# Déséquilibre multi-classes : une approche évidentielle de rééchantillonnage hybride

Fares Grina, Zied Elouedi, Eric Lefevre

# Déséquilibre multi-classes : une approche évidentielle de rééchantillonnage hybride

# Multi-class imbalance: an evidential hybrid resampling approach

Fares Grina[1,2] Zied Elouedi[1] Eric Lefevre[2]

[1] Université de Tunis, Institut Supérieur de Gestion de Tunis, LARODEC, Tunisia

[2] Univ. Artois, UR 3926, Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A), F-62400 Béthune, France

grina.fares2@gmail.com, zied.elouedi@gmx.fr, eric.lefevre@univ-artois.fr

**Résumé :**

La classification, dans le cas d'ensembles de données déséquilibrés, connait un intérêt considérable dans la communauté de l'apprentissage automatique. Dans cet article, nous présentons une méthode de rééchantillonnage hybride évidentiel pour traiter le déséquilibre des classes dans un contexte multi-classes. Cette technique exploite la théorie des fonctions de croyance pour attribuer une étiquette évidentielle à chaque objet. Cette représentation fournit plus d'informations sur la région de chaque observation, ce qui améliore la sélection des objets à la fois pour le sous-échantillonnage et le sur-échantillonnage. Un ajustement a également été intégré afin d'éviter un sur-échantillonnage et un sous-échantillonnage excessifs. Des résultats ont montré une amélioration significative des mesures G-Mean et AUC par rapport à d'autres méthodes classiques de rééchantillonnage.

**Mots-clés :**

Rééchantillonnage, Déséquilibre de classes, Théorie des fonctions de croyance, Incertitude

**Abstract:**

Learning from class-imbalanced datasets has gained substantial attention in machine learning community. Yet, there has been little emphasis given to dealing with multi-class imbalance learning. In this paper, we present an evidential hybrid resampling method for dealing with class imbalance in the multi-class setting. This technique uses the belief function theory to assign a soft label to each object. This evidential modeling provides more information about each object's region, which improves the selection of objects in both undersampling and oversampling. An adjustment has also been integrated in order to avoid excessive oversampling and undersampling. Benchmarking results have shown significant improvement of G-Mean of AUC metrics over other popular resampling methods.

**Keywords:**

Resampling, Multi-class imbalance, Belief Function Theory, Uncertainty

## 1 Introduction

Le déséquilibre des classes est une situation très courante dans les problèmes de classification, notamment dans de nombreuses applications réelles telles que la détection des intrusions [5], le diagnostic médical [16] et la détection des fraudes [17]. Formellement, une base de données déséquilibrée contient au moins une classe avec un nombre d'exemples beaucoup plus faible que les autres classes. Les classes sous-représentées sont appelées classes minoritaires, tandis que les autres classes, ayant un plus grand nombre d'exemples, sont appelées classes majoritaires. Dans la plupart des cas, les modèles de classification ont tendance à favoriser les classes majoritaires en raison de leur forte présence, tout en classant incorrectement les instances des classes minoritaires. Cela pose un problème puisque les classes minoritaires ont souvent plus d'importance que les classes majoritaires.

De nombreuses études (comme les approches de coûts ou de rééchantillonnage) ont été menées afin d'améliorer les performances de classification sur des ensembles de données binaires, dans lesquels il n'existe qu'une classe minoritaire et une classe majoritaire. Parmi ces stratégies, le rééchantillonnage est une approche efficace dont le but est de rééquilibrer l'ensemble d'apprentissage au niveau du prétraitement en ajoutant des objets minoritaires synthétiques (sur-échantillonnage), en supprimant des échantillons majoritaires (sous-échantillonnage), ou les deux (rééchantillonnage hybride). Les techniques de rééchantillonnage classiques sont le

sur-échantillonnage aléatoire (ROS) et le sous-échantillonnage aléatoire (RUS). La première sélectionne aléatoirement les objets minoritaires et les reproduit, tandis que la seconde supprime aléatoirement les instances majoritaires.

