, so for each region we choose a number of non-hazardous weather conditions to maintain an equal ratio with the number of hazardous weather conditions [24]. Because the hazardous weather cases depend strongly on the seasons, we choose the same number of non-hazardous weather cases in a month. For example, given 5 cases of heavy rain in October 2004, we choose 5 non-heavy rain cases for the same time period. If we do not maintain the ratio between hazardous and non-hazardous weather conditions, the prediction model becomes over-fitted and predicts all weather conditions as non-hazardous.

Table 2. True-false table

	Positive	**Negative**
True	True positive (TP)	False negative (FN)
False	False positive (FP)	True negative (TN)

Thus, we make training and test data for the SVMs maintaining a balance between hazardous and non-hazardous weather conditions. We verify the performance of each SVM using the k-fold cross validation method with $k = 5$ based on the collected data. Cross validation is a prediction model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set.

We use Accuracy as the evaluation index based on the true-false table shown in Table 2. Accuracy indicates how often a prediction model predicts correctly.

The evaluation indexes are defined as Eq. (2) shown below:

$$Accuracy = (TP + FN)/(TP + TN + FP + FN). \quad (2)$$

TP (True positive): Model predicted the occurrence of hazardous weather, and hazardous weather occurred

TN (True negative): Model predicted the non-occurrence of hazardous weather, and hazardous weather did not occur

FP (False positive): Model predicted the occurrence of hazardous weather, and hazardous weather did not occur

FN (False negative): Model predicted the non-occurrence of hazardous weather and hazardous weather did occur

5. Experimental Results

We used the following data in the experiments: five attributes, *Height* (Z), *Humidity* (R), *Temperature* (T), *Uwind* (U), and *Vwind* (V), and three isobaric surfaces, 500, 700, and 850 hPa, for each type of hazardous weather and region.

Tables 3 and 4 show the prediction accuracy for heavy rainfall

and heavy snowfall. The prediction results for the other hazardous weather conditions are summarized in Table 5. Tables 3 and 4 consist of *Region*, *Attributes*, A_1, $A_1 + A_2$, $A_1 + A_3$, $A_2 + A_3$, and $A_1 + A_2 + A_3$ columns. *Region* represents the area for which the hazardous weather prediction is made, and *Attributes* represents the three attributes with their isobaric surfaces selected in Step 2 of the modified top-down attribute selection method in Algorithm 1. For example, *V(850)*, *V(700)*, and *R(500)* are selected for prediction of heavy rainfall at Seoul. *V(850)* means *Vwind* at 850 hPa; *V(700)* is *Vwind* at 700 hPa; and *R(500*) is *Humidity* at 500 hPa. The first attribute in the *Attribute* column is A_1, the second is A_2, and the third is A_3. The columns of A_1, A_1+A_2, A_1+A_3, A_2+A_3, and $A_1+A_2+A_3$ indicate the accuracies of the models built with the corresponding attributes. We choose the best results, marked in bold, as the final prediction models. If the prediction performance of two models is the same, we choose the model with the fewest weather attributes.

For heavy rainfall prediction, the models with a single weather attribute show the best performance in 12 regions; the models with 2 weather attributes are the best for 3 regions; and the model with 3 weather attributes is the best in only 1 region. *Vwind* at 700 hPa is used 6 times, so *Vwind* can be considered an effective weather attribute to predict heavy rainfall. The average accuracy of the prediction results across the 16 regions is 79.04%.

In Table 4, on heavy snowfall prediction, single-attribute models show the best performance for only 3 regions, whereas 2-attribute models are best for 6 regions, and 3-attribute models are best for 7 regions. Unlike the heavy rainfall prediction models, the heavy snowfall prediction models are mostly made by combining weather attributes. *Vwind* at 850 hPa is used 9 times, and *Uwind* and *Vwind* at 700 hPa are used 5 times each. Thus, the winds have a greater effect than the other weather attributes when making prediction models for heavy snowfall. The average prediction performance across the 16 regions is 78.86%.

