Real-time Acoustic Emission Classification: New Evolutionary Approach

Eraldo POMPONI, Alexei VINOGRADOV

Laboratory for the Physics of Strength of Materials and Intelligent Diagnostic Togliatti State University, 14 Belorussskaya St., 445667 Togliatti, Russia
E-mail: eraldo.pomponi@gmail.com, alexei.vino@gmail.com

Abstract
A common target of clustering in Acoustic Emission (AE) is to disclose the hidden connection that exists between different source mechanisms, e.g. plastic deformation, crack initiation or corrosion cracking, and features that can be extracted from AE signals. On one hand, well known schemes like $k$-means and fuzzy $c$-means have been proven effective in this regard but both are iterative processes, which make it hardly possible to utilize them in real-time applications. On the other hand, Acoustic Emission is a non-destructive test (NDT) technique that bases its success on the real-time monitoring capability. This lack of algorithms able to integrate directly into the AE acquisition/processing chain has motivate this work. Inspired by sequential $k$-means, i.e, a non-iterative variant of the classic $k$-means, we present a novel classification technique that, although general purpose, it is specially suited for AE and respect to other well known schemes like $k$-means and fuzzy $c$-means, it is specifically thought to be applied in real-time. The number of classes and their elements are both inferred from the data, making this approach completely "unsupervised". Preliminary results obtained comparing this method with the schemes mentioned above, have shown it is effective to capture the dynamics of the underlying physical process.

Keywords: Acoustic Emission (AE), Signal Processing, Pattern Recognition, Real-time clustering, $k$-means.

1. Introduction
In the wide landscape of NDT techniques, Acoustic Emission (AE) stands out against all the others when those features are mandatory:

- Damage localization capability
- Full body scan
- On-line monitoring

Each of them has a different impact depending on the application but what makes AE so appealing is that they come all together “out of the box”. From a technical point of view, those features come with a price: AE inspection demands high performance hardware (HW) and software (SW). It happens mostly because of the high sampling rate, e.g., compared to Vibrometry, and the number of deployed channels that spans from few units to thousands in some special cases (e.g. big vessels). Clustering is computationally expensive and time consuming mostly due to the iterative nature of schemes like $k$-means, $c$-means or SOM (i.e. Self Organizing Map). Another sensible issue regards the amount of a-priori information required by supervised approaches adopting Neural Networks or Support Vector Machine (i.e. a training set that must represent all source mechanisms involved). This actually prevented/discouraged the use of clustering techniques as part of the on-line AE acquisition/processing chain. In this study we present a novel classification algorithm that aims to overcome the above limitations so that it can be integrated in AE systems without increasing high-performance HW demand, i.e., potentially directly incorporated in the existing one, that also address known weakness points and missed features of sequential $k$-means to which it is inspired. The work is organized as follows:
in Section 2, we briefly present the literature about developed clustering procedures and their application in the AE research field as a tool to disclose different source mechanisms. In Section 3, we describe the algorithm structure, its parameters, and their formal definition. Section 4. reports the details about the dataset adopted to check the algorithm and a brief indication of those indexes that have been chosen to quantitative measure its performance and how it compares with established approaches. Obtained results are shown and discussed in Section 5. In the last Section 6, we summarize advantages and drawbacks of the proposed scheme together with the scopes for future developments.

