Volcano-Seismic Events Classification
Using Document Classification Strategies

Manuele Bicego[1]([✉]), John Makario Londoño-Bonilla[2],
and Mauricio Orozco-Alzate[3]

[1] Dipartimento di Informatica, Università degli Studi di Verona,
Ca' Vignal 2, Strada le Grazie 15, 37134 Verona, Italia
manuele.bicego@univr.it
[2] Observatorio Vulcanológico y Sismológico de Manizales, Servicio Geológico
Colombiano, Avenida 12 de Octubre 15–47, Manizales 170001, Colombia
[3] Departamento de Informática y Computación,
Universidad Nacional de Colombia - Sede Manizales,
km 7 vía al Magdalena, Manizales 170003, Colombia

Abstract. In this paper we propose a novel framework for the classification of volcano-seismic events, based on strategies and concepts typically employed to classify documents – subsequently largely employed also in other fields. In the proposed approach, we define a dictionary of "seismic words", used to represent a seismic event as a "seismic document" (i.e. a collection of seismic words). Given this representation, we exploit two well-known models for documents (Bag-of-words and topic models) to derive signatures for seismic events, usable for classification. An empirical evaluation, based on a set of seismic signals from Galeras volcano in Colombia, confirms the potentialities of the proposed scheme, both in terms of interpretability and classification accuracies, also in comparison with standard approaches.

Keywords: Bag-of-words · Mel frequency cepstral coefficients · Seismic volcanic signal classification · Topic models

1 Introduction

The analysis and the classification of seismic signals play a vital role in volcano monitoring. In the literature, several techniques have been proposed to address this challenge, each one using different representations and exhibiting different interpretability features, performances and computational requirements – see [21] for a comprehensive review of the literature.

In this paper a novel approach to the classification of volcano-seismic events is proposed, based on a set of tools and concepts introduced in the text processing community. In particular, our framework is based on two effective and largely applied tools, namely the bag-of-words approach [13,17] and the topic models [5]: after their introduction in the text mining community, such models have been successfully exported to many other scenarios, such as — just to cite a few — Computer Vision [7,22], Bioinformatics [4,8,23], and Audio Analysis [11,14,16]. To the

best of our knowledge, their usefulness in the seismic scenario has never been inves-
tigated: this paper represents a first effort in this direction. It seems very appeal-
ing to investigate the capabilities of these models — which, in many applications,
demonstrated powerful classification characteristics as well as interesting interpre-
tation properties [4, 7, 8, 11, 14, 16, 22, 23] — in the seismic scenario: we can inter-
pret every event as a "document", which employs particular "words", and which
can focus on one or more "topics". The same topic can be present in different classes
of events, maybe because it is related to a shared geophysical cause; going further
with our reasoning, two events can "speak" about the same set of topics, but using
a different dictionary, maybe only because the signals have been acquired from
different stations.

Clearly, in order to apply such models, we should define the concept of "seismic
documents" and "seismic words". The first step is therefore to define a dictionary
of words, containing the constituting elements of a document: in our approach, sim-
ilarly to what is done in other contexts — e.g. image analysis [25] or audio process-
ing [14, 16] — the dictionary is built by first extracting some meaningful features
from all the signals, subsequently applying a vector quantization / clustering app-
roach to derive the words. As features we used the classical Mel Frequency Cepstral
Coefficients (MFCCs), extracted from subsequent frames of the seismic signals –
this represents a standard preprocessing in seismic event recognition [2, 10]. Given
the dictionary, every event is now characterized as a sequence of words, which is
encoded using either the bag-of-words (BoW) approach or a topic model – here we
employ the Latent Dirichlet Allocation (LDA – [6]).

For classification, we fed the BoW representation directly to a classifier; for
topic models, we made an extra-step: actually, we set up a hybrid generative-
discriminative classification scheme [12, 15], in particular by following the so-called
generative embedding strategy: the trained topic model is used to map the sig-
nals into a feature space, in which a discriminative classifier is employed. Different
mappings have been proposed in recent years (see [22] and the references therein
included): in our approach we employed the recent FESS scheme [22], which has
been proven to be a highly informative description in different applications.

The proposed framework has been thoroughly tested with a set of pre-
triggered signals (divided into 4 classes) coming from Galeras volcano in Colom-
bia, investigating the effect of different parameters. A comparison with reference
approaches showed that the proposed method represents a valid alternative to
standard seismic classification techniques.

