EasyChair Preprint

№ 2898

Comparative Analysis Of Machine Learning Algorithms On Landslide

Deeba Kannan, Trayambak Kumar and Sagar Krishna Kashyap

March 8, 2020

COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS ON LANDSLIDE

Mentor Name : Mrs. K. Deeba

Department of Computer Science and Engineering

SRM Institute of Technology, KTR

Trayambak Kumar

Department of Computer Science and Engineering

SRM Institute of Technology, KTR

Sagar Krishna Kashyap

Department of Computer Science and Engineering

SRM Institute of Technology, KTR

ABSTRACT

Landslide is one of the continuous geological disorders during rainy season, which make property damage and economic losses in all part of the world. In worldwide natural disorders landslide take the responsible is 17%. The frequent landslide occurrence has been increased by global climate change which causes losses and damages are increased. Therefore, automatic and accurate prediction of landslide occurrence is important to reduce the damages and losses of property. Since a lot studies have been carried out on landslides and their reduction, this field has seen a lot of positive progress. Landslide incidences are taken as dataset along with associated triggers. We apply Decision Tree (DT) and Random Forest Classifier (RF) and through our experimental evaluations find the suitable algorithm for the proposed work.

Keywords: Landslide prediction, Decision Tree, Random Forest, machine learning.

I. INTRODUCTION

The natural risks managed inside the Engineering Geology scope, influence both, the social and the monetary parts of human lives. Unsafe marvels inconvenience at various scales, with various interims and perseverance, leaving the various results. The techniques of their administration need to treat the current hotspots, however in any case to manage possibly new ones by foreseeing their conduct, volume and seriousness, before their latent capacity activating. In this, one of the most across the board unsafe marvels is to be considered. This tends to landslides and the same mass developments for example their susceptibility.

Landslide's assessment had been illustrated in versatile techniques in various case studies, yielding more or less reliable results depending on the complexity of the approach. The central idea of all the studies implies the processing of input geo-parameters into a single final model through various weighting and interpolating methods. However, the latter is characterised as quite accurate when combined with other techniques, and the closest to the original geotechnical assessment. Expert based techniques combined with machine learning have been shown to yield better solutions for the regional problems. In regional studies, certain generalisation is necessary, so direct modelling could be extremely

time-consuming and inefficient unlike machine learning trained over an expert-based model. Hence, a proper reconstruction of the final model is possible with sparse geo-inputs.

The decontrolled urbanisation process in the huge urban areas forces changes in the ecological nature balance. Numerous zones close to the inclines are involved by a huge piece of populace and these occupations advance the deforesting, soil vegetable spread wrecking, trash aggregation, and so on.

This paper proposes the two machine learning models i.e. Decision Tree and Random Forest for predicting the landslide; finally presents the quantitative evaluation i.e. accuracy, MSE, MAE, R-squared and RMSE for analysing the different models.

II RELATED WORKS

In [1] Marjanovic et al proposed the support vector machine (SVM) and K-nearest neighbour models for assessment of landslide susceptibility. Next Analytical Hierarchy Process (AHP) for weighting influences of different input parameters. Finally, landslide features is given to SVM and KNN algorithms. Al last, landslide susceptibility assessment by SVM with Gaussian kernel achieved the accuracy is 88%.

In [2] Kadavi et al presented landslide susceptibility mapping based on various ensemble based machine learning models such as Adaboost, LogitBoost, Multiclass Classifier and Bagging models. Thus maps are calculated based on their models and validated by area under the curve (AUC) model. Finally, multiclass classifier achieved AUC is 85.9 %, Bagging achieved AUC is 85.4 %, LogitBoost based AUC is 84.8% and AdaBoost achieved AUC is 84.0 %.

In [5] Chao Shen et al proposed genetic algorithm (GA) and support vector regression (SVR) to predict the displacement of the rainfall-induced landslide, which is to provide some reference value for landslide prediction.

III PROPOSED WORK

This paper presents the two different machine algorithms i.e. DT and RF for prediction or classification of landslide based on data attributes. Fig.1 shows the flow of process for landslide prediction follow as,

i. Landslide dataset is collect from the Global Landslide Catalog.

ii. The dataset has some NaN values; it is needed to replace by numerical values. This process is done in preprocessing step.

iii. After that, our dataset is split into training and testing data for validation.

iv. Finally, machine learning models such as Decision Tree and Random Forest are proposed to build a predictive model and make comparison between two machine learning algorithms based on parameters such as MAE, MSE, R-squared, RMSE and accuracy.

