Optimisation de la segmentation de données d'émission acoustique à l'aide d'un algorithme génétique

Arnaud Sibil, Nathalie Godin, Mohamed R'Mili, Gilbert Fantozzi

10ème Congrès Français d'Acoustique
Lyon, 12-16 Avril 2010

Optimisation de la segmentation de données d'émission acoustique à l'aide d'un algorithme génétique.

A. Sibil[1], N. Godin[1], M. R'Mili[1], G. Fantozzi[1]

[1] MATEIS, INSA Lyon, 7 av. Jean Capelle F-69621 Villeurbanne, arnaud.sibil@insa-lyon.fr

La segmentation des données d'émission acoustique – ou partition des données – est un des enjeux actuels permettant d'accéder à des analyses plus approfondies de l'endommagement des matériaux. Parmi les méthodes de classification non supervisées (pour lesquelles on ne connaît pas à priori le nombre de classes présentes dans le jeu de données), une des plus utilisées reste la méthode des K-moyenne [1]. La faiblesse majeure de cette méthode réside dans l'initialisation aléatoire des centres de classes. Afin d'améliorer la robustesse de la méthode, un algorithme génétique a été optimisé. Les résultats du traitement de jeux de données issus de données provenant d'essais sur matériaux permettent de mettre en valeur les apports d'un tel algorithme. Cet algorithme à capable de distinguer des classes aux populations déséquilibrées ou aux paramètres extrêmes. Par ailleurs, l'importance du critère d'évaluation de la qualité de la segmentation et de son utilisation est soulignée par l'introduction d'un critère alternatif : les silhouettes. Enfin, les limites de la méthode sont discutées.

1 Introduction

La segmentation des données d'émission acoustique – ou partition des données – est un des enjeux actuels permettant d'accéder à des analyses plus approfondies de l'endommagement des matériaux par la discrimination des signaux provenant de différents mécanismes ou sources. Dans cette démarche, deux hypothèses principales sont à énoncer. On suppose tout d'abord que différents mécanismes sources conduisent à des libérations d'énergie différentes et donc des formes d'ondes différentes pour les signaux correspondants [2]. On considérera également que si les signaux sont différents à la source alors les signaux reçus par les capteurs seront également différents. En d'autres termes, la fonction de convolution comprenant la propagation des signaux et leur acquisition est donc supposée agir de manière analogue pour tous les signaux. Parmi les méthodes de classification non supervisées [3] (pour lesquelles on ne connaît pas à priori le nombre de classes présentes dans le jeu de données), une des plus utilisées reste la méthode des K-moyennes [1]. Les différents signaux y sont regroupés autour de centres de classe par un processus itératif; le nombre de classes est à définir par l'utilisateur. La qualité de la partition obtenue est évaluée par un critère reposant sur des calculs de similarité / dissimilarité entre les signaux de ces classes, critère qu'il convient en général de minimiser.

La faiblesse majeure [4] de cette méthode réside dans l'initialisation aléatoire des centres de classes pouvant aboutir à un minimum local plutôt que global du critère. Afin d'y palier, un algorithme génétique [5] a été utilisé. Ce dernier remplit deux fonctions principales : il permet tout d'abord un balayage plus large de l'espace des solutions et garantit ensuite la convergence de la solution par son processus évolutif guidé vers la performance [5,6]. Dans ce cas précis, il s'agit de minimiser une fonction dite « objectif » : le critère d'évaluation.

Dans ce travail, les résultats du traitement de jeux de données contenant quatre classes de signaux issus d'essais sur matériaux permettent de mettre en valeur les apports d'un tel algorithme génétique par rapport à un traitement des données qui en est dépourvu. L'importance du critère d'évaluation de la qualité de la segmentation et de son utilisation est soulignée par l'introduction d'un critère alternatif : les silhouettes. Enfin, la robustesse de l'algorithme est éprouvée sur des jeux de données dont la population d'une classe est progressivement réduite, indiquant sa faculté à traiter des jeux de données aux populations déséquilibrées.