Outre ces méthodes aléatoires, la méthode qui est considérée comme une référence en sur-échantillonnage et qui est la plus utilisée est nommée SMOTE [6]. Contrairement à ROS, SMOTE crée, par interpolation, de nouvelles instances minoritaires, entre les objets minoritaires qui sont proches les uns des autres. Néanmoins, de nombreux travaux [7, 10, 13] ont proposé d'autres versions de SMOTE, car cette approche peut potentiellement provoquer une amplification du bruit et des chevauchements déjà présents dans les données.

En ce qui concerne le sous-échantillonnage, les contributions se sont concentrées sur la sélection intelligente des échantillons majoritaires à éliminer, plutôt que de les supprimer de manière aléatoire. Ainsi, les méthodes ENN (Editing Nearest Neighbors) [23] et TL (Tomek Links) [15] ont été proposées pour le sous-échantillonnage.

La combinaison du sur-échantillonnage et du sous-échantillonnage est également une stratégie utile pour rééquilibrer l'ensemble des données. Traditionnellement, les méthodes combinent le sur-échantillonnage de SMOTE avec des techniques de filtrage par sous-échantillonnage. Dans [4], par exemple, les auteurs ont proposé d'appliquer ENN et TL, afin de créer deux approches d'échantillonnage hybride : SMOTE-ENN et SMOTE-TOMEKLINKS.

En comparaison avec le cas binaire, l'apprentissage à partir d'ensembles de données déséquilibrés multi-classes a été moins traité dans la littérature. Cela s'explique par la complexité des relations multi-classes. Généralement, des approches binaires sont utilisées pour traiter un problème multi-classes

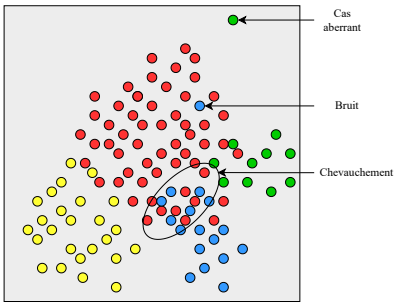


FIGURE 1 – Illustrations des problèmes rencontrés dans le cas de données multi-classes déséquilibrées.

qui aura été préalablement décomposé en problème binaire avec des approches comme : un-contre-un [14], un-contre-tous [19] et code correcteur d'erreurs (ECOC) [9]. Même si ces stratégies de décomposition sont des approches simples et directes, certaines peuvent conduire à ignorer l'apprentissage de certaines régions. Plus particulièrement, lorsqu'il y a des données incertaines, comme l'ambiguïté créée par un chevauchement élevé ou le bruit [2].

Afin de remédier à ces inconvénients, cet article présente une approche de rééchantillonnage hybride évidentiel (MC-EVHS). Nous proposons d'utiliser une méthode de rééchantillonnage dans le cadre de la théorie des fonctions de croyance [12] pour traiter spécifiquement les ensembles de données déséquilibrés avec plusieurs classes. L'utilisation de cette théorie permettra également de pallier les difficultés liées aux données comme le chevauchement de classes, le bruit des étiquettes et les valeurs aberrantes. Ces différents aspects sont illustrés sur la Figure 1. Les fonctions de croyance nous fournissent plus d'informations afin de mieux choisir les emplacements des objets à générer et les instances majoritaires à supprimer. Ainsi, nous appliquons une version évidentielle de SMOTE sur les classes minoritaires, et un sous-échantillonnage évidentiel sur les classes majoritaires. Notre approche présente également un mécanisme permettant d'identifier les classes à considérer pour le sur-échantillonnage ou le

sous-échantillonnage, ainsi qu'un contrôle de la quantité de rééchantillonnage à effectuer pour chaque classe. Cela offre un comportement adaptatif à notre approche.

