UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Volcanic Seismic Events Classification using Unsupervised Learning Models

Kevin Alexandro González Castro

Ingeniería en Sistemas

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero en Sistemas

Quito, 7 de mayo de 2020

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Volcanic Seismic Events Classification using Unsupervised Learning Models

Kevin Alexandro González Castro

Calificación:

Nombre del profesor, Título académico **Noel Pérez, Ph.D.**

Firma del profesor: _____

Quito, 7 de mayo de 2020

DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Firma del estudiante: _____

Nombres y apellidos: Kevin Alexandro González Castro

Código: 00118375

Cédula de identidad: 1715437669

Lugar y fecha: Quito, 07 mayo de 2020

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

Este documento explora el uso de seis diferentes clasificadores basados en clustering, para categorizar dos diferentes eventos sísmico-volcánicos y encontrar posibles señales solapadas que pueden ocurrir al mismo tiempo o inmediatamente después de la aparición de eventos sísmicos. De acuerdo con el espacio de clasificadores explorado, el spectral-clustering con k=2 fue escogido como el mejor modelo, alcanzando una precisión del 92%. Este resultado representa un desempeño satisfactorio y competitivo en cuanto a clasificación, comparado con los métodos señalados en el estado de arte. Además, el clasificador CURE con k=3 alcanzó una precisión del 87%, la misma que es considerada también como un desempeño razonable. Este modelo fue el más eficiente en la detección de señales solapadas en los eventos sísmico-volcánicos. Considerando los resultados obtenidos, es posible establecer que la exploración propuesta, basada en clustering fue efectiva en proveer modelos competitivos para la clasificación de eventos sísmico-volcánicos y la detección de señales solapadas.

Palabras clave: categorización de eventos sísmico-volcánicos, k-means, BFR, CURE, BIRCH, Expectation-maximization, spectral-clustering, métodos de clustering, aprendizaje no supervisado.

ABSTRACT

This paper explores the use of six different clustering-based classifiers to categorize two different volcanic seismic events and to find possible overlapping signals that could occur at the same time or immediately after seismic events occurrence. According to the explored classifiers space, only one out of 27 models was selected using the first selection criteria. Afterward, the Spectral Clustering classifier with *k=2* was chosen as the best model, reaching an accuracy score of 92%. This result represents a satisfactory and competitive classification performance when compared to the state of art methods. The CURE classifier with *k=3* attained an accuracy value of 87%, enabling it as the only model to detect seismic events with overlapped signals. Therefore, the proposed clustering-based exploration was effective in providing competitive models for seismic events classification and overlapped signal detection

Keywords: volcanic seismic event categorization, k-means, BFR, CURE, BIRCH, Expectation Maximization, Spectral Clustering, clustering methods, unsupervised learning.

TABLA DE CONTENIDO

Introduction **10**

Materials and methods **12**

Volcano seismic event dataset **12**

Clustering-based classifiers **12**

k-Means method 13

BFR method 14

CURE method 15

BIRCH method 15

Expectation-maximization method 16

Spectral-clustering method 16

Experimental setup **17**

Dataset normalization 17

Model configuration 18

Assessment metrics 19

Selection criteria 19

Results and discussion **20**

Performance of explored models **20**

State of the art-based comparison **22**

Conclusion and future work **24**

Acknowledgement **24**

References **25**

ÍNDICE DE TABLAS

Table 1. ACC-based Performance Results for Explored Models.....21
Table 2. Comparison based on the ACC between previous works available in the literature and the selected best model.....23

ÍNDICE DE FIGURAS

Figure 1. Data visualization using the t-SNE technique of the cluster with k=2 from left to right: k-means, BFR, CURE (top row) and BIRCH, Expectation-maximization, Spectral-clustering (bottom row) classifiers. 21
Figure 2. Data visualization using the t-SNE technique of the cluster with k=3 from left to right: k-means, BFR, CURE (top row) and BIRCH, Expectation-maximization, Spectral-clustering (bottom row) classifiers. 22

INTRODUCTION

Volcanic eruptions have been responsible for thousands of deaths since the year 1500 (Tilling, 1996). Historical records show that between 1986 and 2019, approximately 7670 deaths were reported from direct and indirect volcanic activity worldwide. There are many highly populated cities around the world where people reside within a 30km radius to volcanoes (Siebert, Simkin, & Kimberly, 2011) (Phillipson, Sobradelo, & Gottsmann, 2013) such as Quito (Ecuador) near to Cotopaxi (last active in 2012), Guagua Pichincha (last active in 2000), and Reventador (last active in 2002) volcanoes, Mexico City (Mexico) near to Popocatepetl volcano, Tokyo (Japan) near to Mt. Fuji, Naples (Italy) close to Vesuvius, Seattle (USA) close to Mount Rainier among others (Schmincke, 2004). Currently, volcanic observatories worldwide use seismic monitoring as the most effective tool for forecasting eruptions (Schmincke, 2004). However, most of these methods involve manual classification of seismic events which could lead to delays and errors due to human subjectivity.

