STORM SURGE FORECAST MODEL USING GENETIC PROGRAMMING

Chu Thi Quyen*, Nguyen Thi Cam Ngoan, Ngo Thi Thanh Hoa

Abstracts: *Stormsurge is a typical genuine fiasco coming from the ocean. Therefore, an accurate forecast of surges is a vital assignment to dodge property misfortunes and decrease the chance of tropical storm surges. Genetic Programming (GP) is an evolution-based model learning technique that can simultaneously find the functional form and the numeric coefficients for the model. Moreover, GP has been widely applied to build models for predictive problems. However, GP has seldom been applied to the problem of storm surge forecasting. In this paper, a new method to use GP for evolving models for storm surge forecasting is proposed. Experimental results on data-sets collected from the Tottori coast of Japan show that GP can become more accurate storm surge forecasting models than other standard machine learning methods. Moreover, GP can automatically select relevant features when evolving storm surge forecasting models, and the models developed by GP are interpretable.*

Keywords: Storm surge; Genetic programming; Typhoon; Surge deviation; White-box forecasting.

1. INTRODUCTION

Storm surge, a phenomenon of rising water commonly associated with low atmospheric pressure and strong winds weather system (typically called a tropical typhoon), is a severe natural disaster coming from the sea. The wind causes the water to pile up higher than the ordinary sea level, and low pressure at the center can have a small secondary effect. Storm surges are especially harmful when they happen at a high tide, combining the surge and tide impacts. According to [1], within over 600 million people living in low-lying coastal zones, coastal surges can destroy societal impacts. A number is approximated that over 0.8 to 1.1 million people each year are flooded globally. This number is reported by the flooding of Holland and the United Kingdom in 1953. Also, in the United States, the most remarkable recorded storm surge was produced by 2005's Storm Katrina, which created a storm surge 9 meters tall within the town of Cove St. Louis, Mississippi, and resulted in over 2,000 fatalities. The overall misfortune as a result of Katrina was calculated to surpass $100 billion [2]. Whereas high-impact occasions will, without a doubt, happen within the future, the approach and assist the advancement of the forecast of storm surge may enormously reduce the misfortune of lives and possibly lessen the sum of property harm.

A conventional approach to storm surge determining is to utilize process-based numerical forecast models, firstly in the 1950s, which are computationally seriously. An alternative approach is to use machine learning algorithms such as artificial neural networks (ANN) for predicting the relationship between storm surge levels and the corresponding features such as sea surface levels, winds, sea level pressures, and tropical cyclone (TC) characteristics. For illustration, a productive and vigorous artificial intelligent model was created to anticipate the peak storm surge utilizing the tropical storm parameters: central pressure, span to most extreme winds, forward velocity, and storm track. ANN has been applied with critical effects on the prediction of tides, sea surface levels, waves, tsunami, breakwaters, and especially storm surge [3]. The artificial neural arrangement method appeared to outflank support vector machines for extraordinary storms. A surrogate model for storm surge forecast was created in utilizing an artificial neural network with the measured tidal level. The created surrogate demonstrates fulfilled high-accuracy and high-speed for anticipating the storm surge based on an artificial intelligence method and a grid-free framework. In comparison with ordinary models, neural network models have preferences of quick computational time in some tens of minutes to estimate output information after the model is trained. Despite these advantages, neural network models are black-box, so it is hard to interpret or explain them. Moreover, neural network models often can not achieve good estimation in peak points, vitality in forecasting storm surge levels.

Genetic Programming (GP), which continues the initial idea ofthe problem of representation in Genetic Algorithm [4] by raising the complexity of the structures undergoing adaptation. GP is an evolutionary procedure to create solutions within the representation of computer programs for a problem. GP's ability to learn the definition of a function itself from sample information makes it an excellent choice for symbolic regression. The group of primitive functions used typically includes operations, mathematical functions, and domain problem environments. Therefore, GP has been widely used to build regression models for many real applications. For example, based on predictions of stock prices using GP, a possibly profitable trading strategy was proposed in the financial market analyst system [5]. Not only the application of GP to the financial market but also in weather conditions forecasting. Gaur and Deo [6] proposed GP's application to build a model for real-time wave forecasting by two locations of the United States coast. GP's estimates proved that it could be respected as a promising tool for future applications of oceanic forecasts. Next, Azamathulla and Ghani [7] provide a GP-based method that results in an accurate estimation for river pipeline scour. And the last GP performance was more effective than both the results of regression equations and artificial neural systems modeling in anticipating the explore depth around channels. Despite having remarkable results in real-life issues, including natural phenomena, GP has not been constructed or researched to build models for forecasting storm surge levels. Indeed, this paper proposes the first attempt to investigate GP's power to construct a white-box model for forecasting storm surge levels.