Ce document est structuré de la manière suivante. Tout d'abord, les notions de base de la théorie des fonctions de croyance sont rappelées dans la Section 2. La Section 3 détaille chaque étape de notre méthode. L'évaluation expérimentale est présentée dans la Section 4. Nous terminons notre article par une conclusion et nos travaux futurs dans la Section 5.

## 2 Théorie des fonctions de croyance

La théorie des fonctions de croyance [8, 20, 21], également appelée théorie de Dempster-Shafer (DST), est un cadre adapté pour la représentation et la combinaison d'informations incertaines. Le cadre de discernement, qui est défini comme un ensemble fini de $M$ événements possibles exclusifs (ex. les étiquettes possibles pour un objet à classer), est noté $\Omega = \{w_1, w_2, ..., w_M\}$. La connaissance partielle sur $\Omega$ est représentée par une fonction de masse (*bba*) $m : 2^\Omega \rightarrow [0, 1]$ telle que :

$$\sum_{A \in 2^\Omega} m(A) = 1 \tag{1}$$

Chaque masse $m(A)$ quantifie la croyance allouée à un événement $A$ de $\Omega$. Un élément focal est tout sous-ensemble $A$ de $2^\Omega$ tel que $m(A) > 0$. La fonction de plausibilité est une autre représentation de cette connaissance définie comme suit :

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \quad \forall \ A \in 2^\Omega \tag{2}$$

$Pl(A)$ s'interprète comme la part de croyance qui pourrait potentiellement être allouée à $A$.

## 3 Approche évidentielle de rééchantillonnage hybride

L'approche que nous proposons, MC-EVHS, est une méthode de rééchantillonnage qui combine le sur-échantillonnage et le sous-échantillonnage pour rééquilibrer les ensembles de données multi-classes. Tout d'abord, une fonction de masse est créée pour chaque objet de l'ensemble de données. Cela nous permet de représenter l'incertitude des données. Les *bba*s calculées sont ensuite utilisées pour sélectionner les échantillons majoritaires non désirés pour le sous-échantillonnage, et pour choisir les bonnes régions pour le sur-échantillonnage des classes minoritaires.

Dans cet article, nous considérons comme minoritaire chaque classe qui a un nombre d'objets inférieur au nombre d'objets moyen $s$ par classe. Pour toutes les classes dont le nombre d'objets est inférieur à la moyenne $s$, nous utilisons les appartenances évidentielles afin d'effectuer un sur-échantillonnage dans les frontières de la classe minoritaire. Ainsi, notre version du sur-échantillonnage s'adapte à chaque classe et génère des instances minoritaires synthétiques uniquement aux endroits voulus, sans dépasser la moyenne $s$. Pour toutes les classes dont le nombre d'objets est supérieur à la moyenne $s$, les appartenances attribuées sont utilisées pour effectuer un sous-échantillonnage adaptatif. Toutefois, après ce sous-échantillonnage, la taille de chaque classe majoritaire ne doit pas être inférieure à la moyenne calculée $s$.

### 3.1 Création des *bba*s

L'approche que nous proposons détermine les centres de chaque classe et méta-classe (la région de chevauchement), puis crée une *bba* reposant sur la distance entre chaque objet et chaque centre de classe. Les centres de classe sont calculés en utilisant la valeur moyenne de l'ensemble d'apprentissage dans la classe correspondante. En ce qui concerne les régions de chevauchement représentées par des méta-classes, les centres sont définis par le barycentre des centres des classes concernées, comme suit :

$$C_U = \frac{1}{|U|} \sum_{\omega_i \in U} C_i \tag{3}$$

où $\omega_i$ sont les classes de $U$ (qui représente la méta-classe), et $C_i$ est le centre correspondant.