Machine learning classifiers with supervised or unsupervised learning have been employed during the last decade to different application contexts. Successfully, supervised learning approaches employed to face the problem of seismic events classification are artificial neural networks (ANN) (Lara-Cueva et al., 2016), random forest (Rodgers et al., 2016), hidden Markov models (Benitez et al., 2007), Gaussian mixture models (Venegas et al., 2019) and support vector machine methods (Curilem et al., 2014). On the other hand, unsupervised learning methods, which have been applied to different problems as well (Krishna & Murty, 1999), intend to form structured groups or clusters in datasets without prior knowledge of any class labels (Zheng et al., 2017). Some studies reported in the literature include: principal component analysis (PCA) (Unglert, Radić, & Jellinek, 2016), mixtures of Gaussian (Hammer,

Beyreuther, & Ohrnberger, 2012), hidden Markov models (Bebbington, 2007) and self-organizing map (SOM) (Kuyuk et al., 2011).

Approaches focusing on volcanoes and their seismic activities have been less explored, but the SOM models seeing to be the most popular. In (Köhler, Ohrnberger, & Scherbaum, 2010), a SOM model focused on volcanic wavefield patterns was used to analyze the Mount Merapi (Indonesia), classification errors of 6% and 26% were obtained for volcano-tectonic and rockfall events, respectively. However, when both events were combined into one cluster class, the error value was significantly reduced to 12%. In (Reyes & Mosquera, 2017), SOM and k-means models were used to classify volcanic signals recorded from the Tungurahua volcano (Ecuador), attaining accuracy (ACC) values of 91 % and 86% for noise and infra-sound signals, respectively. In (Messina & Langer, 2011), SOM and clustering-based models were integrated to build the KKAnalysis software, a tool that takes less than a minute to classify events, reaching an ACC value of 90%. Despite the several developed approaches, the problem of volcano seismic event classification remains an open challenge.

This paper aims to explore six different clustering-based classifiers in the context of volcano seismic events classification and overlapped signals detection. The employed models belong to the unsupervised learning models and have the advantage of being trained without knowing the output label of input instances, making it a real-life solution. The main drawback is that they are less accurate than supervised learning models.

The remainder of this paper is organized as follows: The Materials and Methods section, presents the experimental volcano seismic event dataset, the selected clustering-based classifiers and the experimental setup design used in this work. The Results and Discussion section presents an exploratory comparison based on the ACC scores obtained for each method and against the state of art-based methods. Finally, Conclusions and future work are drawn in the last section.

MATERIALS AND METHODS

Volcano seismic event dataset

A public dataset (SeisBenchV1) from the ESeismic repository, which is the first annotated Ecuadorian volcano seismic repository with several samples recorded at the Cotopaxi volcano (Benítez, et al., 2020), was used for this work. For convenience, the SeisBenchV1 dataset was provided by courtesy of the Instituto Geofísico of Escuela Politécnica Nacional (IGEPN) and collaborators, available at: http://www.igepn.edu.ec/eseismic_web_site/index.php.

The SeisBenchV1 dataset is composed of a total number of 668 already computed feature vectors distributed in 587 and 81 samples of long-period (LP) and volcanic tectonic (VT) event classes, respectively. Each vector contains a set of 84 features, including: 13 features from the time domain, 21features from the frequency domain, and 50 features from the scale domain. Since this dataset is a real-life one, it also contains samples with signal overlapped (signals that could occur at the same time or immediately after the event occurrence). Thus, some LP and VT events were recorded in conjunction with, for example, a rockfall or an icequake occurrence. This effect produced a mixed signal in the seismometer used to record the event

Clustering-based classifiers

Clustering is a term used for the process of data grouping. Data are represented as points in a multidimensional space and are placed in different clusters according to a given metric, commonly, distance measures (Pandove & Goel, 2015). We considered six different clustering-based models instead of PCA or factor analysis, which are unsupervised learning models as

well, since clustering-based models are not sensitive to the internal data correlation as could be the others. In real-life data, the correlation of features is an inherited problem; thus, the use of non-sensitive models is preferred to avoid data preprocessing steps. A brief description of selected models is presented below:

k-Means method

The k-means algorithm partitions the whole dataset into small number (k) of clusters of data in a way that the resulting intra-cluster similarity is high, but the inter-cluster similarity is low. The cluster similarity is measured regarding the Euclidean distance to the mean value of the samples in a cluster (centroid) (Tamilselvi, Sivasakthi, & Kavitha, 2015). Selecting the right value of *k* is a hard decision due to the unknown class number. Thus, the basic in the k-means model is to optimize the *k* value in a range of possible clusters (Pandove & Goel, 2015). Additionally, k-means is mainly based on the distance computation (see Equation1) between the randomly selected sample (instance to be assigned) and the centroid (cluster mean) of the considered clusters (Oliveira Martins et al., 2009). In the last step, the model recomputes the cluster centroid in which the sample was assigned (Sharma, Bajpai, & Litoriya, 2012). The process is repeated until all the samples are analyzed.

