IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-8, Issue-2, Feb 2020 E-ISSN: 2347-2693

Landslide Type Prediction using Random Forest Classifier

Harish Kumar N.G.[1], Pooventhiran G.[2*], Karthika Renuka D.[3]

[1,2,3]Dept. of Information Technology, PSG College of Technology, Coimbatore – 641 004, India

Corresponding Author: pooventhiran_g@icloud.com

DOI: https://doi.org/10.26438/ijcse/v8i2.711 | Available online at: www.ijcseonline.org

Accepted: 12/Jan/2020, Published: 28/Feb/2020

***Abstract*—** This paper talks about the prediction of types of landslides. It employs Random Forest Classifier technique, the ensemble version of Decision Trees. The results of the experiment show that ensemble techniques provide a better result compared to other algorithms. The dataset used here, in this paper, is Landslides After Rainfall dataset from NASA. This model achieves 59% accuracy without feature selection and 84% accuracy with feature selection.

***Keywords*—**Artificial Intelligence, Machine Learning, Decision Tree, Ensemble Learning, Random Forest Classifier.

I. INTRODUCTION

The term landslide also called, less frequently, landslip refers to mass wasting which includes wide range of ground movement, which could be rock falls, slope failures, debris flows, multi slides or mudflows. Submarine landslides are caused by landslides that occur underwater, coastal or onshore environments. The primary trigger for the occurrence of landslides is the weakest known force i.e. gravity; the other factor is the slope stability [1]. Generally, pre-conditional factors that build up sub surfaces or specific surface conditions that create a slope prone to failure, an external trigger which could generally be rainfall or earthquake is generally required before the actual landslide to occur [2].

II. IDENTIFYING NATURAL DISASTER

Safer environment is expected in a society so that we can live, prosper and sustain future generations. Basically, when thinking about the threats to our wellbeing, the first thing that comes to our mind is the over exploitation of water resources, followed by contaminating the water resources and soil loss. Despite the reason that Natural hazards cannot be easily controllable or avoided or in a lot of cases prediction goes short term process, it has profound influences on our safety, economic security, political stability and social development and even in individuals' overall wellbeing.

Natural hazards relate to the process that drives our planet. Natural hazards or geohazards include events such as earthquakes, volcanic eruptions, landslides and ground collapse, tsunamis, floods and droughts, geomagnetic storms, and coastal storms.

The key aspect of these natural hazards involves understanding and mitigating their impacts, which requires a geoscientist taking a four-sharp approach. It must include a fundamental understanding of the processes that cause the hazard, an assessment of the hazard, monitoring to observe changes in conditions that can be used to determine the status of a potential hazardous event, and maybe most importantly, delivery of information to a broader community to evaluate the need of action.

III. MACHINE LEARNING

Machine Learning is a tool for solving Artificial Intelligence (AI) problems. It does not require explicit programming; instead, it automatically learns and improves from experience. It focusses on developing computer programs that can access data and use it to train and learn by themselves.

It involves observations or data such as direct experience or instruction, so that it looks at the patterns in data and takes better decisions in future based on the examples provided [3,5]. Primary goal is to allow the system to learn automatically without human intervention but to adjust actions according to the situation.

IV. MACHINE LEARNING ALGORITHMS

Machine learning algorithms are majorly classified as supervised or unsupervised.

A. Algorithms based on Supervised Learning

These algorithms apply the learning that has been carried out with the past data using labelled examples to predict upcoming events. Start analysis to know a training dataset, the learning algorithm produces an inferred function to make predictions about the output values. New inputs will be provided with targets after sufficient training. The output of the machine learning algorithm can be compared with the right values, to get desired output and to modify the errors as required [4][7][8].

B. Algorithms based on Unsupervised Learning

These algorithms work fine when the training data are neither classified nor labeled. To develop a meaningful structure which is hidden from unlabeled data, we can use unsupervised learning techniques. The system is not going to generate the right output; instead it will derive all the hidden structures from the unclassified data.

C. Algorithms based on Semi-supervised Learning

These algorithms use both the supervised and unsupervised learning techniques. Both labelled and unlabeled data are used to train the data – generally it uses large amount of unlabeled data and small amount of labeled data. The learning accuracy will be improved for the systems that use this method. Usually when the labelled data require skilled and relevant resources to train or to learn from, we use semi supervised learning techniques.