Therest of the paper is organized as follows. We first present related work on predict storm surge levels in section2. Section 3, 4 proposes the GP method and other machine learning techniques. Wedescribe the experimental settings in section 5, presentingand discussing the results in section 6. We provide overallconclusions in section 7 and highlight directions for futurework.

2. RELATED WORKS

A common approach to estimating storm surge levels is to utilize progressed machine learning models such as artificial neural networks. These models are driven by the connection between storm surge levels and corresponding data likesea surface levels (astronomic tide level and surge level), winds, sea level pressures, and typhoon-characteristics (typhoon location, central atmospheric pressure, and maximum wind speeds near typhoon center). [3]

As mentioned in [8] and [3], a couple of neural network applications were tried to form the storm surge prediction models. The predictedresults of distinctive schedule strategies were compared with the observed ones. The confirmation appeared that the neural network technique might well be effectively applied within the schedule, operational figure service.

A strategy to foresee ocean level varieties related to meteorological occasions that utilize a neural network model was displayed in [4]. The comes about appeared that the model is able to capture the impacts of climatic and maritime intuitive. It can be considereda proficient model for anticipating the nontidal residuals and can viably complement the standard consistent consonant analysis model.

In [9], an artificial neural network model was developed to predict storm surges in all Korean coastal regions, with a particular focus on a regional extension. The combination of cluster neural networks and agglomerative clustering was used to predict storm surge. The exploratory has demonstrated that the model can be connected within the territorial storm surge expectation and had prospects to be used to build an estimate framework.

As Kim and colleagues have described in [3] diverse information, comprising the neighborhood meteorological and hydrodynamic parameters collected and the TC-characteristics, they were connected to the brand new developed after-runner surge prediction model for Sakai Minato on Tottori coast, Japan. It was found that (at Sakai Minato) the combination of surge level, sea-surface level, seal-level pressure and its drop rate, longitude and latitude of TC, ocean surface level, wind speed, and wind direction are the ideal information sets to yield accurate surge level with the lead time of 24h.

Comparing to ordinary methods, neural network models surpass them by preferences of quick computational time, within the processing time of over ten minutes to estimate. However, coming with the advantage of speed, these models are black-box, so it is challenging to account for or analyze them. And there are other issues, ANN models sometimes can not achieve good estimation in peak points, which is a crucial factor in forecasting storm surge levels.

3. RESEARCH METHODOLOGY

In this section, we describe the proposed procedure for addressing the problem of storm surge.Genetic Programming (GP) is one of the foremost well-known evolutionary algorithm methods inspired by a natural mechanic to evolve issue solutions within the frame of computer programs [10, 11]. In the 1990s, GP was primarily applied to general fundamental issues since it was very computationally seriously. Along with the development of executing power in computers, the GP frameworks are also changed and applied to various world issues broadly. The applications of GP incorporate electronic design, diversion playing, quantum computing, and sorting.

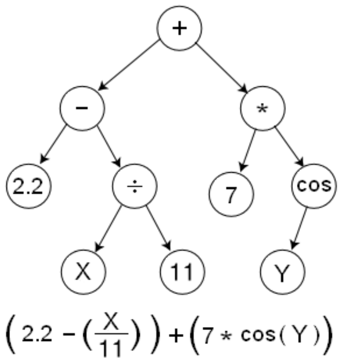


Figure 1. *GP representation.*

3.1. Population

The GP procedure begins with a population that contains random individuals. Thesize of the initial population can be determined based on the problem. Each population is generated by the well-known method ramped-half and-half.

3.2. Fitness function

The fitness used for assessment will be based on the root mean squared error andthe coefficient of correlation (CC) given by:

$$\text{fitness} = \text{NRMSE}(1 + \frac{1}{CC}) \tag{1}$$

where NRMSE and CC are calculated by:

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{obs,i} - y_{pre,i})}}{(y_{obs,max} - y_{obs,min})} \tag{2}$$

$$CC = \frac{\sum_{i=1}^{n}(y_{obs,i} - \bar{y}_{obs})(y_{pre,i} - \bar{y}_{pre})}{\sqrt{\sum_{i=1}^{n}(y_{obs,i} - \bar{y}_{obs})^2}\sqrt{\sum_{i=1}^{n}(y_{pre,i} - \bar{y}_{pre})^2}} \tag{3}$$

where n is the length of the training set, $y_{pre,i}$ represents the predicted surge leveland $y_{obs,i}$ represents the actual surge level for the i-th data point (time index).

3.3. Elitism

Elite selection is usually carried out to preserve the fittest chromosomes. In addition, it is mainly used to ensure that a good-quality chromosome does not miss upon the next generation when the population is updated. We use the elitism strategy to preserve our best fitness scores. Once the elitism step is completed, we move on to the next phaseto create a new population again.

