Heavy rainfall classification using Genetic Programming

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

Hanoi University of Industry - No 298 Cau Dien Street, Bac Tu Liem district, Hanoi.
*Corresponding author: chuthiquyen_cntt@haui.edu.vn
Received 11 October 2021; Revised 12 January 2022; Accepted 14 February 2022.
DOI: https://doi.org/10.54939/1859-1043.j.mst.77.2022.150-160

ABSTRACT

Classification of overwhelming heavy rainfall is a critical issue in the field of meteorology because it has extraordinary effects on people's lives and economies. Every year, a huge number of people all over the world suffer serious consequences from overwhelming precipitation events such as flooding and disease spreads. In this paper, we use Genetic Programming (GP) to predict if there will be heavy rain the next day. GP is an evolution-based machine learning methodology that can identify the model's functional form as well as its numerical coefficients. Our model was trained and evaluated on a data set collected from 17 stations in Vietnam's provinces. The experimental results show that models constructed using GP can perform better for heavy rain classification than models using other popular machine learning methods.

Keywords: Genetic Programming; Imbalance classification; Heavy rain classification.

1. INTRODUCTION

The phenomenon of heavy rains is a consequence of several special types of weather such as storms, tropical depressions or tropical convergence strips, and strong winds converging on many floors, cold fronts, and dotted lines. The combination of these factors at the same time will be more dangerous than cause heavy rain, strong winds, thunderstorms, and hails for a long time over a wide range of lands [1].

Among the meteorological factors in Vietnam, the rain element is the most different. The rainfall depends much on the terrain factors, and rainfall events are different in the amount and duration of rain. In Vietnam, there is currently a lack of ground observation stations, the number of automatic observation stations is even rare. This causes difficulties for forecasting and research work on rainfall [2].

In 2008, it was recorded the heaviest rain event over the last 100 years in the North and Northern areas of Vietnam. This unforecastable rainfall lasted for many days and caused a historic flood in Hanoi-the capital of Vietnam-and widespread flooding in the northern central provinces. Consequently, a lot of people died, and the loss of property is estimated at VND 3,000 billion or US $150 million. Therefore, early prediction of heavy rains is highly important.

There are various techniques that are commonly used when dealing with data classification tasks, such as K-nearest neighbor, artificial neural networks, naïve bayes. An alternative approach for this problem is genetic programming (GP). GP is a powerful evolutionary approach and a promising machine learning and searches technique for classification. It has successfully solved a wide range of classification problems [3]. In addition, the cost adjustment in GP can be applied to the fitness function [4]. Our approach focuses on modifying GP algorithms to overcome the problem of the imbalanced data class. Other machine learning (ML) techniques such as Support Vector Machine (SVM), Multilayer Perceptron (MLP), K-nearest Neighbor (kNN), and Random Forest were also implemented in order to compare to GP in rainfall classification problem.

The rest of this paper is sorted out as it takes after. Section 2 states the problem clearly that explains what we will address in this work. Section 3 provides a detailed literature review on

imbalanced data problems and discusses the advantages and disadvantages of each solution. Section 4 gives a proposed method on GP and ML methods utilized in this paper. Our experiment design and parameter setting are depicted in section 5. The results and analyses are given in section 6. Section 7 states conclusions and future work.

2. PROBLEM DESCRIPTION

In this paper, we have taken the daily rainfall series of several cities in Vietnam. Each series contains rainfall recorded from 2016 to 2020. Our aim is to forecast if there is rainfall on the next day in those cities.

Construction of input/output pairs: Let $\{x_1, x_2, \ldots, x_N\}$ stands for a rainfall time series. It can be reconstructed into a series of delay vectors as

$$X_t = x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(k-1)\tau},$$

where $X_t \in \mathbb{R}^k$, τ is the delay time as a multiple of the sampling period, and k is the embedded dimension. Suppose that X_t is a vector, then the available historical data can be summarized as a set of pairs $\{X_t, x_{t+T+(k-1)\tau} : t = 1, .., n\}$.

Table 1 illustrates the classifications for heavy rainfall provided by National Centre for Hydro - Meteorological Forecasting for 12-h accumulated rainfall. In our study, we consider all cases with rainfall greater than 8mm as heavy rainfall, otherwise, it is no heavy rainfall.

Another important thing is that each year, the number of days with heavy rainfall is much smaller than the number of days with no rain or light rain. Therefore, the problem of classification, whether there is heavy rain or not, falls into the case of imbalanced data.