D. Algorithms based on Reinforcement Learning

These algorithms interact with its environment by producing actions based on critics and rewards. The trial and error search and delayed reward are the characteristics of reinforcement learning. This lets the software agents to determine the ideal behavior automatically. This maximizes the performance of the model within the specific context. The agent requires simple feedback to learn the best action; this feedback is termed as the reinforcement signal. Machine learning techniques enables analysis of enormous amount of data. They deliver faster and accurate results in order to identify profitable opportunities and dangerous risks; they also may need additional resources and time to train it properly. Combining AI, and cognitive technologies with machine learning makes it more effective for processing large quantity of data.

V. ENSEMBLE LEARNING

It is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a computational intelligence problem. It is primarily used to improve the performance of a model (classification, prediction, function approximation, etc.), or reduce the likelihood of an unfortunate selection of a poor one. The other applications of ensemble learning include assigning a confidence to the decision made by the model, selecting optimal (or near optimal) features, data fusion, incremental learning, non-stationary learning and error-correction. This article focuses on classification related applications of ensemble learning; all principle ideas described below can be easily generalized to function approximation or prediction type problems as well.

A. Random Forest Classifier

This technique is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. When compared to other process, it is one of the most used algorithms, because of its simplicity and the fact that it can be used for both classification and regression tasks.

B. Working

This technique is a supervised learning algorithm. From the name itself it is clear that, it creates a forest and makes it somehow random. The forest that it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

To say it in simple words: Random forest algorithm builds multiple decision trees by picking input values in random and merges them together to get better, accurate and stable prediction.

The benefit of random forest is, that it can be used for both regression and classification problems, which form the majority of current machine learning systems. Random forest is talked about in classification, since classification is sometimes considered as the building block of machine learning. In fig.1 it can be seen how a random forest would look like with two trees:

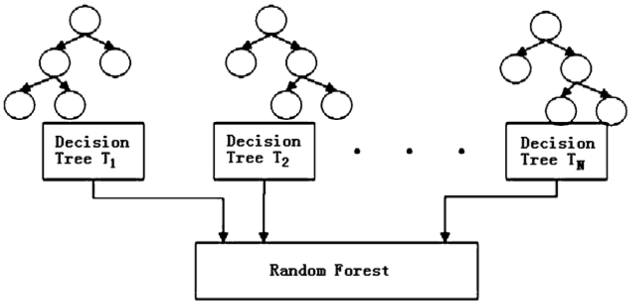


Figure 1. Random Forest Classifier

Figure 1. shows how the random forest algorithm classifies the dataset based on the generated model.

Random Forest is related to hyper parameters as a decision tree or a bagging classifier. Favorably, you don’t have to combine a decision tree with a bagging classifier and can just simply use the classifier-class of Random Forest. Like it is already proposed, with Random Forest, you can also deal with Regression tasks by using the Random Forest Regressor.

Random Forest adds external randomness to the model, while developing the trees. Instead of searching for the most important feature while separating a node, it searches for the best feature among a random subset of features. This results in a wide range that generally results in a better model.

Therefore, in Random Forest, only a random subset of the features is taken into deliberation by the algorithm for separating a node. You can even make trees more random, by externally using random thresholds for each feature rather than searching for the best available thresholds (like how a normal decision tree does).

C. Training Random Forest

For some number of trees *T*:

1. Sample *N* cases are selected in random with replacements to create subsets of the data. The subset must be about at least 66% of the total set.
2. At each node:
   a) For some number 'n', n predictor variables will be selected at random from the predictor variables.
   b) The predictor variable which provides the best split, according to an objective function, is used to do a binary split in the node.
   c) At the next node, choose another 'n' variable at random from all predictor variables for doing the same.

When a new input is entered, it runs down into the system, all of the trees. The result may either be an average or weighted average of the entire terminal nodes reached or in the case of categorical variables it is voted for the majority values.

- With enormous predictors, the eligible predictor will be set quite different from node to node.
- The random forest error rate will be great if the inter-tree correlation is high, so the trees are kept uncorrelated whenever it is possible.
- As n goes down, both inter-tree correlation and the strength of individual tree goes down. So optimal value of *'n'* must be identified.

D. Features

- Accuracy is 84%.
- It is suitable for large datasets.
- It can handle multiple input variables and no need of variable deletion.
- It gives estimation for the variables that were important for classification.
- Various methods are available for balancing error while populating the class in unbalanced datasets.
- Already generated forests can be used for future to use other datasets.
- Computed prototypes provide the information regarding the relationship between variables and classification.
- Experimental methods are offered for detecting the variable iterations.