2 Méthode

2.1 Données traitées

Des jeux de données ont été générés numériquement par création et combinaison de plusieurs classes de signaux. Les signaux de chaque classe sont décrits par un ensemble de paramètres dénommés descripteurs (fig.1) dont les valeurs ont été choisies proches de celles issues de données expérimentales (monitoring sur acier, céramique, fibre de verre, composite,…). Cinq classes ont ainsi été crées (table 1).

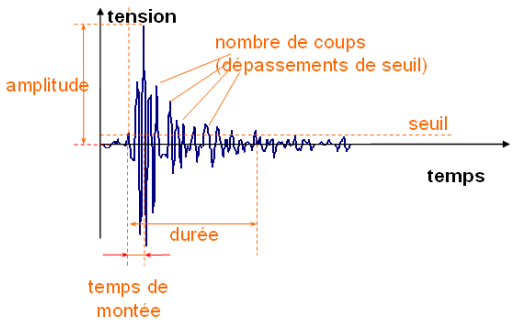
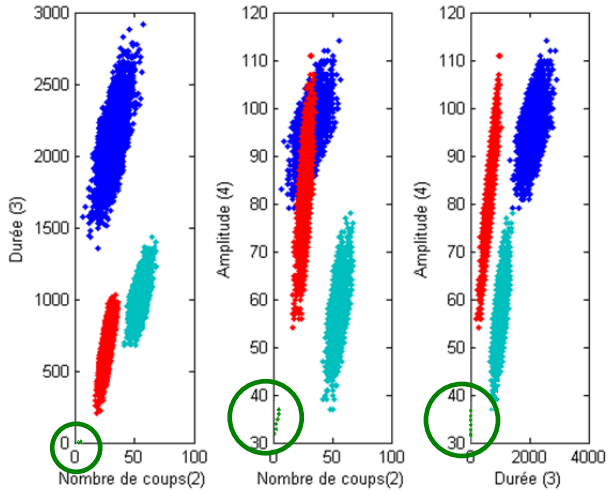


Figure 1. Présentation de quelques descripteurs.

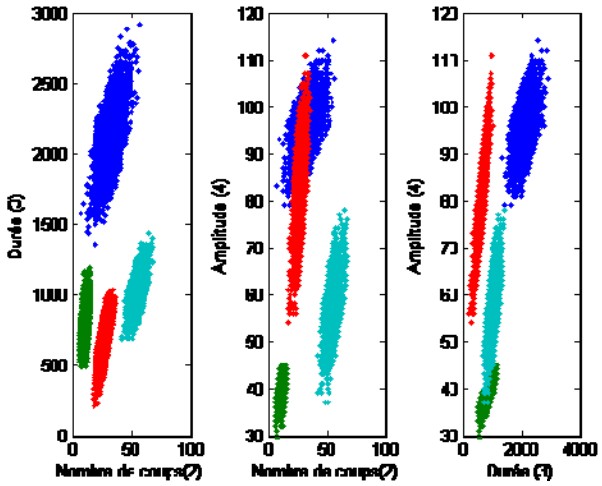
	Classe 1		Classe 2		Classe 3		Classe 4		Classe 5	
	m	σ	m	σ	m	σ	m	σ	m	σ
Temps de montée (µs)	27,4	4,1	21,0	3,0	9,1	2,6	2,0	0,4	8,0	2,3
Nombre de coups	32,3	7,8	54,9	4,3	26,7	3,3	2,5	0,6	10,6	1,9
Durée (µs)	2122,8	250,5	1024,3	124,1	643,5	133,4	3,5	0,8	821,9	122,8
Amplitude (dB)	95,6	5,7	57,5	6,7	82,2	9,2	35,0	0,9	38,5	2,5
Nombre de coups au pic	10,5	1,8	17,0	2,9	3,5	0,5	1,0	0,1	3,5	0,6
Energie absolue (aJ)	554,9	161,6	227,7	67,2	967,6	271,0	0,1	0,0	10,0	2,6

Table 1. Valeurs moyennes et écarts types des descripteurs des différentes classes.

