Journal of Volcanology and Geothermal Research 316 (2016) 34–49

Contents lists available at ScienceDirect

Journal of Volcanology and Geothermal Research

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

Feature selection of seismic waveforms for long period event detection at Cotopaxi Volcano

R.A. Lara-Cueva [a,c,*], D.S. Benítez [d], E.V. Carrera [a], M. Ruiz [e], J.L. Rojo-Álvarez [b,c]

[a] Departamento de Eléctrica y Electrónica, Universidad de las Fuerzas Armadas ESPE, Av. Gral. Rumiñahui s/n, Sangolquí, Ecuador
[b] Prometeo Program, Departamento de Eléctrica y Electrónica, Universidad de las Fuerzas Armadas ESPE, Av. Gral. Rumiñahui s/n, Sangolquí, Ecuador
[c] Information and Communications Technology Department, Rey Juan Carlos University, Camino del Molino s/n, 28943 Fuenlabrada, Madrid, Spain
[d] Universidad San Francisco de Quito USFQ, Colegio de Ciencias e Ingenierías El Politécnico, Campus Cumbayá, Casilla Postal 17-1200-841 Quito, Ecuador
[e] Instituto Geofísico, Escuela Politécnica Nacional, Ladrón de Guevara E11-253, Quito, Ecuador

ARTICLE INFO

Article history:
Received 1 February 2015
22 February 2016
Accepted 23 February 2016
Available online 3 March 2016

Keywords:
Machine learning
k-NN and decision trees
Feature extraction and selection
Volcano seismic classification
Seismic event detection

ABSTRACT

Volcano Early Warning Systems (VEWS) have become a research topic in order to preserve human lives and material losses. In this setting, event detection criteria based on classification using machine learning techniques have proven useful, and a number of systems have been proposed in the literature. However, to the best of our knowledge, no comprehensive and principled study has been conducted to compare the influence of the many different sets of possible features that have been used as input spaces in previous works. We present an automatic recognition system of volcano seismicity, by considering feature extraction, event classification, and subsequent event detection, in order to reduce the processing time as a first step towards a high reliability automatic detection system in real-time. We compiled and extracted a comprehensive set of temporal, moving average, spectral, and scale-domain features, for separating long period seismic events from background noise. We benchmarked two usual kinds of feature selection techniques, namely, filter (mutual information and statistical dependence) and embedded (cross-validation and pruning), each of them by using suitable and appropriate classification algorithms such as *k* Nearest Neighbors (*k*-NN) and Decision Trees (DT). We applied this approach to the seismicity presented at Cotopaxi Volcano in Ecuador during 2009 and 2010. The best results were obtained by using a 15 s segmentation window, feature matrix in the frequency domain, and DT classifier, yielding 99% of detection accuracy and sensitivity. Selected features and their interpretation were consistent among different input spaces, in simple terms of amplitude and spectral content. Our study provides the framework for an event detection system with high accuracy and reduced computational requirements.

1. Introduction

Volcano monitoring systems have been deployed as an attempt to mitigate risks, to forecast eruptions, and to assess hazards, due to the necessity of safeguarding human lives and resources. These monitoring systems use principled information related to ground deformation (Dzurisin, 1980; Dvorak and Dzurisin, 1997; Voight et al., 1998; Bonaccorso et al., 2006), gas flux (Baubron et al., 1991; Galle et al., 2003; Lewicki et al., 2003), seismicity (McNutt, 1996; Chouet and Matoza, 2013; Sparks, 2003), and other factors, as main monitoring measurements to determine the activity of volcanoes. In this context, seismology is an important and effective tool for monitoring volcanoes, since seismicity is the fastest and most commonly used method in order to detect changes on volcanoes, by assessing earthquakes and other ground vibrations sensed by the seismometers or geophones networks (McNutt, 2000; Sicali et al., 2015; Papadimitriou et al., 2015). The events recorded in these systems present differences in their seismic wave patterns so their seismological signature can be interpreted by analysts to identify different types of events. For instance, most volcanoes present Volcano Tectonic (VT) earthquakes, Long Period (LP) events, Tremors (TRE), and Hybrid (HYB) events. Other non-volcanic originated events, such as Lightnings (LGH), can be occasionally recorded by seismometers (Behnke et al., 2013).