VI. PROPOSED METHOD

Figure 2. explains the design of the proposed system in an elaborated manner. Initially landslide dataset is trained. The dataset will be cleaned and pre-processed. The processed dataset will be given as input to the Random Forest Classifier.

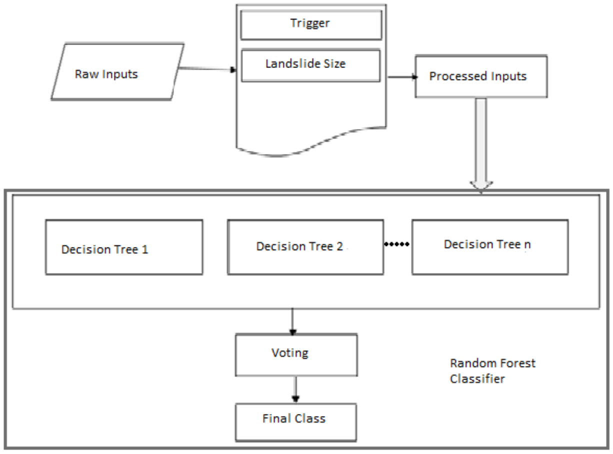


Figure 2. Proposed Model for predicting Landslides

A. Collection of Data

A dataset is a collection of data. Most commonly a dataset corresponds to the contents of single database table, a single statistical data matrix, where each column of the table represents a particular variable and each row represents given member of the dataset in question. The dataset lists values for each of the variable, such as height and weight of an object, for each member of the dataset. Each value is known as datum. The dataset comprises data of one or more members, corresponding to the number of rows. NASA Landslides After Rainfall dataset is collected from kaggle machine learning repository. The various attributes of datasets are

- id
- hazard_type
- landslide_type
- landslide_size
- trigger

Landslide dataset is the dataset taken for training and testing the Random Forest classifier where the classification done using Random Forest is the type of landslide.

B. Data Pre-processing

Data pre-processing is an important step in the data mining process and machine learning projects. Data gathering methods if not controlled strictly, will result in out of range values, impossible data combinations, missing values etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is the first and foremost before running an analysis.

If the data are more inconsistent and irrelevant with lots of noise and unreliable data, knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount in processing time. The various pre-processing steps include cleaning, transformation, normalization, feature extraction and selection etc. The final training dataset is the outcome of the product of pre-processing. In most cases, missing data should be pre-processed so as to allow the whole dataset to be processed by a supervised machine learning algorithm.

Moreover, most of the existing machine learning algorithms are able to extract knowledge from dataset that stores discrete features. If the features are continuous, the algorithms can be integrated with a discretization algorithm that transforms into discrete attributes. Decision tree algorithm forms the base of the Random forest classifier algorithm. It resembles a number of decision trees running again and again in loops since the nodes are calculated for the same dataset. By creating a greater number of decision trees, it is possible to model and develop the forest which is not going to use the same apache to construct decision-based model on the basis of information gain or gini index.

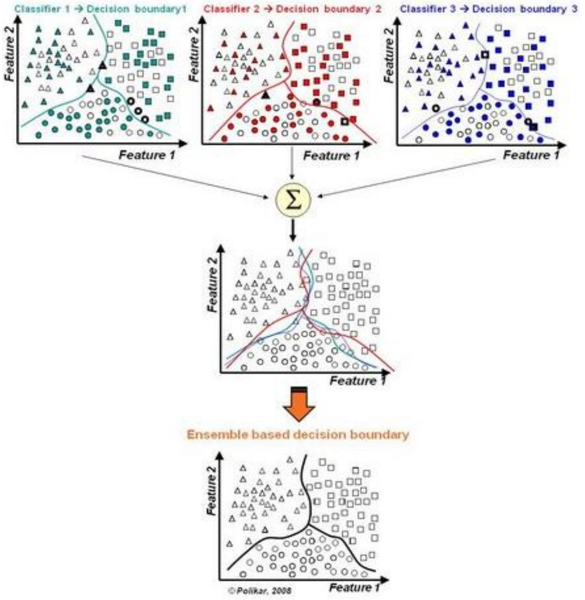

Classifier 1 → Decision boundary1
Classifier 2 → Decision boundary 2
Classifier 3 → Decision boundary 3
Feature 2
Feature 1
Feature 2
Feature 1
Feature 2
Feature 1
Σ
Feature 2
Feature 1
Ensemble based decision boundary
Feature 2
Feature 1
©Polikar, 2006

Figure 3. Decision Tree implemented with Random Forest Classifier

VII. EVALUATION MEASURES

A. Accuracy

Accuracy is the degree of the closeness of measurements of a quantity to that quantity actual value. It is an important measure to evaluate the efficiency of the system.

