Apprentissage statistique: classification automatique de signaux volcano-sismiques

Marielle MALFANTE[1], Mauro DALLA MURA[1], Baptiste BOULLAY[1], Jean-Philippe MÉTAXIAN[2], Jérôme MARS[1]

[1]Univ. Grenoble Alpes, CNRS, GISPA-Lab, 11 rue des mathématiques, 38 000 Grenoble, France

[2]ISTerre, IRD, CNRS, Université Savoie Mont Blanc 73 376 Le Bourget du Lac, France.

prenom.nom@gipsa-lab.fr, metaxian@univ-savoie.fr

Résumé – Dans le contexte de surveillance volcanique, la question de l'analyse automatique des données sismiques enregistrées en observatoire reste sans réponse. La quantité de données à analyser est telle qu'une analyse manuelle n'est plus possible. On considère ici l'utilisation de méthodes d'apprentissage statistique, et en particulier d'apprentissage supervisé pour proposer une architecture de classification automatique des données volcano-sismiques. C'est l'une des premières méthodes d'analyse automatique pour ce type de données, testée et validée sur une large base de données. Le succès du système présenté vient en particulier de la représentation choisie pour les données (espace des descripteurs).

Abstract – Automatic classification of seismic signals recorded in observatories for volcanic monitoring is still an open issue. The amount of data to be analysed is such that manual processing is no longer a viable option. In this paper, we consider machine learning methods to propose supervised architecture for automatic classification of volcano-seismic signals. To the best of our knowledge, this is one of the first functional attempt to automatically process such data. The success of the proposed method lies in the chosen data representation (feature space).

1 Introduction

La communauté scientifique dans son ensemble est intéressée par les méthodes de classification automatique, en particulier depuis l'essor du phénomène dit de *Big Data*. Les méthodes de surveillance environnementale ont été particulièrement touchées par le phénomène ; ainsi en géophysique, le nombre de stations dédiées à l'enregistrement de données a drastiquement augmenté au cours des dernières années. Néanmoins, malgré les besoins des observatoires, très peu d'outils d'analyse automatique leurs sont associés. Les enjeux humains et économiques de ces outils d'analyse sont pourtant immédiats : les conséquences d'événements volcaniques non ou mal anticipés sont catastrophiques.

Nous proposons dans ce papier une nouvelle architecture de classification automatique de signaux sismiques liés à une activité volcanique. Les principales contributions sont : (i) une analyse de plusieurs descripteurs utilisés pour représenter les signaux, issus d'un état de l'art de différents domaines scientifiques. (ii) L'utilisation de ces descripteurs pour caractériser les signaux dans plusieurs espaces de représentation incluant les espaces temporel, fréquentiel et fréquentiel de fréquentiel. (iii) La définition d'une architecture de classification automatique pour l'analyse de grandes bases de données volcano-sismiques.

Ce papier est structuré comme suit : la Section 2 présente un état de l'art sur la représentation de signaux, la Section 3 détaille l'architecture proposée et les résultats expérimentaux sont expliqués dans la Section 4. Enfin, les conclusions et perspectives sur ce travail sont proposées en Section 5.

2 État de l'Art

À notre connaissance, il n'y a pas de procédure établie pour la classification automatique de signaux volcano-sismiques. Néanmoins, la littérature propose plusieurs travaux de classification automatique de signaux transitoires. L'un des points clé de ces méthodes est le choix du domaine de représentation des signaux, via l'extraction de descripteurs. Il s'agit en fait d'extraire l'information apte à discriminer les signaux en leurs classes. Cet espace de représentation a souvent besoin d'être ajusté aux données considérées [3]. Nous récapitulons ici les différentes représentation des signaux utilisées pour la classification automatique de signaux sismiques [9], acoustiques (environnementaux, bio-acoustiques [11, 5, 7], anthropiques [10]), de parole et de musique [4, 6].

Les signaux sismiques sont représentés dans [9] par quelques descripteurs classiques comme la durée, des descripteurs statistiques (skewness, kurtosis) et la fréquence fondamentale. Pour la classification de signaux sonars, [10] propose d'utiliser des descripteurs de forme du signal, des moments statistiques et des descripteurs d'énergie. Certains de ces descripteurs sont également utilisés en bio-acoustique, on peut en particulier mentionner [11, 5] pour distinguer les sons de bateaux de baleines ou pour l'identification d'espèces d'oiseaux. Des descripteurs basés sur l'entropie des signaux peuvent également être utilisés, par exemple dans [7] pour la classification de sons de grenouille ou dans [4] pour la discrimination de genres musicaux. Pour l'identification automatique d'instruments de musique, on retrouve également des mesures de bande fréquentielle ou de

ratio de silence dans [6].