2   The Proposed Approach

In this section we discuss how to construct the dictionary, how to characterize
seismic events as documents, and how to classify them. Schemes of the first two
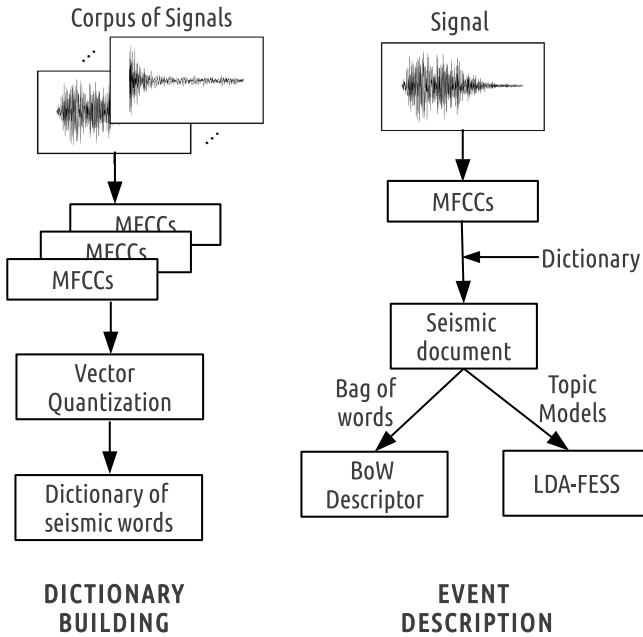phases are shown in Fig. 1.

Fig. 1. Block diagram of the proposed approach.

2.1   Dictionary Building

Before applying text mining tools to seismic events, we should define what documents and words are in the specific context. Intuitively, we can associate a seismic document to a seismic event: the seismic document represents a collection of seismic words. To define seismic words, we take inspiration from the audio modelling community [14,16], adopting the following strategy: seismic signals are parametrized with the conventional MFCCs, and the seismic words are then derived using Vector Quantization. As done by [14], we use a frame-based analysis —with a fixed length— in order to represent the time varying properties of the seismic signal. MFCCs are a popular choice in the seismic community [2,10], able to provide a spectral parametrization of the signal considering human auditory properties. In this work we used frames of 256 sample values (2.5 s) with overlap of 50%, considering 13 coefficients together with their derivatives (as in [2,10]).

2.2   Event Description

Once given the dictionary, every event can be described as a sequence of words, namely a document. In order to characterize the documents, here we employed two techniques: the bag-of-words and the LDA topic model.

Bag-of-Words. The Bag-of-words approach represents a straightforward description of a document, still being really descriptive and useful. In particular, given a

dictionary of W words, the BoW descriptor of a document d is a vector of length W which, in the entry j, measures the number of times the j-th word appears in such document. Therefore, every event is described by a vector of length W. It is important to note that this representation (as well as the one derived from the LDA model) does not consider the order in which the words appear in the document. This is a well known problem of this class of approaches, which are known to somehow destroy the structure of the object (the order of the words, in this case). Even if alternatives have been recently proposed (e.g. [9]), these basic descriptors are still widely and successfully applied in many fields [4,7,8,11,14,16,22,23], due to their excellent discriminative capabilities: actually, the vectorial representation permits to completely exploit powerful discriminative vector-based classifiers, such as Support Vector Machines.

Topic Models: Latent Dirichlet Allocation. Topic models represent a powerful extension of the BoW approach, able to take into consideration the context in order to disambiguate the meaning of the words. In particular, these methods aim at characterizing each document with the presence of one or more topics (e.g. economics, fashion, finance), each one inducing the presence of some particular words. From a probabilistic perspective, we can see the document as described via a mixture of topics, each one giving a probability distribution over words. Such distributions are learnt by analysing word co-occurrences in the training data. The characterization of documents and words with these probabilistic technique allows the individual interpretability of each topic, since it provides a probability distribution over words that extracts a coherent group of correlated terms. In our approach we employed the Latent Dirichlet Allocation (LDA - [6]), one of the first and most famous topic models[1].