Several techniques have been developed for automatic identification of events, such as stochastic processes analysis, mathematical modeling, and signal processing in time, frequency, and scale domains (Scarpetta et al., 2005). The latter refers to the use of the wavelet transform for a time-scale domain analysis of signals with fast changing spectral contents, using wavelets it is possible to represent a signal in a time and frequency response scale where any event contained in the signal will mark an entire region in the time-scale plane, solving therefore the

* Corresponding author at: Departamento de Eléctrica y Electrónica, Universidad de las Fuerzas Armadas ESPE, Av. Gral. Rumiñahui s/n, Sangolquí, Ecuador.
E-mail addresses: ralara@espe.edu.ec (R.A. Lara-Cueva), dbenitez@usfq.edu.ec (D.S. Benítez), evcarrera@espe.edu.ec (E.V. Carrera), mruiz@igepn.epn.edu.ec (M. Ruiz), joseluis.rojo@urjc.es (J.L. Rojo-Álvarez).
URL: http://www.wicom.espe.edu.ec (R.A. Lara-Cueva).

http://dx.doi.org/10.1016/j.jvolgeores.2016.02.022
0377-0273/

resolution problem presented by the Fourier transform in the frequency domain. The stated advantage of using wavelets compared to Fourier transform is the trade-off between frequency and time resolution at different frequencies. Special attention has been put in classification techniques from Machine Learning Theory, since these methods are capable of describing in detail patterns from different types of events, as it will be summarized latter in Section 2. However, to the best of our knowledge, no comprehensive and principled study has been conducted in previous works to compare the influence of the many different sets of features that have been used to approximate the input space, defined in terms of the possible values that the input parameter can have.

The aim of this work is to present an automatic recognition system for volcano seismicity, by considering all the stages of the process including feature extraction, feature selection, and event classification, and in order to provide us with a high-fidelity event detection, as a first step to an automatic detection system in real-time. Our primary hypothesis is that a carefully designed feature extraction with a suitable and appropriate machine learning technique will reduce the processing time and will avoid overfitting. We address here the optimal design of a detection system, based on classification techniques, for separating LP seismic events from seismic background noise with high accuracy, since LP events often precede volcanic eruptions and they are accordingly used in forecasting (Chouet and Matoza, 2013; Chouet, 1996; Trombley and Toutain, 2005; Lyons et al., 2014; Bean et al., 2014; Cusano et al., 2015; Syahbana et al., 2014). The classification among the other types of events is beyond our scope at this stage, but it could be specifically addressed with the proposed approach. We benchmarked two commonly used feature selection techniques, namely, filter and embedded, each of them in a suitable and appropriate classification algorithm, namely, *k* Nearest Neighbors (*k*-NN) and Decision Trees (DT).

Our study refers to Cotopaxi, an active volcano, located in the so-called Ring of Fire at Ecuador, in which a permanent monitoring system (24 h/7 days a week) has been previously deployed (Ortiz Erazo, 2013). Its activity produced over 100 MB of data per day during 2009 and 2010, which means an average of 21 and 17 events per day, respectively. There are 16 seismological stations deployed at Cotopaxi Volcano as highlighted in Section 3, therefore, expert scientists must daily analyze the vertical component of 16 seismograms of volcanic signals by visual inspection in order to label and classify the events. Currently, its activity is increasing and presents 130 events on average per day, too many records are generated during periods of high volcanic activity; therefore data assessment can become an extremely slow process, which can also cause damming of information (Newman and Jain, 1995; Mery and Medina, 2004).

The rest of the paper is organized as follows. Section 2 summarizes previous works and results about the automatic classification of seismic events by using machine learning techniques. Section 3 describes the dataset used in this work, which were collected at Cotopaxi Volcano. Section 4 describes the proposed approach and the experimental study including feature extraction and event detection. Section 5 presents the results obtained in automatic classification and detection processes for different segmentation windows and with different classifiers. Finally, Section 6 presents the discussion and some concluding conclusions.