Après la création des centres, nous attribuons à chaque exemple une *bba* sur le cadre de discernement $\Omega = \{\omega_1, ..., \omega_M, \omega_0\}$, où les $M$ classes sont représentées. La proposition $\omega_0$ est incluse dans le cadre de discernement pour représenter l'aberration, c'est-à-dire l'affectation des objets qui sont éloignés de toute classe. Comme appliqué dans [18], le centre de la méta-classe devrait être plus proche des centres des classes impliquées que des centres des autres classes. Si ce n'est pas le cas, la méta-classe n'est pas considérée comme élément focal. Soit $x_s$ un objet appartenant à l'ensemble d'apprentissage. L'idée est que chaque centre de classe ou de méta-classe représente un élément de preuve de l'appartenance de $x_s$. Par conséquent, les valeurs de masse pour chaque élément focal en ce qui concerne les appartenances de $x_s$ devraient dépendre de $d(x_s, C)$, c'est-à-dire de la distance entre le centre de la classe/méta-classe $C$ et de $x_s$. Plus le centre est éloigné, plus la valeur de masse pour la classe/méta-classe correspondante est faible. Par analogie, plus $x_s$ est proche d'un centre de classe/méta-classe, plus il est probable qu'il appartienne à cette classe/méta-classe. Par conséquent, les masses initiales non normalisées doivent être représentées par des fonctions décroissantes en fonction de la distance. La distance de Mahalanobis est employée afin de traiter les ensembles de données anisotropes [18]. Les masses non normalisées sont calculées en conséquence :

$$\hat{m}(\{\omega_i\}) = e^{-d(x_s, C_i)} \tag{4}$$

$$\hat{m}(U) = e^{-\gamma\ \lambda\ d(x_s, C_U)}, \quad for\ \ |U| \geq 1 \tag{5}$$

$$\hat{m}(\{\omega_0\}) = e^{-t} \tag{6}$$

où $\lambda = \beta\ |U|^{\alpha}$. La valeur de $\alpha$, qui permet de pénaliser les méta-classes avec de grandes cardinalités, est fixée à 1 comme recommandée pour obtenir de bons résultats en moyenne, et $\beta$ est un paramètre tel que $0 < \beta < 1$. Ce dernier permet de contrôler le nombre d'objets dans la région de chevauchement. La valeur de $\gamma$ est égale au rapport entre la distance maximale de $x_s$ aux centres dans $U$ et la distance minimale. Elle est utilisée pour mesurer le degré de distinguabilité entre les classes de $U$. Plus $\gamma$ est petit, plus le degré de distinguabilité entre les classes de $U$ est faible pour $x_s$. La classe aberrante $\omega_0$ est prise en compte afin de traiter les objets éloignés de toutes les classes, et sa valeur de masse est calculée selon un seuil d'aberration $t$. Enfin, la fonction de masse $\hat{m}$ est normalisée comme suit :

$$m(A) = \frac{\hat{m}(A)}{\sum_{B \subseteq \Omega} \hat{m}(B)} \tag{7}$$

### 3.2 Sous-échantillonnage évidentiel adaptatif

Cette partie consiste à sous-échantillonner les classes majoritaires. Les *bba*s créées sont utilisées ici pour déterminer si un objet est nécessaire ou non pour la phase d'apprentissage. Notre idée est d'écarter les instances majoritaires qui ont une grande incertitude, c'est-à-dire les observations qui présentent une difficulté relativement plus élevée à être classer correctement. Ce type d'instances implique une forte ambiguïté (échantillons chevauchant les classes), des valeurs aberrantes et du bruit d'étiquette. La fonction de masse est utilisée pour détecter ces échantillons.