Comme évoqué précédemment, la classification automatique de signaux volcano-sismique est un domaine encore très récent et peu exploré. Les données volcano-sismiques sont en effet particulièrement difficiles à classifier, même pour les experts du domaines. La difficulté d'accès des milieux étudiés rend la connaissance physique des phénomènes volcaniques incomplète. Dès lors, les classes de signaux sismiques sont régulièrement mises à jour avec l'évolution de la compréhension des mécanismes physiques de production des données. De plus, les outils de classification automatique existants dans d'autres domaines ne peuvent pas être directement appliquées aux données sismiques : typiquement, même si les signaux sont par certains aspects similaires à des signaux de parole, ils restent incompatibles avec les descripteurs du domaine (leurs contenus spectraux étant disjoints). Néanmoins, la nature transitoire des signaux sismiques rend plausible l'utilisation de descripteurs issus d'autres domaines où des données transitoires doivent être classifiées.

3 Architecture Proposée

L'architecture proposée dans ce papier est basée sur l'apprentissage statistique supervisé que nous présentons ici rapidement [8]. L'apprentissage statistique est une branche de l'Intelligence Artificielle visant à construire des modèles pouvant séparer des données en plusieurs *classes*. Ces méthodes sont utilisées dans de nombreux domaines d'application : par exemple en traitement de parole ou d'images [2]. Parmi les algorithmes d'apprentissage supervisé particulièrement performants, on retiendra en particulier Random Forest (RF) ou Support Vector Machine (SVM).

Nous détaillons ci-après les différentes étapes nécessaires à la création et la validation d'un modèle de prédiction pour données volcano-sismiques.

Mise en forme des signaux - Aux vues des quantités de données à traiter et du nombre élevé de classes considérées, nous proposons de formater les données afin d'obtenir un nombre N_i d'observations par classe i du même ordre de grandeur (voir Tableau 2 pour plus de détails). Le base de données ainsi constituée servira à l'apprentissage du modèle ainsi qu'à sa validation. Chaque observation de la base de donnée est donc une portion d'enregistrement qui a manuellement été assignée à sa classe par les experts. Chaque observation est également normalisée en énergie (une même classe pouvant contenir des signaux d'énergie très variable).

Extraction des descripteurs - Cette étape va permettre de représenter les observations dans un nouvel espace de représentation, à savoir l'espace des descripteurs. Ce changement de représentation se justifie d'une part (i) par la réduction de dimension des données. Les algorithmes d'apprentissage étant sujet au *fléau de la dimension* [1], il sera possible de construire un modèle à partir d'un faible nombre N d'observations lorsque ces observations sont de dimension d faible devant N. D'autre part (ii), des algorithmes tels que RF ou SVM ne peuvent pas modéliser des représentations ordonnées telles que les représentations temporelle ou spectrale. Des informations telles que la forme du signal, sa dérivée ou son accélération seraient perdues si le signal était directement utilisé par le modèle. Au contraire, utiliser des descripteurs qui quantifient explicitement ces informations permet une meilleure représentation et donc classification des données. (iii) Rappelons enfin que le modèle ne voit les observations qu'à travers ses descripteurs. Les résultats de classification sont donc directement liés à la capacité de l'espace des descripteurs à séparer les données en leurs classes respectives. L'enjeu principal du travail présenté dans ce papier réside dans la pertinence de l'espace de représentation choisi.

Pour notre étude, nous proposons d'utiliser les descripteurs généraux présentés dans le Tableau 1. Ces descripteurs vont décrire le signal et ses caractéristiques. On calcule en particulier des entropies, des descripteurs de forme et des moments statistiques comme l'écart type, le skewness ou le kurtosis qui décrivent respectivement l'étalement, l'asymétrie ou l'aplatissement d'une observation. L'une des nouveauté présentée par ce papier est d'extraire les descripteurs à partir de plusieurs domaines de représentation de l'observation considérée. En particulier, d'une observation $x(t)$, les descripteurs proposés sont extraits :

- du signal temporel $x(t)$ afin de décrire la forme d'onde du signal,
- du domaine fréquentiel $X(f) = \mathcal{TF}\{x(t)\}$ pour décrire le contenu spectral du signal.
- du domaine des fréquences de fréquences, aussi dénommé domaine *quéfrentiel* en traitement de parole. Ce signal quéfrentiel $\mathcal{X}(q) = \mathcal{TF}\{X(f)\}$ permet de visualiser d'éventuelles périodicités dans le spectre du signal et ainsi de souligner l'harmonicité d'un signal.