Table 1. *Classifications for heavy rainfall for a station.*

S.no	Terminology	Rainfall range (mm)
1	Heavy rainfall (HR)	8–25
2	Very heavy rainfall (VHR)	26-50
3	Extremely heavy rainfall (EHR)	Greater or equal to 50
4	No heavy rainfall	Less than 8

3. RELATED WORKS

A significant number of studies on classification using GP have been carried out in recent years [3]. To the best of our knowledge, an approach for GP based class pattern categorization is presented in Kishore et al.'s paper [5]. In this paper, a genetic programming classifier expression (GPCE) is evolved as a discriminant function for each class to describe the provided-class problem as a two-class problem. The GPCE has been programmed to detect samples from its own class and reject samples from other classes. For each GPCE, the strength of association (SA) metric is calculated to demonstrate how well it can detect samples from its own class. The results of the experiments show that GP can be used to classify multicategory patterns, and the results are determined to be satisfactory when compared to the MLC (Maximum Like Classifier).

In [6], the authors propose a new GP-based strategy for classifying multi-class microarray data sets. The structure of the proposed solutions gives its uniqueness. Every solution that deals with a c-class problem consist of c sub-ensembles (SEs), each of them has k trees. Every individual is made up of $c\ x\ k$ trees in this way. The SEs' outputs are determined via a weighted voting procedure that employs the outputs of the associated SE's k trees as arguments and the classification accuracies of the trees (on training data) as weights. They discovered that every SE learns with a positive to negative ratio of $\frac{1}{c-1}$, assuming an equal number of training points in each class.

In recent literature, researchers recast the receiver operating curve convex hull (ROCCH) maximization problem using multi-objective GP to achieve binary classification [7, 8]. In these studies, authors use an evolutionary multi-objective optimization framework, in which the true positive rate was maximized while the false positive rate was minimized [8]. They also study the performance of various evolutionary multi-objective optimization techniques in [9]. On the other hand, in [8], the disadvantage of non-dominated sorting in EMOAs for ROCCH maximization is also examined, and a novel convex hull-based sorting scheme without redundancy is presented to overcome the difficulties of issues. This paper also introduces an area-based selection strategy that maximizes the convex hull's area.

4. PROPOSED METHOD

In this section, we propose a recommended approach for tackling the challenges associated with imbalanced datasets to early predict heavy rain. Because of the imbalanced distribution of data, most classification judgments are biased towards the negative class, resulting in the misclassification of positive class samples. The major goal of this study is to increase the positive class classification accuracy by avoiding the disadvantages of existing approaches, as discussed in the previous section. Our proposed method improves GP based on analysis of the imbalance dataset.

a. Population

The GP technique starts with a population made up of random people. The problem can be used to determine the size of the beginning population. The well-known approach of ramped-half and-half is used to generate each population.

b. Fitness function

Our proposed fitness function for assessment is based on a weighted – average (the coefficient of correlation (CC)) of classification accuracies of the minority and majority classes:

$$fitness = CC * \frac{TP}{TP + FN} + (1 - CC) * \frac{TN}{TN + FP} \tag{1}$$

In equation (1), CC is calculated as:

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

where n is the length of the training set, $y_{pre,i}$ represents the predicted value and $y_{obs,i}$ represents the actual value for the i-th data point (time index). From equation 1, the $fitness$ denotes class membership in the range of [0, 1]. Naturally, values tending towards '0' indicating no heavy rain class, and values tending toward '1' indicate heavy rain class.

c. Elitism

The goal of elite selection is to maintain the fittest chromosomes. Furthermore, it is primarily utilized to ensure that high-quality chromosomes are not missed in the next generation when the population is updated. To keep our best fitness scores, we employ the elitism technique. After the elitist step is accomplished, we move to the next phase-generating, a new population.

d. Creating a new population

After each individual's fitness has been assessed, the next stage is to create a new population. Selection, crossover, and mutation are the three operations that make up the process of creating a new population. The operations for selecting parents and offspring generation are briefly explained as follows:

- *Selection* The tournament selection method is used to choose parents for the offspring

generation. To begin, we constructed a group consisting of a number of persons randomly selected from the current population. In addition, the best individual based on a fitness assessment was picked as the first parent to breed a new individual, and the second parent was chosen using the same approach. The purpose of parental selection is to differentiate between individuals and offer preference to the best of them as future parents.