3.4. Creating a new population

The next step is to generate a new population after each individual fitness isevaluated. The process involves three operations to create a new population: selection, crossover, and mutation. Below is a brief explanation of these operationsfor selecting parents and offspring generation.

a) Selection: For the offspring generation, the tournament selection method is usedfor the selection of parents. First, we created a group of the number of individualsselected randomly from the current population. Furthermore, the best individual basedon a fitness assessment was chosen as the first parent to breed a new individual; the same process selects a second parent. The role of parental selection is to distinguish between individuals and give preference to the best individual as next-generation parents.

b) Crossover: The objective of the crossover operator is to create a new offspring from a parent pair. In this work, we have used a uniform crossover 0.6 as a crossover rate.

c) Mutation: To maintain the genetic diversity in the population, we have employed a random replacement of subtree with 0.3 as the mutation rate in this work.

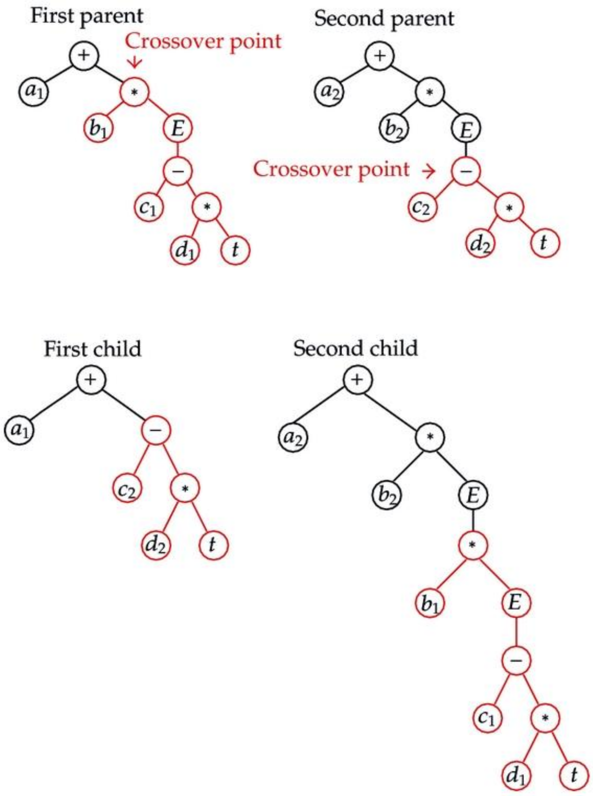


Figure 2. *Crossover operator.*

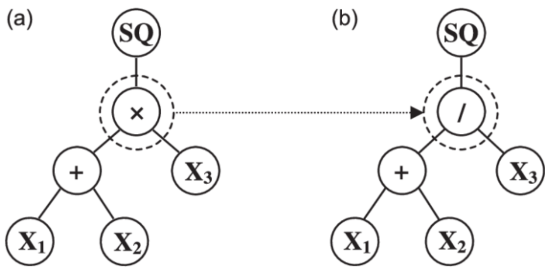


Figure 3. *Mutation operator: (a) Before (b) After.*

3.5. Terminating conditions

We used the following termination conditions for our designed algorithm. If one ofthe following conditions is met algorithm terminates the search: 1) The fitness value has reached an optimal global value. 2) The number of total generations have reached their limits. The GP returns the best individual with the best fitness score on training data; after the whole procedure is completed, we then compared the performance with other modern algorithms using the same training and testing data-set.

3.6. The pseudo-code for GP

Now, it is time to put together the components of SSGP (GP for Storm Surge) and present the pseudo-code for the whole Algorithm 1. SSGP starts with the initialization randomly of the population. It then evaluates the individuals based on the fitness value of each individual. SSGP selects in dividuals for the application of genetic operators for the next-generation using the tournament selection. This loop will continue until the stopping criteria are met.

Algorithm 1: *SSGP*

```
Initialise population P_0)
Evaluate population P_0
while stopping criteria not met do
    Ascertain individual's fitness
    Select one or two individual(s) from the population
    P_t with a probability based on fitness to participate
    in genetic operations (see Section ).
    Create new individual by applying genetic operations
    with specified probabilities (see Section ).
    Evaluate new population P_(t + 1)
    P_t = P_(t + 1)
end while
return the best individual and adjust the best
```

4. OTHER MACHINE LEARNING METHODS

Besides GP, we have some other approaches to solve the storm surge problem. In this current section, there are four other machine learning methods tried out in our experiments. They are the Multi-Layer Perceptron (MLP), k-Nearest Neighbors (k-NN), Decision Trees (DCT), and Support Vector Regression (SVR), which are ordinarily called data-driven models, since of the capacity to capture the relationship between input and output variables without studying directly the natural rules that dominate the mechanism of storm surge and these models are entirely based on information obtained from the collected data.