Plusieurs jeux de données sont testés successivement. Le jeu n°1 (fig. 2a) est constitué de 2000 signaux des trois premières classes et de 1937 signaux de la classe n°4 présentant des valeurs très extrêmes pour ces descripteurs. Les jeux de données n°2 (fig. 2b) testés dans la dernière partie sont également composés de 2000 signaux des trois premières classes puis d'un nombre décroissant de signaux de la classe 5 (de 1500 à 100).



(a)



(b)

Figure 2. Représentation des jeux de données. (a) jeu n°1. (b) jeu n°2. Durée en µs, amplitude en dB.

2.2 Algorithmes de partition des données et critères d'évaluation utilisés

Issu d'une série de travaux [7,8], l'algorithme initial est un algorithme de classification basé sur l'algorithme des K-moyennes (fig. 3).

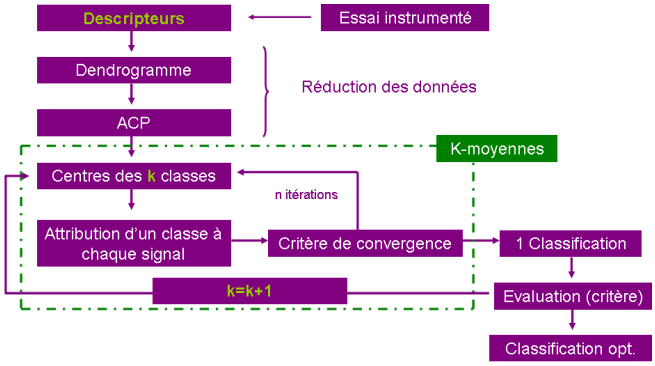


Figure 3. Algorithme de classification initial.

Il se décompose en plusieurs sections. La première tâche de l'algorithme est de normaliser les descripteurs ce qui permet d'éviter qu'un descripteur prenne plus d'importance que les autres en raison des ses valeurs plus élevées (énergie absolue notamment). Il s'agit ensuite de pouvoir décrire les signaux dans un espace plus réduit par une réduction du nombre de descripteurs à prendre en compte. Cela présente deux intérêts : les temps de calcul sont considérablement réduits et les signaux sont décrits dans un espace plus apte à les séparer en classes. Cela s'effectue par le biais d'un dendrogramme puis d'une analyse en composante principale (ou ACP [9]). Le nombre de classes désirées est un paramètre à indiquer. Les centres de classes correspondants sont alors initialisés de manière aléatoire. Le regroupement des signaux se fait à chaque itération par le biais du calcul des distances aux différents centres. Dans le cadre de ce travail, des distances euclidiennes pondérées par les valeurs propres ont été utilisées. Les centres de classe sont recalculés à la fin de chaque itération, ils prennent la valeur du barycentre des signaux qui leur sont rattachés. Afin de garantir la convergence pour un certain nombre d'itérations, des essais de répétabilité doivent être conduits. A l'issu de la classification des données, on procède à son évaluation par un critère. Le critère retenu initialement est le critère de Davis et Bouldin [10] calculé comme suit :

$$DB = \frac{1}{k}\sum_{i=1,k}^{k} \max\left[\frac{(\delta_i + \delta_j)}{\delta_{ij}}\right] \qquad (1)$$

où k est le nombre de classe, δ_i et δ_j les distances moyennes signaux-centres respectivement des classes i et j et δ_{ij} la distance entre les centres des classes i et j. Une valeur faible de DB indique ainsi une bonne partition des données, on cherche donc à en minimiser la valeur. La comparaison de cette valeur pour plusieurs partitions permet de déterminer la partition optimale.

Une optimisation de l'algorithme (fig. 4) précédent a été réalisée en insérant un algorithme génétique dans l'algorithme des K-moyennes.

