DATA IMBALANCE IN LANDSLIDE SUSCEPTIBILITY ZONATION: UNDER-SAMPLING FOR CLASS-IMBALANCE LEARNING

S. K. Gupta1,*, Muskan Jhunjhunwala2, A. Bhardwaj1, D. P. Shukla1

1 School of Engineering, Indian Institute of Technology Mandi, India (*sharadgupta27@gmail.com)
2 Department of Computer Science, National Institute of Technology Hamirpur, India

KEY WORDS: Landslide susceptibility zonation (LSZ), Imbalanced learning, Under-sampling, Artificial neural network (ANN), Fisher Discriminant Analysis (FDA), Logistic Regression (LR)

ABSTRACT:

Machine learning methods such as artificial neural network, support vector machine etc. 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 amount of non-landslide pixels (negative class pixels) are enormous in numbers. This under-represented data and severe class distribution skew create a data imbalance for learning algorithms and suboptimal models, which are biased towards the majority class (non-landslide pixels) and have low performance on the minority class (landslide pixels).

In this work, we have used two algorithms namely EasyEnsemble and BalanceCascade for balancing the data. This balanced data is used with feature selection methods such as fisher discriminant analysis (FDA), logistic regression (LR) and artificial neural network (ANN) to generate LSZ maps. The results of the study show that ANN with balanced data has major improvements in the preparation of susceptibility maps over imbalanced data, whereas the LR method is ill-effect by data balancing algorithms. The FDA does not show significant changes between balanced and imbalanced data.

1. INTRODUCTION

Landslides are amongst the most devastating natural disasters, which 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/absence of landslides is of utmost importance. Numerous studies have contributed to reduce the damage from landslides through modelling and production of susceptibility maps. (Roodposhti et al., 2019, Jhunjhunwala et al., 2019, Gupta et al., 2018, Shukla et al., 2016). The susceptibility mapping can be a crucial tool for a wide range of end-users, from both private and public sectors, aimed at hazard mitigation purposes at both local and international levels. Landslide susceptibility zonation (a.k.a. LSZ) maps give approximate information about the occurrence of landslides. The susceptibility mapping requires data of various factors responsible for slope instability. In this work, we have considered seven causative factors such as aspect, elevation, plan curvature, profile curvature, slope, tangential curvature, topographic wetness index.

In recent years, there is an increasing application of machine learning techniques to complex real-world problems. The application ranges from daily life problems to nation’s security, processing of the information to decision making support system and from micro-scale analysis of data to macro-scale discovery of knowledge (Stumpf et al., 2012). Most standard machine learning algorithms presume or expect balanced class distributions or equal misclassification costs (He, G arcia, 2009) in the analysis. Data imbalance refers to a scenario where majority classes dominate or overpower minority classes. In simple words, there is disproportionate distribution of observations in each class. It leads to the classifier b eing m ore b ased t owards the dominating class (Pradhan et al., 2014). Generally, the data is imbalanced when the class ratio is of the order of 1:100, 1:1000 or 1:10000 etc. i.e. number of points in one-class are 100 times or 1000 times or 10000 times less than that in another class (Liu et al., 2009). The imbalance level can be as high as 1010. In this research, the class ratio is 1:300, i.e. for each landslide pixel we have more than 300 non-landslide pixels. When we use various machine-learning approaches for the generation of LSZ maps then the algorithms do not classify the landslide pixels correctly. Therefore, it is necessary to reduce the imbalance in the susceptibility mapping data. There are two major data balancing techniques, which are oversampling of a minority class and under-sampling of majority class (He, Garcia, 2009). The minority oversampling cannot be applied, as it will create false landslide pixels. We under-sample the majority class (i.e., non-landslide pixels) using EasyEnsemble and BalanceCascade methods. Some of the techniques used by various authors to overcome data imbalance are random oversampling and under-sampling, informed under-sampling, Synthetic sampling with data generation etc. (Hai xiang et al., 2017, Stumpf et al., 2014, Stumpf et al., 2012, Galar et al., 2012, Chawla, 2010, Liu et al., 2009, He, Garcia, 2009).