4.1. k-nearest neighbor (k-NN)

k-NN is one of the simplest supervised-learning algorithms in Machine Learning. This non-parametric strategy bases its forecast on the target yields of the k-nearest neighbors of the given data point (Y. Song, J. Liang, J. Lu, and X. Zhao, 2017). For more clearly, given a set of data points, we determine the value of Euclidean measure between that point and all points within the training set. After that, we choose the closest k preparing information focuses. And we set the prediction value as a medium of the target output values for k points, which are k nearest training data points. By the k-NN method, the prediction x_t is defined by:

$$K = \frac{1}{k}\sum_{t\in S(X,m)} x_t \qquad (4)$$

where $S(X, m)$ expresses the k-nearest neighbor index set to the $X(m)$ attribute vector. The forecast x_t^F in equation 4 is the sample average of the output surge level of the k-nearest neighbors to $X(m)$, intuitively.

4.2. Support vector machine

Support vector machines can be applied to both classification and regression problems. When applied to a regression problem, it will be called support vector regression (SVR). SVR (R. M. Balabin and E. I. Lomakina, 2011) is a successful technique of sanctioning the ensuing complexity by adding the penalty term to the error function and an alternative approach for ANN. In consideration of a linear model for illustration, and the forecast is formulated by:

$$f(x) = w^T x + o \tag{5}$$

where w, o and x denote respectively are the weight vector, the bias, and the input vector. Let x_m is m-th training input vector and y_m is the target output, $m = 1, \dots, M$. The following formula to calculate the error function:

$$J = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}|y_m - f(x_m)|_\epsilon \tag{6}$$

In the above error function, the first term is a term that penalizes the complexity of the model. The remaining term is called the ϵ-insensitive loss function.

It does not penalize errors below epsilon, allowing the parameters to transfer some wiggle room to reduce the sophistication of the model. It could be clarified that the solution is given by equation 6, in which the error function is minimized.

$$f(x) = \sum_{m=1}^{M}(\alpha_m^* - \alpha_m)x_m^T x + o \tag{7}$$

where α_m and α_m^* are Lagrange multipliers. The training vectors that provide nonzero Lagrange multipliers are referred to as support vectors, which is a critical concept in SVR's theory. Non-support vectors indirectly contribute to the solution. The number of support vectors is the degree of the model's complexity(R. M. Balabin and E. I. Lomakina, 2011). This model is expanded to the nonlinear case by yielding the kernel principle K.

$$f(x) = \sum_{m=1}^{M}(\alpha_m^* - \alpha_m)K(x_m^T x) + o \tag{8}$$

The widely used Gaussian kernel is employed in this article. Its width, δ_K, is the standard deviation of the Gaussian function.

4.3. Multi-layer perceptron networks

The multi-layer perception network is by far the most common ANN model, which usually uses the error propagation technique to train the configuration of the network. The multi-layer perception network is by away the most commonplace ANN model, which generally uses the error propagation method to learn the network structure. The ANN architecture is made up of several hidden layers and several neurons in the input layer.

In the output layer, either multiple or single neurons come up. ANNs with one hidden layer are usually utilized in hydrologic modeling because those networks are taken into consideration to provide sufficient complexity to simulate the nonlinear properties of the hydrologic process accurately. The ANN model for forecasting is defined as:

$$x_t^F = f(X_t, w, \theta, m, h) = \theta_0 + \sum_{k=0}^{n} w_j^{out}\, \emptyset\left(\sum_{i=1}^{m} w_{ji} x_t + \theta_j\right) \tag{9}$$

where:

- φ indicates transfer functions;
- w_{ji} are the weights of the link between the input layer's i-th node and the hidden layer's j-th node;
- θ_j are biases related to the j-th node of the hidden layer;
- w_j out are the weights linked to the association between the j-th node of the hidden layer and the node of the output layer;
- and θ_0 is the bias at the output node.

A suitable training algorithm is needed to optimize w and θ to apply equation 9 to surge level expectations.

4.4. Decision Tree - REFTree

Reduces Error Pruning (REP) Tree Classifier could be a decision-making tree learning algorithm based on the rule of calculating the data pick up with entropy information gain and minimizing the mistake emerging from variance (J. R. Quinlan, 2014)(W. N. H. W. Mohamed, M. N. M. Salleh, and A. H. Omar, 2012). REP Tree applies regression tree logic. In altered iterations, it produces several trees. It chooses one of all spawned trees best afterward. This algorithm uses variance and knowledge gain to construct the regression/decision tree. This algorithm also prunes the tree with a reduced-error pruning process. It sorts numeric attribute values once at the start of the model preparation. As in the C4.5 algorithm (J. R. Quinlan, 2014), this algorithm also tackles the missing values by separating the related instances into parts.