We summarize the results of all the prediction models for the rest of the hazardous weather conditions in each region in Table 5. Table 5 contains the best result attributes and accuracy values. For example, *Uwind* at 500 hPa, *U(500)*, shows the best prediction result at Seoul for heat wave prediction, and its prediction accuracy is 84.83%, whereas the combination of *Vwind* at 850 hpa, *V(850)*, *Vwind* at 700 hPa, *V(700)*, and *Uwind* at 850 hPa *U(850)* show the best prediction result for lightning at Seoul. Except for wind waves prediction, 3 weather attributes (*Uwind*, *Vwind*, *Humidity*) are used most often to predict most hazardous weathers in most regions.

Table 6 represents the effectiveness of each weather attribute in predicting hazardous weather. As shown in Table 6, attributes tend to be selected more often as they approach the ground. Thus, weather attributes close to the ground can be considered effective for predicting hazardous weather. *Temperature* and *Height* are rarely used to make prediction models. *Temperature* is used only 1 time, and *Height* is used 6 times. However, *Height* is mostly used when predicting wind waves, which means that *Height* is the most effective weather attribute when building wind wave prediction models.

Table 7 compares the average results of the single-attribute models, combined-attribute models, and final selected models for all hazardous weather conditions. If the prediction models are made using only the best attribute, the average accuracy for all hazardous weather conditions is 73.60%. The combined-attribute models show performance almost equal to or lower than the single-attribute models on average. However, using the modified top-down selection method, we can achieve an accuracy of 79.61%, an improvement of about 8% over the best single-attribute models.

Table 8 shows our analysis of all hazardous prediction models: the number of the final models together with their attributes. For example, in the case of heavy rainfall, the final models for 11 regions have a single attribute (A_1), and the final models for the other 5 regions have combined attributes. We build 86 hazardous prediction models in total. Among them, 36 models have a single attribute, 34 models have two attributes ($A_1 + A_2$, $A_1 + A_3$ or $A_2 + A_3$), and 16 models have three attributes. Seven models use A_2 and A_3, which the traditional top-down selection method cannot find.

To select the final models for the 7 types of hazardous weather for all regions, we build and evaluate 1,634 models (7 hazardous weathers $\times$ 4–16 regions $\times$ 19 candidate models), whereas the traditional top-down attribute selection method requires 3,612 experiments (7 hazardous weathers $\times$ 4–16 regions $\times$ 42 candidate models). Our proposed method decreases the number of models by about 45% compared to the traditional top-down attribute selection method.

To summarize, we select optimal weather attributes to efficiently build regional hazardous weather prediction models with fewer experiments than required by the traditional method. The average prediction result is 79.61% for all prediction models, and that result can help forecasters decide whether hazardous weather will occur for their region.

Table 3. Prediction results for heavy rainfall (unit, %)

Region	Attributes	A_1	$A_1 + A_2$	$A_1 + A_3$	$A_2 + A_3$	$A_1 + A_2 + A_3$
Seoul	*V(850), V(700), R(500)*	**87.39**	86.59	77.39	77.39	76.56
Jeju	*R(500), R(700), U(850)*	70.50	**75.75**	73.17	66.75	75.75
Gangneung	*V(850), R(500), V(700)*	**74.97**	71.76	72.75	71.76	72.87
Gwangju	*R(700), R(500), V(850)*	72.55	**77.06**	73.73	69.02	77.06
Daegu	*V(500), R(500), V(700)*	**85.71**	83.21	85.71	83.21	83.21
Daejeon	*V(700), V(850), V(500)*	**80.00**	80.00	76.00	78.00	80.00
Mokpo	*V(850), R(700), V(500)*	77.12	78.49	73.79	75.00	**80.30**
Busan	*R(700), V(700), V(500)*	**82.35**	82.32	81.48	80.58	82.32
Andong	*T(700), T(500), V(700)*	**53.33**	53.33	42.67	42.67	42.67
Yeosu	*V(700), V(850), V(500)*	**90.00**	90.00	86.25	83.75	86.25
Ulsan	*V(700), V(850), V(500)*	81.03	77.82	81.16	**82.57**	81.03
Incheon	*V(700), R(500), V(850)*	**80.76**	71.52	79.43	72.86	71.52
Jeonju	*V(850), V(700), U(700)*	**74.73**	71.27	67.09	67.27	69.27
Chuncheon	*V(700), V(500), V(850)*	**85.81**	83.46	83.46	83.46	83.38
Chungju	*R(500), V(700), V(850)*	**80.15**	78.34	78.34	78.49	78.34
Seosan	*V(700), V(850), R(500)*	**73.74**	73.74	70.59	70.59	70.59