- *Crossover* The crossover operator's goal is to produce a new offspring from a parent pair. It produces new children that are created from parts embodied in each parent; hence, crossover makes variation within the population. In the execution of standard crossover, firstly, two parents are chosen according to a selection strategy. After that, one subtree is randomly chosen in each parent. If the two chosen subtrees are compiled to the requirements (depth of the resultant children, syntactic closure property, etc.), the crossover operation is swapping them. At that point, the new offspring are included in the next generation [10].

- *Mutation* To start, a mutation point is randomly chosen. After that, the subtree rooted at the mutation point is expelled. A randomly created subtree is used to replace the outgoing one. The mutation operator is used in order to maintain genetic diversity in the population.

The crossover rate is still high, whereas the mutation rate is still low. The lower mutation rate keeps the population unpredictable and prevents chromosome recurrence, while the greater crossover rate keeps the optimum local solution to not converging too soon.

e. Terminating conditions

The following termination conditions were employed in our algorithm design. The search step is terminated if one of the following conditions is met:

- The fitness value has reached a global optimum;
- The total number of generations has reached its maximum.

After the search step is completed, the GP returns the best individual with the best fitness value on training data. The classification model based on the best individual is then compared to other models based on other modern classification algorithms in terms of their performance.

5. EXPERIMENTAL SETUP

In this section, an experimental environment is established for the proposed approach to demonstrate its performance on imbalanced datasets, keeping in mind that we used the original datasets in our experiments. 17 datasets collected from 17 stations were used to evaluate the performance of our proposed GP in comparison to 4 other algorithms: SVM, kNN, MLP, and REFTree. Evaluation measures such as Acc, Recall, GM, FM, MCC, and AUC were used to evaluate the effectiveness and efficiency of classification models. To assess whether there are significant differences in performance between models based on different algorithms, we performed non-parametric statistical tests such as Bonferroni–Dunn post-hoc test [11], Friedman test, and Wilcoxon Signed rank test. The implementation of GP within the explore employments the ECJ framework [12]. Table 2 shows the setting parameters for the proposed methods. Thirty independent runs of the GP were performed for each figure, yielding diverse solutions at each run. Independent runs imply a diverse seed of the pseudo-random number generator. We even utilize elitism in our evolutionary process, which is a duplicate of the most excellent individual to put into the next generation.

In machine learning, learning algorithms are frequently categorized into decision trees such as REFTree [13], instance-based classifiers such as k nearest neighbor (kNN) [14], and function-based classifiers such as a multilayer perceptron (MLP) and SVM [15]. Hence, four algorithms (REFTree, kNN, MLP, and SVM) are utilized to intercompare with each other in this problem. The primary four algorithms are executed beneath the WEKA [16], so we use SMO instead of SVM. The parameters after tuning these algorithms are shown in table 3.

Table 2. *GP parameter settings.*

Parameter	Value
Function set	+, -, x, / (protected division), sin, cos, SQRT, LOG
Variable terminals	all features
Constant terminals	Random float values
Population size	1024
Initialization	Ramped half-and-half
Generations	200
Crossover probability	60%
Mutation probability	30%
Reproduction rate	10%
Selection type	Tournament (size=7)

Table 3. *Optimal parameters using weka for the four models: SVM, k-NN, MLP, and RF for heavy rainfall classification.*

SVM		**kNN**	
SVM Type	epsilon-SVR	K	7
Cost	9.6974	Distance Function	Euclidean
Gamma	6.8399		
Kennel type	RBF		
Epsilon	0.001		
MLP		**REFTree**	
HiddenLayers	5	Minimum number of instances	2
LearningRate	0.3	Unpruned	no
Momentum	0.2	Number of folds	3
Epochs	500		

Table 4. *IR and total instances of imbalanced datasets.*

No.	**Dataset name**	**Code**	**IR(%)**	**No.**	Dataset name	Code	IR(%)
1.	BAC QUANG	48836	15.17	10.	BA TO	48/52	15.1
2.	QUANG HA	48838	11.32	11.	CUA ONG	48/83	9.76
3.	TAM DAO	48846	10.8	12.	HA TINH	48/86	9.49
4.	CAT TIEN	48852	15.17	13.	HUE	48/91	10.89
5.	KY ANH	48863	10.7	14.	QUANG NGAI	48/92	10.13
6.	A LUOI	48895	14.46	15.	MONG CAI	48/93	10.4
7.	NAM DONG	48899	13.20	16.	DONG PHU	48/94	12.29
8.	TAM KY	48/34	11.19	17.	THU DAU MOT	48/95	10.43
9.	TRA MY	48/50	17.17				

We test with the k values according to 100 days (equivalent to 3 months).