5. EXPERIMENTAL SETTINGS

In our experiments, we use data from two countries are Vietnam and Japan. Firstly, based on the data represented in the (S. Kim, Y. Matsumi, S. Pan, and H. Mase, 2016), Tottori Coast, Japan. As described in (S. Kim, Y. Matsumi, S. Pan, and H. Mase, 2016) the typhoons that induced storm surges in Sakai Minato, Kim and his colleagues selected three typhoons: Maemi 2003, Songda 2004 and Megi 2004, of which surge levels are found almost corresponding to the surge level of 100 year return period (reported by Japan Meteorological Agency). The surge levels of Maemi and Megi events are especially close to the 100 year return period level. They are regarded as the representative and most intensive events on the Tottori coast, which caused high after-runner surge levels in Sakai Minato. Secondly, we collected data from 12 storms: Haiyan 2013, Rammasun 2014, Kujira 2015, Mujigae 2015, Mirinae 2016, Nida 2016, Dianmu 2016, Sarika 2016, Haima 2016, Hato 2017 and Talim 2017at HonDau station in Vietnam.

The following meteorological and hydrodynamic parameters were used for three typhoons from the observation stations as seen in:

- Meteorological parameters: wind speed (m/s), wind direction (degree), sea-level pressure (hPa), and drop of sea-level pressure from average sea-level pressure (¼1013 hPa) at five stations with black circle

- Hydrodynamic parameters: sea surface level (m) (¼ the atmospheric tidal level-the surge level) and surge level (m) (¼ observed sea surface level-atmospheric tidal level).
- Typhoon-characteristic parameters: longitude and latitude (degree), central atmospheric pressure (hPa), and highest wind speed near the typhoon center (m/s)

As mentioned before, in this paper, we will reuse the data-sets (including training and testing) from the research of Kim (S. Kim, Y. Matsumi, S. Pan, and H. Mase, 2016). Following that, the traditional storm surge forecast models depend on a parametric wind and pressure model or an atmospheric general circulation model in some ordinary methods. For that, the accuracy of these methods is significantly dependent on the quality of the meteorological input, but also generally, these input data are not entirely correct. To decrease the ANN forecast model's uncertainty in [3], the hourly measurement of meteorological and hydrodynamic data is collected by the observation stations on the Tottori coast, Japan. Moreover, there are some variables from typhoon characteristics such as typhoon position (longitude and latitude), central atmospheric pressure, and the highest wind speed near the typhoon center.

6. RESULTS AND DISCUSSION

In this section, the paper will conduct training and validate the correctness of building a GP prediction model. For each data set corresponding to the 5h, 12h, and 24h lead, we created 3 separate models for each milestone. First, it will be necessary to clearly identify some of the formulas used to evaluate the results of the method obtained with other methods, where the two are in which NRMSE calculates the deviation between observed data and predicts data. At the same time, CC is the correlation coefficient between actual and predicted numbers.

Table 1. *GP parameter settings.*

Parameter	Value
Function set	+, -, x, / (protected division), sin, cos
Variable terminals at	all features
Constant terminals	Random double values range from -1 to 1
Population size	500
Initialization	Ramped half-and-half
Generations	150
Crossover probability	60%
Mutation probability	30%
Reproduction rate	10%
Selection type	Tournament (size=3)

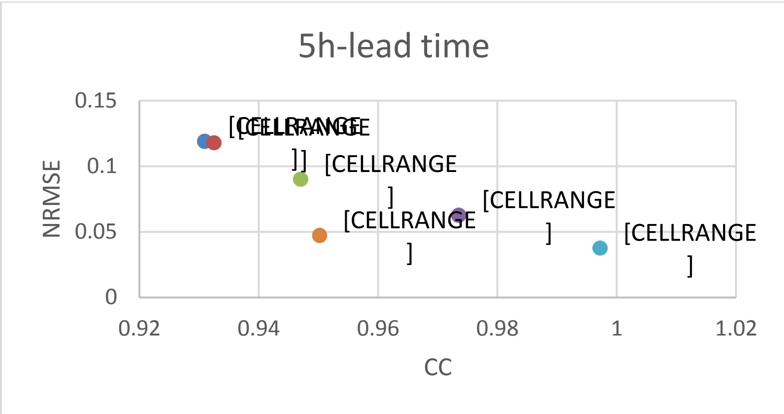
As mentioned, we will compare the results of GP with other methods such as Support Vector Machines, k Nearest Neighborhood, Multilayer Perceptron, Decision Tree, or more specifically SMOreg, Ibk, MLP, and REPTree in Weka (M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, 2009). GP will be implemented by using ECJ (S. Luke, L. Panait, G. Balan, S. Paus, Z. Skolicki, J. Bassett, R. Hubley, and A. Chircop.). Along with that, applying the results obtained from Kim's method, we will also compare the feature sets B3 with 5h lead, B1 with 12h lead, and B4 with 24h lead. Specific parameters for each method are as follows in table 1 and table 2.