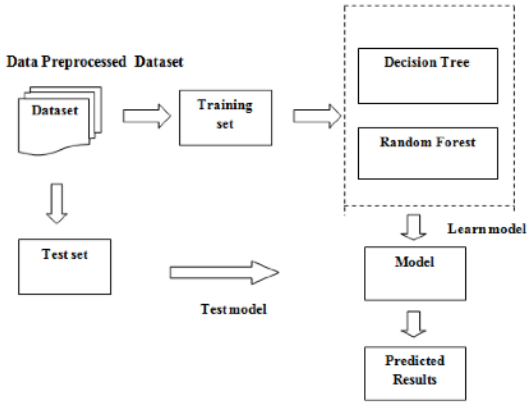

Fig.1 Overall Process for Landslide Prediction

1. Dataset Collection

The landslide dataset is obtained from the Global Landslide Catalog; it is published by NASA. This dataset contains many landslide incidences with corresponding trigger. This dataset has total 22 features, in that we choose 11 features for our work.

Feature	Description
event_month	The calendar month of the landslide incident
event_time	Time at which the landslide event took place
landslide_category	The type of landslide movement – slide (rock slide, debris slide, earth slide), creep, debris_flow, earth_flow, snow_landslide, lahar, rock_fall, earth_fall, complex (combination of two or more of the categories)
landslide_size	The general size of the landslide (small, medium, large, and very_large)
fatality_count	The number of fatalities due to the landslide
injury_count	The number of injuries that took place due to the landslide
country_name	Name of the country where the landslide occurred
population	The population count of the location at which the landslide took place
longitude	Exact longitude of the landslide location
latitude	Exact latitude of the landslide location
landslide_trigger	The trigger that caused the landslide – rain, construction, earthquake flooding, freeze-thaw, mining, monsoon, snow, and tropical cyclone

Table.1 Dataset Features Details

i. Data Visualisation

A large amount of information represented in graphic form is easier to understand and analyse. In our approach, the detection rates of landslide is shown as data visualisation part.

Fig.2. Data Visualisation of Country wise Landslide

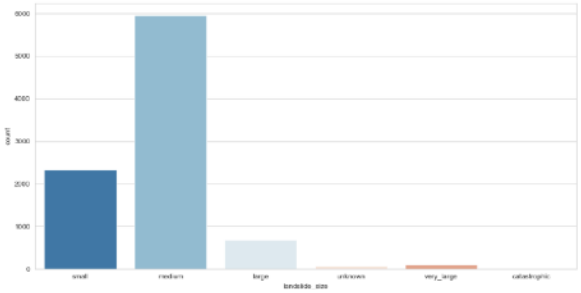

Fig.3. Data Visualisation of dataset based on landslide size statistics

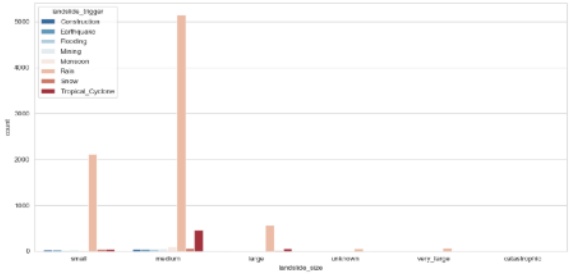

Fig.4. Data Visualisation of dataset based on landslide trigger statistics

2. Data Preprocessing

The dataset contains original attributes and NaN values. In programming, we cannot the process the NaN values so these values are transform into another value i.e. numerical value. NaN values are replaced by the mean value of columns.

3. Data Splitting

The splitting step is used for creating the training and testing data to analysing process. In that, our whole dataset is divide into training and testing data; use 80% of data for training and 20% of data for testing.

4. Prediction Model

In prediction, split training and testing data are evaluated based on machine learning models. First, training data was trained by using two different machine learning models such as Decision Tree and Random Forest. After that testing data are validated based on trained data with high classification accuracy rate. Two different algorithms are explained details given as follows,

A. Decision Tree

Decision Tree is one of the supervised learning algorithms. Mostly classification problems are solved by using decision tree. It easily performs with continuous and categorical attributes. Based on significant predictors, the population is dividing into two or more similar set in DT. The first step of DT is calculating entropy for each and every attribute. Next, based on the variables/ predictors the dataset is split with high information gain or less entropy. Above two steps are followed to remaining attributes.

$$Entropy\,(E) = \sum_{k=1}^{l} -q_k \log_2 q_k \qquad (2)$$

where l is refers to response variable modules count, q_k is the ratio of the count of the kth class procedures to a whole count of models.