$$S = \sum_{j=1}^{k} \sum_{i=1}^{n} \left|\left|x_i^j - c_j\right|\right|^2$$

where $\left|\left|x_i^j - c_j\right|\right|^2$ is the distance from any sample x_i^j to the centroid c_j; *k* is the total number of clusters *n* is the number of samples in the dataset and *S* is the similarity value of the i^{th} sample respect to the *k* clusters.

BFR method

BFR stands for Bradley, Fayad, and Reina, who developed a variant of the k-means algorithm, which is mainly used for clustering large amounts of data (Pandove & Goel, 2015). The BFR algorithm assumes that clusters are typically distributed around centroids in a Euclidean space. On its first iteration, the whole data is read and loaded to memory. Then, it computes some simple statistic variables such as the number of points *N*, vector *SUM* and *SUMSQ* (Daoudi & Meshoul, 2017) that will serve to avoid memory full load in the next iterations. The initial *k* centroids are also estimated in the first iteration, usually by taking a random sample, picking up random points (instance of data), and then taking *k-1* more points (far as possible from the previous ones). There are three classes of points that are using to represent the data and to perform the inclusion of a given point to a cluster (Daoudi & Meshoul, 2017):

- Discard set (DS): the points that are close to a known centroid can be discarded for further iterations.
- Compression set (CS): the points that are close together, but not really close to any *k* centroid, are summarized but not assigned to any existing cluster.
- Retained set (RS): the isolated points that do not belong to any cluster and need to be retained in the buffer, waiting to be assigned.

Once the DS, CS, and DS sets are conformed (in the first iteration), the BFR iterates over the CS and RS to assign their points to a specific cluster. Before each inclusion, the data dispersion (using the Mahalanobis distance) is calculated among the internal elements of the cluster with the highest probability of hosting the new point (Aletti & Micheletti, 2017). After including a new point, the internal distances of the cluster are recalculated.

CURE method

CURE (clustering using representatives) is a specialized model used to cluster the data in non-spherical shapes (Guha, Rastogi, & Shim, 1998), usually ring or S-shape, and its main application is related to process large amounts of data (big data). The clusters formation starts by considering a group of representative points instead of centroids like the other methods do (Pandove & Goel, 2015). CURE treats each sample in the data as an individual class. Then, the closest samples (without taking into consideration the class) are merged until reach the number of desired clusters. After that, the samples are multiplied by an appropriate shrinkage factor to make them closer to the center of the cluster and to diminish the misleading effect of noise (Min & Li, 2015). CURE is the most robust model for outliers and size variances.

BIRCH method

The balanced iterative reducing and clustering using hierarchies (BIRCH) is an algorithm designed for clustering large amounts of numerical data by combining hierarchical clustering with iterative partitioning (Han & Kamber, 2005). It provides two strengths over other agglomerative clustering algorithms, such as solving the size scalability issues of the dataset, and it can undo operations that were made in previous steps (Han & Kamber, 2005).

BIRCH applies the principles of the clustering feature to summarize the cluster and the clustering feature tree to describe the cluster hierarchy (Parimalam & Sundaram, 2017). For any given dataset, regardless of the number of features per object, its clustering feature will be always a three-dimensional vector that summarizes the information of the objects within the dataset. Besides, this vector is used to calculate the centroid, radius, or diameter of the cluster, being the radius and diameter two measures of tightness (Han & Kamber, 2005). On the other hand, the clustering feature tree principle is a height-balanced tree containing the clustering

features according to the hierarchy criterion. This tree has the branching factor and threshold parameters to indicate the maximum number of children per internal node (is not leaves) and to represent the maximum diameter possible for storing subclusters as leaf nodes of the tree, respectively (Han & Kamber, 2005). Particularly, BIRCH performs data exploration by assuming they are not uniformly distributed; therefore, data points are not equally important (Zhang, Ramakrishnan, & Livny, 1996).

Expectation-maximization method

The Expectation-maximization (EM) method is a class of iterative algorithm to find maximum likelihood or maximum a posterior estimation (McLachlan & Krishnan, 2007), where the model depends on unobserved latent variables, in clustering problems with unlabeled data. (Nigam et al., 2011). The EM algorithm uses an initial conjecture based on a covariance matrix to estimate the model parameters iteratively. Each iteration consists of an expectation step, which finds the distribution of unobserved variables given the known values of observed ones and the currently estimated parameters. The maximization step re-estimates the model parameters according to the maximum likelihood of the previously found distribution in the expectation step, assuming it is correct. It continuously iterates between the expectation and maximization steps until reaching the threshold convergence. These iterations have been demonstrated to improve the true likelihood (Neal & Hinton, 1998) (Moon, 1996).

