Increasing the reliability of seismic classification: A comparison of strategies to deal with class size imbalanced datasets

Chantal van Dinther, Marielle Malfante, Yoann Cano, Pierre Gaillard

EGU23-16410
EGU General Assembly 2023

Increasing the reliability of seismic classification: A comparison of strategies to deal with class size imbalanced datasets.

Chantal van Dinther[1], Marielle Malfante[1], Yoann Cano[2] and Pierre Gaillard[3]
[1] Univ. Grenoble Alpes, CEA, List, F-38000 Grenoble, France
[2] Université Paris-Saclay, CEA, DIF, F-91297, Arpajon, France
[3] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

Recent employment of large seismic arrays and distributed fibre optic sensing cables leads to an overwhelming amount of seismic data. As a consequence, the need for reliable automatic processing and analysis techniques increases. Therefore, the number of machine learning applications for detection and classification of seismic signal augments too.

A challenge however, is that seismic datasets are highly class imbalanced, i.e. certain seismic classes are dominant while others are underrepresented. Unfortunately, a skewed dataset may lead to biases in the model and thus to higher uncertainties in the model predictions. In the machine learning literature, several strategies are described to mitigate this problem. In presented work we explore and compare those approaches.

For our application, we use event catalogues and seismic continuous recordings of the RD network in France [RESIF, 2018]. Using a simple 3-layered convolutional neural network (CNN) we aim to differentiate between six seismic classes, which are based on hand-picked catalogues. The training set we obtained is highly skewed with earthquakes as the majority class, containing 77% of the samples. The remaining classes (quarry blasts, marine explosions, suspected induced events, noise and earthquakes with unquantifiable magnitude) represent 2.1 - 7.5% of the dataset, respectively.

We compare four strategies to deal with an imbalanced datasets for a multi-class classification problem. The first strategy is to resample the dataset (in our case we chose to reduce the size of the majority class). Another approach is the adaptation of the loss function by weighting the classes when penalizing the loss (i.e. increasing the weight of the minority classes). Those class weights can be adjusted either w.r.t. the reciprocal of class frequency [inspired by King and Zeng, 2001] or w.r.t. the effective number of samples [Cui et al., 2019]. Lastly, we have explored the use of a focal loss function [Lin et al., 2020].

Using balanced accuracy as a metric while minimizing the loss, we found that in our case adjusting the class weights in the loss function according to the reciprocal of the class frequency provides the best results.

References:

- RESIF, 2018: https://doi.org/10.15778/RESIF.RD
- King, G., & Zeng, L. (2001). Logistic regression in rare events data. Political analysis, 9(2), 137-163.
- Lin et al. (2020), Focal Loss for Dense Object Detection, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 42, NO. 2, FEBRUARY 2020
- Cui et al. (2019), Class-Balanced Loss Based on Effective Number of Samples, https://doi.org/10.48550/arXiv.1901.05555