2. Clustering Overview

The use of "batch" clustering procedures as tools to identify groups of AE events originating by the same source mechanism has been a common practice in the recent years (e.g. [1, 4, 10, 12, 19, 20], etc.) with results not always satisfactory. This problem is ill-posed due to a lack of a-priori information about active source mechanisms and the unavoidable presence of noise so that multiple solutions can be expected [1, 6, 7] without a general rule of preference. Attempt to develop robust clustering procedure, e.g., [3] can be found in AE literature and in many other research fields but adopting them in a real-time applications has not been a priority for most algorithm designers. Being on-line monitoring a key feature of AE inspection technique makes inadequate to leave the clustering to a batch process: That is what motivated us to propose a procedure capable of operating in real-time and designed specifically for AE purposes by taking into account actual testing conditions and acquisition features. It has been inspired by the well known $k$-means algorithm and its non-iterative variation sequential $k$-means. $K$-means is in one hand appealing due to its high speed paired with simplicity (despite the recently disclosed complexity increase in the worst case [2]). On the other hand, it suffers of two major shortcomings: 1) The number of clusters $k$ must be stated a-priori and 2) it is an iterative procedure. The first issue is that there is no way to guess the number of clusters a-priori in real life NDT inspections as well as in many laboratory tests. The second issue prevents the real-time application when the acquisition occurs at a "high speed", which is common in AE tests where sampling is done often at 5 Msamples/s or faster. A flexible approach with less a-priori information requirements has already been envisioned by MacQueen in his seminal paper [11]. He introduced two extra parameters, $C$ for 'coarsening' and $R$ for 'refinement', that should control the former, when two clusters are merged (i.e. coarsening) and the latter, when a new cluster is created (i.e. refinement), both operation depending on the distance between clusters centroid (i.e. See reference). Unfortunately, adding this two parameters doesn’t solve the issue, rather transfer it to a proper choice of $C$ and $R$. The proposed method tries to solve that problem discovering "natural" clusters [13] induced by the chosen metric as they appear, "on the fly" and estimating $C$ and $R$ directly from their compactness and distribution in the feature space. We shall show that this method has been proved effective in clustering AE signals in a way consistent with the underlying physical process. For a sake of simplicity AE signals have been represented through their Power Spectral Density (PSD) and, as dissimilarity measures, we adopted the correlation distance leaving more sophisticated setup to further investigation.

3. Method Description

The algorithm differ from the sequential $k$-means at which it is inspired in terms of the information that must be stored for each cluster and due to ”coarsening” (C) and ”refining” (R)
processes that will take place depending on the clusters properties. Avoiding special requirements but just relying on common practices, it is also a peculiarity of this method to use as much as possible the information available about the background noise always present during an AE inspection; information collected in the form of a set of \( N \) noise realizations \( y_1, y_2, \ldots, y_N \) acquired before any real AE signal \( x_1, x_2, \ldots \). Here, \( x_i \) and \( y_i \) are the features vectors for the signal and the noise in the multi-dimensional feature space, respectively. Assuming that the distance \( D(\cdot, \cdot) \) has been chosen to quantify the dissimilarity between two generic data points in the dataset, the minimum descriptive information must be attributed to each cluster \( k_i \) as:

- Centroid or Mean \( m_i \)
- Number of elements \( n_i \)
- “Evolutionary” intra-cluster distance (EICD):
  \[ \delta_{ki} = [\delta_1, \delta_2, \ldots, \delta_n] \text{ where } \delta_j = D(x_j, m_j) \]
- Point-wise centroid drift vector:
  \[ \gamma_{ki} = [\gamma_1, \gamma_2, \ldots, \gamma_n] \text{ where } \gamma_1 = 0 \text{ and } \gamma_{j,j\neq 1} = D(m_j, m_{j-1}) \]

The point-wise centroid drift vector is used as weighting coefficients for EICD so that it is possible to quantify cluster homogeneity and separability in a reasonably accurate way.

To have a lower boundary \( R_{noise} \) for the coefficient \( R \) we calculate the mean distance \( \bar{D} \) between all possible noise realizations \( D(y_i, y_j) \) and its standard deviation \( \sigma_{yy} \). In terms of these two quantities \( R_{noise} \) reads as:

\[ R_{noise} = \bar{D} + \tau \sigma_{yy} \quad (1) \]

where \( \tau \) is constant and \( 0 \leq \tau \leq 3 \) and commonly equals to 3 when the hypothesis of normality for the noise distribution holds. If there is no noise available \( R_{noise} = 0 \). The \( R \) value is then computed as

\[ R = \max (\tilde{R}, R_{noise}) \quad (2) \]

where

\[ \tilde{R} = \frac{1}{2} \max_{i,j} D(m_i, m_j) \quad (3) \]

is half the maximum distance between each pair of clusters centroids. This quantity is updated at each iteration. If the distance between the current point \( x \) (i.e. acquired signal) to the closest mean \( m_j \) is greater that \( R \), a new cluster is created. To avoid serious instability, this process of ”refinement” must be balanced by a corresponding ”coarsening” that, in this case, is based simply on this assumption: If two clusters consistently ”overlap” in a measure that we will specify, they can be merged together. To estimate the amount of overlap between different clusters, we need a way to measure their ”aperture” (e.g. surface if we are in a 2-dim space). This has been obtained considering the hyper-sphere that can be associated to each cluster in the chosen feature space. To simplify its radius evaluation, it has been assumed that the data-points are normally distributed around their specific centroid so that we can consider it proportional to the distribution’ variance. If the distance between the centroids of two clusters is less than a quantity proportional to that radius, they are merged together. Formally cluster \( k_i \) and \( k_j \) are merged together if:

