Ecological Informatics 51 (2019) 177–184

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Towards better volcanic risk-assessment systems by applying ensemble classification methods to triaxial seismic-volcanic signals

Mauricio Orozco-Alzate[a], John Makario Londoño-Bonilla[b,c], Valentina Nale[d], Manuele Bicego[d,*]

[a] Universidad Nacional de Colombia - Sede Manizales, Departamento de Informática y Computación, km 7 vía al Magdalena, Manizales 170003, Colombia
[b] Servicio Geológico Colombiano, Observatorio Vulcanológico y Sismológico de Manizales, Av. 12 de Octubre No. 15 - 47, Manizales 170001, Colombia
[c] Universidad Católica de Manizales, Faculty of Engineering and Architecture, Prevention, Reduction and Attention of Risk Disaster Program, Cra. 23 No. 60 - 63, Manizales 170002, Colombia
[d] Università degli Studi di Verona, Dipartimento di Informatica, Ca' Vignal 2, Strada le Grazie 15, Verona 37134, Italy

ARTICLE INFO

Keywords:
Classification ensembles
Environmental monitoring
Risk assessment
Seismic signals
Triaxial recordings
Volcanoes

ABSTRACT

Analysis of seismic data is the most important method for volcano monitoring. Such data typically consists in digital signals acquired with an arrangement of triaxial seismic sensors which are strategically deployed on the volcano and its surrounding areas. Very rich measurements of the underlying phenomena are obtained from the arrangement because each sensor, through their corresponding three sensing axes, uninterruptedly acquires data at a relatively high sampling rate. Such an uninterrupted acquisition, however, turns manual classification of seismic signals into an inefficient and often error-prone task. As a solution, several systems for automated classification of seismic-volcanic signals have been proposed. All these systems, however, are limited to the usage of only one direction of acquisition; typically the vertical one. In this paper we make a step forward, exploring the potential benefit of using information from the three axes of the signals gathered by a single sensor. Integration is performed by classifier combining techniques, applied at different levels: this permits to take into account in the classification all the three orthogonal orientations —vertical, East-West and North-South— of the phenomenon. Preliminary experimental results on a set of volcanic signals gathered at Nevado del Ruiz volcano in Colombia confirmed the richness of this information.

1. Introduction

One of our major concerns, nowadays, is how to both sustainably and safely benefit from the environment such that its ecological functions are preserved and human life does not get threatened by the Earth's natural processes. Both issues are addressed by *environmental geology* which, according to (Keller, 2011), includes the study of five fundamental concepts, namely: population growth, sustainability, earth systems, natural hazards and scientific knowledge. Even though all of them are important for environmental managers, the two latter are particularly relevant for the prevention of the potentially catastrophic consequences of sudden events such as volcanic eruptions, earthquakes, tsunamis, landslides, floods and hurricanes. Scientific knowledge of natural hazards —typically based on the collection and analysis of sensor data— allows environmental decision makers to better understand natural processes with the aim of enhancing risk-assessment and hopefully taking early responses such as evacuations, relocations and other mitigation strategies.

Scientific knowledge of volcanic activity is based on the acquisition and analysis of geophysical and geochemical measurements: the so-called *volcano monitoring*. Among the first type of measurements, seismic data is considered the main source of information because most physical processes in a volcano may trigger ground movements and earthquakes. In order to undertake the monitoring, volcano observatories have deployed dense and telemetered networks of digital signal acquisition systems, which are placed in strategic locations (called *stations*) and typically use triaxial seismic sensors. Triaxial means that sensors provide measurements of the seismic phenomenon in three orthogonal orientations: vertical axis, East-West axis and North-South axis. Since signals are uninterruptedly acquired —in several stations, along three spatial axes and at a relatively high sampling rate— seismic phenomena are richly represented but the volume of seismic data to be processed is always increasing and might overload the observatory personnel.

The primary and very time-demanding duty in seismic-based volcanic risk-assessment systems is the classification of seismic signals into

[*] Corresponding author.
E-mail addresses: morozcoa@unal.edu.co (M. Orozco-Alzate), jmakario@sgc.gov.co, jmakario@ucm.edu.co (J.M. Londoño-Bonilla), manuele.bicego@univr.it (M. Bicego).

https://doi.org/10.1016/j.ecoinf.2019.03.004
Received 26 June 2018; Received in revised form 11 February 2019; Accepted 15 March 2019
Available online 19 March 2019
1574-9541/

predefined categories. Several proposals —based on pattern recognition and machine learning (Bishop, 2006) techniques— to automate this task are found in the literature; see for instance the ones reviewed in (Orozco-Alzate et al., 2012) and (Malfante et al., 2018) as well as other recent studies by Bicego et al. (2013), Cárdenas-Peña et al. (2013), Cortés et al. (2014), Orozco-Alzate et al. (2015), Bicego et al. (2015), Curilem et al. (2016), Cortés et al. (2016), Lara-Cueva et al. (2016) and Soto et al. (2018). Almost all of those studies consider one axis per signal, namely the vertical one which is said to be the most discriminative. Orozco-Alzate et al. (2015), for instance, say that "even though seismic sensors deliver three components, only registers from the vertical one are considered". Curilem et al. (2016), similarly, restrict themselves to use the vertical axis claiming that "it provides a better signal-to-noise (S/N) ratio in most events". However, there is the possibility that the other two directions encode interesting (and complementary) information. Consequently, the main goal of this paper is exactly to investigate this possibility, trying to understand if it is possible to improve classification accuracy by integrating the three axes of observation.

This investigation seems reasonable also from a seismic point of view: actually, different characteristics of the volcanic events are encoded along different directions. More in detail, in seismology, it is a well-known fact that P-waves[1] are observed more clearly at the vertical axis, while S-waves at the horizontal ones. To integrate the information contained in the three axes, we resort to the research field of ensemble classification methods (Kuncheva, 2014), which are aimed at combining different pattern recognition systems to improve the performances of single separate ones. It has been shown in this field that the combination is particularly suited when sources of information (also known as *modalities*) are complementary – this seeming exactly the case of the triaxial seismic records.

Please note that ensemble classification methods have been already shown to be useful in seismic-volcanic analysis, e.g. by Duin et al. (2010) and Curilem et al. (2016). However, previous attempts have been limited to consider one axis per signal to combine either different classifiers or several recording stations. In this paper we experiment ensemble methods at all levels: feature-level (combining feature representations), score-level (combining classifier matching scores), and decision-level (combining classifier decisions) by using a subset of triaxial seismic signals including examples of volcano-tectonic (VT) events and long-period (LP) events, all of them recorded at Nevado del Ruiz volcano, Colombia and digitized with a 16-bit analog-to-digital converter; consequently, signal amplitudes (counts) may take values in the range $[-32768, 32767]$. Illustrative samples of each class are shown in Fig. 1. Experiments were promising and reveal that exploiting the richness of information available by resorting to the combination —also known as *fusion*— at all levels is indeed advisable.