Given a set of V different words, the LDA mediates the observation of a particular word w_i in a document t through a latent topic variable $z, z \in Z = \{z_1, \ldots, z_Z\}$, which is picked from a multinomial distribution $p(z \mid t) = \theta^t$. The multinomial θ^t represents the topic proportions, peculiar for every document t: intuitively, the θ^t variable describes "how much each topic is spoken in such document". Without entering too much into the details (interested readers are referred to [6]), we can simply say that the probability of observing a given word w_i in a document t is:

$$p(w_i^t) = p(\theta^t \mid \alpha) \sum_k p(z_k|\theta^t)p(w_i^t|z_k)$$

$$= p(\theta^t \mid \alpha) \sum_k \theta_{z_k}^t \cdot \beta_{w_i,z_k} \qquad (1)$$

where β_{w_i,z_k} expresses how much a word w_i is related to the topic z_k: roughly speaking, this distribution describes "how probable is to use the word w_i when

---

[1] Even if many complex topic models have been proposed, here we decided to use this simple – yet powerful – model, in order to investigate the suitability of the coding technique.

the document is speaking about topic z_k". Finally, $p(\theta^t \mid \alpha)$ is a Dirichlet prior over the possible topics' assignments.

As better detailed in [6], the various distributions of the model are learned using a variational Expectation-Maximization (EM), a technique that maximizes the log-likelihood (or its tractable lower bound called Free Energy) by iterating between two steps: the E-step, which computes the posterior over the topics (i.e. θ^t), given the current estimate of the model; the M-step, where the parameters of the models (α and β) are re-estimated, given the current θ^t. Once the model has been trained, it is possible to use the learned parameters α and β to perform inference, estimating the topic proportion $\theta^{t_{new}}$ of an unseen document t_{new}. Since the EM algorithm converges to a local optimum, a proper initialization of the LDA model is of crucial importance to guarantee the convergence to a proper local optima [18]. In our framework the initialization issue has been faced in a standard way, by repeating several times the training procedure, each starting from different random parameters, finally retaining the configuration which leaded to the highest likelihood.

2.3   From Documents to Feature Vectors

Whereas the BoW of a document is already a vectorial representation, for LDA we made an extra step. In particular, we employed a hybrid generative-discriminative classification scheme [12,15], which aims at merging together the best characteristics of both the generative and the discriminative paradigms: the first step is to learn a generative model – suitable to describe the problem at hand – from the data; then the learnt generative model is exploited to define a mapping which projects every object of the problem in a feature space (typically called generative embedding space), in which a discriminative classifier can be used.

In our approach we trained a single LDA model on the whole encoded training set, performing inference to get distributions also on the testing set. Then we employed the very recent Free Energy Score Space (FESS) approach [22] to derive a feature space: without going too much into the details – interested readers are referred to [22] – we can briefly state that the FESS vector is able to capture how well each object of problem fits the different parts of the generative model, modelled via the variational free energy (which represents a lower bound of the negative log-likelihood). It has been shown in [22] that such representation is highly informative for classification, permitting to reach state-of-the-art results in different bioinformatics and computer vision problems.

2.4   Document Classification

In the obtained feature space (the BoW or the LDA-FESS spaces), any classic discriminative classifier can be used. In our experiments we employed the standard k nearest neighbor (kNN) and the linear support vector machine (SVM) classifiers.

3   Experiments

In this section the proposed approach is evaluated. After describing the experimental details, we report some classification accuracies, also in comparison with some other standard techniques. Then, we investigate the impact of the parameters on the performances; finally, we present some intuitions about the interpretability potentialities.

3.1   Experimental Details

A data set of seismic signals from Galeras volcano (Colombia) has been used to test the proposed framework. Galeras is a stratovolcano situated in the Andes mountains, near (7 km W) to the city of Pasto[2]. After around 40 years of silence, such volcano started again its activity in 1988, with several ash and gas emissions as well as some eruptions: the most relevant were on May 1989, on July 1992, on the first semester of 1993 (many different eruptions), on the second semester of 2004, at the end of 2005 and more recently in January 2008, during the whole 2009, and in January and August 2010.

Data used in our experimentation have been gathered with a seismic network composed by seven short-period seismic stations – here we employed the Anganoy station, which is the highest station (4227 m.a.s.l) and the closest to the active crater (0.8 km). After acquisition, signals are telemetered by radio to the Observatory, where they are pre-processed using a 12-bit analog-to-digital converter with a sampling rate of 100.16 samples/s; after that, interesting events are obtained using an automatic detection/segmentation stage; finally, segmented events are stored on a series of servers as files using the Seismic Unified Data System protocol.