Spectral-clustering method

It is a graph-theory based method which finds connected structures (Jia et al., 2014). It involves several techniques to extract all the graph structural properties by using the eigen decomposition of its associated matrix representation (Thrun, 2018). Thus, it comprises several steps, such as finding the affinity matrix of the data that will be clustered, computing

the main *k-eigenvectors* of the affinity matrix, projecting the data into a new space defined by the computed *k-eigenvectors* and the data clustering in the newly transformed space (Han & Kamber, 2005).

Spectral-clustering method also uses the similarity graph during the data clustering. Its final purpose is to find a partition of the graph such that the edges between different groups have very low weights, which means that points in different clusters are dissimilar from each other. The edges within a group have high weights, which means that points within the same cluster are similar to each other (Jia et al., 2014). Due to this, it is effective to analyze high dimensional data and to detect arbitrarily shaped clusters (Han & Kamber, 2005). However, it lacks scalability and robustness when dealing with little spatial separations from cluster to cluster (Han & Kamber, 2005) (Thrun, 2018).

Experimental setup

This section outlines the experimental evaluation carried out with the selected six clustering-based models using the SeisBenchV1 dataset containing feature vectors of LP and VT seismic events. Dataset normalization, model configuration, assessment metrics, and selection criteria are important aspects that are described next.

Dataset normalization

All the values of the dataset were normalized using the min-max method (Jain & Bhandare, 2011) for bringing them into the range between 0 to1and thus, avoiding data dispersion.

Model configuration

The main parameter on clustering-based classifiers is the number of *k* clusters. For all models, the *k* number was optimized in the range from 2 to10 (empirically selection). Other hyperparameters were determined using a brute force-based approach such as random seed, which varied between 0t o10000 units, the number of children per node in the range from 0to the number of features in the dataset (688) and the threshold value from some of the optimal configuration per classifier are briefly described next:

- k-means: the initialization algorithm for centroid selection and the maximum of iterations for each run was set to k-means++ method and 1000 units, respectively.
- BFR: the merge threshold, which determines the approximation of two clusters was set to 2 units, the Mahalanobis factor, which measure the nearness of point and cluster was tuned to 3 units, the Euclidean threshold to determine the closeness of two points in the retained set was tuned to 3 units and the initial number of iterations was 40 units.
- CURE: the affinity metric used to compute the distance between sets was set to the Euclidean distance algorithm.
- BIRCH: The maximum number of children per internal node and the maximum diameter threshold for sub-clusters were set to 53 and 0.75 units, respectively.
- Expectation-maximization: the covariance type was set to tied, which means all components share the same covariance, the maximum iteration was tuned to 1500 units, the converge threshold to 10^{-5} and the random seed for initial covariance matrix to 25 units.
- Spectral-clustering: the random seed for initial eigenvectors decomposition was tuned to 3107 units, the affinity matrix construction was set to the nearest neighbor algorithm with the number of neighbors and eigenvectors equal to 20 and 16 units, respectively.

Assessment metrics

The classification performance of all employed models was based on the accuracy (ACC) metric. The SeisBenchV1 used in this work is a benchmarking dataset and provides all the needed information about the samples, including the class labels required to assess classification performance.

Selection criteria

Since the considered classifiers explore several *k* values, it was mandatory to select the best model using the following criteria: (1) the highest ACC score and, (2) if there is a tie rating in performance, the one with less algorithmic complexity is preferred. Despite not existing a universal rule to select the best classifier, we stated the "rule of gold" for the selection based on the particularity of the experimental SeisBenchV1 dataset. Thus, we ranked the model complexity in an ordered sequence of k-means, BFR, BIRCH, Spectral-clustering, CURE and Expectation-maximization classifiers.

We used the t-SNE (t- Distributed Stochastic Neighbor Embedding) technique (Maaten & Hinton, 2008) to visualize the multidimensional feature space presented in the SeisBenchV1 dataset into a bi-dimensional one. It was always applied after the classification process to avoid transforming the data before feeding the classifiers. The implementation of all classifiers was done in Python language version 3.7.4 (Python Core Team, 2019)with the scikit-learn (Sklearn) library (CURE implementation based on Agglomerative Clustering) (Pedregosa et al., 2011), the BFR implementation posted at (Berglund, 2018) and the PyClustering Library (Novikov, 2019).