**Chevauchement de classes.** Dans notre cadre, les objets qui se chevauchent ont des masses élevées affectées aux éléments focaux de la méta-classe, c'est-à-dire aux propositions non singleton. Par exemple, un échantillon dont la masse maximale est attribuée à $U = \{\omega_1, \omega_2, \omega_3\}$ signifie qu'il est situé dans la région qui croise les trois classes $\omega_1$, $\omega_2$, et $\omega_3$. Cet objet peut être supprimé dans la phase de sous-échantillonnage, afin de réduire l'ambiguïté des données et la taille des classes majoritaires, en même temps. Puisque le chevauchement des classes n'est pas bien défini mathématiquement, il convient de

mettre en place un certain contrôle du nombre d'exemples à supprimer. Par conséquent, les objets sélectionnés pour le sous-échantillonnage sont triés dans un ordre décroissant en fonction de la valeur de masse moyenne attribuée aux éléments non-singletons $\bar{\mu}$. Formellement, pour un objet sélectionné $x_i$ :

$$\bar{\mu}_{x_i} = \frac{\sum_{|A|>1} m(A)}{k}, \quad A \in 2^{\Omega} \qquad (8)$$

où $k$ représente le nombre d'éléments focaux non-singletons. En d'autres termes, les objets les plus ambigus (plus grande imprécision) sont d'abord éliminés jusqu'à ce que la taille de la classe majoritaire atteigne la moyenne $s$. En ce qui concerne les objets majoritaires dont la masse la plus élevée n'est pas affectée à une proposition non-singleton (méta-classe), nous pouvons confirmer que ceux-ci ne sont pas situés dans une région de chevauchement. Cependant, ils pourraient être situés loin de toutes les classes (aberration), ou dans une classe différente (étiquette bruitée). Pour mieux détecter ces types d'échantillons, on utilise la plausibilité maximale $Pl_{max} = max_{\omega \in \Omega} Pl(\{\omega\})$.

**Objet aberrant.** Ce type d'objets est situé loin de toutes les classes. Ainsi, les objets, dont la plausibilité maximale est attribuée à $\omega_0$, en d'autres termes $Pl_{max} = Pl(\{\omega_0\})$, sont éliminés de l'ensemble de données.

**Étiquette bruitée.** Raisonnablement, un objet localisé dans une région sûre devrait avoir le maximum de plausibilité attribué à son étiquette. Sinon, ce dernier pourrait être considéré comme situé dans une autre classe, ce qui pourrait être décrit comme du bruit. Suivant cette logique, chaque objet, dont la plausibilité maximale est affectée à une autre étiquette que la sienne, est éliminé de l'ensemble de données.

### 3.3 Sur-échantillonnage évidentiel adaptatif

Afin de renforcer la présence des classes minoritaires dans l'ensemble de données, une phase de sur-échantillonnage est ajoutée pour rendre les frontières de chaque classe minoritaire plus robustes. Notre objectif, dans cette phase, est de mettre l'accent sur les limites de chaque classe minoritaire, tout comme d'autres techniques de sur-échantillonnage telles que BorderlineSMOTE [13]. Un autre aspect de notre approche consiste à éviter le sur-échantillonnage d'exemples bruités et aberrants.

Les *bbas* calculées précédemment sont utilisées, dans cette phase, pour choisir intelligemment les régions où les objets minoritaires doivent être créés. Les instances minoritaires sont triées en trois catégories : chevauchement, bruit d'étiquette ou aberration. Si un objet ne correspond pas à l'une de ces trois catégories, il est considéré comme situé dans une région sûre et n'est pas sélectionné pour être un nouvel objet. Il en va de même pour les étiquettes bruitées et les valeurs aberrantes. En effet, la sélection d'objets bruités et de valeurs aberrantes pour générer de nouveaux échantillons peut conduire à une amplification de ces problèmes.

Ainsi, seuls les objets dont la masse la plus élevée est engagée vers une région de chevauchement sont sélectionnés pour le sur-échantillonnage. Cette stratégie nous permet de mieux définir les frontières des classes minoritaires.

Comme mentionné ci-dessus, le nombre d'exemples générés est également contrôlé et la taille de chaque classe minoritaire ne doit pas dépasser la moyenne $s$. Ainsi, les objets de la classe minoritaire correspondante sont triés par ordre décroissant en se basant sur l'Eq. 8. L'idée est de donner la priorité aux objets minoritaires avec la plus grande incertitude afin de générer des objets synthétiques dans des endroits difficiles à classer.