The following subsections provide a description of the datasets and evaluation measures used in the experiments.

a. Dataset description

In this study, we used a total of 17 datasets according to rainfall observation data of 17 provinces in Vietnam from 2014 to 2019. On each day, data were recorded 2 times at 0 am and 12 pm. The data was divided into 2 sets: training data (rainfall observations from 2014 to 2018)

and testing data (rainfall observations in 2019). Table 4 shows the dataset sources and their characteristics based on the imbalance ratio (IR).

b. Evaluation criteria

By convention, the data records of the minority class have the positive class label, and the data records of the majority class have the negative class label. Accuracy is used as the main criteria to assess the performance of a classification model. However, this measure is insufficient to deal with the imbalanced datasets. In this work, we used additional evaluation metrics to compare the performance of our proposed method to 4 other approaches. Table 5 shows a confusion matrix used in this paper to construct the relevant evaluation measures.

Table 5. *Confusion matrix.*

Predicted class / Actual class	Positive	Negative	Total
Positive	TP	FP	TP+FP
Negative	FN	TN	FN+TN
Total	TP+FN	FP+TN	N

Accuracy (Acc): This is the most common metric for assessing classification models. It is calculated as the percentage of cases that are correctly classified into the total number of cases.

$$Acc = \frac{\text{Number\_of\_correct\_predictions}}{\text{Total\_number\_of \_correct\_predictions}} \tag{3}$$

In terms of positive and negative accuracy, the following formula can be used:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

Sensitivity (SN) is often known as Recall or True Positive Rate (TPR): SN is the fraction of actual positive records that are accurately labeled to the total number of records classified as positive; a high SN rate implies that the class is successfully recognized.

$$SN = \frac{TP}{TP + FN} \tag{5}$$

Specificity (SP): True Negative Rate is another name for SP (TNR). The specificity refers to the classifier's capacity to recognize bad outcomes. In other words:

$$SP = \frac{TN}{TN + FP} \tag{6}$$

Precision (PR): The percentage of the total number of positive instances properly categorized to the total number of actual positive instances.

$$PR = \frac{TP}{TP + FP} \tag{7}$$

F-Measure (FM): When high performance on both positive and negative classes is required, FM is used.

$$FM = 2 \times \frac{PR \times SN}{PR + SN} \tag{8}$$

G-Mean (GM): When the data set is balanced, GM maximizes the classification accuracy of the total population. To put it another way, GM is only good if the classification accuracy of both negative and positive classes is good.

$$GM = \sqrt{SP \times SN} \tag{9}$$

Matthews Correlation Coefficient (MCC): The MCC is a correlation coefficient between the target and the prediction. It normally ranges from -1 to 1. -1 indicates that the reality and prediction are in perfect disagreement, while 1 indicates that they are in perfect agreement. As the MCC considers all four outputs of the confusion matrix, it can typically provide a more balanced model accuracy evaluation, even for imbalanced data sets.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (10)$$

AUC: AUC - ROC curve is a performance measure for classification problems at various thresholds settings. ROC is a probability curve, and AUC represents the degree or measure of separability. It tells how accurate models distinguish classes. Higher the AUC, better the accuracy of the model. Similarly, higher the AUC, better the model in distinguishing days with heavy rain and no heavy rain. The ROC curve is plotted with TPR against the FPR, where TPR is the y-axis and FPR is the x-axis. An excellent model has an AUC value close to 1, which means it has a good measurement of separability. A poor model has an AUC value close to 0, which means it has the worst measurement of separability. When AUC is 0.5, it means the model has no class separation capacity whatsoever.

6. RESULTS AND ANALYSIS

The experimental results of our experiments and various research studies are presented and discussed in this section. Firstly, we evaluate the performance of our proposed GP and 4 other algorithms on various imbalanced datasets and compare the performance of models using non-parametric statistical tests. Secondly, a discussion is conducted to analyze the results obtained from evaluation metrics.

a. Performance comparison of class imbalance classification algorithms

In this subsection, the performance of these 5 algorithms is examined using six different evaluation metrics. Figure 1 to figure 6 show the boxplot graphs of the Acc, Recall, FM, GM, MCC, and AUC of each method in 17 datasets.