The remaining part of the paper is organized as follows. Related studies are presented in Section 2. Representation and fusion methods are described in Section 3. Experimental results are shown and discussed in Section 4. Finally, our concluding remarks are given in Section 5.

2. Related work

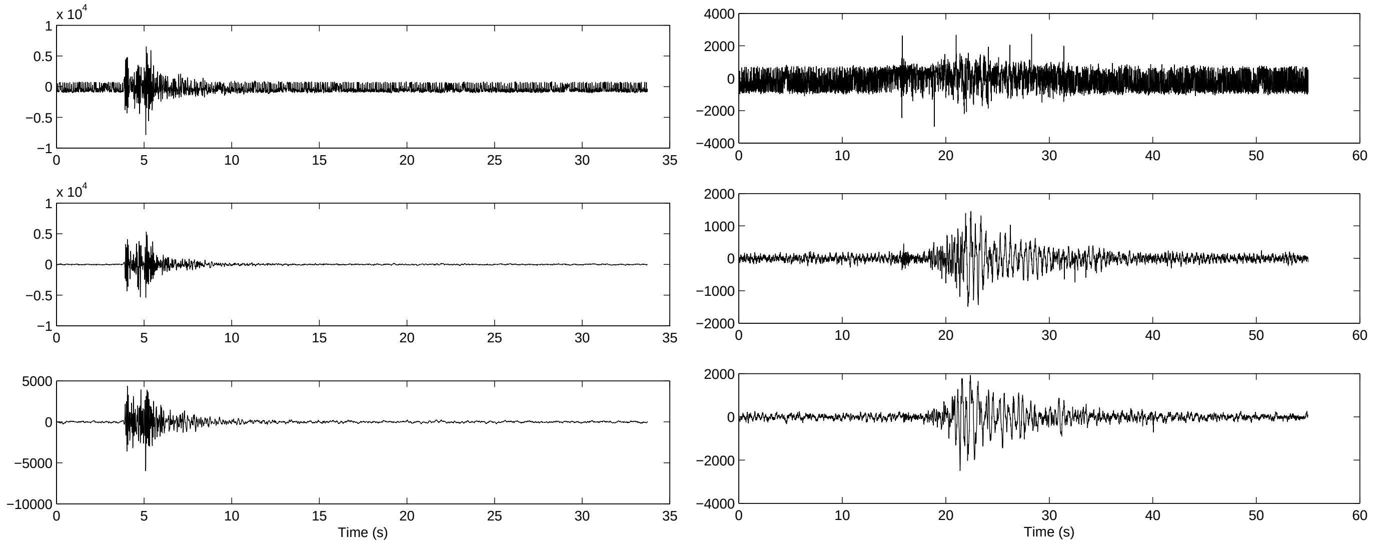
The availability of multiple training sets, recorded for the same seismic events but as seen at different stations, motivated the use of ensemble fusion methods for seismic signal classification. Such methods are expected to perform better than individual systems when either the multiple training sets or the different classifiers are diverse; that is, when the first ones convey non-redundant information and the second ones exhibit different behaviors.

According to that motivation, a study on the combination of signals —equal in length (12,032 time samples) and simultaneously recorded at five stations, along with the combination of separate quadratic classifiers trained per station, was carried out in (Duin et al., 2010). Three classes of seismic signals from Nevado del Ruiz volcano were taken into account in that study, namely volcanic earthquakes, icequakes and tremors. Moreover, two different representations were considered according to the features extracted for the original incoming signals: i) 40-dimensional feature vectors, each one resulting from spectra of half of the length (due to the mirror property of the Fourier transform) of the original seismic signals, followed afterwards by a principal component analysis to project the spectra onto the first 40 principal components; ii) spectrograms having a size of 128 frequency bands and 93 time windows. For each one of the above-mentioned representations, the authors tried three different combination scenarios to be compared against separate classifiers trained and tested on each recording station; namely: i) feature-level combination, by decision templates (Kuncheva et al., 2001; Kuncheva, 2014, p. 173) of the signal representations from the five stations in order to test them on each classifier (per station); ii) score-level combination of the five quadratic classifiers, trained with data from individual stations, also combined by the same strategy and observing data from each station; iii) score-level combination of the five quadratic classifiers trained and testing their corresponding signals: those ones recorded in the same station of the classifier. Notice that the first two scenarios are cross-station in terms of training and test while the second one is not. Their results showed that, overall, the latter is the best combination strategy to improve classification results.

Later on, in (Bicego et al., 2013), a system based on hidden-Markov-model (HMM) embeddings was evaluated in terms of its generalization capability to classify signals recorded at a station different from the one the training signals came from. However, in this case, there was no combination rule but just a cross-station training/test evaluation in a transfer learning scenario (Pan and Yang, 2010). Several generative embeddings were tested; namely Fisher score embedding, log-likelihood embedding, state embedding and transition embedding. Three-class and four-class problems were considered for a set of signals recorded at Galeras volcano. The authors concluded that the HMM-based embeddings outperform the usage of HMMs alone, also in some cases of the challenging cross-station scenario. Soon after, in (Orozco-Alzate et al., 2015), a fusion at the feature-level was explored to combine two and three dissimilarity representations of the one-axis seismic signals recorded at a single station from Nevado del Ruiz volcano; particularly by simply averaging either two or three of the dissimilarity matrices computed from pairwise comparisons of waveforms, spectra and spectrograms. The Euclidean and the Dynamic-Time-Warping (DTW) distances were used to built the dissimilarity matrices. The combination of the dissimilarity matrices by averaging them, however, did not appear to be convenient but, conversely, was worse than using DTW-based dissimilarity matrices computed from the spectrograms alone.