Table 2. *Optimal parameters using Weka for the four models: SVM, k-NN, MLP, and REPTree for surge levels forecasting.*

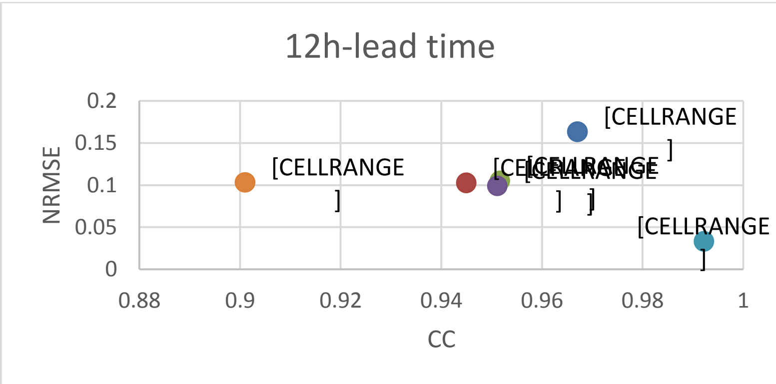
SVM	**k-NN**
SVM Type: epsilon-SVR Kernel Type: RBFKernel cacheSize: 250007 gamma: 0.01 Epsilon:1.0E-12	k:5 distanceFunction: Euclidean
MLP	**REPTree**
hiddenLayers: 5 learningRate: 0.3 Momentum: 0.2 epochs: 500	minimum number of instances: 2.0 noPrunning: false number of folds: 3

The experimental results are represented by the values of CC and NRMSE in each lead time 5h, 12h, 24h.

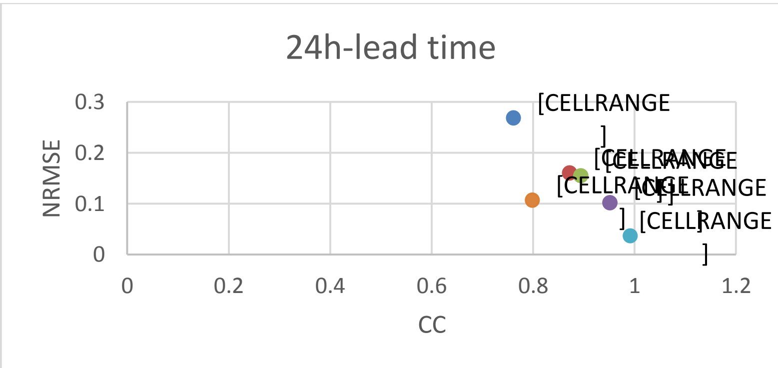
To evaluate our evolved GP models in detail on the 5h-lead time forecasts compared to the other five learning methods, we calculated two records of the correlation coefficient (CC) and the normalized root mean square error (NRMSE) in rate as appeared in fig.4 (a). It is observed that the value of CC inall of the models ranges from 0.93 to 0.99, while that of NRMSE is from 5% to12%. Among these models, the one evolved with GP has the smallest generalizationerror and the largest correlation coefficient. Similarly, in fig.4 (b), the correlation coefficient (CC) and the normalized root mean square error (NRMSE) of all models for 12h-lead time surge level forecasts are depicted. The CC values of the models are mostly concentrated in therange of 0.9 to 0.98. These models also have the NRMSE values from 5% to13%. Among them, GP shows the best performance for 12h-lead time forecasts. In addition, two records of CC and NRMSE are again calculated for all modelsin 24h-lead time forecasts. They are shown in fig.4 (c). The CC values of mostof the models are scattered from 0.76 to 0.99, and the NRMSE values are from 5% to 20%. Even though GP-based model performance deteriorates compared to the 5h- and 12h-lead time cases, it is still the best forecasting model among all, with the values of CC and NRMSE as 0.99 and 6.8%, respectively. Finally, with HonDau's data in fig. 5, the CC and NRMSE model by GP as 0.98 and 6.0% is the best among all models. Overall, these experiments and comparisons show that GP is the most effective method in building a surge level forecasting model.



(a) 5h-lead time



(b) 12h-lead time



(c) 24h-lead time

Figure 4. *Diagram of the correlation coefficient (CC) and the normalized root mean square error (%) for the surge levels forecast made in 3 interval times in advance by 6 models for Japan's stations.*