Accuracy (Acc): The effectiveness of the GP on imbalanced datasets is examined and compared with SMO, MLP, kNN, and REFTree. The proposed GP method is superior to other methods in 7 of 17 datasets and achieves competitive Acc results (REFTree is 6, MLP is 2, SMO is 5). The average values of Acc and average ranks of algorithms in all datasets suggest that: GP is better than other algorithms on some problems. In particular, in terms of the average value of Acc, GP is better than kNN and REFTree, and its average value of about 0.88 is slightly smaller than those of MLP and SMO (0.89). This average accuracy suggests that GP-classifier is good enough to use. In order to statistically compare the effectiveness of the proposed algorithm with the 4 other algorithms, a nonparametric statistical test is conducted. Friedman test establishes an F -distribution value at 20.8 for a significance level of alpha = 0.05. It means that the test rejects the null hypothesis, and therefore, it can be said that there are statistically significant differences between the Acc results of the methods. The result of the Bonferroni–Dunn test for Acc of methods with $alpha = 0.05$ suggests that MLP, REFTree, SMO, and GP have no statistical difference. GP is better than kNN.

Recall: The experimental results show that the proposed GP method is better than other methods in all 17 datasets and achieves competitive recall results. GP approach performs better than the other approaches in terms of the average recall value and average ranking results. The proposed GP method with an average score of 1.0 is ranked first, the kNN method is ranked second with a score of 2.529. Regarding the Friedman test, the F-distribution value is 54.95 with a significance level of alpha = 0.05. The p-value for F(4; 54.95) is 0.0000 < alpha. This shows that the test rejects the null hypothesis, and therefore, it can be said that the recall results of the

compared methods are significantly different on the adopted imbalanced datasets. The results of Bonferroni–Dunn tests shown suggest that GP is better than all 4 machine learning methods.

F-Measure (FM)*:* The proposed GP method is better than the other methods in 16 out of 17 datasets and achieves competitive FM results. The GP approach performs better than the other approaches in terms of the average FM value and average ranking results. Moreover, the proposed GP approach achieves an average FM of 0.603 outperforming other algorithms. These results indicate the generalization characteristics of GP.

The proposed GP method with an average score of 1.059 is ranked first, the kNN method is ranked second with a score of 2.853. Regarding the Friedman test, the F-distribution value is 50.39 with a significance level of $alpha = 0.05$. The p-value for F(4; 50.39) is 0.0000 < alpha. This shows that the test rejects the null hypothesis, and therefore, it can be said that the FM results of the compared methods are significantly different on the adopted imbalanced datasets. The results of the application of the Bonferroni–Dunn test to FM for $alpha = 0.05$ show that all methods perform significantly worse than GP.

GM: The proposed GP method is better than the other methods in 7 out of 17 datasets and achieves competitive GM results. The GP approach performs better than the other approaches in terms of the average GM value and average ranking results. Moreover, the proposed GP approach achieves an average GM of 0.692 outperforming other algorithms. These results indicate the generalization characteristics of GP.

The GP method with an average score of 2.111 is ranked first, the MLP method is ranked second with a score of 2.333. Regarding the Friedman test, the F-distribution value is 14.18 with a significance level of alpha= 0.05. The p-value of F(4; 14.18) is 0.007< alpha. This shows that the test rejects the null hypothesis, and therefore, it can be said that the GM results of the compared methods are significantly different on the adopted imbalanced datasets. The results of the application of the Bonferroni–Dunn test to FM with alpha = 0.05. Observing this figure, the SMO performs significantly worse than GP.