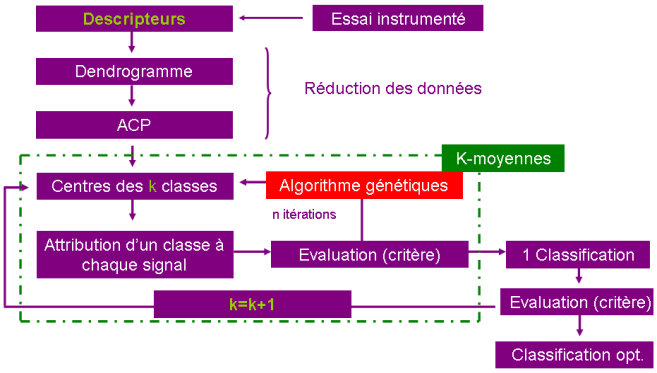


Figure 4. Genetic K-means algorithm.

L'intérêt de cet algorithme est de permettre un balayage de l'ensemble de l'espace des solutions. Dans ce cadre, ce n'est plus un jeu de centres de classes qui est initialisé mais un ensemble (population) de jeux de centres de classes qui sont crées, chaque jeu étant dénommé individu. Le choix du nombre d'individus doit se faire de manière raisonnée afin de ne pas allonger inutilement les temps de traitement. Dans ce travail, une valeur de 100 individus a été retenue. Les générations suivantes sont construites par des opérations qui reprennent la terminologie de la génétique. Les individus les plus performants (pour lesquels le critère d'évaluation présente une valeur élevée) sont conservés d'une génération à la suivante ; ils sont appelés les élites. Leur nombre a été fixé à deux. Les autres « enfants » sont crées par différents processus : sélection, mutation et croisement. La sélection des individus qui vont permettre de construire la génération suivante se fait avec une probabilité proportionnelle à la valeur des coefficients d'évaluation attachés à chaque individu. Cette opération est commune à tous les futurs individus. En fait, une partie de la population est issue de croisement (80%), l'autre partie de mutation (20%). Les croisements s'effectuent entre deux individus sélectionnés, les parents, en mode multipoints : pour chaque caractéristique de l'enfant (gène), les probabilités de transmission du « père » et de la « mère » sont égales. La mutation quant à elle frappe les gènes des autres individus avec une faible probabilité (1%) : la valeur du gène à muter est déterminée de manière aléatoire dans la gamme de valeur occupée par le jeu de données.

Contrairement à l'algorithme précédent, l'arrêt de la classification se fait par convergence du critère d'évaluation. Ici, à partir d'une variation de moins de 1E-7 du coefficient de Davis et Bouldin par génération, on considère que l'algorithme a convergé.

Par ailleurs, un autre critère a également été implémenté, il s'agit des silhouettes [11]. Elles sont déterminées pour chaque signal par :

$$s_{ind} = \frac{b_i - a_i}{\max(a_i ; b_i)} \qquad (2)$$

où a_i est la distance signal i au centre sa classe de rattachement et b_i est la distance entre ce signal et la classe la plus proche. Ce critère est à considérer comme un indice de confiance de l'appartenance du signal à sa classe, une valeur de 1 signifiant une parfaite appartenance tandis qu'une valeur négative indique une mauvais affectation. Ce critère, individuel, permet d'évaluer de manière plus précise la qualité de la classification. Il peut également être utilisé comme critère global par le calcul de S_i, moyenne des silhouettes de tous les signaux.

3 Résultats et discussion

3.1 Apports de l'algorithme génétique ; traitement d'un jeu de données contenant une classe extrême.

Le jeu de données n°1 a été traité successivement par les deux algorithmes. Afin de vérifier la pertinence des critères d'évaluation et d'évaluer la capacité des algorithmes à distinguer les quatre classes qui composent ce jeu, des partitions ont été conduites pour des nombres de classes allant de 2 à 6. Pour chacun des algorithmes et pour chaque classification, la valeur du coefficient de Davies et Bouldin (DB) et la moyenne des silhouettes (Si) ont été calculées (table 2).

