Machine Learning integrated to pipeline monitoring with Distributed Acoustic Sensing

Introduction

Distributed Acoustic Sensing (DAS) is a new and innovative technique that allows to convert fibre optic cables into hundreds to thousands of acoustic sensors. This technology is based on the analysis of a LASER pulse sent from an interrogator unit. Due to asperities intrinsically present along the fibre, each of its points acts as a reflector scattering the light back to the interrogator unit. The phase of this backscattered signal contains the information, in space and time, on the strain along the cable. It then enables the detection of a passing acoustic wave with enough energy to excite the fibre cable (Dean et. al, 2018). Allowing the interrogation of profiles of several tens of kilometres with dense spatial sampling uneasy to obtain with classic geophysical techniques, DAS instrumentation proved its relevance for seismic acquisition and infrastructure monitoring (Mateeva et. al, 2014, Muggleton et. al, 2020). However, the dense spatial and temporal samplings generate large datasets difficult to deal with. The importance of an automation of the data processing, with the use of machine learning techniques, seems important to implement.

When considering infrastructure monitoring, and more precisely pipeline monitoring, using DAS, the source of acoustic vibration must be detected and clearly identified at the structure neighbourhood. The source classification appears as a major leverage for the identification of potential threat and, when appropriate, the alert triggering. It must be fast accurate and robust. Moreover, as explained previously, DAS acquisition can create large amounts of data often difficult to handle. The used machine learning algorithm will then have to demonstrate its ability to deal with such big datasets.

Considering these requirements, the Random Forests machine learning algorithm appears as a good candidate for acoustic sources identification. Random Forests algorithm is indeed able to handle large number of attributes related to temporal and spectral characteristics of the signal and gives good reliability on event classification. This algorithm already proved its efficiency for seismic source identification such as earthquakes, rock falls or icequakes (Hibert et. al, 2017, Maggi et. al, 2017).

The Random Forest algorithm

The Random Forests algorithm has been developed by Breiman in 2001 (Breiman, 2001) and is since generally used in different domains such as biology, remote control or in seismology for the automated classification of tectonic events or nuclear explosions in different contexts (Provost et. al, 2017, Wenner et. al, 2020).

The Random Forests algorithm is a supervised classifier based on the generation of a large number of decision trees constituting a “forest”. Each tree is trained on a selected dataset of well identified events. The decision on the class of a studied event is taken considering the majority of tree’s votes. Trees are composed of nodes, testing an attribute value which will be compared to a threshold, and of leaves, corresponding to the attributed class of the tree. A score of “correct classification”, varying from 0 to 1, is associated with the classification results considering the proportion of trees which voted for the final estimated class.

The Random Forests method is then based on the evaluation of different parameters, named attributes, of the studied signal. These attributes are divided into three families with different tested signal characteristics: the waveform, the spectral content of the signal and the spectrogram. For this study, 57 attributes have been selected split into the three different families. Amongst them we can, for example, count the energy and kurtosis of the signal filtered between various bandwidths, the spectral centroid, or the kurtosis of the maximum of all discrete Fourier transforms as a function of time (Provost et. al, 2017).
The event catalogue

We focus our study on tests lead along a gas pipeline instrumented with fibre-optic cable. Different third-party works have been conducted and recorded using FEBUS A1-R DAS system developed by FEBUS OPTICS. The studied fibre optic section is 190m long. Strain rate, physical parameter delivered by DAS acquisition, is expressed in nm/m/s and is here computed for a spatial resolution of 5m and a corner frequency fixed at 200Hz. We work on the discrimination of six classes of acoustic sources: ambient noise, manual compactor, mechanical excavation, drill, jackhammer, sheet piling and circular saw.

During this work, instead of a “classical” use of a detection phase used in seismology for example (Hibert et. al, 2017, Provost et. al 2017), we prefer an analysis in “flow”. For this method, signals are divided into channels, corresponding to each position where strain rate is recorded along the fibre, and into time windows of 20s moving every 2s (Figure 1). Then at each window position, the extracted signal is identified. For this approach, an additional class, labelled “ambient noise”, needed to be added to other classes. It results in the constitution of a dataset of around 170,000 sub-signals associated with the six different classes cited before (Table 1). We observe that the number of events associated with ambient noise is dominating the overall catalogue.

![Figure 1 Example of spatial and temporal signal sub-division for a manual compactor event.](image)

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of windows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual compactor</td>
<td>13,962</td>
</tr>
<tr>
<td>Mechanical excavation</td>
<td>2,567</td>
</tr>
<tr>
<td>Drill</td>
<td>1,123</td>
</tr>
<tr>
<td>Jackhammer</td>
<td>7,805</td>
</tr>
<tr>
<td>Circular saw</td>
<td>6,368</td>
</tr>
<tr>
<td>Ambient noise</td>
<td>131,962</td>
</tr>
</tbody>
</table>

*Table 1* Number of signals extracted by the windowing in time and space of the recorded strain rate over windows of 20s moving every 2s.
The Random Forest algorithm is applied to events recorded with the FEBUS A1-R DAS and synthetically combined and repeated in space and time. This process enables to test the ability of the algorithm to handle the identification of events succeeding in space and time. The studied dataset is presented in Figure 2.

**Figure 2** Tested dataset synthetically computed with the succession and combination of true events recorded with the FEBUS A1-R. On the left: recorded strain rate (in nm/m/s), on the right: true segmentation (nois: ambient noise, comp: manual compactor, exca: excavation, dril: drilling, ham: jackhammer, palp: sheet piling, saw: circular saw).

The evaluation of the performances of the classification can be achieved using several statistical tools. Among them, we count the confusion matrix. Each column of this matrix corresponds to the true class whereas each row is affected to an estimated class. If we consider a classification with 2 classes, the confusion matrix will be (Table 2):

<table>
<thead>
<tr>
<th></th>
<th>True class: 1 or positive</th>
<th>True class: 0 or negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated class: 1 or positive value</td>
<td>True positive</td>
<td>False negative</td>
</tr>
<tr>
<td>Estimated class: 0 or negative value</td>
<td>False positive</td>
<td>True negative</td>
</tr>
</tbody>
</table>

**Table 2: Confusion matrix in the case of a classification of two classes**

Results of classification

Once the algorithm trained with a portion of the dataset, the studied dataset established and the tools for the evaluation of the algorithm’s performances defined, the classification is implemented in “real time”, with a 20s-window moving every 2s. The classification is then achieved for each temporal window and results in a map of classification (or segmentation map) where each line corresponds to the event identification at a given time t. Considering the size of window chosen in this study, the classification is achieved for a history of 20s, implying a delay of 20s to obtain the first results. This time window can of course be customized considering the needs for quick alert releases. Results of the classification are presented in Figure 3. We observe a very good precision of the estimated class compared to the true segmentation. The confusion matrix (Figure 3.c) reveals the good quality of the classification with an accuracy equal to 96.85%.
Figure 3 Results of the classification: (a) True segmentation, (b) Segmentation estimated using the Random Forest algorithm and (c) Confusion matrix associated with the classification.

Conclusions

We demonstrated the efficiency of the Random Forest algorithm for automatic classification of events recorded with DAS system. The application of the algorithm on a restricted dataset appears very promising for future use of automatic event identification on longer fibre optic cables deployed along pipelines with a constitution of larger training datasets.

References