\[ D(m_i, m_j) \leq \max(wd_{k_i} + \alpha \sigma_{k_i}, \tilde{R}) \quad \text{or} \quad D(m_i, m_j) \leq \max(wd_{k_j} + \alpha \sigma_{k_j}, \tilde{R}) \quad (4) \]

where \( \alpha \) is usually fixed to 2 due to the type of estimator in eq.6 and where:

\[ w_{ki} = \delta_{ki} * \max \left\{ 0, \left(1 - \frac{\gamma_{ki}}{\max{\delta_{ki}} - \min{\delta_{ki}}} \right) \right\} \quad (5) \]
\[ wd_{ki} = \text{mean}(w_{ki}) \]  
\[ \sigma_{ki} = \sqrt{\text{mad}(w_{ki})}/.6745 \]

To summarize, at each iteration the method works like the sequential \( k \)-means plus checks for coarsening and refining the set of classes as described above. It must be noticed that the coarsening procedure is performed not at every iteration but just when a new cluster is created because we assume that just when this happen the clusters structure is changed consistently to foresee a possible coarsening.

4. Experiment and Performance Indexes

The probing experimental dataset was created from a compressive test of a bulk metallic glass, which has been taken as a model material for this work. The mechanism of plastic deformation of metallic glasses at low homologous temperature is associated with formation of very narrow (of 10 -20 nm wide) shear bands (see a comprehensive review by Schuh [15] for details), propagating through the sample at fairly high velocity of several m/s (up to several hundred m/s) [17]. It has been well documented that the shear bands are accompanied by strong burst type AE [18] and that two types of AE due to shear bands in metallic glasses are distinguished: (i) the microscopic shear band nuclei appear as pre-cursors of (ii) major macroscopic shear events [9, 16].

\( Zr_{62.5}Fe_{5}Cu_{22.5}Al_{10} \) (in nominal atomic percents) metallic glass alloy was prepared by arc melting followed by the copper-mold injection and gravity casting. The as-cast 3x3 mm\(^2\) cross-section rods were cut to 6 mm length for compression test and the surfaces were mechanically polished. The samples were examined by X-ray diffractometry and showed no signatures of crystallinity. The samples were subjected to compressive loading at room temperature under a constant crosshead speed corresponding to the nominal strain rate of \( 1 \times 10^{-3} \text{s}^{-1} \) using a screw-driven testing frame equipped with an HBM piezoelectric load cell. The specimens for compression testing were set between two parallel ceramic platens. The miniature (8 mm diameter) broadband AE sensor MSAE-1300WB with a built-in low-noise 27 dB preamplifier and a 30 kHz high-pass cut off filter made by Microsensors AE, Ltd (Russia) has been securely mounted directly on a lower platen and axially centered with the specimen. Vacuum oil was used as a coupling media to ensure a good acoustic contact between the transducer and the platen. To be aware of the “noiseless” performance of the testing rig, the compressive load up to 30 kN was gradually applied to the two ceramics platens without any specimen between them. No signals were recorded in a whole range of applied forces and loading rates during either loading or unloading stage. Of course, this does not exclude the possible appearance of "false" acoustic emissions arising from external sources such as sample rubbing. The PC controlled digital system was used for AE data acquisition. The AE signal from the pre-amplifier was transferred through the MSAE-FA01 filter-amplifier with the frequency band 50-1200 kHz and 40 dB gain and then acquired by means of a 12 bit ADC operating at 6.25 MHz frequency and threshold-triggering mode of operation. The indexes adopted to evaluate the performance of the proposed method and compare them with what achieved by other schemes, are: Silhouette Validation Method [14], Davies-Bouldin Validity Index [5] and Dunn’s Validity Index [8]. For a detailed description see the references. Notice that a higher value signifies a better performance for the Silhouette and the Dunn’s index while it is opposite for the Davies-Bouldin index.
5. Application Results

In what follows we will present the results obtained with the present method working on the data-set described above comparing the outcomes with two reference "batch" scheme, \( k \)-means and fuzzy \( c \)-means, imposing them to produce the same number of cluster as the real-time procedure. Table 1 shows the number of elements associated with each cluster.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>realtime</th>
<th>( c )-means</th>
<th>( k )-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>166</td>
<td>188</td>
<td>159</td>
</tr>
<tr>
<td>2</td>
<td>659</td>
<td>438</td>
<td>318</td>
</tr>
<tr>
<td>3</td>
<td>96</td>
<td>125</td>
<td>150</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>225</td>
<td>349</td>
</tr>
</tbody>
</table>

Table 1: Clusters elements.