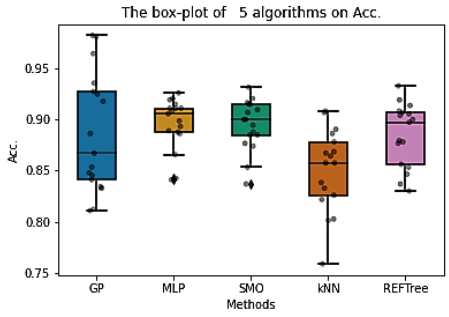


Figure 1. *The box-plot of 5 algorithms on Acc.*

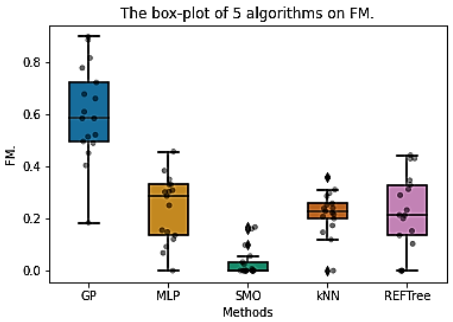


***Figure** 2. The box-plot of 5 algorithms on FM.*

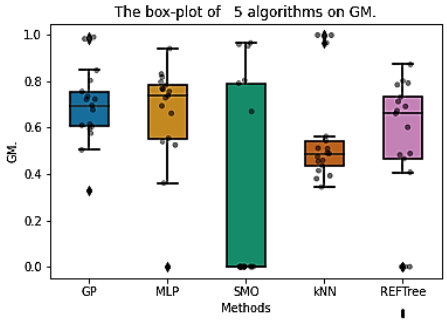


Figure 3. *The box-plot of 5 algorithms on GM.*

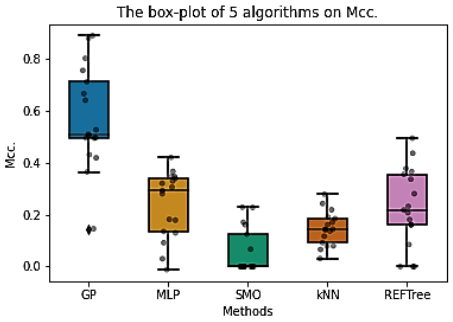
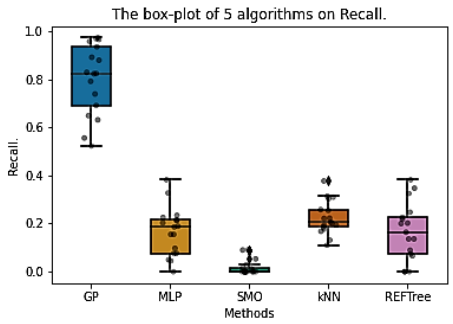


Figure 4. *The box-plot of 5 algorithms on Mcc.*



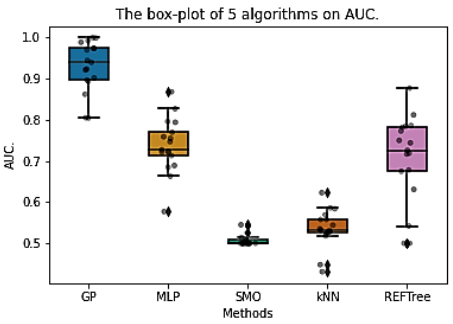


Figure 5. *The box-plot of 5 algorithms on Recall.* ***Figure 6.*** *The box-plot of 5 algorithms on AUC.*

MCC: The proposed GP method is better than other methods in 16 out of 17 datasets and achieves competitive Mcc results. The GP approach performs better than the other approaches in terms of the average Mcc value and average ranking results. Moreover, the proposed GP approach achieves the average Mcc of 0.884 outperforming other algorithms. These results indicate the generalization characteristics of GP. The GP method with an average score of 1.118 is ranked first, the MLP method is ranked second with a score of 2.735. Regarding the Friedman test, the F-distribution value is 47.45 with a significance level of alpha = 0.05. The p-value of F(4; 47.45) is 0.0000 < alpha. This shows that the test rejects the null hypothesis and therefore, it can be said that the Mcc results of the compared methods are significantly different on the adopted imbalanced datasets. The GP performs significantly better than all 4 methods.

AUC: The proposed GP method is better than other methods in all of 17 datasets and achieves competitive AUC results. The GP approach performs better than the other approaches in terms of the average AUC value and average ranking results. Moreover, the proposed GP approach achieves an average AUC of 0.884 outperforming other algorithms. These results indicate the generalization characteristics of GP. The GP method with an average score of 1.0 is ranked first, the MLP method is ranked second with a score of 2.412. Regarding the Friedman test, the F-distribution value is 58.73 with a significance level of $alpha = 0.05$. The p-value for $F(4; 58.73)$ is $0.0000 < alpha$. This shows that the test rejects the null hypothesis and therefore, it can be said that the AUC results of the compared methods are significantly different on the adopted imbalanced datasets. The results of the application of the Bonferroni–Dunn test to AUC with $alpha = 0.05$, GP performs significantly better than other methods.