RESULTS AND DISCUSSION

According to the experimental setup section, a total of 54 clustering-based models were evaluated on the experimental dataset which contains 668 features vectors. The straightforward comparison based on the ACC performance highlighted interesting results for the classification of LP and VT seismic events, as are described next:

Performance of explored models

Regarding the first selection criteria, only one out of 54 models was selected after exploring the whole classification space. According to the results shown in Table 1, the spectral-clustering classifier with k=2 was able to reach the highest ACC value of 92%. Except for the k-means method, the remaining classifiers achieved a slightly less performance when compared to the best model, but still obtained more than 85%, which are considered as good results as well. The BFR and BIRCH methods with k=2 received the second-best ACC value with 88%. The CURE with k=2 and k=3 accomplished the same ACC value of 87%, respectively. The k-means classifier obtained the worst performance, but the ACC value attained with k=2 was the higher among all the presented results of this classifier.

The better performances were obtained with k=2 for all classifiers, this was expected since the experimental dataset contains only LP and VT seismic events. Beyond this fact, the CURE classifier still assigned the same ACC value of 87% to a new cluster (k=3). This situation is related to the internal configuration of the SeisBenchV1 dataset, in which some samples of LP or VT have signals overlapped. Eventually, this situation leads to an incorrect classification when using supervised learning models due to the inaccurate event segmentation and, therefore, the calculation of the wrong features used to feed the classifiers (Pérez et al., 2020). However, the unsupervised learning CURE classifier was able to categorize and

understand this particular data behavior. An approximation of the data clustering at k=2 and k=3 using the t-SNE technique is shown in Fig.1 and 2, respectively. From Fig.2, it is possible to corroborate that the CURE classifier was able to detect most of those samples with overlapped signals, enabling it as a non-sensitive model to be use in real-life environments. However, the spectral-clustering classifier with k=2 constituted the best model selection for the problem under analysis.

Table 1. *ACC-based Performance Results for Explored Models*

Model	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
K-means	58	24	57	34	14	33	8	26	12
BFR	88	71	60	80	29	45	57	45	8
CURE	87	87	85	81	81	16	16	78	81
BIRCH	88	75	73	54	15	38	37	38	24
Spectral-clustering	92	72	66	66	64	50	34	13	26
EM	87	64	10	10	6	7	8	38	4

Note. ACC - accuracy; All values were rounded to the closest integer and are represented in percent (%)

Figure 1. *Data visualization using the t-SNE technique of the cluster with k=2 from left to right: k-means, BFR, CURE (top row) and BIRCH, Expectation-maximization, Spectral-clustering (bottom row) classifiers.*

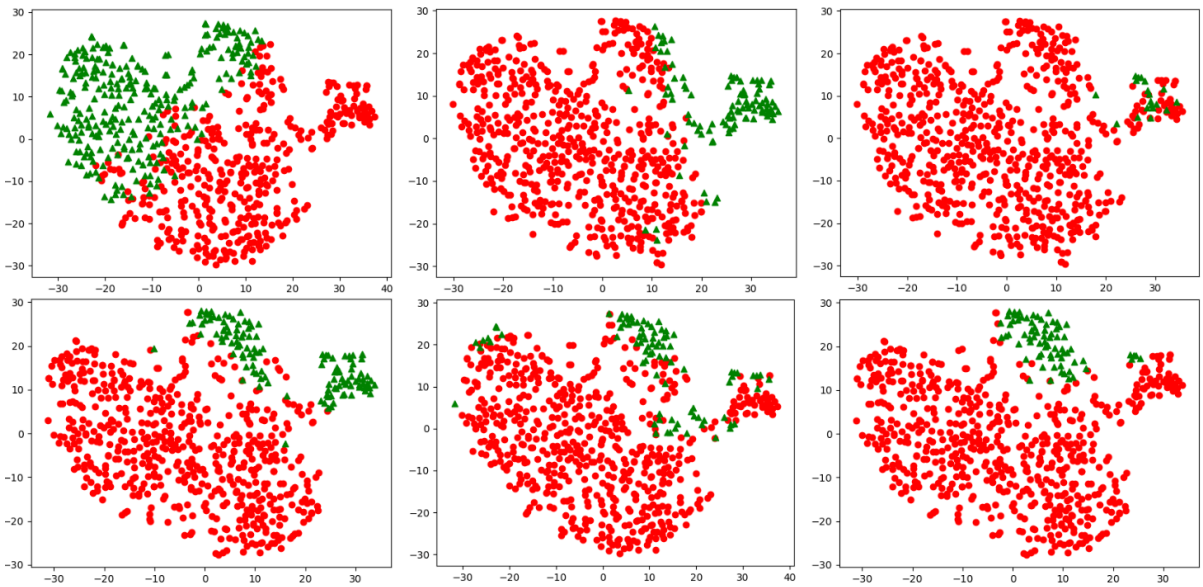
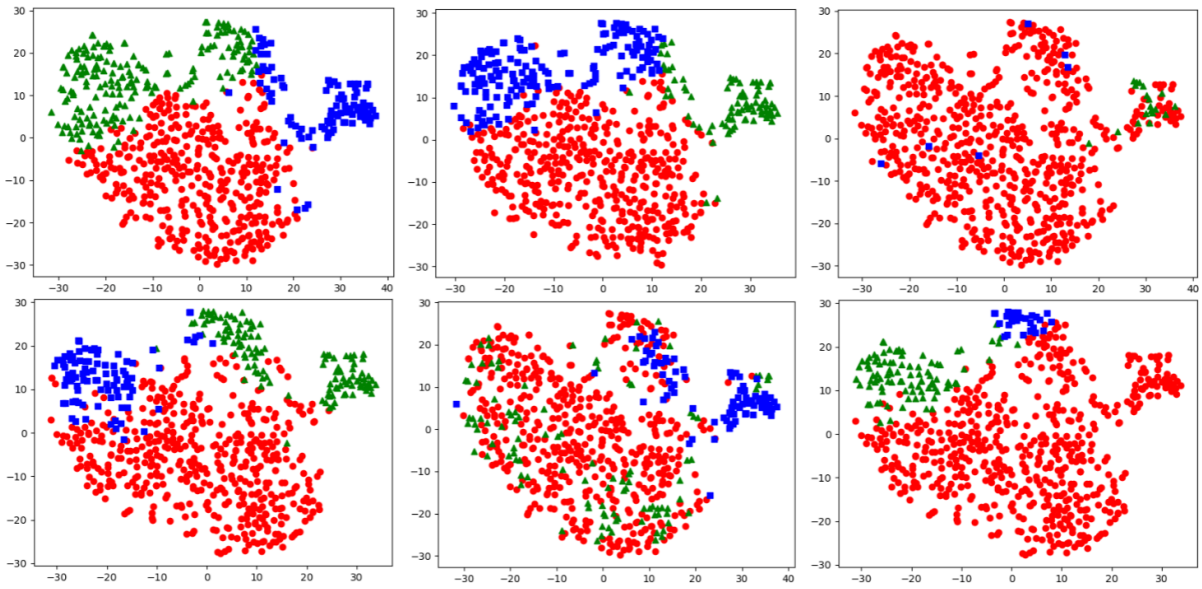