The most recent study on ensemble methods applied to the classification of seismic signals is found in (Curilem et al., 2016). As in the previous cases, only the vertical axis is considered in this study because, as already stated above, the authors claimed that "it provides a better signal-to-noise ratio in most events". These authors considered three seismic recording stations from Llaima volcano and four classes of seismic signals —LP events, VT events, tremors and other events (counterexamples)— represented by five features: three statistical descriptors from the waveforms, the dominant frequency from the spectra and the energy in a specific band of the wavelet transform. The base classifiers used were support vector machines (SVMs) with radial-basis-function kernel. They tried combination scenarios at feature-, score- and decision-level by using the following combination rules for each case, respectively: i) just merging the feature representations from the three stations; ii) applying the so-called Bayes-based confidence

[1] Seismic signals are composed by two body waves: i) Primary or P-waves and ii) secondary or S-waves. The first ones are faster, longitudinal and compressional; the second ones are slower, transverse and elastic.



(a) VT event

(b) LP event

Fig. 1. Sample signals of volcano-tectonic (VT) and long-period (LP) events as registered by the three axes of a seismic sensor located in Olleta station at Nevado del Ruiz volcano, Colombia. In each subfigure, from top to bottom, the recording axes are Vertical, North-South and East-West, respectively. Signal amplitudes are given in counts of a 16-bit analog-to-digital converter.

measure (Becerra Yoma et al., 2005) to the scores given by the separate SVMs and iii) majority vote (Kuncheva, 2014, pp. 113) of the crisp labels assigned by the classifiers. In contrast with the conclusion by Duin et al. (2010), these authors observed that the best option is merging the features instead of the late fusion at either the score or the decision levels, highlighting that the combination at the feature-level is more affine to the way in which the experts assign the labels at the volcano observatory; that is, by looking at the waveform recordings of the several stations and, afterwards, deciding which class label must be assigned.

All these works present a common characteristic: they only analyze signals relative to the vertical direction. Even if this seems justifiable, it is possible that the other directions convey important information or, in a problematic but realistic scenario, that the vertical sensing axis is for some time out of service or corrupted with noise; e.g. as observed for the top signals in Fig. 1. Our main goal is, therefore, to explore this direction, namely whether it is possible to successfully fuse representations and/or classifiers derived from the three recording axes of a single sensor. For the sake of illustration, we used a set of triaxial seismic signals recorded at Nevado del Ruiz volcano, Colombia.

3. Methods

In this section the methods used in our experimental evaluation are presented. In particular, we start with the characterization of the seismic signals; then we briefly present the classifier we used (the K nearest neighbors rule) and, finally, we introduce the classifier combining techniques we employed. As a reference starting point, in Fig. 2 we show the pipeline of a classical classification system for seismic signals: a signal to be classified is first characterized with a set of features, which are then fed into a classifier; such a classifier computes a set of scores (matching scores), one for each class, which indicate for every class the confidence in assigning a given signal to that class.

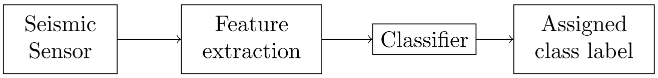


Fig. 2. Classical classification scheme for seismic-volcanic signals.

Finally, on the basis of these scores, a label is assigned (the decision). Typically, the classifier has to be trained using a training set (a set of labeled examples).

3.1. *Feature extraction for seismic-volcanic signals*

Many different alternatives for representing seismic-volcanic signals have been explored during the last years, ranging from morphological features extracted from the signal waveforms, time-frequency features estimated from spectra, spectrograms or wavelet transforms of the original signals, dissimilarities computed from pairwise comparisons of either raw signals or their transforms, as well as other more sophisticated representations such as time-variant features, bag-of-words, topic models and HMM-based generative embeddings; see (Castro-Cabrera et al., 2014) for an experimental comparison of some of the conventional feature representations.

Among all the above-mentioned options, a frequently used and often well-performing one is based on the so-called mel frequency cepstral Coefficients (MFCC). These coefficients are imported from the speech recognition field, where they have proved to be a robust feature representation that mimics the human hearing sense by, in a logarithmic scale, emphasizing the content in some frequency bands while attenuating it in others. Even though seismic signals exhibit a different behavior in the frequency domain, their nature is analog to that one of acoustic signals, both consist in waves traveling through elastic media: earth and air, respectively. Thereby, an MFCC-based representation of the seismic signals allows not just to extract relevant information for specific frequency bands but also to reduce the dimensionality of the problem typically to 26 features including the log-energy of the signal, the first 12 MFCCs and their corresponding first derivatives or 39 features if the second derivatives are also included. The reader is referred to (Álvarez et al., 2012) for a more detailed discussion on the computation of these coefficients for the case of representing seismic-volcanic signals.

Spectrograms are another widely employed method for seismic-signal representation. They consist in the computation of the discrete-time Fourier transform (DFT) in small and typically overlapping time frames, reason why spectrograms are also known as the magnitude of the short-time Fourier transform. Spectrograms allow examining the "evolution" of the frequency content, which is especially useful for non-stationary signals; that is, signals whose energy distribution —in the frequency domain— changes across time as it is precisely the case of the seismic waveforms. In addition to the length of the DFT and the width of the time frames, a percentage of overlapping among the latter must be defined as well as an envelopping window to smooth the frame borders and avoid undesired high-frequency artifacts. Bell-shaped windows with a 50% of overlap are customary.

3.2. Classification of seismic-volcanic signals

Many classification methods have been proposed in the past to classify seismic signals, ranging from simple ones such as the nearest neighbor rule and Bayesian classifiers (Orozco-Alzate et al., 2006) up to complex methods such as Neural Networks (Curilem et al., 2009) and SVMs (Curilem et al., 2014; Soto et al., 2018). Even though we could choose using a state-of-the-art and complex classifier such as a SVM, we preferred to employ the simple K Nearest Neighbors rule (Duda et al., 2001) —Knn— in our experiments in order to fully understand the potentialities of the integration of the three directions.

The Knn rule represents a widely applied and well-known classifier which, when $K = 1$, assigns a test signal to the class of the signal from the training set whose distance to the testing one is minimum, i.e. the nearest neighbor. Despite its simplicity, this classifier is widely applied, since it permits to obtain highly non linear decision boundaries; moreover, by showing the nearest neighbors, it gives to the user a direct explanation of the class label that is assigned (Duin et al., 2014). From our perspective, we chose it also because it does not need density estimation or function optimization as it entirely relies on the user-defined distance measure computed on the given representation; in this sense, therefore, it is very suitable when the goal is to compare different feature representations.