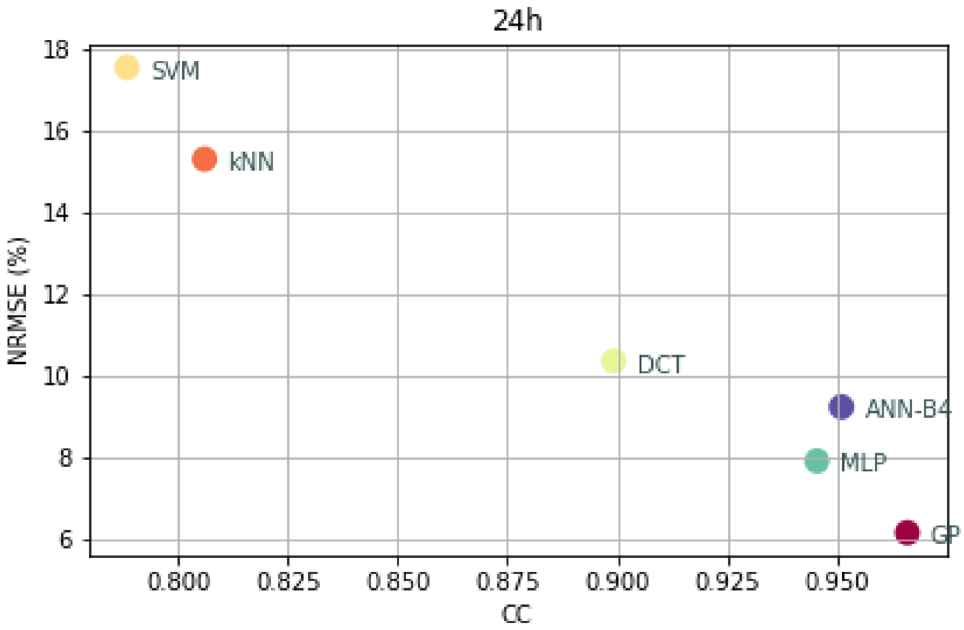


Figure 5. *Diagram of the correlation coefficient (CC) and the normalized root mean square error (%) for the surge levels forecast made in 24h lead time in advance by 6 models for HonDau's station.*

The Evolved Model

The best model evolved by GP are shown following for 24h-lead time:

sqrt(sub(mul(X10,0.921279(log(mul(X3,sub(X8,X9))))),mul(X10,mul(log(add(cos(sqrt(div(sub(X5,sub(X2,log(0.299569(X10)))),X9))),add(cos(sqrt(div(sub(X5, sub(cos(mul(X2,add(X7,X10))),mul(mul(X2,X2),X10))),mul(cos(X10),X9)))),X3))),sqrt(mul(X10,cos(log(X4)))))))).

corresponds to the expression:

$$\sqrt{\begin{array}{c} x_{10} \times 0.921279 \log\big(x_3 \times (x_8 - x_9)\big) \\ -\log(cos\sqrt{\dfrac{(x_5 - x_2 + log0.299569x_{10})}{x_9}} + \\ cos\sqrt{\dfrac{x_5 - \cos\big(x_2 \times (x_7 + x_{10})\big) - x_2^2 \times x_{10}}{cosx_{10} \times x_9}} + x_3) \times \\ \sqrt{\dfrac{x_{10}}{coslogx_4}} \times x_{10} \end{array}}$$

In the above model, the variables are used in the following table 3.

***Table 3.** Mapping table between variables and attributes.*

Variable	Attribute	Name
x_1	SS	the surge level
x_2	SLP	sea-level pressure
x_3	DSLP	the drop of sea-level pressure from average sea-level pressure
x_4	LG	longitude
x_5	LT	latitude
x_6	CAP	the central atmospheric pressure of typhoon
x_7	HWS	the highest wind speed near the typhoon center
x_8	SSL	the sea surface level
x_9	WS	the wind speed
x_{10}	WD	the wind direction

Within x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, to x_{10} respectively stand for the meteorological and hydrological factors like the surge level, sea-level pressure, the drop of sea-level pressure from average sea-level pressure, longitude, latitude, the central atmospheric pressure of typhoon, the highest wind speed near the typhoon center, the sea surface level, the wind speed, the wind direction.

It is clear from the above expression that GP can automatically select features to evolvethe best models. This clearly demonstrates the meaning of the white-box of GP.

7. CONCLUSIONS

Based on the above experimental comparison results, it can be confirmed that the application of the Genetic Programming method to solve the storm surge levels forecasting problem. Not only does the method outperform existing methods in terms of accuracy, but the GP method proposed by the paper is ultimately data-independent, is a white-box technique, and can be applied to forecasting many locations again.

REFERENCES

[1]. S. Muis, M. Verlaan, H. C. Winsemius, J. C. Aerts, and P. J. Ward, *"A global reanalysis of storm surges and extreme sea levels 7:11969,"* Nature Communications, 2016.

[2]. B. Sahoo and P. K. Bhaskaran, *"Prediction of storm surge and inundation using climatological data-sets for the Indian coast using soft computing techniques,"* Soft Computing, vol. 23, p. 12363–12383, 2019.