Extraire les descripteurs de ces trois représentations d'une même observation améliore les résultats de classification.

Apprentissage et validation du modèle - Les observations représentées dans l'espace des descripteurs sont utilisées conjointement avec l'information de leurs classes associée par un algorithme d'apprentissage pour apprendre un modèle de prédiction. Dans ce travail, on utilise SVM pour la consistance de ces performances (des résultats similaires sont obtenus en utilisant RF). Ce modèle va permettre de prédire la classe d'une nouvelle observation à analyser. Des N observations de la base de donnée initiale, $\alpha \cdot N$ avec le taux d'apprentissage $0 < \alpha < 1$ sont utilisées pour apprendre le modèle et les $(1 - \alpha) \cdot N$ restantes sont utilisées pour sa validation. On répète ce procédé jusqu'à obtention de résultats statistiquement stables (validation croisée).

4 Résultats Expérimentaux

4.1 Données et conditions expérimentales

Les données utilisées pour valider la méthode proposée ont été enregistrées sur les flancs du volcan Ubinas, d'altitude 5676m

au sud du Pérou (16 22' S, 70, 54' W). Ubinas est le volcan le plus dangereux du Pérou, et est à ce titre hautement surveillé par l'Instituto Geofisico del Péru. 6 classes de signaux sont considérés pour cette étude. Ces classes ont été définies par les géophysistes, et sont présentés dans le Tableau 2. Pour chacune des classes, 800 observations sont considérées, moins si la classe est faiblement représentée. Les signaux considérés peuvent être de longueur extrêmement variable. On fixe alors une durée maximale de 40s, afin que la seule durée ne soit pas un critère suffisant de discrimination. La base de donnée considérée comporte 3125 observations. SVM est utilisé avec un noyau gaussien de paramètre $\gamma = 0.01$ et un coût $C = 10$. Ces valeurs ont été fixées pour optimiser les résultats sur un sous-ensemble de données et sont constantes au long des divers tests. Tous les résultats considérés sont obtenus via validation croisée sur 50 tests. Une étude préliminaire a montré la stabilité des résultats au delà de 50 tests. Enfin, le taux d'apprentissage α varie au cours des tests afin d'estimer la quantité de données nécessaires à la construction d'un modèle.

TABLE 1 – Descripteurs généraux pour un signal $s[i]_{i=0}^{n}$. E et E_i désignent respectivement les énergies totales et à l'instant i du signal.

Feature	Definition	Ref.
Length	$n = \text{card}(s)$	[10]
Mean	$\mu_s = \frac{1}{n}\sum_i s[i]$	[10]
Standard deviation	$\sigma_s = \sqrt{\frac{1}{n-1}\sum_i (s[i]-\mu_s)^2}$	
Skewness	$\frac{1}{n}\cdot\sum_i \left(\frac{s[i]-\mu_s}{\sigma_s}\right)^3$	[9]
Kurtosis	$\frac{1}{n}\cdot\sum_i \left(\frac{s[i]-\mu_s}{\sigma_s}\right)^4$	[9]
Average	$\frac{\sum s(u)}{n}$	[10]
Central Energy	$\bar{i} = \frac{1}{E}.\sum_i E_i \cdot i$	[10]
RMS bandwidth	$RMS_i = \sqrt{\frac{1}{E}\cdot\sum_i i^2\cdot E_i - \bar{i}^2}$	[10]
Mean skewness	$\sqrt{\frac{\sum_i (i-\bar{i})^3.E_i}{E.RMS_i^3}}$	[10]
Mean kurtosis	$\sqrt{\frac{\sum_i (i-\bar{i})^4.E_i}{E.RMS_i^4}}$	[10]
Shannon entropy	$-\sum_j p(s_j).\log_2\left(p(s_j)\right)$	[4], [7]
Rényi 'entropy'	$\frac{1}{1-\alpha}\cdot\log_2\left(\sum_j p(s_j)^\alpha\right)$	[7]
rate of attack	$\max_i\left(\frac{s[i]-s[i-1]}{n}\right)$	[10]
rate of decay	$\min_i\left(\frac{s[i]-s[i+1]}{n}\right)$	[10]
Specific values	Ratios, min, max, mean, etc.	[10, 9]