Nombre de classes	DB	Si
2	0,179	0,658
3	0,315	0,613
4	0,295	0,567
5	0,260	0,540
6	0,248	0,644

(a)

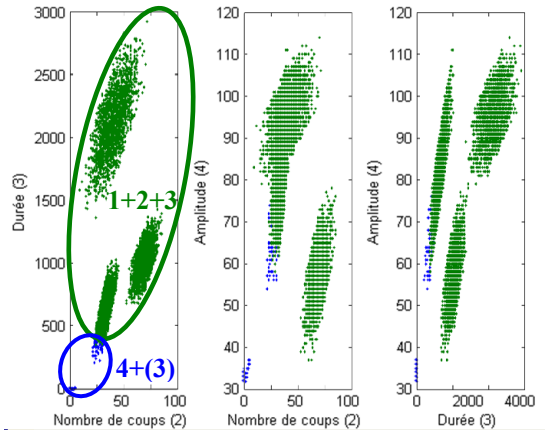
Nombre de classes	DB	Si
2	0,170	0,659
3	0,197	0,597
4	0,168	0,675
5	0,262	0,671
6	0,226	0,552

(b)

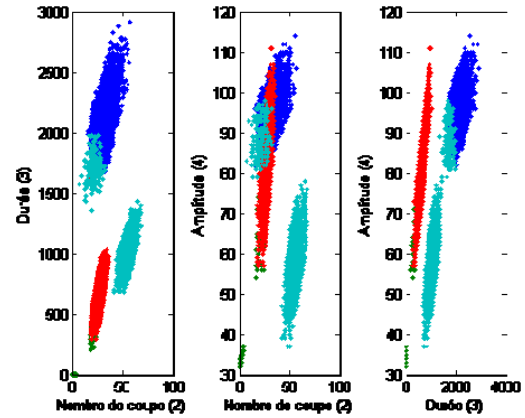
Table 2. Valeurs des critères pour les différentes classifications. Algorithmes (a) initial et (b) génétique.

Ces résultats permettent de mettre en lumière plusieurs éléments. Tout d'abord, les valeurs du coefficient de Davis et Bouldin, coefficient qui sert de fonction « objectif » à l'algorithme génétique, sont systématiquement plus faibles dans le cas de l'utilisation de cet algorithme, témoignant de meilleures partitions des données. L'algorithme optimisé permet donc bien d'obtenir des classifications de meilleure qualité. Le fait que les résultats ne dépendent pas du nombre d'itérations en fait également un algorithme reproductible. Aussi, les valeurs de silhouettes calculées sont corrélées avec les valeurs de DB ; leur maximum coïncide avec le minimum de DB marquant pour chacun des algorithmes la classification optimale. L'algorithme génétique parvient à dissocier les quatre classes composant

le jeu alors que l'algorithme initial présente des valeurs aboutissant à une classification optimale faite de deux classes. Les populations des classes résultant de ces partitions permettent d'évaluer le pourcentage d'erreur réalisé dans chaque cas. Pour l'algorithme initial, la segmentation se compose d'une classe de 2000 signaux et d'une classe de 5937 signaux. Pour l'algorithme génétique, les classes ont des populations respectives de 1817, 1968, 1971 et 2181 signaux. En ce qui concerne ce dernier, seulement 181 signaux ont été classés de manière erronée, ce qui correspond à une faible erreur de l'ordre de 2,3 %. Ce résultat est bien meilleur que celui obtenu par l'algorithme initial qui a regroupé trois classes (classe 1, 2 et 3) et en a isolé une avec quelques parasites (classe 4 et quelques signaux de la classe 3). Pour cet algorithme, l'erreur pour quatre classes culmine à près de 37%. Les partitions des signaux peuvent être visualisées dans des plans composés de deux descripteurs (fig. 5)



(a)



(b)

Figure 5. Visualisation des classifications optimales réalisées avec l'algorithme (a) initial et (b) génétique. Durée en µs, amplitude en dB.

