SMOTE: Apprenons-nous à classifier ou à prédire la nature synthétique des données ?

N. Boudegzdame[1], K. Sedki[1], R. Tsopra[2,3,4], and JB. Lamy[1]

[1]LIMICS, INSERM, Université Sorbonne Paris Nord, Sorbonne Université, France
[2]INSERM, Université de Paris Cité, Sorbonne Université, Cordeliers Research Center, France
[3]HeKA, INRIA, France
[4]Department of Medical Informatics, Hôpital Européen Georges-Pompidou, AP-HP, France

nadaboudegzdame@gmail.com, {karima.sedki, jean-baptiste.lamy}@univ-paris13.fr

Résumé

Les algorithmes de suréchantillonnage sont des prétraitements utilisés dans l'apprentissage automatique dans le cas de données fortement déséquilibrées dans le but de rééquilibrer le nombre d'instances par classe et donc d'améliorer la qualité des modèles appris. Bien que le suréchantillonnage puisse être efficace pour améliorer les performances des modèles de classification sur les classes minoritaires, il peut également introduire plusieurs problèmes. Durant notre travail, nous avons remarqué que les modèles apprennent à detecter le bruit ajouté par les algorithmes de suréchantilloange au lieu d'apprendre les informations pertinentes. Dans cet article, nous définirons le suréchantillonnage et présenterons les techniques les plus courantes, avant de proposer une méthode pour évaluer les algorithmes de suréchantillonnage.

Mots-clés

Données déséquilibrées, suréchantillonnage, SMOTE, augmentation de données, apprentissage automatique, apprentissage profond, réseaux de neurones, données synthétiques.

Abstract

Oversampling algorithms are used as preprocess in machine learning, in the case of highly imbalanced data in an attempt to balance the number of samples per class, and therefore improve the quality of models learned. While oversampling can be effective in improving the performance of classification models on minority classes, it can also introduce several problems. From our work, it came to light that the models learn to detect the noise added by the oversampling algorithms instead of the underlying patterns. In this article, we will define oversampling, and present the most common techniques, before proposing a method for evaluating oversampling algorithms.

Keywords

Imbalanced data, oversampling, SMOTE, data augmentation, machine learning, deep learning, neural networks, synthetic data.

1 Introduction

Le suréchantillonnage est une technique utilisée pour résoudre le problème du déséquilibre des classes dans l'apprentissage automatique. Le déséquilibre des classes se produit lorsque le nombre d'instances d'une classe est beaucoup plus faible que le nombre d'instances de(s) autre(s) classe(s). Ceci génère d'un problème, car le classificateur aura du mal à apprendre à partir de la classe minoritaire. Les techniques de suréchantillonnage génèrent des instances supplémentaires appartenant à la classe minoritaire afin que le classificateur ait une meilleure chance d'apprendre à partir d'eux [12, 1].

Le suréchantillonnage crée de nouvelles instances des classes minoritaires en 1) reproduisant des instances existantes ou 2) en synthétisant des instances. Parmi les techniques les plus répandues, citons le suréchantillonnage aléatoire [5], *SMOTE* [5], *ADASYN* [11], *Borderline SMOTE* [10], *SMOTEN* [5], *Safe-Level SMOTE* [2] , et Minority Oversampling Technique (*MOTE*) [13]. Ces techniques consistent à dupliquer ou à générer des instances synthétiques de la classe minoritaire afin d'augmenter sa représentation dans le jeu de données. Beaucoup de ces techniques sont des variantes ou des extensions de *SMOTE*, une technique de suréchantillonnage largement utilisée qui interpole entre les instances minoritaires existants pour générer de nouvelles instances synthétiques.

