Solving Data Imbalance in Landslide Susceptibility Zonation

Sharad Gupta[1] and Dericks Shukla[1]

[1]Indian Institute of Technology Mandi

November 23, 2022

Abstract

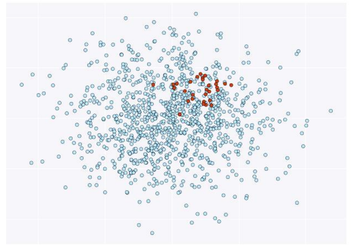
Landslides cause billions of dollars in property damage and thousands of deaths every year worldwide. India has more than 15% of its land area prone to landslides, hence mapping of these areas for the presence of landslides is of utmost importance. Landslide susceptibility zonation maps give approximate information about the occurrence of landslides. There are various factors responsible for slope instability. In this work, 11 causative factors have been considered such as Aspect, Elevation, Geology, Distance from thrusts, Distance from streams, Plan curvature, Profile curvature, Slope, Stream power index, Tangential curvature, Topographic wetness index. Machine learning methods such as artificial neural network, support vector machine require a large amount of training data; however, the number of landslide occurrences are limited in a study area. The limited number of landslides leads to a small number of positive class pixels in the training data. On contrary, the number of non-landslide pixels (negative class pixels) are huge in numbers. This under-represented data and severe class distribution skew create a data imbalance for learning algorithms and sub-optimal models, which are biased towards the majority class (non-landslide pixels) and have low performance on the minority class (landslide pixels). Generally, the data is imbalanced when the class ratio is of the order of 100:1, 1000:1 and 10000:1 (i.e., one-class points are 100, 1000 or 10000 times more than that of another class points). In our work, class ratio is more than 300:1 (i.e. for each one landslide pixel, we have more than 300 non-landslide pixels). Thus, we can clearly say that our data is imbalanced. There are two major data balancing techniques, which are oversampling of a minority class and under-sampling of majority class. The minority oversampling cannot be applied, as it will create false landslide pixels. We have performed under-sampling of non-landslide pixels using various techniques. We will discuss landslide susceptibility zonation with and without using data imbalance technique and show major improvements in accuracy over imbalanced learning.

Solving Data Imbalance in Landslide Susceptibility Zonation

Sharad K. Gupta, Dericks P. Shukla

INTRODUCTION

- Disproportionate ratio of observations in various classes.
- The data is imbalanced when the class ratio is of the order of 1:100, 1:1000 and 1:10000 (i.e., number of points in one-class are 100 times or 1000 times or 10000 times less than that in another class).



Blue – Non-Landslide
Red – Landslide Point

METHODS

Two major data balancing techniques

Over-sampling of a minority class	Under-sampling of majority class

Under sampling Methods (Liu et al. 2009)

Random under-sampling	Tomek links	Cluster centroids	**Balance Cascade**	**Easy Ensemble**

- EasyEnsemble Algorithm
- BalanceCascade Algorithm

} **Data Balancing**

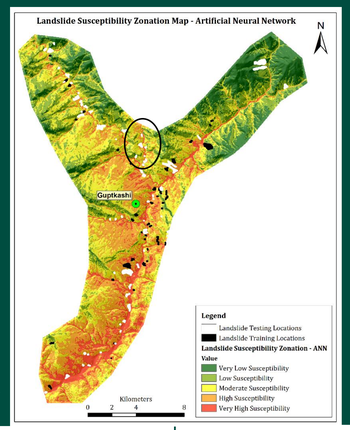
- Fisher Discriminant Analysis
- Logistic Regression
- Artificial Neural Network

} **Weightage Determination**

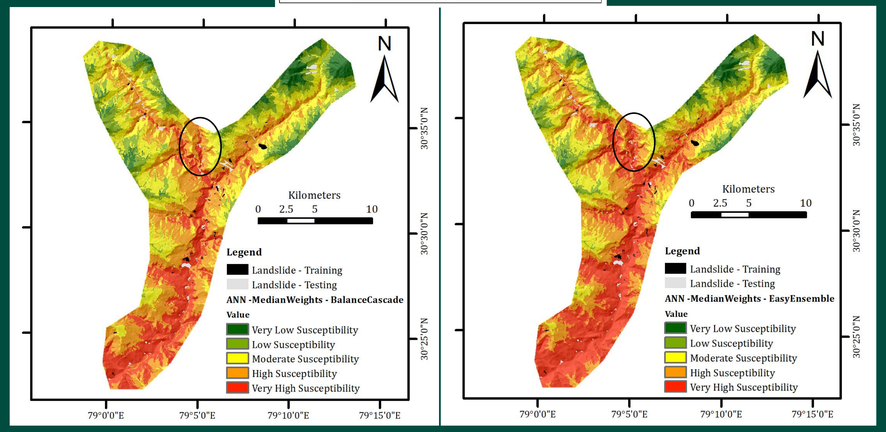
- Heidke Skill Score
- Recall (Sensitivity)

} **Accuracy Assessment**

Machine learning methods require data balancing whereas data driven methods do not need balancing.



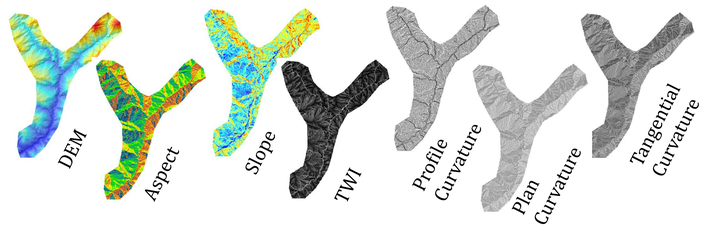
Imbalanced data



Balanced Data

Take a picture to download the **full paper**

DATA PROPERTIES



- The study area comprises of total 122 landslides occurred between 2004 and 2017
- Training - 46 landslides (1203 pixels) occurred from 2004 to 2012, Testing - 76 landslides (2744 pixels) occurred from 2013 to 2017

RESULTS

Table - 1: Statistics for all the three methods **(imbalanced data)**

Method	LR	FDA	ANN
Mean	0.58	0.55	**0.43**
Median	0.58	0.56	**0.42**
Standard Deviation	0.11	0.12	**0.17**

Table 2. Statistics for all the three methods **(Balanced data)**

Balancing Method	Statistical Quantities	LR	FDA	ANN
Easy Ensemble	Mean	0.3834	0.5558	0.5822
	Median	0.3870	0.5604	0.5948
	Std. Dev.	0.0775	0.1163	0.1364
Balance Cascade	Mean	0.2934	0.5518	0.5455
	Median	0.2960	0.5562	0.5582
	Std. Dev.	0.0565	0.1151	0.1268

Decreased No significant change Improved Significantly

CONCLUSIONS

- LR method is not able to model the underlying probability distribution after data balancing.
- The FDA method may or may not show major changes in the results after data balancing.
- Balancing algorithms must be applied before preparation of LSZ maps using machine learning methods. However the data driven methods do not need balancing as seen from the results.

ACKNOWLEDGEMENTS

I would like to heartily thank AGU for providing Austin Travel Grant for attending the fall meeting. I am also very grateful to HIMCOSTE for providing Partial Travel Support for presenting the paper.