3.2 Utilisation des silhouettes

Si des résultats satisfaisants peuvent être obtenus par l'utilisation de l'algorithme génétique, ces résultats restent perfectibles. En particulier, il peut être envisagé de ne plus considérer, après classification, les signaux qui ont un faible niveau d'appartenance à leur classe. Cette évaluation du niveau d'appartenance est réalisée pour chaque signal par le calcul de sa silhouette. Il est donc possible, à l'issu du traitement des données et pour la classification réalisée précédemment à l'aide de l'algorithme génétique sur le jeu de données n°1, de tracer les histogrammes de leurs valeurs pour chaque classe (fig. 6).

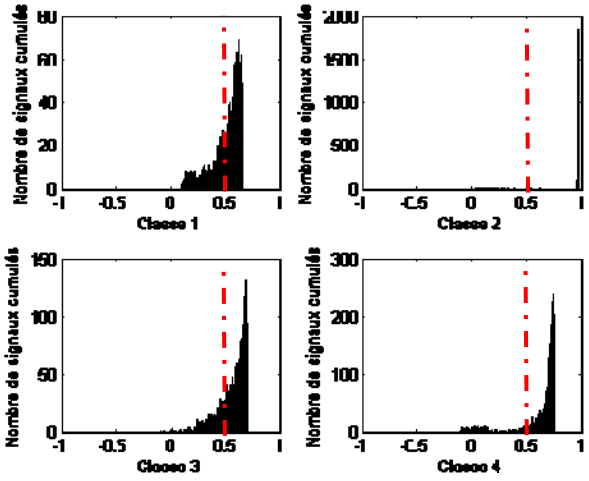


Figure 6. Histogrammes des silhouettes des quatre classes.

Les valeurs proches de zéro et même certaines valeurs négatives correspondent aux signaux qui ont été classés de manière erronée. Afin de diminuer l'erreur commise et d'améliorer la classification (diminuer l'erreur commise sur la partition du jeu de données), on peut considérer que seuls les signaux qui possèdent une silhouette supérieure ou égale à 0,5 doivent être conservés. Ce seuil représente le niveau de confiance à atteindre pour qu'un signal puisse être affecté indiscutablement à une classe. Cette procédure de sélection conduit à conserver 83,4 % des signaux.

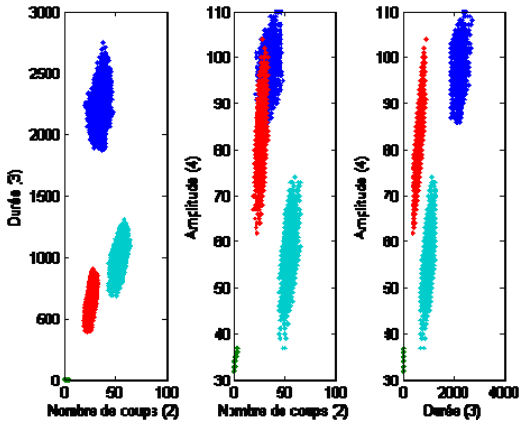


Figure 7. Classification obtenue après sélection par les silhouettes ($s_{ind} \geq 0,5$). Durée en µs, amplitude en dB.

La visualisation de la partition des signaux conservés (fig. 7) permet dans le cas présent de vérifier l'absence de recouvrement de signaux de classes différentes.

3.3 Variation des populations de classes

Un autre défi majeur des algorithmes de classification résident dans leur capacité à distinguer une classe de faible population parmi d'autres classes de plus fortes populations. Pour évaluer les aptitudes des deux algorithmes dans ce domaine, des jeux de données tests ont été créés. Ils se différencient uniquement par la population de la quatrième classe (en vert, fig. 2b) qui décroit progressivement de 1500 signaux à 100 signaux (contre 2000 signaux pour les autres classes).