## 4 Étude expérimentale

Dans cette section, avant de présenter et de commenter les résultats obtenus nous

TABLEAU 1 – Description des bases de données sélectionnées à partir de KEEL.

| Datasets | Distribution | Caractéristiques | Taille | #classe |
|---|---|---|---|---|
| dermatology | 112 ; 61 ; 72 ; 49 ; 52 ; 20 | 34 | 366 | 6 |
| wine | 71 ; 59 ; 48 | 13 | 178 | 3 |
| pageblocks | 492 ; 33 ; 8 ; 12 ; 3 | 10 | 548 | 5 |
| thyroid | 17 ; 37 ; 666 | 21 | 720 | 3 |
| hayes-roth | 51 ; 51 ; 30 | 4 | 132 | 3 |
| contraceptive | 629 ; 511 ; 333 | 9 | 1473 | 3 |
| yeast | 463 ; 429 ; 244 ; 163 ; 51 ; 44 ; 35 ; 30 ; 20 ; 5 | 8 | 1484 | 10 |

présentons la méthodologie avec laquelle nous avons mené nos expérimentations.

### 4.1 Cadre expérimental

**Ensembles de données.** Dans ce travail, sept bases de données déséquilibrées multi-classes ont été sélectionnées dans le référentiel KEEL [3]. Les détails de chaque base de données sont résumés dans le tableau 1.

**Mesures d'évaluation.** Les performances des approches testées ont été mesurées à l'aide de la moyenne géométrique G-Mean $= \sqrt{spécificité \times sensibilité}$, où la spécificité et la sensibilité représentent respectivement les proportions de négatifs et positifs correctement identifiés. Dans le cas multi-classes, la moyenne géométrique standard pour chaque classe est calculée séparément au moyen de la stratégie un-contre-tous. Cette dernière stratégie a également été utilisée pour adapter la mesure AUC, qui consiste à calculer l'aire sous la courbe ROC. Finalement, une analyse statistique a été effectuée à l'aide du test de rang signé de Wilcoxon [22].

**Classifieur de base.** Comme classifieur de base, nous utilisons l'arbre de décision, plus précisément CART. Toutefois, nous pouvons utiliser tout autre classifieur.

**Méthodes comparées et paramètres.** Nous comparons notre méthode aux approches de rééchantillonnage proposées spécifiquement pour les ensembles de données déséquilibrées multi-classes : sur-échantillonnage par distance de Mahalanobis (MDO) [1], Static-SMOTE (S-SMOTE) [11], la version de base de SMOTE associée à la stratégie un-contre-un (SMOTE-MC), ainsi que la méthode de base (BL) sans rééchantillonnage. Les paramètres considérés pour notre méthode MC-EVHS sont : $\alpha$ a été fixé à $1$ comme recommandé dans [18], le paramètre $t$ pour $m(\{\omega_0\})$ a été fixé à $2$ pour obtenir de bons résultats en moyenne, et nous avons testé trois valeurs différentes pour $\beta$ ($0.3$, $0.5$ et $0.7$) et avons sélectionné la plus performante pour chaque ensemble de données, puisque la quantité de chevauchement de classes diffère dans chaque cas. Pour les autres méthodes, nous avons utilisé les paramètres recommandés par leurs auteurs.

### 4.2 Résultats et discussion

Les résultats G-Mean et AUC sont présentés dans le tableau 2. Ces résultats correspondent à la moyenne de 10 exécutions après avoir appliqué la validation croisée stratifiée. Les résultats indiquent que notre approche MC-EVHS produit les meilleurs résultats, en termes de G-Mean et d'AUC. Notre proposition obtient les meilleurs résultats sur 5 des 7 ensembles de données pour G-Mean et sur 6 des 7 ensembles de données pour AUC. Nous pouvons remarquer que toutes les performances se détériorent avec l'augmentation du bruit et du chevauchement présents dans l'ensemble de données. Cependant, notre approche a obtenu de meilleures performances dans les ensembles de données où il y a beaucoup d'objets difficiles à classer. On peut affirmer que notre approche est robuste lorsque cette dernière est appliquée à des ensembles de données complexes.

