Towards harmonized seismic analysis across Europe using supervised machine learning approaches

Riccardo Zaccarelli, Dino Bindi, Fabrice Cotton, and Angelo Strollo
German Research Centre for Geoscience GFZ, Potsdam, Germany (bindi@gfz-potsdam.de)

In the framework of the Thematic Core Services for Seismology of EPOS-IP (European Plate Observing System-Implementation Phase), a service for disseminating a regionalized logic-tree of ground motions models for Europe is under development. While for the Mediterranean area the large availability of strong motion data qualified and disseminated through the Engineering Strong Motion database (ESM-EPOS), supports the development of both selection criteria and ground motion models, for the low-to-moderate seismic regions of continental Europe the development of ad-hoc models using weak motion recordings of moderate earthquakes is unavoidable. Aim of this work is to present a platform for creating application-oriented earthquake databases by retrieving information from EIDA (European Integrated Data Archive) and applying supervised learning models for earthquake records selection and processing suitable for any specific application of interest. Supervised learning models, i.e. the task of inferring a function from labelled training data, have been extensively used in several fields such as spam detection, speech and image recognition and in general pattern recognition. Their suitability to detect anomalies and perform a semi- to fully- automated filtering on large waveform data set easing the effort of (or replacing) human expertise is therefore straightforward. Being supervised learning algorithms capable of learning from a relatively small training set to predict and categorize unseen data, its advantage when processing large amount of data is crucial. Moreover, their intrinsic ability to make data driven predictions makes them suitable (and preferable) in those cases where explicit algorithms for detection might be unfeasible or too heuristic. In this study, we consider relatively simple statistical classifiers (e.g., Naive Bayes, Logistic Regression, Random Forest, SVMs) where label are assigned to waveform data based on "recognized classes" needed for our use case. These classes might be a simply binary case (e.g., "good for analysis" vs "bad") or more complex one (e.g., "good for analysis" vs "low SNR", "multi-event", "bad coda envelope"). It is important to stress the fact that our approach can be generalized to any use case providing, as in any supervised approach, an adequate training set of labelled data, a feature-set, a statistical classifier, and finally model validation and evaluation. Examples of use cases considered to develop the system prototype are the characterization of the ground motion in low seismic areas; harmonized spectral analysis across Europe for source and attenuation studies; magnitude calibration; coda analysis for attenuation studies.