To test the classification potentialities of the proposed framework we used two different classification problems. The first task is composed by 300 signals, divided into three classes: Volcano-tectonic (VT) earthquakes, long-period (LP) events, and tremors (TR), which represent the most important volcano-induced earthquakes. The second classification task is definitely more challenging, since it includes the class of hybrid (HB) events: actually it has been shown in many studies [1,2] that distinguishing between LP and HB earthquakes is challenging; however, this discrimination is rather crucial in the specific application scenario.

In all the experiments, the seismic events were characterized as described before. Bag-of-words and LDA-FESS descriptors were extracted. We made dictionaries of different sizes taking the values 32, 64, 128, 256 and 512, while for LDA we made the number of topics varying from 2 to 20 (step 2). In the generative embedding space we used as classifiers the kNN and linear SVM. In the kNN case, k was automatically estimated with Leave One Out cross validation on the training set. For the SVM, after some preliminary trials, C has been fixed

---

[2] Further details about Galeras volcano activity are at the institutional web site of the Observatory: http://www.sgc.gov.co/Pasto/Volcanes/Volcan-Galeras/Generalidades.aspx

to 1. In all experiments, classification accuracies were computed using Averaged Holdout CV, with results averaged over 20 repetitions. In order to assess of the statistical significance of the results, we computed for every different set of experiments the standard errors of the mean. For the 3-class problem, they were all less than 0.0050, whereas for the 4-class problem they were all less than 0.0025.

3.2   Results and Comparison with other Methods

To have a quick summarizing view of the behaviour of the proposed scheme, in this part we show the accuracies obtained from the best configuration of parameters, leaving comments on the impact of the parameter choice to the following section. We compare our approach with some other well established seismic classification strategies: (a) the same classifiers employed in our tests on the averaged MFCCs (MFCCs averaged over all the frames of a given event – the baseline results); (b) a set of descriptive spectral parameters [24], classified using Back Propagation Neural Network (BP-NNet), Levenberg-Marquardt (LM-NNet) and the SVM with rbf kernel. The topology of the networks has been set as in [24], whereas the σ and C parameter of SVM have been optimized with cross-validation on the training set; (c) a Bayesian approach based on Continuous Gaussian Hidden Markov Models (HMMs) trained on event spectrograms [3] – being HMMs the most widely applied approach in this context. In such case we performed experiments varying the number of states in a proper range, reporting the best result obtained.

Table 1. Comparative classification accuracies.

Method	3-class	4-class
Averaged MFCC + kNN	0.8580	0.7925
Averaged MFCC + linSVM	0.9050	0.8122
Time-frequency feat + BP-NNet	0.9343	0.7082
Time-frequency feat + LM-NNet	0.9277	0.7215
Time-frequency feat + rbf SVM	0.9367	0.6963
Spectrograms + HMMs	0.9150	0.8348
Bag-of-Words (best)	0.9187	0.8177
Topic Model (best)	**0.9410**	**0.8375**

From Table 1, it is evident that the proposed approach represents a valid alternative to other standard and well established classification approaches, well comparing also with advanced tools as those based on HMM + spectrograms (in such case, almost equivalent accuracies were obtained for the 4-class problem).

3.3   Effect of the Parameters

The two most crucial parameters in our framework are the size of the dictionary and the number of topics for topic models. In the former case, even if some clever

strategies and studies have recently appeared in some specific community (e.g. [19]), the problem still remains unsolved, and the typical solution is to make different trials (as done in [14]). In Table 2 we study the different results while varying this parameter (taking the best number of topics): it seems evident that the BoW approach prefers small size dictionaries, with deteriorating performances while increasing the size. On the contrary, when using FESS features extracted from the LDA, this problem is less present.

Table 2. Results when varying the dictionary size (dS).

dS	BoW+kNN	BoW+SVM	TM+kNN	TM+SVM
	3-class Problem			
32	0.9187	0.9160	0.9440	0.9360
64	0.9057	0.8937	0.9390	0.9383
128	0.8760	0.9053	0.9400	0.9380
256	0.8327	0.9017	0.9370	0.9410
512	0.7477	0.9057	0.9387	0.9397
	4-class Problem			
32	0.7960	0.8177	0.8193	0.8290
64	0.7800	0.7882	0.8303	0.8315
128	0.7765	0.7735	0.8305	0.8375
256	0.7410	0.7790	0.8090	0.8282
512	0.6540	0.7848	0.7975	0.8223