Bien que le suréchantillonnage puisse s'avérer efficace pour améliorer les performances des modèles de classification sur les classes minoritaires, il peut également introduire plusieurs problèmes. Dans cet article, nous définirons les problèmes et les défis potentiels liés à l'utilisation du suréchantillonnage. L'un des problèmes les plus importants avec les techniques de suréchantillonnage, en particulier lorsque les données sont très déséquilibrées, est que les données originales représentent une petite fraction du jeu

de données résultant pour la classe minoritaire, la grande majorité des données étant représentée par les données synthétiques. En conséquence, cela peut entraîner une redéfinition du problème d'apprentissage. En fait, après le suréchantillonnage, la classe minoritaire contiendra principalement des données synthétiques, ce qui peut amener le modèle d'apprentissage automatique à apprendre à détecter la nature synthétique des données, c'est-à-dire le bruit ajouté par le suréchantillonnage, plutôt qu'à prédire la classe minoritaire à partir des motifs correspondants. Cela peut entraîner une mauvaise généralisation du modèle, conduisant ainsi à des faibles performances sur les données du monde réel [22, 8, 6, 20].

Par conséquent, nous proposons une méthode d'évaluation des techniques de suréchantillonnage sur un jeu de données précis et nous l'appliquerons à la prédiction des hémorragie causée par les médicaments. La méthode consiste à essayer d'apprendre un nouveau modèle capable de prédire le statut synthétique de l'instance ; les performances de la technique de suréchantillonnage sont inversement proportionnelles aux performances de ce modèle. Enfin, nous mettrons à l'épreuve les techniques de suréchantillonnage les plus courantes et évaluerons leur efficacité dans un exemple de cas d'utilisation.

2 Travaux antérieurs

Plusieurs techniques de suréchantillonnage ont été mises au point pour remédier au déséquilibre des classes dans les jeux de données. Voici quelques-unes des plus populaires :

- ***Suréchantillonnage aléatoire [5] :*** Cette technique simple consiste à dupliquer aléatoirement des instances de la classe minoritaire pour augmenter sa représentation dans le jeu de données.
- ***Synthetic Minority Oversampling Technique (SMOTE) [5] :*** crée de nouvelles instances synthétiques en interpolant entre des instances minoritaires existants. Elle sélectionne deux instances similaires et crée une nouvelle instance le long de la ligne qui les relie. De nombreuses modifications et extensions ont été apportées à la méthode *SMOTE* depuis sa proposition.
- ***Adaptive Synthetic (ADASYN) [11] :*** génère des instances synthétiques basés sur la densité de la distribution des données. Elle crée plus des instances synthétiques pour la classe minoritaire qui sont plus difficiles à apprendre et s'appuie sur la méthodologie de *SMOTE*.
- ***Borderline SMOTE [10] :*** cette technique est une variante de *SMOTE* et se concentre sur des instances qui se situent à la frontière entre la classe minoritaire et la classe majoritaire. Elle génère des instances synthétiques uniquement pour les instances situées à la frontière.
- ***SMOTEN [5] :*** est une extension de la technique de suréchantillonnage *SMOTE* qui peut traiter des jeux de données avec des attributs nominaux et continus. Elle utilise une métrique de distance adaptée aux types d'attributs mixtes afin de générer des instances synthétiques pour les attributs nominaux.
- ***Safe-Level SMOTE[2] :*** cette technique combine à la fois le *SMOTE* et le suréchantillonnage aléatoire. Elle génère des instances synthétiques pour la classe minoritaire, mais ne génère pas d'instances pour ceux qui entraîneraient un chevauchement avec la classe majoritaire.
- ***Minority Oversampling Technique (MOTE) [13] :*** est une variante de *SMOTE* qui sélectionne uniquement les instances mal classées par le modèle actuel et génère des instances synthétiques uniquement pour ces instances.