There is immediately a clear distinction between the proposed method and the two reference schemes: for the \( k \)-means and the fuzzy \( c \)-means the four clusters are almost homogeneous in terms of number of elements while in the realtime clustering there are two that are dominant (cl.1 and cl.2). This last situation seems in agreement with the expected number of source mechanisms that have been outlined in Section 4. An impression of the performance of the proposed clustering process can be obtained from Fig.1 where the population of AE signals is represented as a bi-dimensional distribution in "median frequency - energy" coordinates: the points naturally fall into different groups with small overlapping for the realtime procedure while they are often superposed for the other two methods. A quantitative performance measure is given in Table 2. The Silhouette index is the highest for the proposed method. Hence, this is a clear indication that the realtime clustering performs better then others. The consistent gap that exists between the \( c \)-means and the \( k \)-means performance suggests that the \( c \)-means is far from being optimized with the present setup. Looking at the other two indexes and bearing in mind that for the Davies-Bouldin index a lower value corresponds to a better performance while for the Dunn index the picture is opposite, it is confirmed that the solution proposed by the realtime clustering is quantitatively better than that proposed by the \( k \)-means or the fuzzy
Table 2: Performance indexes.

<table>
<thead>
<tr>
<th></th>
<th>realtime</th>
<th>c-means</th>
<th>k-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silhouette</td>
<td>0.569</td>
<td>0.389</td>
<td>0.450</td>
</tr>
<tr>
<td>Davies-Bouldin</td>
<td>0.483</td>
<td>0.609</td>
<td>0.624</td>
</tr>
<tr>
<td>Dunn</td>
<td>2.467</td>
<td>0.921</td>
<td>0.989</td>
</tr>
</tbody>
</table>

c-means. Of course, one cannot expect that the proposed scheme preforms always better than other clustering techniques, but it can be stated that it preforms reasonably well for real time applications.

6. Conclusions

A novel general purpose clustering algorithm capable to work in realtime has been proposed. In has been described the specific setup to be adopted when working during AE inspection. It can, potentially, be included in existing AE systems without a consistent increase in the demand of computational power. Preliminary results in AE applications, one of which has been reported in Section 5., have shown that the method seems to be capable of capturing the dynamics of the underlying physical process that is correctly reflected in the clusters temporal and spatial distribution. Most relevant features of to the proposed method are:

1. Number of clusters not specified *a-priori* but data driven.
2. Non-iterative procedure
3. Efficiency in terms of computational cost and memory consumption (i.e. the same order of efficiency as the sequential *k*-means).

In terms of drawbacks, the most serious one of this method (i.e. the same suffered by the whole category of *k*-means “inspired” clustering procedure) is a strong dependency of the clustering results on the particular initial point (i.e. signal) where the procedure has started. This is the reason why it is highly suggested to start clustering the \(N\) noise realization as the first \(N\) signals. In this way, the algorithm has a chance to stabilize (i.e. noise is supposed to be stable) before the first AE real signal is processed so that the stability of the solution will be improved. Moreover, a consistent uncertainty is due to the plethora of different “feature space/dissimilarity measure” combinations that are available. To choose the right one for the specific class of cases under study is of primary importance because it affects completely the obtained results. Unfortunately, there is no specific “rule” on how to do this choice. It is up to the end user with his experience and expertise to discover/assume which “feature space/dissimilarity measure” will provide an acceptable sparsity of the data-points that will lead the algorithm to find a feasible solution. In conclusion, further tests must be performed in order to verify the suitability of this method to address real-life AE inspections but laboratory test are clearly showing a promising behavior and performance of the same level or even better than some of the most common batch clustering scheme like *k*-means and fuzzy c-means.

References