The choice of the second parameter (number of topics) represents a classic model selection problem, for which some techniques already appeared in the literature: hold-out likelihood [23], cross-validation, a priori knowledge or general probabilistic model selection methods. Another option, which has been used in the microarray scenario [4], appeared to be a straightforward but effective rule to select such number. Actually, authors in [4] started from the consideration that topic models have originally been designed to perform clustering, i.e. to discover groups of documents; therefore, fixing the number of topics to be proportional to the number of classes seems reasonable. Despite the simplicity of this rule-of-thumb, obtained results were very satisfactory. In Fig. 2(a) we plot the classification accuracies of the LDA-FESS descriptor when varying the number of topics (linear SVM as classifier, best dictionary size). In the figure, the two dashed horizontal lines represent the baseline results, obtained with linear SVM on the averaged MFCCs. The dotted vertical lines highlight the number of topics corresponding to two times the number of classes. From the figure, it can be seen that i) there is a quite large range of values for which the accuracies are over the baseline and ii) selecting two times the number of classes represents a reasonable choice.
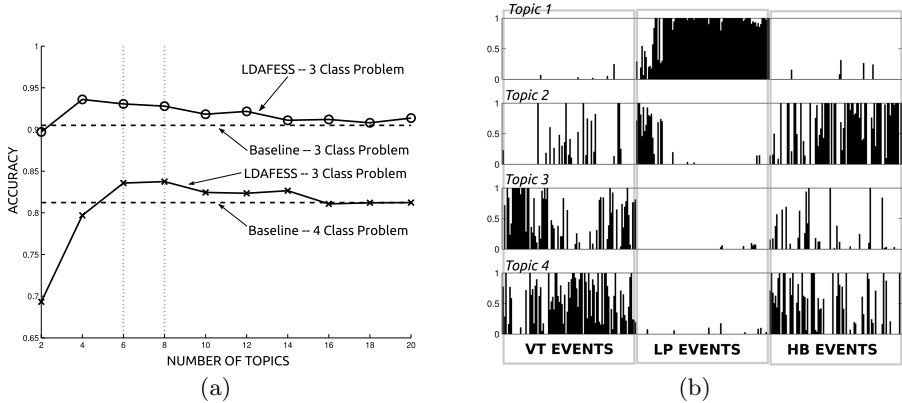
Fig. 2. (a) Results varying the number of topics for the two problems. (b) Interpretation of Seismicity.

3.4 Interpretability

In order to illustrate the interpretation capabilities of these tools – shown in many different contexts [4,8] – we trained a LDA with 4 topics on a dataset composed by VT, LP and HB events. Topic proportions for all the events belonging to the different classes are displayed in Fig. 2(b). From the plot we can firstly observe that the topics which are more representative for the LP events (the first and the second) are mainly different from those related to the VT events (the third, the fourth – and, partially, the second). This is expected, since it is a well-known fact in volcano seismology [20] that LP events have in general spectral energies concentrated at lower frequencies, whereas VT events display a relatively high-frequency spectral content. Topic models capture co-occurrences of words, which, for the two classes of events, are reasonably different. An even more interesting observation derives by considering the HB events: from a theoretical point of view, we know that such events are defined as a mixture between VT and LP events. Actually this is partially reflected in the plots: HB events are "active" mainly in topics 2, 3 and 4. The last two are the topics mainly "spoken" in VT events, whereas the second topic explains those LP events which are not explained by the first. From these plots we can: i) confirm that HB events are a mixture of VT and LP events; ii) hypothesise that in Galeras volcano, HB events are mainly dominated by features of VT events at the Anganoy station; this can be attributed either to a dominant fracture process related with the local geology at Anganoy station, or to a site effect related with the local geology at the Anganoy station. Further tests by using other stations are needed to clarify this issue.

4 Conclusions

This paper represents a first step towards the application of document classification tools for the classification of seismic events. After defining a dictionary

of "seismic words", events are characterized as documents, subsequently modelled exploiting Bag-of-Words and topic models. Experimental results confirm the potentialities of the proposed scheme, both in terms of interpretability and classification accuracies.