Depuis l'introduction du premier *SMOTE*, il y a plus de 20 ans, en 2002, de nombreuses nouvelles techniques ont été mises en œuvre pour l'améliorer. La première amélioration, *Borderline SMOTE*, a résolu le problème de surapprentissage qui peut survenir avec *SMOTE* lorsque les instances synthétiques générées sont uniquement pour les instances de classe minoritaire situées près de la frontière de décision. *ADASYN* a été la suivante, elle a ajusté la distribution de densité de l'espace des attributs pour générer plus d'instances synthétiques pour les instances de la classe minoritaire difficiles à apprendre. *Safe-Level SMOTE* a été développée pour réduire le risque de mauvaise classification en ne générant des instances synthétiques que pour les instances de la classe minoritaire situées à proximité d'une classe majoritaire sûre. La plus récente amélioration est *SMOTEN*, qui peut gérer des jeux de données avec des attributs nominaux et continus, en utilisant une approche différente pour générer des instances synthétiques pour les attributs nominaux.

Ces techniques de suréchantillonnage ont des forces et des faiblesses différentes et peuvent donner des résultats différents en fonction du jeu de données et du problème de classification. Pour résoudre les problèmes restants, plusieurs approches ont été proposées, telles que la combinaison du suréchantillonnage et du sous-échantillonnage, l'utilisation de techniques d'échantillonnage synthétique plus avancées ou l'ajustement du seuil de classification. Cependant, chaque approche a ses propres avantages et limites, et une attention particulière est nécessaire pour sélectionner la méthode de suréchantillonnage appropriée pour un jeu de données et un problème de classification particuliers.

3 Problème du Oversampling

La technique de suréchantillonnage la plus courante [12] et la plus efficace [1] est connue sous le nom de *SMOTE*

: Synthetic Minority Over-sampling Technique. Elle fonctionne en générant des instances synthétiques de la classe minoritaire par interpolation entre les instances existantes de la classe minoritaire et leurs k plus proches voisins dans l'espace des attributs, augmentant ainsi la représentation de la classe minoritaire sans dupliquer les instances existantes. Cette technique est souvent utilisée pour remédier au déséquilibre des classes, un problème courant dans de nombreux jeux de données du monde réel où une classe est considérablement sous-représentée. Bien que le suréchantillonnage puisse être efficace pour améliorer les performances des modèles de classification sur les classes minoritaires, il peut également introduire plusieurs problèmes.

On peut classer les problèmes de *SMOTE* dans les six catégories suivantes :
1. Tout d'abord, l'un des problèmes les plus courants associés au suréchantillonnage est qu'il peut introduire un biais en faveur de la classe minoritaire [22, 8]. Lorsque le suréchantillonnage est appliqué, la classe minoritaire est artificiellement gonflée en créant de nouvelles instances synthétiques, ce qui peut amener le modèle à prédire trop fréquemment cette classe au détriment de la classe majoritaire. En conséquence, le modèle peut avoir une grande précision sur les données d'apprentissage, mais mal fonctionner sur les données du monde réel, car la classe minoritaire est beaucoup moins fréquente.

2. Le suréchantillonnage peut par ailleurs entraîner des incohérences dans les types de données, puisque les points de données synthétiques peuvent générer des valeurs qui se situent en dehors de la plage typique de la variable ou dans un format différent. Par exemple, si les données originales ne contiennent que des nombres entiers pour l'âge, le suréchantillonnage peut générer des nombres décimaux qui ne sont pas présents dans les données du monde réel.

3. Les instances synthétiques créées par suréchantillonnage sont supposés appartenir à la classe minoritaire, mais cela peut ne pas être vrai. Il peut également produire des instances mal étiquetées appartenant à la classe majoritaire, ainsi que des instances "bruits" absurdes et ne correspondant à aucune classe ou réalité, telle qu'un patient de 3 ans et pesant 100 kg.

4. La distribution des données peut également être modifiée par des instances synthétiques. Par exemple, si la classe minoritaire comprend 50% d'enfants, mais que les données synthétisées n'en comprennent que 20% alors la distribution n'est pas la même.

5. Le suréchantillonnage peut réduire la diversité du jeu de données en créant des instances synthétiques très similaires aux instances existantes. Cela peut entraîner un surapprentisssage et avoir un impact négatif sur la capacité du modèle à généraliser à de nouvelles données. Le jeu de données suréchantillonné peut ne pas refléter avec précision la véritable diversité du problème.