This work aims at first, balancing the data using two different methods (i.e., EasyEnsemble and BalanceCascade). This balanced data is used for computing the weights using various methods such as fisher discriminant analysis (FDA), logistic regression (LR) and artificial neural network (ANN) to generate LSZ maps. Furthermore, visual analysis, statistical quantities, Heidke Skill Score (HSS) and Recall is used to assess the quality of susceptibility maps. Based on these observations, we can make assertions as how accuracy is affected when balancing techniques are applied to our data w.r.t. imbalanced data (Jhunjhunwala et al., 2019).

2. STUDY AREA & DATA RESOURCES

A small part of Mandakini river basin of Garhwal Himalaya in Uttarakhand has been considered for the study as shown in Figure-1. Mandakini river originates from the Chorabari Glacier near Kedarnath in Uttarakhand, India. The study area covers about 275.60 sq. km area and lies between 30°19’00"N to 30°49’00"N latitude and 78°49’00"E to 79°21’13"E longitude. The study area falls in the Survey of India toposheet no. 53J and 53N. This region is highly prone to landslides during the monsoon season. This region has highly rugged topography, deep

*Corresponding author
gorges, high peaks where higher areas are mostly snow-
covered forming the U-shaped wide valleys of glacial land-
scape. The study area is highly prone to landslide as
every year many landslides occur in this area. In this re-
search we have used 30m shuttle radar topography mission
(SRTM) DEM, which was downloaded from EarthExplorer

The following data balancing algorithms have been applied
to the initial experimental data set to obtain a balanced
data set.

3.1 Data Balancing Algorithms

3.1.1 EasyEnsemble This method works on samples of
majority class. It reduces the number of observations from
majority class to make the data set balanced. This method
is best to use when the data set is huge and reducing the
number of training samples helps to improve run time and
storage troubles. EasyEnsemble is an example of informed
under sampling as it explore subsets of majority class by
independent replacement of subsets (He, Garcia, 2009, Liu
et al., 2009). This method is an unsupervised learning al-
gorithm as it explores subsets of majority class by inde-
pendent random sampling with replacement.

Algorithm - EasyEnsemble (Liu et al., 2009)

i Suppose, we have a majority class training data set
M and minority class data set as N, where |N| < |M|.
We can have T different randomly sampled
subsets of majority class as M1, M2,....., MT which
can be represented as M, where |Mi| << |M| &
|M| = |N|.

ii Easy Ensemble uses different weak classifiers Ci, the
number of classifiers being same as the number of
subsets of majority class. The classifier Ci is trained by
N minority class samples and M, majority class samples.
The decisions generated by weak classifi-
cers Ci are combined using AdaBoost ensemble (Ht)
using the Eq-1 (Liu et al., 2009)

\[ H_t(x) = \text{sgn} \left( \sum_{j=1}^{T} \alpha_{i,j} C_{i,j}(x) - \beta_i \right) \]  \hspace{1cm} (1)

where, \( I_t \) is the number of iterations, \( C_{i,j} \) is the weak
classifier/learner, \( \alpha_{i,j} \) is the weights of \( c_{i,j} \), \( \beta_i \) is
the ensemble’s threshold.

iii This uses boosting algorithm which is an repetitive
technique to tune weights of an observation based on
previous classification. This is useful in reducing the
bias error in machine learning models. Then instead
of voting the decisions of all the classifiers to select
one, boosting is used to combine the results of all T
classifiers and make an improved final decision (Liu
et al., 2009). The output ensemble is obtained as
Eq-2.

\[ H(x) = \text{sgn} \left( \sum_{t=1}^{T} I_t \sum_{j=1}^{T} \alpha_{i,j} C_{i,j}(x) - \sum_{t=1}^{T} \beta_t \right) \]  \hspace{1cm} (2)

3.1.2 BalanceCascade In this method we create several
subsets of data which are balanced, and a weak classi-
fer is trained for each subset. This method reduces the
majority class training sets at every step by removing all the
examples that are correctly classified. It is different from
Easy Ensemble in two steps. Primarily the weights are
adjusted based on false positive rates that a classifier have
to achieve. Second, the samples are removed which are
correctly classified. This sequential dependence mainly
focusses on reducing the redundant information in majority
class.

In Balance Cascade, the training can finally be stopped.
when size of majority class (|M|) is less than size of minority class (|N|), as size of majority class is getting shrunk at every iteration. The main advantage of Balance Cascade is that it generate the restricted sample space to extract as much useful information possible (Liu et al., 2009).