As described above, the nearest neighbor rule requires the definition of a distance. Even if many different distances can be used, here we resort to the simplest choice, namely the Euclidean distance, which, for two vectors $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^d$, with $\mathbf{x} = \{x_1 \cdots x_d\}$, $\mathbf{y} = \{y_1 \cdots y_d\}$, is defined as: $d_E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$. Let $\mathbf{y}$ be a test sample and $\mathscr{D} = \{(\mathbf{x}_1, \theta_1), \ldots, (\mathbf{x}_N, \theta_N)\}$ a labeled training set with N objects, the 1nn rule assigns a label $\widehat{\theta}$ to y as follows:

$$\widehat{\theta} = \theta_{n^*}, \text{ where } n^* = \underset{i=1,\ldots,N}{\arg\min}\ d_E(\mathbf{x}_i, \mathbf{y}). \tag{1}$$

For $K > 1$, the rule simply extends to majority vote among the labels of the K nearest neighbors.

3.3. Ensemble fusion methods

The conventional approach to profit from multiple recordings is to resort to ensemble fusion methods (Re and Valentini, 2012) – also referred to as "combining classifier theory" – which, in brief, combine in some way either multiple and different representations of the same data or multiple scores or decisions taken by a number of classifiers. The ultimate aim of the fusion is, of course, obtaining a more accurate classification system. Several taxonomies to group fusion methods have been proposed. The most frequent one is based on the level at which the fusion is performed (Ross and Jain, 2004). Typically, three main classes are distinguished: feature-level fusion, score-level fusion and decision-level fusion. The fusion occurs before the classifier in the feature-level while, in contrast, it occurs after the classifiers in both the score-level and the decision-level. Therefore, according to some authors —e.g. by Morvant et al. (2014)— the first fusion scheme is also called *early fusion* and the other two are also known as *late fusion*. Block diagrams of each scheme are shown in Figs. 3 and 4 for our particular case of fusing information from three axes of a single seismic sensor. The three levels of fusion are briefly presented below, particularly focusing on the combination strategies of each level that we use in this paper.

- *Feature-level fusion*: in this case the combination is performed at the representation level, i.e. by combining different representations of the same object (e.g. by concatenation of features). Then, a single classifier is trained on the combined representation.

In our experiments we investigated three common fusion strategies: i) feature concatenation ii) a vectorial summation of the feature vectors of the three axes and iii) a vectorial product of the three feature vectors. The first approach is a classic scheme, ubiquitously used in other fields like biometrics (Rattani et al., 2006; Ross and Govindarajan, 2005); the second and the third represent alternative schemes inspired by the know-how of the seismologists involved in our analysis. In all cases, features are extracted from all the three directions: with feature concatenation, such features are concatenated in a single vector (which is of dimensionality three times the dimensionality of the original features); with the second the three feature vectors are summed (in a vectorial sense), whereas with the last the three vectors are multiplied (again in vectorial sense) – in these two last cases the dimensionality remains the same.

- *Score-level fusion*: in this case, there is one classifier for each modality, based on the features extracted from the three directions. Given a signal, each classifier assigns a set of scores (such as posterior probabilities), which are then fused together via simple rules like max/min/mean. In this way a fused score is obtained, which is then used to assign the label.
- *Decision-level fusion*: also in this case we have one classifier per modality, each one computing its set of scores; then, each classifier is taking its decision based on its own set of scores. The fusion is performed afterwards by combining the decisions, for example with the majority voting rule.

More than these simple schemes, we also investigated a complex classifier combining scheme, called "trained combiners" (Duin, 2002; Kuncheva, 2014). In this scheme the scores of the different classifiers are considered themselves as features, and are used to train another classifier which "learns" how to combine them. In our case, we used the basic scheme proposed by Kuncheva et al. (2001) —which uses a nearest mean classifier over the scores— as well as other combiners that are mentioned below; see Sec. 4. For more information on trained combiners please refer to (Duin, 2002; Kuncheva, 2014; Kuncheva et al., 2001).

4. Experiments and discussion

In this section, the proposed approach of applying ensemble classification methods to triaxial seismic signals is empirically investigated. In particular, we employed a dataset containing 200 seismic events recorded at Nevado del Ruiz volcano, including triaxial examples recorded in 2008 and belonging to the two main classes of volcanic events: 100 from VT class and 100 from the LP events, all of them having different lengths depending on the duration of the corresponding seismic events; refer again to Fig. 1. These classes are typically considered the main ones because they are associated to the most important volcanic processes: fracture of rocks due to internal pressure and transport of fluids such as magma and gases, respectively. For the sake of simplicity but without loss of generality, we restrict ourselves to triaxial signals from a single sensor located at Olleta station, which is considered the reference one by the experts and acquires the signals at a sampling rate of 100 Hz.

As summarized in the previous section, the seismic signals have been characterized using two methodologies: in the first (MFCC) we used mel Frequency Cepstral Coefficients, whereas in the second (Avg. Spect.) we used averaged spectrograms. More in detail, the former

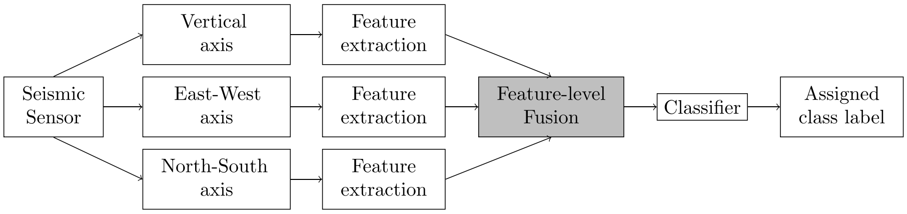


Fig. 3. Feature-level (early) fusion scheme for triaxial seismic-volcanic signals.

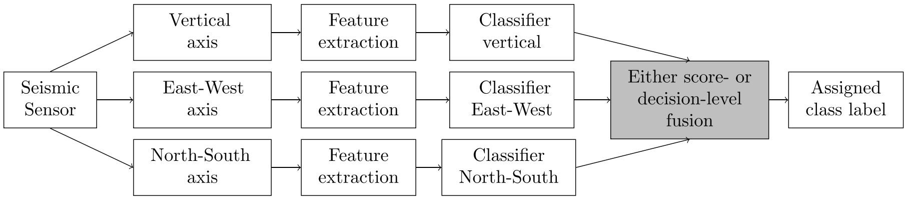


Fig. 4. Late fusion scheme: Combination at either score- or decision-level for triaxial seismic-volcanic signals.

Table 1
A summary of the shortest and longest events in the dataset, along with the resulting sizes of the corresponding spectrograms.

	LP events	VT events
Average Seq Length	7746.07	8582.88
Min Seq Length	2581.00	1970.00
Max Seq Length	24000.00	36000.00
Average Spectrogram length	153.51	170.27
Min Spectrogram length	50.00	38.00
Max Spectrogram length	479.00	719.00

Table 2
Detailed results per method and feature representation. All the reported values correspond to LOO classification errors. The best value of K in the range $\{1,3,5, \ldots ,25\}$ is shown between parentheses next to the reported classification errors.