Table 4. Prediction results for heavy snowfall (unit, %)

Region	Attributes	A_1	$A_1 + A_2$	$A_1 + A_3$	$A_2 + A_3$	$A_1 + A_2 + A_3$
Seoul	*R(700), V(850), T(700)*	58.02	**73.00**	46.08	46.08	42.64
Jeju	*V(700), V(500), U(850)*	86.66	80.00	66.67	80.00	**86.67**
Gangneung	*V(850), R(700), U(850)*	71.00	76.99	**79.57**	74.39	74.36
Gwangju	*V(700), V(850), R(850)*	78.40	**85.08**	80.84	81.16	81.82
Daegu	*R(850), V(500), U(500)*	**45.00**	13.33	13.33	13.33	36.66
Daejeon	*V(700), V(850), U(850)*	76.91	**78.89**	70.00	72.22	73.33
Mokpo	*U(500), R(850), R(500)*	75.76	83.89	75.00	81.11	**86.66**
Busan	*V(850), U(700), R(500)*	**100.0**	86.67	20.00	26.67	63.33
Andong	*R(700), R(850), U(700)*	82.00	81.82	83.64	81.82	**85.60**
Yeosu	*V(850), V(700), V(500)*	**100.0**	100.00	100.00	100.00	100.00
Ulsan	*V(850), R(700), R(500)*	86.66	**100.00**	33.33	73.33	21.66
Incheon	*R(500), U(500), V(500)*	46.66	66.15	**67.69**	53.84	50.64
Jeonju	*U(700), V(700), U(500)*	74.16	81.31	79.70	79.16	**81.67**
Chuncheon	*R(700), R(500), U(700)*	59.60	76.81	79.09	63.96	**81.19**
Chungju	*V(850), U(700), V(700)*	61.11	68.97	73.60	73.75	**75.22**
Seosan	*R(850), V(850), U(700)*	73.02	79.05	79.05	70.95	**79.66**

6. Conclusion

We proposed a modified top-down method to find the optimal weather attributes to efficiently build regional hazardous

Table 5. Prediction results for 5 hazardous weather conditions (unit, %)

Region	Heat wave	Lightning	Cold wave	Strong winds	Wind waves
Seoul	*U(500)*	*V(850), V(700), U(850)*	*U(700)*	-	-
	84.83	76.84	80.66		
Jeju	*U(850)*	*V(850), V(700)*	-	*V(700), V(500), Z(700)*	-
	69.17	68.02		79.92	
Gangneung	*U(700)*	*V(850)*	*V(850)*	-	-
	75.01	64.05	76.66		
Gwangju	*U(500), U(700)*	*V(700)*	*V(850)*	-	-
	76.55	63.84	60.00		
Daegu	*U (500)*	*R(700), R(500), R(850)*	*V(700)*	-	-
	76.45	73.05	40.00		
Daejeon	*U(850), U(700)*	*U(850), R(850)*	*V(850)*	-	-
	79.29	75.75	71.40		
Mokpo	*U(700), U(500)*	*R(700), U(850), R(850)*	-	*V(850), V(700)*	-
	88.00	74.72		88.02	
Busan	*U(500)*	*V(850), U(850)*	*U(500)*	*U(850), V(500)*	-
	84.28	73.43	80.00	91.43	
Andong	*U(500), V(850)*	*R(700), U(700)*	*R(700), V(700), V(500)*	-	-
	81.54	72.97	69.77		
Yeosu	*R(850)*	*V(850), U(700)*	*V(700), V(500)*	*V(850), V(700), V(500)*	-
	50.00	77.92	80.00	88.92	
Ulsan	*U(700), U(850)*	*R(700), U(850)*	*R(700)*	-	-
	79.20	70.03	73.33		
Incheon	*U(700)*	*V(700), V(850), U(850)*	*U(500)*	-	-
	90.00	79.19	69.72		
Jeonju	*U(500)*	*V(850), U(850)*	*V(700), U(850)*	-	-
	82.37	74.19	62.67		
Chuncheon	*U(500), U(850)*	*U(850), V(850)*	*V(850), R(850)*	-	-
	77.40	72.30	78.67		
Chungju	*U(500), U(850)*	*U(850), V(850)*	*R(850), V(850)*	-	-
	81.59	75.64	77.40		
Seosan	*U(700), U(850)*	*V(850), U(850)*	*R(500)*	-	-
	86.00	76.23	73.50		
Geomundo	-	-	-	-	*Z(850)*
					85.71
Geojedo	-	-	-	-	*Z(500)*
					85.71
Deokjeokdo	-	-	-	-	*Z(850), Z(700), Z(500)*
					78.06
Chilbaldo	-	-	-	-	*V(700), V(850)*
					97.42