Algorithm - BalanceCascade (Liu et al., 2009)

i Suppose, we have a majority class training data set M and minority class data set as N, where |N| < |M|. We can have T different randomly sampled subsets of majority class as M1, M2, ..., Mt which can be represented as M1, where |Mt| << |Mt| & |Mt| = |N|.

ii Let the false positive rate f (the error rate of misclassifying majority class example to the minority class that H1 should achieve), be $r^{-1}$ ($\sqrt{\frac{T}{|M|}}$).

iii BalanceCascade uses different weak classifiers $C_i$, the number of classifiers being same as the number of subsets of majority class. The classifier $C_i$ is trained by N minority class samples and Mt minority class samples. The decisions generated by weak classifiers $C_i$ are combined using AdaBoost ensemble ($H_t$) using the Eq-3 (Liu et al., 2009).

\[ H_t(x) = \text{sgn}\left( \sum_{j=1}^{t} \alpha_{i,j} C_{i,j}(x) - \beta_i \right) \]  

where, $I_t$ is the number of iterations, $C_{i,j}$ is the weak classifier/learner, $\alpha_{i,j}$ is the weights of $C_{i,j}$, $\beta_i$ is the ensemble's threshold.

iv We adjust $\beta_i$ such that $H_t$’s false positive rate is f.

ev Similar to EasyEnsemble method, the output ensemble is obtained as Eq-4

\[ H(x) = \text{sgn}\left( \sum_{i=1}^{t} \sum_{j=1}^{t} \alpha_{i,j} C_{i,j}(x) - \sum_{i=1}^{T} \beta_i \right) \]  

3.2 Methods of Weights Assignment

The following filter and wrapper methods have been used for finding the weights of factors.

3.2.1 Fisher Discriminant Analysis (FDA): This method is used in pattern recognition, statistics and machine learning for finding the linear combination of features to distinguish two or more classes of events or objects (Jhunjhunwalla et al., 2019). In this method, all the factors/ layers are projected in one dimension corresponding to landslide occurrences. A multiplicative factor is required for projection of the data. This multiplicative factor is used for giving weights to all the thematic layers (Gupta et al., 2018).

3.2.2 Logistic Regression (LR): It is a special case of linear regression and predicts the probability of the occurrence of an event by using logit function (Gupta et al., 2018). In this method, the probability of presence or absence of a binary outcome (1 = landslide and 0 = no landslide) is modelled based on the values of predictor variables (Shukla et al., 2016). The independent variable can be interval or categorical while the dependent variable can be multinomial or binary. The LR coefficients are used for giving weights to all the factors.

3.2.3 Artificial Neural Network (ANN): ANN is a computational system, which is inspired by the human brain. The network consists of set of neurons, an input layer, an output layer and few hidden layers. The layers are interconnected and input data is passed to output layer by means of the hidden layers. There can be one or more hidden layers depending on the complexity of data. It is useful when a complex relationship exists between the data and the responses such as between landslide factors and landslide occurrences (Jhunjhunwalla et al., 2019).

ANN require several architectural and training parameters to be selected prior to analysis. The optimal number of hidden layers and the number of neurons per hidden layer are not known apriori. These parameters are empirically determined through rigorous experimentation and examination of different parameter settings (Blackard, Dean, 1999).

3.3 Landslide Susceptibility Index Computation and Susceptibility Mapping

The weights obtained in previous section are used for computation of landslide susceptibility index (LSI). LSI is calculated using Weighted linear combination method (Jhunjhunwalla et al., 2019, Gupta et al., 2018, Michael, Samanta, 2016). LSI can be calculated as given in Eq-5

\[ LSI = \sum \text{attributes} \times \text{weights} \]  

LSI can be classified into five different zones i.e LSZ (from very high susceptibility to very low susceptibility) based on natural break in the data. Results obtained with/without data balancing are compared using statistical quantities and visual quality analysis. The methodology used in the current study has been shown in Figure-2.

The EasyEnsemble and BalanceCascade methods were applied on total 30 subsets, which were randomly taken from the non-landslide pixels by the algorithm in python. After the data balancing, the FDA, LR and ANN methods were applied on these 30 subsets of data, which resulted in 30 LSI images for each of the methods. The mean and median of these 30 LSI images for each method was taken to generate mean image and median image for all the methods.