	MFCC		Avg. Spect.	
Method	Not scaled	Scaled	Not scaled	Scaled
Monomodal Vertical	0.325(7)	0.390(11)	0.310(17)	0.345(1)
Monomodal North-South	0.250(25)	0.300(25)	0.205(9)	0.235(3)
Monomodal East-West	0.225(23)	0.270(13)	0.185(1)	0.200(3)
Feature level: concatenation	0.200(11)	0.285(11)	0.225(1)	0.305(9)
Feature level: vect summation	0.200(13)	0.230(23)	0.205(1)	0.285(7)
Feature level: vect product 1	0.340(15)	0.425(3)	0.265(9)	0.335(13)
Feature level: vect product 2	0.315(5)	0.405(11)	0.230(5)	0.360(1)
Score level: mean	0.205(25)	0.260(25)	0.135(1)	0.225(5)
Score level: median	0.215(25)	0.255(23)	0.175(1)	0.230(3)
Score level: prod	0.205(25)	0.260(25)	0.135(1)	0.225(5)
Score level: max	0.215(17)	0.265(25)	0.135(1)	0.250(1)
Score level: min	0.215(17)	0.265(25)	0.135(1)	0.250(1)
Decision level: majority voting	0.215(25)	0.255(23)	0.175(1)	0.230(3)
Trained Combiners (nmc)	0.220(9)	0.245(9)	0.145(1)	0.225(7)
Trained Combiners (ldc)	0.205(17)	0.250(7)	0.170(1)	0.205(3)
Trained Combiners (knnc)	0.215(13)	0.265(17)	0.150(1)	0.240(1)
Trained Combiners (ologc)	0.210(17)	0.260(7)	0.170(1)	0.210(3)
Trained Combiners (rbsvc)	0.210(7)	0.255(7)	0.160(1)	0.230(5)
Trained Combiners (1nn)	0.265(17)	0.315(17)	0.235(11)	0.300(5)

representation consists of a base vector of 13 coefficients (12 cepstral coefficients and the frame log-energy) plus their first order time derivatives, calculated to take into account the frame information. The latter representation consists in computing the spectrogram for each signal by means of the FFT, computing then the average. In particular, we used 1-s frames, a 128-point FFT, a 64-point Hamming window and an overlap of 50%. At the end, the feature vector generated by this representation corresponds to the mean value of each frequency band of the spectrogram across time. After the representation stage, only the values of the first half of this vector, along with the value in the middle of it, are required due to the mirror property of the FFT; thereby, the effective length of the feature vector for the classification stage is 65. A summary of the shortest and longest events in the dataset, along with the resulting sizes of corresponding spectrograms, is presented in Table 1. Remember that the sampling rate is 100 Hz; therefore, the shortest event lasts 19.7 s and the longest one 6 min.

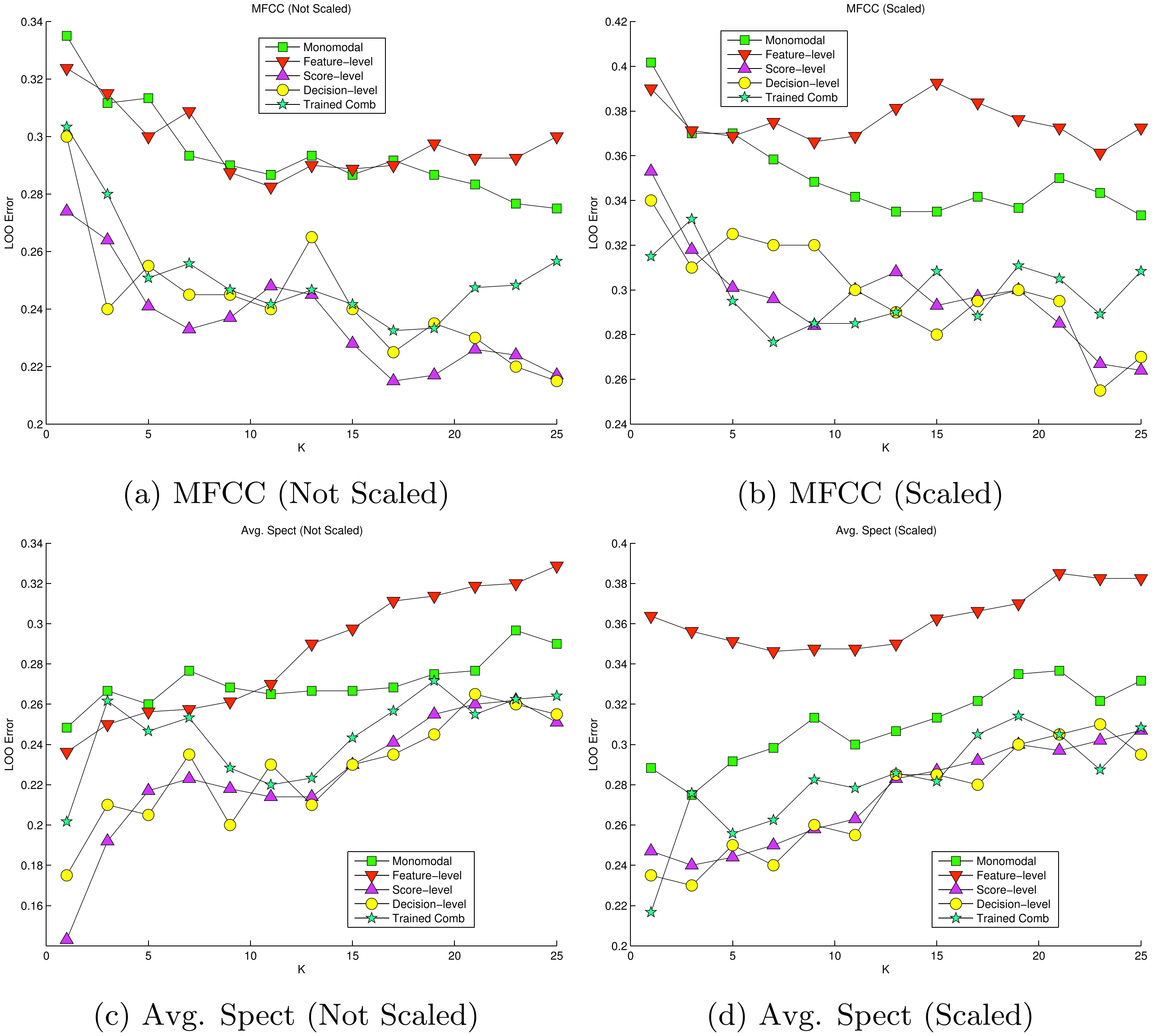
In some cases, space standardization is fundamental to get proper accuracies. Here we investigate its impact, by analyzing both unnormalized spaces as well as spaces normalized with a z-score standardization (to every feature we subtract the mean, dividing then by the standard deviation; in this way all directions of the space have zero mean and unit variance). Summarizing, in total we have four configurations: MFCC scaled, MFCC not scaled, averaged spectrograms scaled and averaged spectrograms not scaled.