Les résultats du test par paire de Wilcoxon sont présentés dans le tableau 3. $R+$ représente la somme des rangs en faveur du MC-EVHS, $R-$, la somme des rangs en faveur des méthodes utilisées pour la comparaison, et les valeurs $p$ sont calculées pour

chaque comparaison. Toutes les comparaisons par paires peuvent être considérées comme statistiquement significatives avec un niveau de $5\%$ puisque toutes les valeurs $p$ sont inférieures au seuil de $0.05$. Ainsi, nous pouvons sans risque affirmer que notre méthode est nettement plus performante que MDO, S-SMOTE et SMOTE-MC.

## 5 Conclusion

L'objectif de cet article est de développer une approche pour traiter les ensembles de données déséquilibrées dans un cadre multi-classes. Notre méthode hybride MC-EVHS exploite la théorie des fonctions de croyance pour déterminer au mieux les emplacements pour le sur-échantillonnage des classes minoritaires, et améliorer la sélection des objets à éliminer dans la phase de sous-échantillonnage. Les expériences menées ont confirmé que la méthode MC-EVHS offre de meilleurs résultats comparativement à ceux obtenus par les méthodes classiques. Ainsi, la combinaison du sous-échantillonnage et du sur-échantillonnage, associée à des appartenances évidentielles, peut réduire le problème du déséquilibre multi-classes.

En terme de perspective, nous proposons d'appliquer notre approche MC-EVHS sur des ensembles de données beaucoup plus larges et des niveaux extrêmes de déséquilibre.

## Références


[1] Abdi, L., Hashemi, S. : To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE transactions on Knowledge and Data Engineering **28**(1), 238–251 (2015)

[2] Agrawal, A., Viktor, H.L., Paquet, E. : Scut : Multi-class imbalanced data classification using smote and cluster-based undersampling. In : 7Th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3k). vol. 1, pp. 226–234. IEEE (2015)

[3] Alcala-Fdez, J., Fernàndez, A., Luengo, J., Derrac, J., Garcia, S., Sanchez, L., Herrera, F. : Keel data-mining software tool : Data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic and Soft Computing **17**, 255–287 (2010)

[4] Batista, G., Prati, R., Monard, M.C. : A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations **6**, 20–29 (2004)

[5] Bedi, P., Gupta, N., Jindal, V. : I-siamids : an improved siam-ids for handling class imbalance in network-based intrusion detection systems. Applied Intelligence **51**(2), 1133–1151 (2021)

[6] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. : Smote : Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)

[7] Dablain, D., Krawczyk, B., Chawla, N.V. : Deepsmote : Fusing deep learning and smote for imbalanced data. IEEE Transactions on Neural Networks and Learning Systems (2022)

[8] Dempster, A.P. : A generalization of bayesian inference. Journal of the Royal Statistical Society : Series B (Methodological) **30**(2), 205–232 (1968)

[9] Dietterich, T.G., Bakiri, G. : Solving multiclass learning problems via error-correcting output codes. Journal of artificial intelligence research **2**, 263–286 (1994)

[10] Douzas, G., Bacao, F., Last, F. : Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information Sciences **465**, 1–20 (2018)

[11] Fernández-Navarro, F., Hervás-Martínez, C., Gutiérrez, P.A. : A dynamic over-