Figure 2. *Data visualization using the t-SNE technique of the cluster with k=3 from left to right: k-means, BFR, CURE (top row) and BIRCH, Expectation-maximization, Spectral-clustering (bottom row) classifiers.*



State of the art-based comparison

Concerning the classification performance, it is not possible to make a statistically direct comparison against previously developed methods in the literature. However, we aimed to carry out the comparison based on the ACC scores reported by the state of art methods, as shown in Table 2.

From Table 2, it is possible to notice that the ACC value of 92% reached by the spectral-clustering classifier was better than the SOM (Köhler et al., 2010) and KKAnalysis models (Messina & Langer, 2011), and inferior to the PCA method (Unglert et al., 2016). The superior performance demonstrated by the PCA method could be linked to the employed datasets; while more and better distributed are the samples, better intraclass variation will have the model during the training process and, therefore, a more accurate classification can be achieved. Nevertheless, in volcano real-life environments, the likelihood of having balanced datasets is very low. For example, although LP and VT are the main types of events recorded at Cotopaxi, the occurrence of LP events is higher than VT events (Molina et al., 2008). On the other hand,

the reported ACC score of 90% by the KKAnalysis model (Messina & Langer, 2011) was slightly inferior to our best model, even when this approach is based in the combination of a SOM artificial neural network and clustering methods. Therefore, the spectral-clustering classifier with k=2 emerges as a generalizable model for further application.

Table 2. *Comparison based on the ACC between previous works available in the literature and the selected best model*

Method	Number of samples	Balanced dataset	Number of features	ACC* (%)
PCA (Unglert et al., 2016)	672	yes	57	99
SOM (Köhler et al., 2010)	40	no	26	88
KKAnalysis (Messina & Langer, 2011)	5464	yes	62	90
Spectral-clustering	668	no	84	92

Note. ACC – accuracy; All values were rounded to the closest integer and are represented in percent (%)*

CONCLUSION AND FUTURE WORK

In this work, we made an ACC based exploration of six different unsupervised learning classifiers within the context of volcano seismic events classification. According to the experimental setup, the spectral-clustering classifier with k=2 was chosen as the best model, reaching an ACC score of 92%. This score represented a satisfactory and competitive classification performance when compared to the state-of-the-art methods. The CURE classifier with K=3 attained an ACC value of 87%. This performance was slightly lower than the selected best model. However, it was the only classifier able to detect LP or VT seismic events with overlapped signals. Therefore, the proposed clustering-based exploration was effective in providing competitive models to classify LP and VT seismic events and to detect signals with overlapping.

As future work, we plan to experiment with the considered clustering-based classifiers in this work on a dataset containing other types of events such as tremors, icequakes, hybrid, and lightning, which were not considered in the current experimental dataset. Also, the development of a new clustering classifier to improve the classification performance obtained in this work.