On top of these four representations we tested different schemes:

- *Monomodal*: this represents the baseline, i.e. the system is designed employing a signal derived from a single recording axis. In particular we analyzed the Vertical, the North-South (N-S) and the East-West (E-W) directions.
- *Feature-level fusion*: here we investigated the three approaches described in the previous section, namely Feature Concatenation (feature vectors of Vertical, N-S and E-W are concatenated), Vectorial Summation (feature vectors of Vertical, N-S and E-W are summed), and Vectorial Product (feature vectors of Vertical, N-S and E-W are multiplied). In this last case, since the vectorial product may change depending on the order of association, we investigated two variants. In particular, given

Table 3
A summary of the best results per representation and mode; namely, single recording axis or *monomodal* and multiple recording axes or *multimodal*.

Representation	Best monomodal	Best multimodal
MFCC (Not Scaled)	0.225 (E-W)	0.200 (Feature level: conc., vect sum)
MFCC (Scaled)	0.270 (E-W)	0.230 (Feature level: vect summation)
Avg. Spect (Not Scaled)	0.185 (E-W)	0.135 (Score level: mean, prod, max, min)
Avg. Spect (Scaled)	0.200 (E-W)	0.205 (Trained Combiners (ldc))



(a) MFCC (Not Scaled)

(b) MFCC (Scaled)

(c) Avg. Spect (Not Scaled)

(d) Avg. Spect (Scaled)

Fig. 5. Leave-one-out classification errors, for each feature representation, as varying the parameter K of the Knn classifier in the range {1,3,5, …,25}.

three vectors A, B, C, we have that: $A \times (B \times C) = B(A \cdot C) - C(A \cdot B)$ and $(A \times B) \times C = -C \times (A \times B) = -A(B \cdot C) + B(A \cdot C)$, where × denotes the vectorial product, whereas · denotes the dot product.

- *Score-level fusion*: here we investigated some classical approaches for combining the posteriors of the classifiers trained on the three signals: the mean of the scores, that is, the final score is the average of the scores given by the individual classifiers and, similarly, other four approaches: the median, the maximum, the minimum and the product of the scores; respectively. All these rules have a clear theoretical interpretation, linked to aspects like complementarity, accuracy of single classifiers and so on; for more information see (Kittler et al., 1998).
- *Decision-level fusion*: here we investigated the classical majority voting rule.
- *Trained combiners*: here we investigated different classifiers built in the score spaces, such as the nmc which stands for the nearest mean classsifier – as in the original scheme introduced by (Kuncheva et al., 2001), the ldc (Linear Bayes classifier assuming normal densities with equal covariance matrices), the Knnc (the K-nearest neighbor scheme, where K is estimated through cross-validation in the training set), the loglc (Linear classifier by maximizing the likelihood criterion using the logistic function), the rbsvc (Support vector machine with rbf kernel whose parameter has been set via cross-validation on the training set) and 1nn (the nearest neighbor). Regarding all these classifiers, please refer to (Duda et al., 2001) for methodological explanations and to (Duin et al., 2007) for implementation details.

Classification errors were computed using the classic Leave-One-Out (LOO) cross-validation procedure (Bramer, 2016) which, according to (Wong, 2015), should be adopted when the number of instances in a dataset is small. With this procedure, the classifier is trained with all the signals except one, which is then used for testing. Then the procedure is repeated by leaving out the second and so on, until all signals have been tested. In this way the test set is always separated from the training set (this permits to measure generalization capabilities), whereas the size of the training set is maximized (this permits to have good classifiers). Two additional advantages of LOO cross-validation are that (i) it does not involve a randomness mechanism and, therefore, research reproducibility is allowed and, (ii) it is approximately unbiased for the expected error (Hastie et al., 2009, p. 242) and, moreover, truly almost unbiased for the nearest neighbor methods which have been proven to be stable[2] by Elisseeff and Pontil (2003).

All the results are shown in Table 2, for the different schemes, different feature representations and the optimal value for the parameter K of the Knn classifier. A summary of the best results, for the different representations, is presented in Table 3. Regarding the classifier combining rules, for implementing score-level fusion we need scores (e.g. posteriors), which in the nearest neighbor case can be computed by exploiting the distance of the nearest neighbor of each class.

From the summarizing table we can immediately note that for all feature representations there is always an improvement when combining the three axes, which is in some cases very relevant. Please note that for all experiments we compute the variance of the LOO error, following the formula given in (Kohavi, 1995): for a test involving N objects, and given the LOO error e, the variance is $\frac{e(1-e)}{N}$. Considering all experiments, the largest variance was 0.0012, thus making the reported differences significant. As a second comment, we can also observe that for every representation we have a different "best" fusion method, thus confirming that finding a proper fusion scheme which works well in all situations can not be so trivial.

Some additional observations can be derived from the full Table 2. First, we can observe that, when considering a single direction, the vertical axis is not the one leading to the best accuracies in contrast with the choices made by authors in this field in the past. Nonetheless, remember that in our dataset the vertical axis exhibits some noise. The E-W axis is, in general, the best performing one; however, it is important to notice also that the accuracies obtained in the N-S direction are not so far. This may suggest that the information contained in the N-S axis is indeed useful for classification since it is possible that the hypocenters location of some events are just to the north of the recording station. This is, therefore, another reason to use three components instead of only one, that is, the location of the earthquake with respect to the station can influence the waveform and all its attributes.

When considering fusion strategies, we can observe that in general late fusion methods performed better than early fusion schemes, even if the obtained accuracies strictly depend on the adopted score-level rule (or the adopted classifier in the score space for trained combiners). In particular ldc in the score space seems to be the best, whereas 1nn represents the worst choice. Concerning the feature representations, we can observe that the best one is by far the averaged spectrogram, especially in the not scaled version: this confirms the findings reported in (Castro-Cabrera et al., 2014). It is also important to note that scaling has always a bad effect on the accuracies; this suggests that absolute differences in magnitudes are also relevant for the discrimination.