4.2 Résultats

Les résultats de classification en 6 classes des données volcanosismiques issus de Ubinas sont présentés dans cette section. Les résultats validant la méthodologie sont présentés dans le Tableau 3. En particulier, on compare l'influence des descripteurs issus de la représentation temporelle $x(t)$, de la représentation fréquentielle $X(f)$ et de la représentation quéfrentielle $\mathcal{X}(q)$. Une matrice de confusion sur ces résultats est également proposée dans le Tableau 4. Enfin, l'influence du taux d'apprentissage α est présentée en Figure 1.

4.3 Analyse des Résultats et Discussion

Le Tableau 3 montre que l'architecture proposée atteint une précision de 90%, validant ainsi la méthodologie et le choix des descripteurs utilisés. On note en particulier l'influence des descripteurs utilisés : la précision ne dépasse pas 86.1%, 83.0% et 79.4% en utilisant les descripteurs calculés sur un seul domaine de représentation, respectivement $x(t)$, $X(f)$ et $\mathcal{X}(q)$. L'information discriminante pour la classification est disséminée dans divers espaces de représentation et c'est en combinant les caractéristiques de ces différents domaines que l'on obtient les meilleurs résultats. Cette observation est consistante avec l'avis des experts qui utilisent diverses représentations (temporelle et spectrale) d'un signal avant de le classifier.

Le Tableau 4 présente la matrice de confusion lorsque tous les descripteurs sont utilisés, avec $\alpha = 70\%$. On peut en particulier y lire les précisions moyennes par classe, mais aussi analyser la répartition des erreurs de prédiction. Deux sources d'erreur principales sont ici à analyser : (i) les tremors et longues périodes peuvent être confondus et (ii) les hybrides confondus avec des volcanos-tectoniques ou des longues périodes. Il est intéressant de noter que les signaux qui se confondent ont des mécanismes de production similaires : tremors et longues périodes sont associés à des fluides en mouvement dans la cheminée du volcan. De même, les hybrides sont des signaux croisés entre les volcano-tectoniques et les longues périodes. Les erreurs du modèle vont également dans le sens des experts qui ré-analysent manuellement les données et proposent encore aujourd'hui des classifications mises à jour.

En analysant l'influence du taux d'apprentissage α sur les performances, on observe dans la Figure 1 le phénomène de plateau pour chacune des classe. La valeur α pour laquelle le plateau est atteint dépend de la classe considérée, et est en particulier du nombre d'occurences N_i d'observations de la classe. Les courbes sont ainsi ordonnées selon les N_i (détaillés dans le Tableau 2). Pour les données considérées, on atteint une stabilité dans l'apprentissage lorsqu'une centaine d'observations est considérée par classe.

5 Conclusion et Perspectives

En conclusion de ce papier, nous insistons sur la nécessité du développement d'outils dédiés à l'analyse des données volcanosismiques. Les observatoires sont demandeurs de ces outils pour aider à l'analyse des grandes bases de données. Nous proposons ici une architecture de classification automatique adaptée à ce type de données. L'architecture est basée sur des techniques d'apprentissage supervisé, et s'appuie sur une description complète des signaux en temps, en fréquences, et en fré-

quences de fréquences. On obtient alors un précision de 90%, validée sur 3125 signaux enregistrés sur les flancs du volcan Ubinas. Les perspectives considérées pour ce travail sont d'associer l'outils à des méthodes de détection afin et traiter des données continues, ainsi que le développement de modèles basés sur l'apprentissage semi-supervisé. Les travaux exposés dans ce papier sont financés par la DGA et le LabEx@OSUG2020. Les auteurs remercient également Dr Orlando Macedo de l'Instituto Geofisico del Péru.

TABLE 2 – Base de données sismique d'Ubinas. Pour chaque classe on donne son nom (reference), une description des signaux, le nombre d'occurrences N_i, la fenêtre temporelle $\Delta_{t,i}$ nécessaire à l'enregistrement des N_i observations (en jours) et leur longueur moyenne.