ACKNOWLEDGEMENT

The authors thank to the Applied Signal Processing and Machine Learning Research Group, USFQ, for providing the computing infrastructure (NVidiaDGX workstation) to implement and execute the developed source code. The seismic data used in this study was provided by Instituto Geofísico, EPN.

REFERENCES

Aletti, G., & Micheletti, A. (2017). *A clustering algorithm for multivariate data streams with correlated components*. Journal of Big Data, 4(1), 48.

Bebbington, M. S. (2007). *Identifying volcanic regimes using hidden markov models.* Geophysical Journal International, 171(2), 921–942.

Benitez, M. C., Ramirez, J., Segura, J. C., Ibanez, J. M., Almendros, J., Garcia-Yeguas, A., & Cortes, G. (2007, Jan). *Continuous hmm-based seismic-event classification at deception island, Antarctica.* IEEE Transactions on Geoscience and Remote Sensing, 45(1), 138-146. doi:10.1109/TGRS.2006.882264

Berglund, J. (2018). *Clustering with BFR.* Retrieved fromhttps://github.com/jeppeb91/bfr

Curilem, M., Vergara, J., San Martin, C., Fuentealba, G., Cardona, C., Huenupan, F., ... Yoma, N. B. (2014). *Pattern recognition applied to seismic signals of the llaima volcano (chile): An analysis of the events features.* Journal of Volcanology and Geothermal Research, 282, 134–147.

Daoudi, M., & Meshoul, S. (2017). *Revisiting bfr clustering algorithm for large scale gene regulatory network reconstruction using map reduce*. In Proceedings of the 2nd international conference on big data, cloud and applications (pp. 1–5).

Guha, S., Rastogi, R., & Shim, K. (1998). *Cure: an efficient clustering algorithm for large databases.* ACM Sigmod record, 27(2), 73–84.

Hammer, C., Beyreuther, M., & Ohrnberger, M. (2012). *A seismic-event spotting system for volcano fast-response systems.* Bulletin of the Seismological Society of America, 102(3), 948–960.

Han, J., & Kamber, M. (2005). *Data mining: concepts and techniques.* San Francisco: Kaufmann. Retrieved fromhttp://www.amazon.com/Data-Mining-Concepts-Techniques-Management/dp/1558604898

Jain, Y. K., & Bhandare, S. K. (2011). *Min max normalization based data perturbation method for privacy protection.* International Journal of Computer & Communication Technology, 2(8), 45–50.

Jia, H., Ding, S., Xu, X., & ru, N. (2014, 06). *The latest research progress on spectral clustering.* Neural Computing and Applications, 24. doi:10.1007/s00521-013-1439-2

Köhler, A., Ohrnberger, M., & Scherbaum, F. (2010). *Unsupervised pattern recognition in continuous seismic wavefield records using self-organizing maps.* Geophysical Journal International, 182(3), 1619–1630.

Krishna, K., & Murty, M. N. (1999). *Genetic k-means algorithm.* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 29(3), 433–439.

Kuyuk, H., Yildirim, E., Dogan, E., & Horasan, G. (2011). *An unsupervised learning algorithm: application to the discrimination of seismic events and quarry blasts in the vicinity of Istanbul.* Nat. Hazards Earth Syst. Sci, 11(1), 93–100.

Lara-Cueva, R., Carrera, E. V., Morejon, J. F., & Benítez, D. (2016, April). *Comparative analysis of automated classifiers applied to volcano event identification.* In 2016 ieee Colombian conference on communications and computing (colcom) (p. 1-6). doi: 10.1109/ColComCon.2016.7516377

Maaten, L. v. d., & Hinton, G. (2008). *Visualizing data using t-sne.* Journal of machine learning research, 9(Nov), 2579–2605.

McLachlan, G. J., & Krishnan, T. (2007). *The em algorithm and* extensions (Vol. 382). John Wiley & Sons.

Messina, A., & Langer, H. (2011). *Pattern recognition of volcanic tremor data on mt. etna (Italy) with KKAnalysis—a software program for unsupervised classification.* Computers & Geosciences, 37(7), 953–961. Retrieved fromhttps://www.overleaf.com/project/5de17b1735faa10001aceb85

Min, Y., & Li, Y. (2015). *Vehicles recognition based on the size characteristics and the cure clustering algorithm*. In 2015 ieee international conference on signal processing, communications and computing (icspcc) (pp. 1–5).

Molina, I., Kumagai, H., García-Aristizábal, A., Nakano, M., & P.Mothes.(2008). *Source process of very-long-period events accompanying long-period signals at Cotopaxi volcano, Ecuador*. J. Volcanol. Geothermal Res., 176(1), 119–133.

Moon, T. (1996, 12). *The expectation-maximization algorithm.* Signal Processing Magazine, IEEE, 13, 47 - 60. doi: 10.1109/79.543975

Neal, R. M., & Hinton, G. E. (1998). *A view of the em algorithm that justifies incremental, sparse, and other variants.* In Learning in graphical models (pp. 355–368). Springer.