3.4 Accuracy Assessment

3.4.1 Heidke Skill Score: The accuracy of the LSZ maps is measured using Heidke Skill Score (HSS). It is the measure of skill of prediction and lies between 0 and 1. HSS can be defined as given in the following Eq-6 (NDFD Verification Score Definitions, 2017, Hyvärinen, 2014).

\[ HSS = \frac{NC - E}{T - E} \]  

where NC is the number of correct predictions, i.e. number of times the prediction and observation match, E is the number of predictions expected to verify based on chance, i.e. incorrect predictions and T is the total number of observations.

3.4.2 Recall: Recall can be defined as the ratio of relevant instances (landslides) predicted correctly by the model to actual number of relevant instances. It is also known as "sensitivity". This gives us a measure of how accurately the model predicts with respect to the actual instances of that class. We can compute recall using the expression in Eq-7

\[ \text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \]
The susceptibility to very low susceptibility (as shown in Figure-3, classified into five different zones i.e. from very high susceptibility to low susceptibility (as shown in Figure-3). The LSI values are normalized between 0 and 1 and correspond to the causative factor layer to obtain the LSI values.

The weights obtained in Table-1 are multiplied with the corresponding causative factor layer to obtain the LSI values. The LSI values are normalized between 0 and 1 and classified into five different zones i.e. from very high susceptibility to very low susceptibility (as shown in Figure-3).

The mean and median of LSI values for all the methods (except ANN) is observed to be near 0.55 (refer Table-2) so 0.55 is set as a threshold for classification.

The landslides with LSI greater than 0.55 are considered to be correctly classified and below 0.55 are considered to be falsely classified. As we can see from Table-2 that ANN has mean/median values significantly less than 0.55 and hence the susceptibility maps generated by ANN have maximum area lying in low or very low susceptibility zones. Hence, after data balancing, the mean/median value of LSI for all the methods (except ANN) is observed to be near 0.55 (refer Table-2) so 0.55 is set as a threshold for classification.

The landslides with LSI greater than 0.55 are considered to be correctly classified and below 0.55 are considered to be falsely classified. As we can see from Table-2 that ANN has mean/median values significantly less than 0.55 and hence the susceptibility maps generated by ANN have maximum area lying in low or very low susceptibility zones. Hence, after data balancing, the mean/median value of LSI for all the methods (except ANN) is observed to be near 0.55 (refer Table-2) so 0.55 is set as a threshold for classification.
Figure 4. LSZ map obtained using weights from FDA, without data balancing (a) and with data balancing (b).

Figure 5. LSZ map obtained using weights from ANN, without data balancing (a) and with data balancing (b). The sample location with significant changes are marked with black circle.

Table 3. Statistics (mean, median and standard deviation) for all the three methods using balanced data.

<table>
<thead>
<tr>
<th>Balancing Method</th>
<th>Statistical Quantities</th>
<th>LR</th>
<th>FDA</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean Image</td>
<td>Median Image</td>
<td>Mean Image</td>
</tr>
<tr>
<td>Easy Ensemble</td>
<td>Mean</td>
<td>0.2125</td>
<td>0.3834</td>
<td>0.5368</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.2141</td>
<td>0.3870</td>
<td>0.5684</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>0.0348</td>
<td>0.0775</td>
<td>0.1140</td>
</tr>
<tr>
<td>Balance Cascade</td>
<td>Mean</td>
<td>0.2337</td>
<td>0.2934</td>
<td>0.5568</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.2358</td>
<td>0.2960</td>
<td>0.5614</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>0.0438</td>
<td>0.0565</td>
<td>0.1149</td>
</tr>
</tbody>
</table>

also shows comparable HSS and recall values. The zonation of LSI obtained using LR can not prepared due to the smaller mean values. The natural break limits for different zones can not be applied on it. The results obtained in this study validates the fact that LR does not require the data balancing, given that we have sufficient samples in positive class (Crone, Finlay, 2012, King, Zeng, 2001). The FDA method may or may not show major changes in the results with/without data balancing, which also agree with the results of few studies (Xue, Hall, 2015, Xue, Titterington, 2008).