Ref.	Description	N_i	$\Delta_{t,i}$	Durée moyenne
LP	Longue période	800	201j	40s
TR	Tremor	800	8j	27min48s
EXP	Explosion	154	1396j	51s
VT	Volcano-tectonique	800	1958j	24s
HIB	Hybride	466	1647j	34s
TOR	Tornillo	105	632j	40s

TABLE 3 – Influence du choix des descripteurs utilisés pour représenter les observations ($\alpha = 70\%$)

Descripteurs	Dimension	Précision
Temporels	32	$86.1 \pm 0.9\%$
Fréquentiels	32	$83.0 \pm 1.0\%$
Quéfrentiels	32	$79.4 \pm 1.0\%$
Tous	96	$90.1 \pm 0.9\%$

TABLE 4 – Matrice de Confusion, $\alpha = 70\%$, l'ensemble des descripteurs est utilisé pour représenter les données. La précision moyenne sur toutes les classes est de 90.4%

		Classe réelle (vérité terrain)					
		LP	TR	EXP	VT	HIB	TOR
Classe prédite	LP	**218**	16	0	1	6	1
	TR	17	**223**	0	1	1	0
	EXP	0	0	**40**	5	0	0
	VT	1	0	6	**219**	11	7
	HIB	3	1	0	12	**121**	0
	TOR	0	0	0	1	0	**24**
Précision :		90.9%	92.8%	86.4%	91.4%	86.8%	75.1%

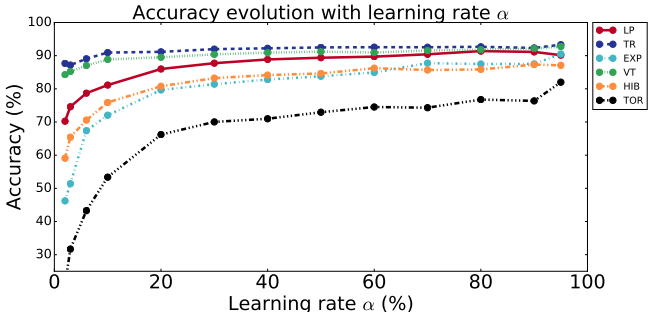

FIGURE 1 – Évolution de la précision de chacune des 6 classes en fonction du taux d'apprentissage α.

Références

[1] R. Bellman. Dynamic Programming and Lagrange Multipliers. *Proceedings of the National Academy of Sciences of the United States of America*, 42(10) :767–769, 1956.

[2] P. Cotret, S. Chevobbe, and M. Darouich. Reconnaissance faciale basée sur les ondelettes robuste et optimisée pour les systèmes embarqués. In *GRETSI*, Lyon, France, Sept. 2015.

[3] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley. Sons Eds, 2001.

[4] S. Esmaili, S. Krishnan, and K. Raahemifar. Content based audio classification and retrieval using joint time-frequency analysis. In *ICASSP-2004, IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 9–12, 2004.

[5] S. Fagerlund. Bird Species Recognition Using Support Vector Machines. *EURASIP Journal on Advances in Signal Processing*, 2007(1), 2007.

[6] I. Fujinaga and K. MacMillan. Realtime recognition of orchestral instruments. In *Proceedings of the International Computer Music Conference (ICMC2000)*, volume 141, pages 141–143, 2000.

[7] N. C. Han, S. V. Muniandy, and J. Dayou. Acoustic classification of Australian anurans based on hybrid spectral-entropy approach. *Applied Acoustics*, 72(9) :639–645, 2011.

[8] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. *Elements*, 1 :337–387, 2009.

[9] N. Langet. *Détection et caractérisation massives de phénomènes sismologiques pour la surveillance d ' événements traditionnels et la recherche systématique de phénomènes rares*. PhD thesis, University of Strasbourg, 2014.

[10] S. Tucker and G. J. Brown. Classification of transient sonar sounds using perceptually motivated features. *IEEE Journal of Oceanic Engineering*, 30(3) :588–600, 2005.

[11] S. Zaugg, M. Van Der Schaar, L. Houégnigan, C. Gervaise, and M. André. Real-time acoustic classification of sperm whale clicks and shipping impulses from deep-sea observatories. *Applied Acoustics*, 71(11) :1011–1019, 2010.