Finally, we explored the influence of the parameter K on the performance of the Knn rule. The five types of methodologies (monomodal plus four types of fusion) were tested while varying the parameter in the range $\{1,3,5, \ldots, 25\}$, then, LOO errors were plotted as a function of K; see Fig. 5. In particular, for all the methodologies, we made the average: e.g. for monomodal we averaged the results of Vert, N-S and E-W; similarly, for Score-level, we averaged the results of mean, median, prod, max and min; and so on. Notice that there is a general clear improvement when using late fusion schemes, whereas for feature level fusion there is no improvement. This was not evident from the results in the tables where, for MFCC scaled, the best result was obtained with a feature level method. Moreover there is a mixed behavior for what concerns K: for MFCC the larger the better, for spectrograms the lower the better.

5. Conclusion

The three recording axes of the seismic sensors convey discriminant and non-redundant information of the seismic events. Such a diversity can be effectively exploit by using ensemble classification methods that might combine the information at different levels, namely at the representation/feature-level, the classifier score-level or the classifier decision-level. In this paper we tried all of them for several alternatives of both feature representations and combination rules. We observed that, overall, the best representation corresponds to non-scaled averaged spectrograms together with score-level rules to combine posteriors of classifiers independently trained by each recording axis. By using all these modalities, better automated classification systems can be designed and, thereby, a more accurate tool to help with the complex volcanic risk-assessment systems would be available at the volcano observatories. Further subsequent studies on ensemble classification methods for the classification of seismic-volcanic signals must face the challenging task of taking into account, simultaneously, all the available information from multiple recording stations and the three orthogonal axes of each seismic sensor.

Acknowledgments

The authors thank Paola Alexandra Castro-Cabrera for her assistance with the extraction of the subset of seismic signals used in the experiments. Part of this work was done while the the first author was visiting the Computer Science Department at Università degli Studi di Verona, Italy, as one of the activities of his Sabbatical Year granted by Universidad Nacional de Colombia - Sede Manizales.

References

Álvarez, I., Garca, L., Cortés, G., Bentez, C., De la Torre, Á., mar 2012. Discriminative feature selection for automatic classification of volcano-seismic signals. IEEE Geosci. Remote Sens. Lett. 9 (2), 151–155. https://doi.org/10.1109/lgrs.2011.2162815.

Becerra Yoma, N., Carrasco, J., Molina, C., 2005. Bayes-based confidence measure in speech recognition. IEEE Signal Process. Lett. 12 (11), 745–748. https://doi.org/10.1109/lsp.2005.856888.

Bicego, M., Acosta-Muñoz, C., Orozco-Alzate, M., jun 2013. Classification of seismic volcanic signals using hidden-Markov-model-based generative embeddings. IEEE Trans. Geosci. Remote Sens. 51 (6), 3400–3409. https://doi.org/10.1109/TGRS.2012.2220370.

Bicego, M., Londoño-Bonilla, J.M., Orozco-Alzate, M., 2015. Volcano-seismic events classification using document classification strategies. In: Murino, V., Puppo, E. (Eds.), Image Analysis and Processing - ICIAP 2015. 18th International Conference, Genoa, Italy, September 7–11, 2015, Proceedings, Part I. Vol. 9279 of Lecture Notes in Computer Science. GIRPR/IAPR, Springer, Cham, Switzerland, pp. 119–129.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, New York, NY.

Bramer, M., 2016. Ch. 7: Estimating the Predictive Accuracy of a Classifier. In: Principles of Data Mining, 3rd Edition. Undergraduate Topics in Computer Science. Springer, London, UK, pp. 79–92. https://doi.org/10.1007/978-1-4471-7307-6_7.

Cárdenas-Peña, D., Orozco-Alzate, M., Castellanos-Domnguez, G., feb 2013. Selection of time-variant features for earthquake classification at the Nevado-del-Ruiz volcano. Comput. Geosci. 51, 293–304. https://doi.org/10.1016/j.cageo.2012.08.012.

Castro-Cabrera, P.A., Orozco-Alzate, M., Adami, A., Bicego, M., Londoño-Bonilla, J.M., Castellanos-Domnguez, C.G., 2014. A comparison between time-frequency and cepstral feature representations for seismic-volcanic pattern classification. In: Bayro-Corrochano, E., Hancock, E. (Eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Proceedings of the 19th Iberoamerican Congress on Pattern Recognition, CIARP 2014. IAPR, Springer, Berlin Heidelberg, pp. 440–447. https://doi.org/10.1007/978-3-319-12568-8_54. Vol. 8827 of Lecture Notes in Computer Science.

Cortés, G., Garca, L., Álvarez, I., Bentez, C., De la Torre, Á., Ibañez, J., 2014. Parallel system architecture (PSA): an efficient approach for automatic recognition of volcano-seismic events. J. Volcanol. Geotherm. Res. 271, 1–10. URL. https://doi.org/10.1016/j.jvolgeores.2013.07.004.

Cortés, G., Bentez, M.C., Garca, L., Álvarez, I., Ibáñez, J.M., jan 2016. A comparative study of dimensionality reduction algorithms applied to volcano-seismic signals. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sensing 9 (1), 253–263. https://doi.org/10.1109/jstars.2015.2479300.

Curilem, G., Vergara, J., Fuentealba, G., Acuña, G., Chacón, M., 2009. Classification of

[2] A classification method is said to be stable if removing one training sample does not significantly change the classification outcomes, even when the size of the dataset is small. Elisseeff and Pontil (2003) showed that Knn methods are stable while other popular ones —for instance support vector machines (SVMs)— are not.

seismic signals at Villarrica volcano (Chile) using neural networks and genetic algorithms. J. Volcanol. Geotherm. Res. 180 (1), 1–8. https://doi.org/10.1016/j.jvolgeores.2008.12.002.

Curilem, M., Huenupan, F., San-Martn, C., Fuentealba, G., Cardona, C., Franco, L., Acuña, G., Chacón, M., nov 2014. Feature analysis for the classification of volcanic seismic events using support vector machines. In: Gelbukh, A., Castro Espinoza, F., Galicia-Haro, S. N. (Eds.), Nature-Inspired Computation and Machine Learning: Proceedings of the 13th Mexican International Conference on Artificial Intelligence, MICAI Springer, Cham, pp. 160–171.