Pour chaque jeu test et pour chaque algorithme, le nombre de classe optimale et le taux d'erreur ont été calculés. Ce taux d'erreur est déterminé par le rapport entre le nombre de signaux correctement classés sur le nombre total de signaux. Les résultats (table 3) montrent la robustesse de l'algorithme génétique qui parvient dans tous les cas à retrouver un nombre de classe optimale de quatre donc en conformité avec ce qui est attendu. Le taux d'erreur reste relativement faible, en dessous de 0,5 %. Au contraire, l'algorithme initiale souffre de son défaut d'initialisation et ne conduit que dans deux cas sur 6 à une classification à quatre classes. Le taux d'erreur reste faible lorsque les quatre classes attendues sont retrouvées mais peut s'envoler à près de 30%.

	Algorithme initial		Algorithme génétique	
Population minoritaire	Nb. Classes	Erreur (%)	Nb. Classes	Erreur (%)
1500	4	0,35	4	0,33
1000	5	28,2	4	0,48
750	5	25,7	4	0,35
500	4	0,4	4	0,24
250	3	31,6	4	0,03
100	6	37,8	4	0,002

Table 3. Résultats des deux algorithmes.

L'algorithme génétique présente donc bien un intérêt important pour la classification des données puisque les mécanismes sources d'émission acoustique ne sont que très rarement activés de manière équilibrée dans les matériaux étudiés ce qui conduit à enregistrer des populations de signaux de tailles très variables.

4 Conclusion

L'algorithme initial permet d'aboutir dans certains cas à la classification optimale. Cependant, il souffre de faible reproductibilité et nécessite d'être répéter un certain nombre de fois pour tenter de se soustraire au problème bien connu d'initialisation des K-Moyennes. Il est également très sensible aux valeurs extrêmes. L'insertion d'un algorithme génétique dans le process de traitement des données permet de palier à ces faiblesses. Tout d'abord, l'algorithme génétique apporte une amélioration systématique de la qualité des partitions réalisées, amélioration soulignée par les critères d'évaluation. Ensuite, cet algorithme conduit à une meilleure pertinence du nombre de classes optimal indiqué. Il est par ailleurs beaucoup moins sensible aux valeurs extrêmes. Enfin, l'utilisation simultanée des silhouettes permet d'améliorer d'avantage la qualité des classifications en réduisant leur taux d'erreur.

Remerciements

Les auteurs tiennent à remercie le Ministère de l'Economie, des Finances et de l'industrie ainsi que Saint Gobain CREE pour leur support financier (ANR NOREV) et leur collaboration.

Références

[1] J. B. MacQueen. Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", (1967), 1:281-297

[2] AG Beattie. Acoustic emission, principles and instrumentation, J. Acoust. Emission (1983), 95–128.

[3] S. Kotsiantis, P. Pintelas. Recent Advances in Clustering: A Brief Survey, WSEAS Transactions on Information Science and Applications (2004), 1:73-81.

[4] P. S. Bradley , U. M. Fayyad. Refining Initial Points for K-Means Clustering, Proc. of 15th Int. Conf. on Machine Learning (1998), 91-99.

[5] John H. Holland. Adaptation in Natural and Artificial Systems. An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence (1975).

[6] David Goldberg, Addison Wesley. Algorithmes génétiques (1994).

[7] N. Godin, S. Huguet, R. Gaertner. Integration of the Kohonen's self-organising map and k-means algorithm for the segmentation of the AE data collected during tensile tests on cross-ply composites, NDT & E International 38 (2005), 299-309

[8] M. Moevus, N. Godin, M. R'Mili, D. Rouby, P. Reynaud, G. Fantozzi, G. FarizyA. Analysis of damage mechanisms and associated acoustic emission in two SiCf/[Si–B–C] composites exhibiting different tensile behaviours. Part II: Unsupervised acoustic emission data clustering, Composites Science and Technology 68 (2008), 1258-1265

[9] C. Ding, X. He. K-means Clustering via Principal Component Analysis, Proc. of Int'l Conf. Machine Learning (2004), 225–232.

[10] D.L. Davies, D.W. Bouldin. A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intelligence (1979), 1:224–227.

[11] P. Rousseeuw. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis, J. Comput. Appl. Math. 20 (1987), 53-65.