Table 5 shows the comparisons of GP and the 4 machine learning methods on the six evaluation measures. In the table, the "1" symbol refers that GP is better than the method to be compared with, whereas the "0" symbol indicates that GP is equivalent to the compared method. From this table, we notice that there is almost no obvious difference between GP and **4** methods. However, the values in the Recall column and the AUC column confirm that the GP classifier can classify more properly in the case of minority classes. This indicates the superiority of our proposed model on highly imbalanced datasets.

Table 5. *Comparisons between GP and the 4 other machine learnings on 6 assessment metrics.*

	Acc	**Recall**	**FM**	**GM**	**Mcc**	**AUC**
MLP	0	1	0	0	1	1
kNN	1	1	0	0	1	1
SMO	0	1	1	1	1	1
REFTree	0	1	0	0	1	1

7. CONCLUSION AND FUTURE WORKS

This paper proposes a method that uses GP to build heavy rainfall classifiers. The proposed method was designed and evaluated on the datasets collected from the 17 observation stations in Vietnam. The experiments compared GP with other common machine learning methods on six evaluation metrics: Acc, Recall, FM, GM, Mcc, and AUC.

The results show that the proposed GP approach performs better than the other methods in terms of efficiency regarding chosen evaluation metrics, especially with measures that emphasize the classification in minority classes. Hence, the GP model classifier provides very good guidance for the weather forecaster to issue heavy rainfall warnings. However, in the future more cases, including experiments with more k values and more data, need to be investigated.

In order to evolve more accurate models, the future work also includes the investigation of advanced Genetic Programming such as semantic Genetic Programming to evolve heavy rainfall classifier models.

Acknowledgment: *The research presented in this paper was supplied data by Vietnam National Centre for Hydro - Meteorological Forecasting Center Hanoi, Vietnam.*

REFERENCES

[1]. IFRC, *"World Disaster Report 2020,"* International Federation of Red Cross and Red Crescent Societies, 2020.

[2]. V. D. B. D. N. L. N. L. Duc, *"Research and quantitative ranifall forecasting from HRM and GSM model products,"* Vietnam Journal of Hydrometerorology, vol. 592, pp. 17-27, 2010.

[3]. S. V. a. F. H. P. G. Espejo, *"A Survey on the Application of Genetic Programming to Classification,"* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 2, pp. 121-144, 2010.

[4]. N. S. A. B. A. Kumar, *"A novel fitness function in genetic programming for medical data classification,"* Journal of Biomedical Informatics, vol. 112, 2020.

[5]. L. M. P. V. M. a. V. K. A. J. K. Kishore, *"Application of genetic programming for multicategory pattern classification,"* IEEE Transactions on Evolutionary Computation, vol. 4, no. 3, pp. 242-258, 2000.

[6]. C. K. H. Liu, *"A genetic programming-based approach to the classification of multiclass microarray datasets,"* Bioinformatics, vol. 25, no. 3, 2009.

[7]. K. T. T. W. E. T. X. Y. P. Wang, *"Multiobjective genetic programming for maximizing ROC performance,"* Neurocomputing, vol. 125, pp. 102-118, 2014.

[8]. M. E. R. L. K. T. T. B. a. X. Y. P. Wang, *"Convex Hull-Based Multiobjective Genetic Programming for Maximizing Receiver Operating Characteristic Performance,"* IEEE Transactions on Evolutionary Computation, vol. 19, no. 2, pp. 188-200, 2015.

[9]. A. P. S. A. a. T. M. K. Deb, *"A fast and elitist multiobjective genetic algorithm: NSGA-II,"* IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182-197, 2002.

[10]. J. R. Koza, *"Genetic Programming"*. On the Programming of Computers by Means of Natural Selection, Massachusetts: MIT Press, Cambridge, 1992.

[11]. O. J. Dunn, *"Multiple comparisons among means,"* J. Amer. Stat. Assoc., pp. 52-64, 2012.

[12]. S. a. P. L. a. B. G. a. P. S. a. S. Z. a. B. J. a. H. R. a. C. A. Luke, *"Ecj: A java-based evolutionary computation research system,"* Downloadable versions and documentation can be found at the following url: http://cs. gmu. edu/eclab/projects/ecj, 2006.

[13]. J. R. Quinlan, *"C4. 5: programs for machine learning"*, Elsevier, 2014.

[14]. Y. a. L. J. a. L. J. a. Z. X. Song, *"An efficient instance selection algorithm for k nearest neighbor regression,"* Neurocomputing, vol. 251, pp. 26-34, 2017.