5. CONCLUSIONS

Data balancing methods improves accuracy for machine learning based methods such as ANN, Support vector machine etc. The EasyEnsemble method coupled with mean of weights seems to overpredict the high susceptibility zones whereas BalanceCascade method with median of
Table 4. Accuracy assessment of FDA and ANN on balanced data using Heidke Skill Score for total 76 landslides from 2013 to 2017

<table>
<thead>
<tr>
<th>Balancing Method</th>
<th>Methods</th>
<th>Correctly Classified Landslides (LSI ≥ 0.55)</th>
<th>Wrongly Classified Landslides (LSI &lt; 0.55)</th>
<th>HSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy Ensemble</td>
<td>FDA Mean</td>
<td>73</td>
<td>3</td>
<td>0.9589</td>
</tr>
<tr>
<td></td>
<td>FDA Median</td>
<td>70</td>
<td>6</td>
<td>0.9143</td>
</tr>
<tr>
<td></td>
<td>ANN Mean</td>
<td>73</td>
<td>3</td>
<td>0.9589</td>
</tr>
<tr>
<td></td>
<td>ANN Median</td>
<td>73</td>
<td>3</td>
<td>0.9589</td>
</tr>
<tr>
<td>Balance Cascade</td>
<td>FDA Mean</td>
<td>69</td>
<td>7</td>
<td>0.8986</td>
</tr>
<tr>
<td></td>
<td>FDA Median</td>
<td>69</td>
<td>7</td>
<td>0.8986</td>
</tr>
<tr>
<td></td>
<td>ANN Mean</td>
<td>72</td>
<td>4</td>
<td>0.9444</td>
</tr>
<tr>
<td></td>
<td>ANN Median</td>
<td>68</td>
<td>8</td>
<td>0.8824</td>
</tr>
</tbody>
</table>

Table 5. Accuracy assessment of FDA and ANN on balanced data using Recall for total 2743 pixels in 76 landslides from 2013 to 2017

<table>
<thead>
<tr>
<th>Balancing Method</th>
<th>Methods</th>
<th>Correctly Classified Pixels (LSI ≥ 0.55)</th>
<th>Wrongly Classified Pixels (LSI &lt; 0.55)</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy Ensemble</td>
<td>FDA Mean</td>
<td>2192</td>
<td>551</td>
<td>0.7991</td>
</tr>
<tr>
<td></td>
<td>FDA Median</td>
<td>2114</td>
<td>629</td>
<td>0.7707</td>
</tr>
<tr>
<td></td>
<td>ANN Mean</td>
<td>2254</td>
<td>489</td>
<td>0.8219</td>
</tr>
<tr>
<td></td>
<td>ANN Median</td>
<td>2198</td>
<td>545</td>
<td>0.8013</td>
</tr>
<tr>
<td>Balance Cascade</td>
<td>FDA Mean</td>
<td>2125</td>
<td>618</td>
<td>0.7747</td>
</tr>
<tr>
<td></td>
<td>FDA Median</td>
<td>2074</td>
<td>669</td>
<td>0.7561</td>
</tr>
<tr>
<td></td>
<td>ANN Mean</td>
<td>2205</td>
<td>538</td>
<td>0.8039</td>
</tr>
<tr>
<td></td>
<td>ANN Median</td>
<td>2011</td>
<td>732</td>
<td>0.7331</td>
</tr>
</tbody>
</table>

weights generated by ANN gives most appropriate LSZ map based on visual analysis. Using statistical quantities, the LSZ generated from both the data balancing methods show mean and median value greater or very near to 0.55, which also justifies the importance of data balancing before using ANN. The HSS and recall value shows the superiorities of LSZ maps prepared using mean of weights in ANN with the data balanced using Easy Ensemble method. The Balanced data do not show good results with Logistic regression as the LR method is not able to model the underlying probability distribution. Balanced data does not effect FDA method and it has approximately same accuracy as before. The landslide data is highly imbalanced in nature, hence 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.

REFERENCES


NDFD Verification Score Definitions, 2017.


This contribution has been peer-reviewed.

https://doi.org/10.5194/isprs-archives-XLII-3-W11-51-2020 © Authors 2020. CC BY 4.0 License.