$$Gain\,(E,G) = Entropy\,(E) - \sum_{v1\epsilon\ Values(G)} \frac{|E_{v1}|}{E} Entropy\,(E_{v1})$$

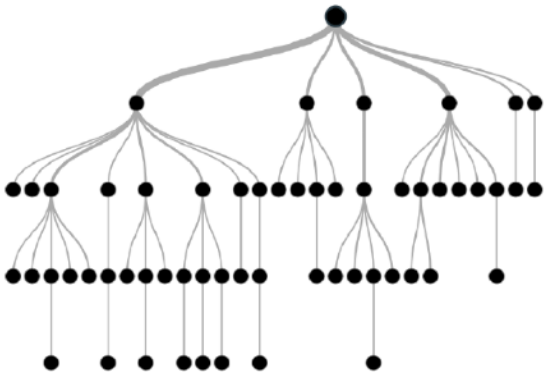

Fig.5. Decision Tree Structure

B. Random Forest

In supervised machine learning models, Random Forest is also one popular model. RF is worked for both classification and regression but it gives only better result to classification. In random forest, before getting the output or result many decision trees used. So, random forest is refers the combining of many decision trees. High number of trees would make the good result in RF. Voting system is used for classification and then decides the class whereas in regression it makes the mean prediction for all the outputs of each and every decision trees. Random forest easily and effectively worked with more number of dataset with high dimensionality.

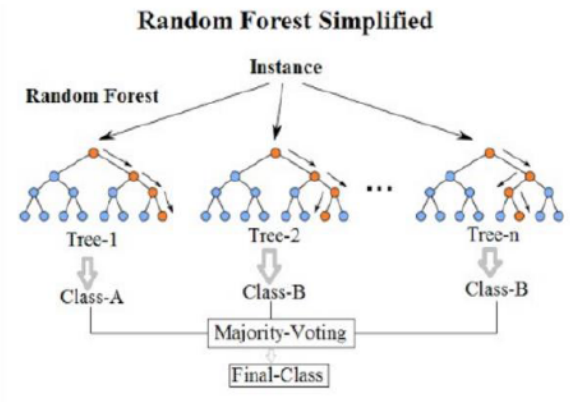

Fig.6. Random Forest Working Structure

IV EXPERIMENTAL RESULTS

In this part, we show the prediction result from various prediction models. We used different parameters for make comparison with two different models; the parameters i.e. Accuracy, R-squared value, Root Mean Squared Value (RMSE), Mean Absolute Value (MAE) and Mean Squared Error (MSE) Value.

The training dataset of landslide of all over world is considered with 11 features. Machine learning algorithm is applied such as decision tree and Random forest. We used two machine learning algorithm and identified Landslide trigger.

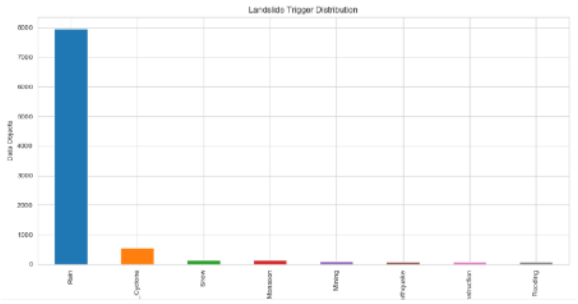

Fig.7. shows the distribution of landslide trigger

Table 2: Quantitative Evaluation with two different models

Algorithm	Accuracy (%)
Decision Tree	84.5
Random Forest	90.47

From the prediction result in table 2 we identify the random forest model give more accuracy rate is 90.47% than the decision tree.

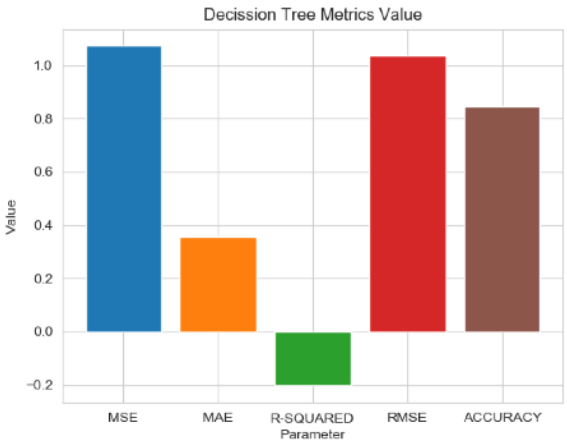

Fig.8. Result metrics of Decision Tree Algorithm

Fig.8 shows the performance analysis to Decision Tree.