2. Detecting seismic events

One of the main goals of volcano monitoring institutions around the world is to understand the behavior of volcanoes, in order to forecast a possible or imminent eruption for safeguarding lives, which is achievable by sensing and distinguishing the increase of volcanic events. In relation to this, the analysts visually identify seismic events received from remote seismometers by actually using digital signal processing techniques both in the time and the frequency domain to make this process more efficient, so event data, such as timestamps of start and end, time duration, and arrival times, can be stored for future reference.

In this context, several techniques have been developed to support vulcanologist in the automatic classification process, and some authors have used supervised (Falsaperla et al., 1996; Langer et al., 2003, 2006; Curilem et al., 2009) or unsupervised (Messina and Langer, 2011; Esposito et al., 2008; Ohrnberger, 2001) learning techniques, in order to distinguish among two or more classification groups of events.

In Ibáñez et al. (2009), for example, Etna and Stromboli volcanoes were studied in terms of VT and TRE events for the first volcano, and background noise and Very Long Period (VLP) events for the second one. Authors found a total of 39 data parameters of main temporal and spectral characteristics, including coefficients of the time evolution of the signal and the energy in a frequency band, by using Hidden Markov Models (HMM) as classifier, yielding 86% and 84% of successful classification rates, respectively.

Meanwhile in Álvarez et al. (2012), a 1 to 25 Hz band-pass filter, first using a time windowing of 4 s and then extended it to 8 s was used to extract temporal and spectral characteristics from data obtained at Colima Volcano in Mexico, yielding two proposed feature vectors with 39 and 84 features, extended the feature vector defined in Ibáñez et al. (2009) by considering the presence or absence of harmonics and the spectral envelope. Discriminative Feature Selection (DFS) based on the Minimum Classification Error (MCE) criterion, and a Gaussian Mixture Model (GMM) were used, which reduced the original sets to 14 and 10 features, respectively. The misclassification percentage was improved for the first feature vector set from 24% to 16%, and for the extended feature vector from 28% to 14%. However, the main features, which improved the results, were not mentioned in such work.

A feature extraction and a subsequent feature selection steps were developed for Nevado del Ruiz Volcano in Colombia (Cárdenas-Peña et al., 2013), considering VT, LP, TRE, and HYB events. In this work, a feature selection strategy was developed based on the relevance of time variant features, i.e., the most significant set of features or those with the greatest contribution to the event, and the results were compared to the use of HMM and *k*-NN. With this approach, the classification error rate was improved from 22% to 12% when using *k*-NN instead of HMM. Another system was defined in Cortés et al. (2014), which used the GMM classifier, which obtained a baseline recognition rate of 92% by using the feature vector with the main features via DFS at Deception Volcano Island in Antarctica.

Although previous works have demonstrated the possibilities of using machine learning techniques in this setting, the literature lacks of supportive evidence about which are the main design parameters to be considered in each stages for signal preprocessing, feature extraction, feature selection, classification, and detection, especially for real-time or near-real-time detection systems.

3. Dataset description

The *Instituto Geofísico de la Escuela Politécnica Nacional* (IGEPN) is the institution responsible for monitoring the seismic activity in Ecuador. IGEPN has installed a high quality seismometer network, covering near 70% of the country. The monitoring stations collect data every day and continuously transmit them to a volcano observatory, which is located 40 km from the Cotopaxi Volcano. These data are transmitted by using radio links in the UHF band.

Cotopaxi is a snow-capped volcano located at latitude 00° 41′ 05″ S and longitude 78° 25′ 54.8″ W in the Andean mountain region of Ecuador. We have chosen this particular volcano for our study given its high hazard and risk status of future eruptions. Cotopaxi has experienced 5 eruptive cycles with 13 significant eruptions since 1534, and past eruptions have produced pyroclastic flows, ash and lapilli falls, lava flows, and far reaching lahars (Hall and Mothes, 2008). The Cotopaxi Volcano is located 40 km from Quito and near to the city of Latacunga, and a potential eruption will directly affect about 800,000 people living in the surrounding area of the volcano. However, being also so close to Quito, a city with a population of over two million inhabitants, the number of people that