[15]. R. M. a. L. E. I. Balabin, *"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, pp. 1703-1712, 2011.

[16]. M. a. F. E. a. H. G. a. P. B. a. R. P. a. W. I. H. Hall, *"The WEKA data mining software: an update,"* ACM SIGKDD explorations newsletter, vol. 11, pp. 10-18, 2009.

[17]. H. M. Doucette J., *"GP Classification under Imbalanced Data sets: Active Sub-sampling and AUC Approximation,"* in Genetic Programming. EuroGP 2008. Lecture Notes in Computer Science, Heidelberg, Springer, Berlin, Heidelberg, 2008, pp. 266-277.
[18]. N. A. ArvindKumar, *"A novel fitness function in genetic programming for medical data classification,"* Journal of Biomedical Informatics, vol. 112, no. 103623, pp. 1-6, 2020.
[19]. H. B. M. B. K. S. T. &. L. S. W. Madsen, *"Data assimilation in rainfall-runoff forecasting,"* in Hydroinformatics 2000, 4th International Conference of Hydroinformatics, 23–27 July 2000, Cedar Rapids, Iowa, USA, 2000.
[20]. J. P. &. M. H. Drécourt, *"Role of domain knowledge in datadrivenmodeling,"* in Proceedings 4th DHI Software Conference & DHI Software Courses. 6–8 June 2001, DHI, Helsingør, Denmark, 2001.
[21]. P. A. &. C. P. F. Whigham, *"Modelling rainfall-runoff using genetic programming,"* Math. Comput. Modell., vol. 33, pp. 707-721, 2001.
[22]. S. T. K. E. C. &. P. O. Khu, *"An evolutionary-based real-time updating technique for an operational rainfall-runoff forecasting model,"* Complexity and Integrated Resources Management, Trans., vol. 1, pp. 141-146, 2004.
[23]. J. R. P. J. S. J. &. R. D. Rabuñal, *"Determination of the unit hydrograph of a typical urban basin using genetic programming and artificial neural networks,"* Hydrol. Process, vol. 21, no. 4, p. 476–485, 2004.
[24]. K. Rodríguez-Vázquez, *"Genetic programming in time series modelling: an application to meteorological data,"* in Proceedings 2001 Congress on Evolutionary Computation, Seoul, Korea, 2001.
[25]. Guven, *"Linear genetic programming for time-series modelling of daily flow rate,"* J. Earth Sci., p. 137–146., 2009.
[26]. W. N. P. K. R. E. &. F. F. D. Banzhaf, *"Genetic Programming: An Introduction.",* California: Morgan Kaufmann, 1998.
[27]. S.-Y. G. T. R. K. S. T. B. V. K. M. &. M. N. Liong, *"Genetic programming: a new paradigm in rainfall–runoff modelling,"* J. AWRA 38, pp. 705-718, 2002.
[28]. D. A. W. G. A. &. D. J. Savic, *"A genetic programming approach to rainfall–runoff modelling,"* Wat. Res. Mngmnt., p. 219–231, 1999.

TÓM TẮT

Sử dụng lập trình di truyền phát hiện mưa lớn ở một số tỉnh Việt Nam

Bài toán phát hiện mưa lớn là một bài toán quan trọng trong lĩnh vực khí tượng học vì nó ảnh hưởng lớn đến các bài toán khác và đời sống và kinh tế của con người. Mỗi năm, hàng triệu người ở nhiều nơi trên thế giới phải chịu đựng những thiệt hại to lớn của mưa lớn như nước dạng làm bệnh tật lây lan, thiệt hại kinh tế, v.v. Lập trình di truyền (GP) là một phương pháp học máy dựa trên sự tiến hóa có thể xây dựng mô hình dưới dạng hàm của các thuộc tính. Bài báo đã thử nghiệm mô hình này trên tập dữ liệu 17 trạm đo ở các tỉnh của Việt Nam. Kết quả thử nghiệm cho thấy rằng lập trình di truyền có thể tiến hóa ra mô hình chính xác hơn so với các phương pháp máy học phổ biến khác như mạng nơ-ron, k láng giềng gần nhất, máy hỗ trợ véc-tơ, rừng ngẫu nhiên khi xác định có xảy ra mưa lớn hay không.

Từ khóa: Lập trình di truyền; Phân lớp với dữ liệu không cân bằng; Phân lớp mưa lớn.