Nigam, B., Ahirwal, P., Salve, S., & Vamney, S. (2011). *Document classification using expectation maximization with semi supervised learning.* arXiv preprint arXiv:1112.2028.

Novikov, A. (2019, apr). *PyClustering: Data mining library.* Journal of Open Source Software, 4(36), 1230. Retrieved fromhttps://doi.org/10.21105/joss.01230doi: 10.21105/joss.01230

Oliveira Martins, L. d., Braz Junior, G., Corrêa Silva, A., Cardoso de Paiva, A., & Gattass, M. (2009). *Detection of masses in digital mammograms using k-means and support vector machine.* ELCVIA: Electronic Letters on Computer Vision and Image Analysis, 8(2), 039–50.

Pandove, D., & Goel, S. (2015). *A comprehensive study on clustering approaches for big data mining.* In 2015 2nd international conference on electronics and communication systems (icecs) (pp. 1333–1338).

Parimalam, T., & Sundaram, K. M. (2017). *Efficient clustering techniques for web services clustering.* In 2017 ieee international conference on computational intelligence and computing research (iccic) (pp. 1–4).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel,O., ... Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825–2830.

Pérez, N., Benítez, D., Grijalva, F., Lara-Cueva, R., Ruiz, M., & Aguilar, J. (2020*). Eseismic: Towards an Ecuadorian volcano seismic repository.* Journal of Volcanology and Geothermal Research, 106855.

Pérez, N., Venegas, P., Benítez, D., Lara-Cueva, R., & Ruiz, M. (2020). *A new volcanic seismic signal descriptor and its application to a dataset from the Cotopaxi volcano.* IEEE Transactions on Geoscience and Remote Sensing.

Phillipson, G., Sobradelo, R., & Gottsmann, J. (2013). *Global volcanic unrest in the 21st century: An analysis of the first decade.* Journal of Volcanology and Geothermal Research, 264, 183–196. doi: 10.1016/j.jvolgeores.2013.08.004

Python Core Team. (2019). Python 3.7.4: *A dynamic, open source programming language.* [Computer software manual]. Retrieved fromhttps://www.python.org/.

Reyes, J. A., & Mosquera, C. J. J. (2017). *Non-supervised classification of volcanic-seismic events for Tungurahua-volcano Ecuador*. In 2017 IEEE second Ecuador technical chapters meeting (etcm) (pp. 1–6).

Rodgers, M., Smith, P., Pyle, D., & Mather, T. (2016). *Waveform classification and statistical analysis of seismic precursors to the July 2008 vulcanian eruption of soufrière hills volcano, Montserrat*. In Egugeneral assembly conference abstracts (Vol. 18).

Schmincke, H.-U. (2004). *Volcanic hazards, volcanic catastrophes, and disaster mitigation*. In Volcanism (pp. 229–258). Springer. doi:10.1007/978-3-642-18952-4_13

Sharma, N., Bajpai, A., & Litoriya, M. R. (2012). *Comparison the various clustering algorithms of weka tools*. facilities, 4(7), 78–80.

Siebert, L., Simkin, T., & Kimberly, P. (2011). *Volcanoes of the world.* Univ. of California Press.

Tamilselvi, R., Sivasakthi, B., & Kavitha, R. (2015). *A comparison of various clustering methods and algorithms in data mining.* Int. J. Multidiscip. Res. Dev, 2(5), 32–98.

Thrun, M. C. (2018). *Projection-based clustering through self-organization and swarm intelligence: combining cluster analysis with the visualization of high-dimensional data*. Springer.

Tilling, R. I. (1996). *Hazards and climatic impact of subduction-zone volcanism: A Global and Historical Perspective*. Washington DC American Geophysical Union Geophysical Monograph Series, 96, 331-335. doi: 10.1029/GM096p0331

Unglert, K., Radić, V., & Jellinek, A. M. (2016). *Principal component analysis vs. self-organizing maps combined with hierarchical clustering for pattern recognition in volcano seismic spectra.* Journal of Volcanology and Geothermal Research, 320, 58–74.

Venegas, P., Pérez, N., Benítez, D., Lara-Cueva, R., & Ruiz, M. (2019). *Combining filter-based feature selection methods and gaussian mixture model for the classification of seismic events from Cotopaxi volcano.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(6), 1991-2003. doi: 10.1109/JSTARS.2019.2916045

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). *Birch: an efficient data clustering method for very large databases.* ACM Sigmod Record, 25(2), 103–114.

Zheng, Y., Jeon, B., Sun, L., Zhang, J., & Zhang, H. (2017). *Student's t-hidden markov model for unsupervised learning using localized feature selection.* IEEE Transactions on Circuits and Systems for Video Technology, 28(10), 2586–2598.