Table 6. Effectiveness of each weather attribute

	V(850)	V(700)	V(500)	U(850)	U(700)	U(500)	R(850)	R(700)	R(500)	T(850)	T(700)	T(500)	Z(850)	Z(700)	Z(500)
Heavy rainfall	5	5	3					4	3		1				
Heat wave	1			6	7	9	1								
Heavy snowfall	9	5	2	2	5	2	4	4	3						
Lightning	10	4		10	2		3	4	1						
Cold wave	5	4	2	1	1	2	2	2	1						
Strong winds	2	3	3	1										1	
Wind waves	1	1											2	1	2
Total	**33**	**23**	**10**	**20**	**15**	**13**	**10**	**14**	**8**	**0**	**1**	**0**	**2**	**2**	**2**

Table 7. Comparison results between single- and combined-attribute models (unit, %)

	Avg. of A_1	Avg. of $A_1 + A_2$	Avg. of $A_1 + A_3$	Avg. of $A_2 + A_3$	Avg. of $A_1 + A_2 + A_3$	Avg. of **final**
Heavy rainfall	78.13	77.17	75.19	73.96	75.70	79.04
Heat wave	77.02	73.19	73.13	72.81	71.94	78.86
Heavy snowfall	73.44	77.00	65.47	66.99	70.07	81.62
Lightning	67.70	70.41	70.09	69.65	71.65	73.01
Cold wave	67.39	55.33	55.11	49.03	49.98	70.98
Strong winds	74.39	81.23	83.10	81.31	82.22	87.07
Wind waves	77.14	79.06	79.06	79.06	79.21	86.73
Average	**73.60**	**73.64**	**71.59**	**70.40**	**71.54**	**79.61**

Table 8. Analysis of final selected models

Hazardous weather	No. of A_1	No. of $A_1 + A_2$	No. of $A_1 + A_3$	No. of $A_2 + A_3$	No. of $A_1 + A_2 + A_3$	Total no. of models
Heavy rainfall	11	2	1	1	1	16
Heat wave	8	1	4	3	0	16
Heavy snowfall	3	4	2	0	7	16
Lightning	2	4	3	3	4	16
Cold wave	10	3	0	0	1	14
Strong winds	0	1	1	0	2	4
Wind waves	2	1	0	0	1	4
Total	**36**	**16**	**11**	**7**	**16**	**86**

weather prediction models. Our proposed method reduced the number of experiments by 45% compared with the traditional top-down attribute selection method. Not only did we decrease the number of experiments, but we also obtained competitive performance from our prediction models. The average performance for the 7 types of hazardous weather in all regions is 79.61%, so the prediction models can help forecasters decide whether hazardous weather will occur. The prediction models in this paper are currently being used by the Korea Meteorological Administration to predict hazardous weather.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

This work was supported by the ICT R&D program of MSIP/IITP (B0101-15-0559, Developing On-line Open Platform to Provide Local-business Strategy Analysis and User-targeting Visual Advertisement Materials for Micro-enterprise Managers). Also, this research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2014M3C4A7030503).