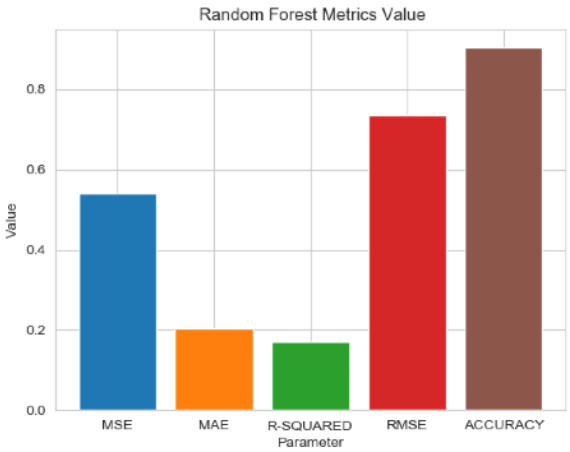

Fig.9. Result metrics of Random Forest Algorithm

Fig.9. shows the performance analysis to Random Forest.

V. CONCLUSION

In this study, we proposed landslide prediction as our major study with machine learning classification algorithm such as Decision Tree and Random forest. The real world dataset of landslide is taken for the study. There are number of parameters affecting the landslide, thus the trigger is considered as the prediction problem. The experimental results shown that random forest outperform in prediction accuracy than decision tree algorithm.

REFERENCES

[1] M. Marjanovic, B. Bajat, M. Kovacevic, "Landslide Susceptibility Assessment with Machine Learning Algorithms" 2009 International Conference on Intelligent Networking and Collaborative Systems, Barcelona, 2009, pp. 273-278.

[2] Mapping, Kadavi, Prima & Lee, Chang-Wook & Lee, Saro, "Application of Ensemble-Based Machine Learning Models to Landslide Susceptibility Mapping,"(2018), Remote Sensing. 10. 1252. 10.3390/rs10081252.

[3] Chae, Byung-Gon & Park, Hyuck Jin & Catani, Filippo & Simoni, Alessandro & Berti, Matteo. "Landslide prediction, monitoring and early warning: a concise review of state-of-the-art"(2017), Geosciences Journal. 21. 1033-1070. 10.1007/ s12303-017-0034-4.

[4] Li, Yupeng & Chen, Gengyun & Tang, C & Zhou, G & Zheng, Linjiang. "Rainfall and earthquake-induced landslide susceptibility assessment using GIS and Artificial Neural Network" (2012), Natural Hazards and Earth System Sciences. 12. 2719-2729. 10.5194/nhess-12-2719-2012.

[5] Chao Shen and Shengjun Xue, "Displacement Prediction of Rainfall-induced Landslide Based on Machine Learning" Journal of Coastal Research Special Issue 83 - Advances in Sustainable Port and Ocean Engineering: pp. 272 – 276, 2018.

[6] Pham, B. T.,Bui, D.T.,Dholakia, M.B., Prakash, I., Pham, H.V., Mehmood, K. and Le, H.Q., "A novel ensemble classifier of rotation forest and Naïve Bayer

for landslide susceptibility assessment at the Luc Yen district, Yen Bai Province (Viet Nam) using GIS," Geomatics, Nat. Hazards Risk, vol. 8, no. 2, pp. 649-671,2017.

[7] Pham, B. T. , Shirzadi, A., Tien Bui, D. , Prakash, I. and Dholakia, M. B., "A hybrid machine learning ensemble approach based on a Radial Basis Function neural network and Rotation Forest for landslide susceptibility modelling: A case study in the Himalayan area, India," Int. J. Sediment Res., vol. 33, no. 2, pp. 157-170,2018.

[8] Government of Sri Lanka, World Bank, UN Sri Lanka, and Global Facility for Disaster Reduction and Recovery, Sri Lanka Rapid Post Disaster Needs Assessment: Floods and Landslides, May 2017, no. May. 2017.

[9] Kadavi, P. R. , Lee, C. W., and S. Lee, "Application of ensemble-based machine learning models to landslide susceptibility mapping," Remote Sens., vol. 10, no. 8, pp. 1-18, 2018.

[10] Gariano, S. L. and Guzzetti, F., "Landslides in a changing climate," Earth-Science Rev., vol. 162, no. August 2016, pp. 227-252, 2016.[6] Pham, B. T. and Prakash, I., "A novel hybrid model of Bagging-based Naïve Bayes Trees for landslide susceptibility assessment," Bull. Eng. Geol. Environ., pp. 1-15, 2017.