B. Precision

Precision is defined as the degree to which repeated measurements under unchanged conditions show the same results.

C. Recall

Recall is also called as sensitivity which measures the proportion of the actual positives which are correctly identified. Sensitivity of 100% means that the test recognizes all actual positives.

D. Confusion Matrix

Performance of a classifier is evaluated by the confusion matrix. In this matrix the number of incorrectly classified instances is sum of diagonals in the matrix.

		Predicted Label	
		Positive	**Negative**
Known Label	**Positive**	True Positive(TP)	False Negative(FN)
	Negative	False Positive(FP)	True Negative(TN)

For simplicity, the assumption is that each instance can only be assigned one of the classes: Positive or Negative. Each instance has a known label, and a predicted label. Some method used to make predictions on each instance. Each instance increments one cell in the confusion matrix. It gives a clear representation of the values which are used to evaluate the performance of classifiers through various measures such as accuracy, precision and recall respectively.

VIII.RESULTS

The proposed system has obtained many intermediate results and a final result. As a reference for these results, snapshots have been attached in this section.

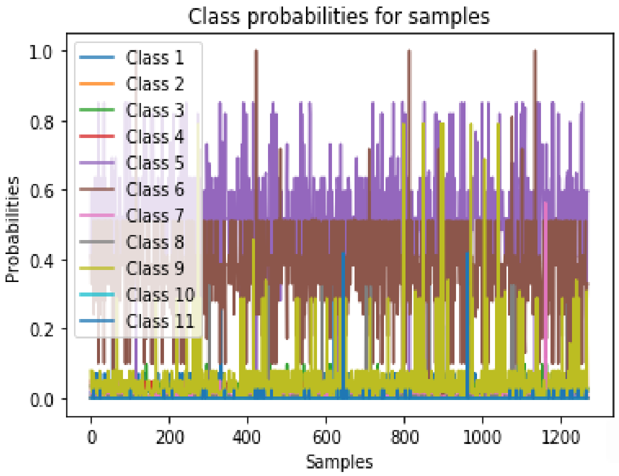
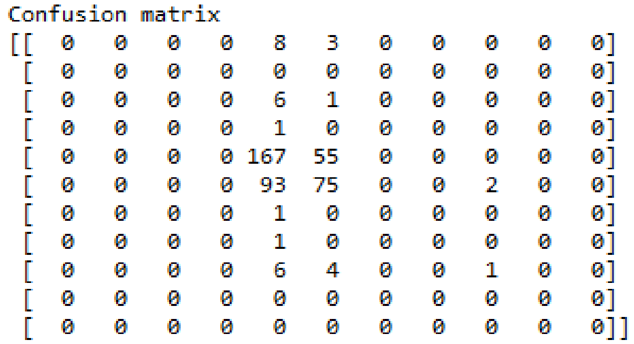


Figure 4. Obtained Probabilistic Curve

Figure 4. shows how the probability of each instance is distributed over the target classes. Each instance is assigned a class with which it shows the higher probability.

The model implemented, achieved an accuracy of about 84% when feature selection is enabled. By that way, only important features can be involved in classification. This is intuitive when there are a lot number of features since only the important features are selected and allowed in classification.

Without feature selection, as all features take part in classification, there is a probability that the unimportant (irrelevant) features affect the performance of the model. It can be seconded by our experiment without feature selection that achieved accuracy of only about 59% which is a significant drop.



```
Confusion matrix
[[  0   0   0   0   8   3   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   6   1   0   0   0   0   0]
 [  0   0   0   0   1   0   0   0   0   0   0]
 [  0   0   0   0 167  55   0   0   0   0   0]
 [  0   0   0   0  93  75   0   0   2   0   0]
 [  0   0   0   0   1   0   0   0   0   0   0]
 [  0   0   0   0   1   0   0   0   0   0   0]
 [  0   0   0   0   6   4   0   0   1   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0]]
```

Fig.5 Confusion Matrix

From Fig. 5, it can be seen how the model has predicted the classes with TP, FN, FP and TN. It also shows the distribution of prediction.