References

1. Trombley, R.B.: The Forecasting of Volcanic Eruptions. iUniverse (September 2006)
2. Benítez, M.C., Ramírez, J., Segura, J.C., Ibáñez, J.M., Almendros, J., García-Yeguas, A., Cortés, G.: Continuous HMM-based seismic-event classification at Deception Island, Antarctica. IEEE Transactions on Geoscience and Remote Sensing **45**(1), 138–146 (2007)
3. Bicego, M., Acosta-Muñoz, C., Orozco-Alzate, M.: Classification of seismic volcanic signals using hidden-Markov-model-based generative embeddings. IEEE Transactions on Geoscience and Remote Sensing **51**(6), 3400–3409 (2013)
4. Bicego, M., Lovato, P., Perina, A., Fasoli, M., Delledonne, M., Pezzotti, M., Polverari, A., Murino, V.: Investigating topic models' capabilities in expression microarray data classification. IEEE/ACM Transactions on Computational Biology and Bioinformatics **9**(6), 1831–1836 (2012)
5. Blei, D.M.: Probabilistic topic models. Communications of the ACM **55**(4), 77–84 (2012)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
7. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
8. Brelstaff, G., Bicego, M., Culeddu, N., Chessa, M.: Bag of peaks: interpretation of NMR spectrometry. Bioinformatics **25**(2), 258–264 (2009)
9. Du, L., Buntine, W., Jin, H., Chen, C.: Sequential latent Dirichlet allocation. Journal of Knowledge and Information Systems **31**, 475–503 (2012)
10. Ibáñez, J.M., Benítez, C., Gutiérrez, L.A., Cortés, G., García-Yeguas, A., Alguacil, G.: The classification of seismo-volcanic signals using Hidden Markov Models as applied to the Stromboli and Etna volcanoes. Journal of Volcanology and Geothermal Research **187**(3–4), 218–226 (2009)
11. Ishiguro, K., Yamada, T., Araki, S., Nakatani, T., Sawada, H.: Probabilistic speaker diarization with bag-of-words representations of speaker angle information. IEEE Transactions on Audio, Speech, and Language Processing **20**(2), 447–460 (2012)
12. Jaakkola, T.S., Haussler, D.: Exploiting generative models in discriminative classifiers. In: Kearns, M.S., Solla, S.A., Cohn, D.A. (eds.) Advances in Neural Information Processing Systems (NIPS), vol. 11, pp. 487–493 (1999)
13. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)

14. Kim, S., Georgiou, P., Narayanan, S.: Latent acoustic topic models for unstructured audio classification. APSIPA Transactions on Signal and Information Processing **1**, 1–15 (2012)
15. Lasserre, J.A., Bishop, C.M., Minka, T.P.: Principled hybrids of generative and discriminative models. In: Proc. of Int. Conf. on Computer Vision and Pattern Recognition (CVPR06), vol. 1, pp. 87–94, June 2006
16. Lee, K., Ellis, D.P.W.: Audio-based semantic concept classification for consumer video. IEEE Transactions on Audio, Speech, and Language Processing **18**(6), 1406–1416 (2010)
17. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. Journal of Machine Learning Research **2**, 419–444 (2002)
18. Lovato, P., Bicego, M., Murino, V., Perina, A.: Robust initialization for learning latent dirichlet allocation. In: Proc. Int. Workshop on Similarity-Based Pattern Analysis and Recognition (2015)
19. Mairal, J., Bach, F., Ponce, J.: Task-driven dictionary learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(4), 791–804 (2012)
20. McNutt, S.R.: Volcanic seismology. Annual Review of Earth and Planetary Sciences **33**(1), 461–491 (2005)
21. Orozco-Alzate, M., Acosta-Muñoz, C., Londoño-Bonilla, J.M.: The automated identification of volcanic earthquakes: concepts, applications and challenges. In: D'Amico, S. (ed.) Earthquake Research and Analysis - Seismology, Seismotectonic and Earthquake Geology, chap. 19, pp. 345–370. InTech, Rijeka (2012)
22. Perina, A., Cristani, M., Castellani, U., Murino, V., Jojic, N.: Free energy score spaces: Using generative information in discriminative classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(7), 1249–1262 (2012)
23. Rogers, S., Girolami, M., Campbell, C., Breitling, R.: The latent process decomposition of cDNA microarray data sets. IEEE/ACM IEEE/ACM Transactions on Computational Biology and Bioinformatics **2**(2), 143–156 (2005)
24. Ibs-von Seht, M.: Detection and identification of seismic signals recorded at Krakatau volcano (Indonesia) using artificial neural networks. Journal of Volcanology and Geothermal Research **176**(4), 448–456 (2008)
25. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision **73**(2), 213–238 (2007)