TABLEAU 2 – Résultats de G-Mean et AUC en utilisant CART

| | *AUC* | | | | | *G-Mean* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Ensembles de donnée | BL | SMOTE-MC | S-SMOTE | MDO | MC-EVHS | BL | SMOTE-MC | S-SMOTE | MDO | MC-EVHS |
| dermatology | 0.917 | 0.950 | 0.933 | 0.947 | **0.955** | 0.921 | 0.937 | 0.904 | 0.931 | **0.952** |
| wine | 0.888 | 0.938 | 0.894 | 0.888 | **0.950** | 0.887 | **0.936** | 0.919 | 0.887 | 0.920 |
| thyroid | 0.983 | 0.983 | **0.989** | 0.983 | 0.985 | 0.933 | 0.919 | 0.933 | 0.933 | **0.981** |
| hayes-roth | 0.796 | 0.801 | 0.757 | 0.796 | **0.811** | 0.804 | 0.788 | 0.779 | 0.804 | **0.820** |
| contraceptive | 0.470 | 0.464 | 0.476 | 0.470 | **0.477** | 0.442 | 0.441 | 0.448 | 0.442 | **0.460** |
| pageblocks | 0.852 | 0.943 | 0.938 | 0.952 | **0.954** | 0.492 | 0.460 | **0.553** | 0.492 | 0.542 |
| yeast | 0.664 | 0.674 | **0.715** | 0.664 | **0.715** | 0.599 | 0.635 | 0.642 | 0.599 | **0.699** |

TABLEAU 3 – Comparaisons statistiques des scores AUC et G-Mean en utilisant le test de Wilcoxon

| | *G-Mean* | | | *AUC* | | |
|---|---|---|---|---|---|---|
| Comparaisons | $R+$ | $R-$ | $p$ | $R+$ | $R-$ | $p$ |
| MC-EVHS vs BL | 28.0 | 0.0 | 0.0078125 | 28.0 | 0.0 | 0.0078125 |
| MC-EVHS vs SMOTE-MC | 26.0 | 2.0 | 0.0234375 | 28 | 0.0 | 0.0078125 |
| MC-EVHS vs S-SMOTE | 26.0 | 2.0 | 0.0234375 | 19.0 | 9.0 | 0.0373677 |
| MC-EVHS vs MDO | 28.0 | 0.0 | 0.0078125 | 28.0 | 0.0 | 0.0078125 |


sampling procedure based on sensitivity for multi-class problems. Pattern Recognition **44**(8), 1821–1833 (2011)

[12] Grina, F., Elouedi, Z., Lefèvre, E. : Uncertainty-aware resampling method for imbalanced classification using evidence theory. In : European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty. pp. 342–353 (2021)

[13] Han, H., Wang, W.Y., Mao, B.H. : Borderline-SMOTE : A new over-sampling method in imbalanced data sets learning. Lecture Notes in Computer Science pp. 878–887 (2005)

[14] Hastie, T., Tibshirani, R. : Classification by pairwise coupling. Advances in neural information processing systems **10**, 507–513 (1997)

[15] Ivan, T. : Two modifications of cnn. IEEE transactions on Systems, Man and Communications, SMC **6**, 769–772 (1976)

[16] Khushi, M., Shaukat, K., Alam, T.M., Hameed, I.A., Uddin, S., Luo, S., Yang, X., Reyes, M.C. : A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access **9**, 109960–109975 (2021)

[17] Li, Z., Huang, M., Liu, G., Jiang, C. : A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection. Expert Systems with Applications **175**, 114750 (2021)

[18] Liu, Z.g., Pan, Q., Dezert, J., Mercier, G. : Credal classification rule for uncertain data based on belief functions. Pattern Recognition **47**(7), 2532–2541 (2014)

[19] Rifkin, R., Klautau, A. : In defense of one-vs-all classification. Journal of machine learning research **5**(Jan), 101–141 (2004)

[20] Shafer, G. : A mathematical theory of evidence, vol. 42. Princeton university press (1976)

[21] Smets, P. : The Transferable Belief Model for Quantified Belief Representation, pp. 267–301. Springer Netherlands, Dordrecht (1998)

[22] Wilcoxon, F. : Individual comparisons by ranking methods. In : Breakthroughs in statistics, pp. 196–202. Springer (1992)

[23] Wilson, D.L. : Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics (3), 408–421 (1972)