REFERENCES

[1] Gwo-Fong Lina, Ming-JuiChanga, Ya-ChiaoHuanga, b and Jui-Yi Hob, *"Assessment of susceptibility to rainfall-induced landslides using improved self-organizing linear output map, support vector machine, and logistic regression",* Elsevier., vol. **224**, pp. **62-74**, **May 2017**.

[2] J.N. Goetz, A. Brenning, H. Petschko and P.Leopold, *"Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling"*, Elsevier., vol. **81**, pp. **1-11**, **August 2015**.

[3] M.T. Brunettia, M. Melilloa, S. Peruccaccia, L. Ciabattaa, b, L. Broccaa, *"How far are we from the use of satellite rainfall products in landslide forecasting?"*, Elsevier., vol. **210**, pp. **65-75**, **June 2108**.

[4] MilošMarjanović, MilošKovačević, BranislavBajat, VítVoženílek, *"Landslide susceptibility assessment using SVM machine learning algorithm"*, Elsevier., vol. **123**, pp. **225-234**, **November 2011**.

[5] Milošmarjanovic;branislavbajat; miloškovacevic, *"Landslide Susceptibility Assessment with Machine Learning Algorithms"*, International Conference on Intelligent Networking and Collaborative Systems, pp. **273-278**, **December 2009**.

[6] Cheng Lian, Zhigang Zeng, Wei Yao and Huiming Tang, *"Performance of Combined Artificial Neural Networks for Forecasting Landslide Displacement"*, International Joint Conference on Neural Networks (IJCNN), pp. **418-423**, **July 2014**.

[7] Cheng Lian, C. L. Philip Chen, Zhigang Zeng, Wei Yao, and Huiming Tang, *"Prediction Intervals for Landslide Displacement Based on Switched Neural Networks"*, IEEE Transactions on Reliability., vol. **65**, pp. **1483-1495**, **September 2016**.

[8] Yu Huang, Lu Zhao, *"Review on landslide susceptibility mapping using support vector machines"*, Elsevier., vol. **165**, pp. **520-529**, **June 2018**.

[9] MunirahRadinMohd Mokhtar, Abdul Nasir Matori and Khamaruzaman, *"Observing Landslide Occurrence by GIS Approach"*, IEEE 4th Control and System Graduate Research Colloquium., pp. **21-26**, **August 2013**.

[10] Gege Jiang, Yuan Tian and Chenchao Xiao, *"GIS-based Rainfall-Triggered Landslide Warning and Forecasting Model of Shenzhen"*, 21st International Conference on Geoinformatics, **October 2013**.

AUTHORS PROFILE

Harish Kumar NG is a PG scholar pursuing Masters in Information Technology at PSG College of Technology, Coimbatore, India. He got graduated in Electronics and Communication Engineering from Anna University, Chennai, India He is also a freelance trainer and has handled sessions in various Engineering colleges across India. His areas of interest include Internet of Things, Machine learning.

Pooventhiran G is an Undergrad in Information Technology at PSG College of Technology, Coimbatore, India. He is a curious learner who craves learning and highly interested in research. He has published a paper on "Evolutionary Models in Software Engineering" to start off his research. His areas of interest include Machine learning and Graph Algorithms.

Dr. D. Karthika Renuka (Dhanaraj Karthika Renuka) working in Department of Information Technology since 2004. Area of specializations includes Data Mining, Evolutionary Algorithms, Soft Computing, Machine Learning and Deep Learning, Client Server Computing, Computer Networks, Information Security. Organized an International Conference on Innovations in Computing Techniques Jan 22- 24, 2015 (ICICT-2015) and National Conference on "Information Processing and Remote Computing" 27th and 28th February 2014 (NCIPRC 2014). Reviewer for Computers and Electrical Engineering, Elsevier, Wiley Book chapter, Springer Book Chapters on "Knowledge Computing and its Applications".

Ongoing Research Projects

- UGC-Minor Research Project (MRP) titled "Multiple Classifier for Email Spam Classification in a Distributed Environment for reducing Network Traffic and improving the Network Performance" for an amount of Rs.3,90,000.
- AICTE –MODEROBS (Data Analytics Tool) Rs. 2,00,000/-
- "Design and Implementation of E-Learning System using Deep Learning Based on Audio-Video Speech Recognition for Hearing Impaired in Native Language" funded by DST - ICPS, for an amount of 70,00,000/-