[3]. S. Kim, Y. Matsumi, S. Pan, and H. Mase, *"A real-time forecast model using artificial neural network for after-runner storm surges on the Tottori coast, Japan,"* Ocean Engineering, vol. 122, p. 44–53, 2016.

[4]. R. Poli, W. B. Langdon, N. F. McPhee, and J. R. Koza, *"A field guide to genetic programming,"* 2008.

[5]. M. A. Kaboudan, *"Genetic Programming Prediction of Stock Prices,"* Computational Economics, vol. 16, no. 3, p. 207–236, 2000.

[6]. S. Gaur and M. Deo, *"Real-time wave forecasting using genetic programming,"* Ocean Engineering, vol. 35, no. 11-12, p. 1166–1172, 2008.

[7]. H. M. Azamathulla and A. A. Ghani, *"Genetic programming to predict river pipeline scour,"* Journal of Pipeline Systems Engineering and Practice, vol. 1, no. 3, pp. 127-132, 2010.

[8]. N. B. Thuy, S. Kim, D. D. Chien, V. H. Dang, H. D. Cuong, C. Wettre, and L. R. Hole, *"Assessment of Storm Surge along the Coast of Central Vietnam,"* Journal of Coastal Research, vol. 33, no. 3, p. 518–530, 2016.

[9]. S. H. You and J.-W. Seo, *"Storm surge prediction using an artificial neural network model and cluster analysis,"* Natural hazards, vol. 51, no. 1, p. 97–114, 2009.

[10]. J. R. Koza, *"Genetic programming as a means for programming computers by natural selection,"* Statistics and Computing, vol. 4, no. 2, p. 87–112, 1994.

[11]. J. R. Koza., *"Genetic programming: on the programming of computers by means of natural selection,"* MIT Press, vol. 1, 1992.

[12]. Y. Song, J. Liang, J. Lu, and X. Zhao, *"An efficient instance selection algorithm for k nearest neighbor regression,"* Neurocomputing, vol. 251, p. 26–34, 2017.

[13]. R. M. Balabin and E. I. Lomakina, *"Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data,"* Analyst, vol. 136, no. 8, p. 1703–1712, 2011.

[14]. J. R. Quinlan, *"C4.5: Programs for Machine Learning"*, Morgan Kaufmann, 2014.

[15]. W. N. H. W. Mohamed, M. N. M. Salleh, and A. H. Omar, *"A comparative study of reduced error pruning method in decision tree algorithms,"* IEEE International conference on control system, computing and engineering, p. 392–397, 2012.

[16]. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, *"The weka data mining software: an update,"* ACM SIGKDD explorations newsletter, vol. 11, no. 1, p. 10–18, 2009.

[17]. S. Luke, L. Panait, G. Balan, S. Paus, Z. Skolicki, J. Bassett, R. Hubley, and A. Chircop., *"Ecj: A java-based evolutionary computation research system.,"* [Online]. Available: http://cs. gmu. edu/eclab/pro.

TÓM TẮT

SỬ DỤNG LẬP TRÌNH DI TRUYỀN DỰ BÁO NƯỚC BIỂN DÂNG DO BÃO

Nước dâng bão là hiện tượng dâng lên của mực nước biển cao hơn mực thủy triều vốn có bởi do tác động của bão, vì thế, việc dự báo chính xác mực nước dâng là nhiệm vụ quan trọng để tránh thiệt hại về tài sản và con người do nước dâng gây ra. Lập trình di truyền (Genetic Programming – GP) là một kỹ thuật học máy có thể giúp ta tìm được mô hình ở dạng công thức toán học. Tuy nhiên, trước đây GP hầu như chưa được áp dụng triệt để cho bài toán dự báo nước biển dâng do bão, cho nên, trong bài báo này, nhóm tác giả đề xuất phương pháp sử dụng GP để phát hiện các mô hình dự báo nước biển dâng do bão. Kết quả thực nghiệm trên dữ liệu nước biển dâng do bão tại trạm Hòn Dáu của Việt Nam cho thấy, phương pháp này có thể đưa ra các mô hình dự báo nước dâng do bão chính xác hơn một số phương pháp học máy phổ biến thường sử dụng. Hơn nữa, GP đưa ra mô hình dự báo dễ hiểu hơn các mô hình mà được xây dựng bằng các phương pháp khác (hộp đen) như là mạng nơ–ron. Ngoài ra, mô hình dự báo do GP đưa ra sẽ giúp ta phát hiện các đặc trưng ảnh hưởng trực tiếp khi phát triển các mô hình dự báo nước biển dâng do bão.

Từ khóa: Lập trình di truyền; Nước biển dâng do bão; Mô hình hộp trắng; Bão; Nước dâng.

Received 14th February 2020
Revised 10th April 2020
Accepted16thNovember 2020

Author affiliations:

Hanoi University of Industry.
*Corresponding author: chuthiquyen_cntt@haui.edu.vn.