References

[1] L. Al-Matarneh, A. Sheta, S. Bani-Ahmad, J. Alshaer, and I. Al-oqily, "Development of temperature-based weather forecasting models using neural networks and fuzzy logic," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 9, no. 12, pp. 343-366, 2014. http://dx.doi.org/10.14257/ijmue.2014.9.12.31

[2] E. T. Al-Shammari, M. Amirmojahedi, S. Shamshirband, D. Petkovic, N. T. Pavlovic, and H. Bonakdari, "Estimation of wind turbine wake effect by adaptive neuro-fuzzy approach," *Flow Measurement and Instrumentation,* vol. 45, pp. 1-6, 2015. http://dx.doi.org/10.1016/j.flowmeasinst.2015.04.002

[3] S. Al-Yahyai, Y. Charabi, and A. Gastli, "Review of the use of numerical weather prediction (NWP) models for wind energy assessment," *Renewable and Sustainable Energy Reviews*, vol. 14, no. 9, pp. 3192-3198, 2010. http://dx.doi.org/10.1016/j.rser.2010.07.001

[4] M. S. K. Awan and M. M. Awais, "Predicting weather events using fuzzy rule based system," *Applied Soft Computing*, vol. 11, no. 1, pp. 56-63, 2011. http://dx.doi.org/10.1016/j.asoc.2009.10.016

[5] F. Babic, P. Bednar, F. Albert, J. Paralic, J. Bartok, and L. Hluchy, "Meteorological phenomena forecast using data mining prediction methods," in *Proceedings of Third International Conference (ICCCI 2011),* Gdynia, Poland, 2011, pp. 458-467. http://dx.doi.org/10.1007/978-3-642-23935-9_45

[6] S. S. Badhiye, P. N. Chatur, and B. V. Wakode, "Temperature and humidity data analysis for future value prediction using clustering technique: an approach," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 1, pp. 88-91, 2012.

[7] V. B. Nikam and B. B. Meshram, "Modeling rainfall prediction using data mining method: A Bayesian approach," in *Proceedings of 5th International Conference on Computational Intelligence, Modelling and Simulation (CIMSim)*, Seoul, Korea, 2013, pp. 132-136. http://dx.doi.org/10.1109/CIMSim.2013.29

[8] 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*, vol. 4, no. 1, pp. 51-59, 2012. http://dx.doi.org/10.5815/ijieeb.2012.01.07

[9] A. L. Pyayt, I. I. Mokhov, B. Lang, V. V. Krzhizhanovskaya, and R. J. Meijer, "Machine learning methods for environmental monitoring and flood protection," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 5, no. 6, pp. 549-554, 2011.

[10] Y. Radhika and M. Shashi, "Atmospheric temperature prediction using support vector machines," *International Journal of Computer Theory and Engineering*, vol. 1, no. 1, 55-58, http://dx.doi.org/10.7763/IJCTE.2009.V1.9

[11] K. Rasouli, W. W. Hsieh, and A. J. Cannon, "Daily streamflow forecasting by machine learning methods with weather and climate inputs," *Journal of Hydrology*, vol. 414-415, pp. 284-293, http://dx.doi.org/2012.10.1016/j.jhydrol.2011.10.039

[12] L. A. S. Romani, A. M. H. Avila, J. Zullo, C. Traina, and A. J. M. Traina, "Mining relevant and extreme patterns on climate time series with CLIPSMiner," *Journal of Information and Data Management*, vol. 1, no. 2, pp. 245-260, 2010.