Curilem, M., Huenupan, F., Beltrán, D., San-Martn, C., Fuentealba, G., Franco, L., Cardona, C., Acuña, G., Chacón, M., Khan, M.S., Yoma, N.B., 2016. Pattern recognition applied to seismic signals of Llaima volcano (Chile): an evaluation of station-dependent classifiers. J. Volcanol. Geotherm. Res. 315, 15–27. URL. https://doi.org/10.1016/j.jvolgeores.2016.02.006.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, 2nd edition. John Wiley & Sons, Inc.

Duin, R.P.W., 2002. The combining classifier: To train or not to train? In: Proceedings of the 16th International Conference on Pattern Recognition. Vol. 2. IEEE Computer Society, Los Alamitos, CA, USA, pp. 765–770. https://doi.org/10.1109/icpr.2002.1048415.

Duin, R.P.W., Juszczak, P., Pekalska, E., de Ridder, D., Tax, D.M.J., Verzakov, S., 2007. PRTools 4.1: A Matlab Toolbox for Pattern Recognition. Delft University of Technology, The Netherlands.

Duin, R.P.W., Orozco-Alzate, M., Londoño-Bonilla, J.M., aug 2010. Classification of volcano events observed by multiple seismic stations. In: Proc. of the 20th Int. Conf. On pattern recognition (ICPR2010). IEEE computer Society, pp. 1052–1055. https://doi.org/10.1109/ICPR.2010.263.

Duin, R. P. W., Bicego, M., Orozco-Alzate, M., Kim, S.-W., Loog, M., aug 2014. Metric learning in dissimilarity space for improved nearest neighbor performance. In: Fränti, P., Brown, G., Loog, M., Escolano, F., Pelillo, M. (Eds.), Structural, Syntactic and Statistical Pattern Recognition: Proceedings of the Joint IAPR International Workshop, S + SSPR 2014. Vol. 8621 of Lecture Notes in Computer Science. IAPR, Springer, Berlin Heidelberg, pp. 183–192.

Elisseeff, A., Pontil, M., 2003. Leave-one-out error and stability of learning algorithms with applications. In: Suykens, J.A.K., Horvath, G., Basu, S., Micchelli, C., Vandewalle, J. (Eds.), Advances in Learning Theory: Methods, Models and Applications. Vol. 190 of NATO Science Series, III: Computer and Systems Sciences. IOS Press, pp. 111–130. URL. https://www.esat.kuleuven.be/sista/natoasi/.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, USA.

Keller, E.A., 2011. Environmental Geology, 9th edition. Prentice Hall, Upper Saddle River, New Jersey.

Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., 1998. On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20 (3), 226–239. https://doi.org/10.1109/34.667881.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95. Vol. 2. Morgan Kaufmann, pp. 1137–1143.

Kuncheva, L.I., 2014. Combining Pattern Classifiers: Methods and Algorithms, 2nd edition. Wiley, Hoboken, NJ. https://doi.org/10.1002/9781118914564.

Kuncheva, L.I., Bezdek, J.C., Duin, R.P.W., 2001. Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recogn. 34 (2), 299–314. https://doi.org/10.1016/S0031-3203(99)00223-X.

Lara-Cueva, R.A., Bentez, D.S., Carrera, E.V., Ruiz, M., Rojo-Álvarez, J.L., sep 2016. Automatic recognition of long period events from volcano tectonic earthquakes at Cotopaxi Volcano. IEEE Trans. Geosci. Remote Sens. 54 (9), 1–11. https://doi.org/10.1109/tgrs.2016.2559440.

Malfante, M., Dalla Mura, M., Métaxian, J.-P., Mars, J.I., Macedo, O., Inza, A., mar 2018. Machine learning for volcano-seismic signals: challenges and perspectives. IEEE Signal Process. Mag. 35 (2), 20–30. https://doi.org/10.1109/msp.2017.2779166.

Morvant, E., Habrard, A., Ayache, S., aug 2014. Majority vote of diverse classifiers for late fusion. In: Fränti, P., Brown, G., Loog, M., Escolano, F., Pelillo, M. (Eds.), Structural, Syntactic and Statistical Pattern Recognition: Proceedings of the Joint IAPR International Workshop, S + SSPR 2014. Vol. 8621 of Lecture Notes in Computer Science. IAPR, Springer, Berlin Heidelberg, pp. 153–162.

Orozco-Alzate, M., Garca-Ocampo, M.E., Duin, R.P.W., Castellanos-Domnguez, C.G., Dec 2006. Dissimilarity-based classification of seismic volcanic signals at Nevado del Ruiz volcano. Earth Sci. Res. J. 10 (2), 57–65.

Orozco-Alzate, M., Acosta-Muñoz, C., Londoño-Bonilla, J.M., 2012. The automated identification of volcanic earthquakes: concepts, applications and challenges. In: D'Amico, S. (Ed.), Earthquake Research and Analysis - Seismology, Seismotectonic and Earthquake Geology. InTech, Rijeka, Croatia, pp. 345–370. https://doi.org/10.5772/27508.

Orozco-Alzate, M., Castro-Cabrera, P.A., Bicego, M., Londoño-Bonilla, J.M., 2015. The DTW-based representation space for seismic pattern classification. Comp. Geosci. 85, 86–95.

Pan, S.J., Yang, Q., oct 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359. https://doi.org/10.1109/tkde.2009.191.

Rattani, A., Kisku, D.R., Bicego, M., Tistarelli, M., 2006. Robust feature-level multibiometric classification. In: Proceedings of the IEEE biometrics symposium (BSYM06), pp. 1–6.

Re, M., Valentini, G., 2012. Ensemble methods: A review. In: Way, M.J., Scargle, J.D., Ali, K.M., Srivastava, A.N. (Eds.), Advances in Machine Learning and Data Mining for Astronomy. Data Mining and Knowledge Discovery Series. CRC Press, Boca Raton, FL, pp. 563–593.

Ross, A., Govindarajan, R., mar 2005. Feature level fusion using hand and face biometrics. In: Proc. SPIE: Biometric Technology for Human Identification II. vol. 5779. pp. 196–205.

Ross, A., Jain, A. K., sep 2004. Multimodal biometrics: An overview. 12th European signal processing conference, 1221–1224.

Soto, R., Huenupan, F., Meza, P., Curilem, M., Franco, L., jun 2018. Spectro-temporal features applied to the automatic classification of volcanic seismic events. J. Volcanol. Geotherm. Res. 358, 194–206. https://doi.org/10.1016/j.jvolgeores.2018.04.025.

Wong, T.-T., 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recogn. 48 (9), 2839–2846. https://doi.org/10.1016/j.patcog.2015.03.009.