[13] D. P. Solomatine and K. N. Dulal, "Model trees as an alternative to neural networks in rainfall-runoff modelling," *Hydrological Sciences Journal*, vol. 48, no. 3, pp. 399-411, 2003. http://dx.doi.org/10.1623/hysj.48.3.399.45291

[14] E. Tsagalidis and G. Evangelidis, "The effect of training set selection in meteorological data mining," in *Proceedings of 14th Panhellenic Conference on Informatics (PCI)*, Tripoli, Libya, 2010, pp. 61-65. http://dx.doi.org/10.1109/PCI.2010.37

[15] D. Wang, X. Zhao, and H. Zhang, "Abnormal weather prediction: A new method combining rough set, BP neural network and temporal association rules," *Journal of Information & Computational Science*, vol. 9, no. 12, pp. 3477-3485, 2012.

[16] M. Yesilbudak, S. Sagiroglu, and I. Colak, "A new approach to very short term wind speed prediction using *k*-nearest neighbor classification," *Energy Conversion and Management*, vol. 69, pp. 77-86, 2013. http://dx.doi.org/10.1016/j.enconman.2013.01.033

[17] Z. Zeng, W. W. Hsieh, W. R. Burrows, A. Giles, and A. Shabbar, "Surface wind speed prediction in the canadian arctic using non-linear machine learning methods," *Atmosphere-Ocean*, vol. 49, no. 1, pp. 22-31, 2011. http://dx.doi.org/10.1080/07055900.2010.549102

[18] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159-175, 2003. http://dx.doi.org/10.1016/S0925-2312(01)00702-0

[19] X. Zhu, J. Cao, and Y. Dai, "A decision tree model for meteorological disasters grade evaluation of flood," in *Proceedings of 4th International Joint Conference on Computational Sciences and Optimization (CSO)*, Yunnan, China, 2011, pp. 916-919. http://dx.doi.org/10.1016/10.1109/CSO.2011.26

[20] J. Lee, S. Hong, and J. H. Lee, "An efficient prediction for heavy rain from big weather data using genetic algorithm," in *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication (ICUIMC'14)*, Siem Reap, Cambodia, 2014. http://dx.doi.org/10.1145/2557977.2558048

[21] S. Fan, L. Chen, and W. J. Lee, "Short-term load forecasting using comprehensive combination based on multimeteorological information," *IEEE Transactions on Industry Applications*, vol. vol. 45, no. 4, pp. 1460-1466, 2009. http://dx.doi.org/10.1109/TIA.2009.2023571

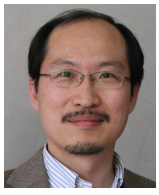
[22] A. M. Foley, P. G. Leahy, A. Marvuglia, and E. J. McKeogh, "Current methods and advances in forecasting of wind power generation," *Renewable Energy*, vol. 37, no. 1, pp. 1-8, 2012. http://dx.doi.org/10.1016/j.renene.2011.05.033

[23] L. Ingsrisawang, S. Ingsriswang, S. Somchit, P. Aungsuratana, and W. Khantiyanan, "Machine learning techniques for short-term rain forecasting system in the northeastern part of Thailand," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 2, no. 5, pp. 1422-1427, 2008.

[24] K. Napierala and J. Stefanowski, "BRACID: a comprehensive approach to learning rules from imbalanced data," *Journal of Intelligent Information Systems*, vol. 39, no. 2, pp. 335-373, 2012. http://dx.doi.org/10.1007/s10844-011-0193-0

[25] R. Nayak, P. S. Patheja, and A. Waoo, "An enhanced approach for weather forecasting using neural network," in *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS2011)*, Roorkee, India, 2011, pp. 833-839. http://dx.doi.org/10.1007/978-81-322-0491-6_76

Jaedong Lee received his B.S. in computer engineering from Dankook University, Cheonan, Korea, in 2011. He is currently pursuing his Ph.D. in computer engineering at Sungkyunkwan University. His research interests include intelligent system and machine learning.
E-mail: ultrajaepo@skku.edu



Jee-Hyong Lee received his B.S., M.S., and Ph.D. in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1993, 1995, and 1999, respectively. From 2000 to 2002, he was an international fellow at SRI International, USA. He joined Sungkyunkwan University, Suwon, Korea, as a faculty member in 2002. His research interests include fuzzy theory and application, intelligent systems, and machine learning.
E-mail: john@skku.edu