Il est important d'examiner attentivement l'impact du suréchantillonnage sur la distribution et la diversité du jeu de données pour s'assurer que le modèle résultant reflète avec précision la véritable nature du problème.

6. Le suréchantillonnage peut augmenter le coût d'apprentissage d'un modèle, car il nécessite de générer des points de données supplémentaires pour la classe minoritaire [6, 20]. Travailler sur de grands jeux de données, implique que la génération de données synthétiques peut prendre beaucoup de temps et de ressources.

En outre, plus un jeu de données est déséquilibré, moins le jeu de données suréchantillonné reflète avec précision la véritable nature du problème [12]. Comme expliqué ci-dessus, l'algorithme de suréchantillonnage ajustera la distribution de classe du jeu de données. Ainsi, plus un jeu de données est déséquilibré, plus de données sont nécessaires pour ajuster la distribution des classes, ce qui entraînera plus de données synthétiques dans le jeu de données suréchantillonné.

Cela peut être particulièrement difficile lors de la manipulation des jeux de données de détection d'anomalies, car elles ont tendance à avoir des distributions de classe très déséquilibrées, car la survenue d'événements ou de conditions rares est peu fréquente par rapport à la population globale. Les jeux de données médicales et de détection de fraude sont des exemples courants d'ensembles de données très déséquilibrés où la détection d'anomalies est essentielle, mais ces anomalies sont rares dans l'occurrence [4].

Les jeux de données médicales sont extrêmement difficiles à suréchantillonner, ce sont souvent les jeux de données de distribution de classe les plus déséquilibrés en raison de la nature des données médicales. Dans les données médicales, la survenue de certaines maladies ou conditions médicales peut être rare par rapport à l'ensemble de la population. Par exemple, une maladie peut n'affecter qu'un petit pourcentage de la population, tandis que la majorité de la population peut être en bonne santé. En conséquence, le jeu de données aura une distribution de classe très déséquilibrée, la classe minoritaire étant la condition médicale d'intérêt [17].

De plus, le coût et la difficulté de la collecte de données dans le domaine médical peuvent également contribuer au déséquilibre de la répartition des classes. La collecte de données médicales nécessite souvent des procédures coûteuses et chronophages, telles que des tests médicaux ou des examens d'imagerie, qui peuvent être difficiles à réaliser sur une population nombreuse et diversifiée. Par conséquent, les données collectées peuvent être biaisées en faveur de certains groupes ou données démographiques, entraînant des distributions de classes déséquilibrées.

Dans la section suivante, nous illustrerons certains problèmes rencontrés avec le suréchantillonnage à partir d'un exemple médical.

4 Méthodologie

4.1 Description de la tâche initiale d'apprentissage automatique

Notre objectif initial était de prédire le risque d'hémorragie à partir des prescriptions médicales des patients de MIMIC, une base de données publique de dossiers médicaux électroniques dépersonnalisés pour les patients admis dans des unités de soins intensifs (USI) aux États-Unis. Elle contient des données cliniques complètes sur plus de 40 000 patients en soins intensifs, y compris des données démographiques, des diagnostics, des résultats de tests de laboratoire, des informations sur les médicaments et des signes vitaux [15].

L'objectif était d'identifier les patients à risque d'hémorragie en raison de certains médicaments, doses et antécédents médicaux. Les patients labellisés comme étant à risque d'hémorragie sont ceux qui ont subi une hémorragie. Il est crucial de les identifier car certains médicaments, doses et antécédents médicaux d'un individu peuvent augmenter le risque d'hémorragie, ce qui peut mettre en danger la vie du patient. Les médicaments couramment connus pour augmenter le risque d'hémorragie comprennent les anticoagulants tels que la warfarine, le dabigatran et l'apixaban, ainsi que les agents antiplaquettaires comme l'aspirine et le clopidogrel. D'autres médicaments, tels que les anti-inflammatoires non stéroïdiens (AINS) et les inhibiteurs sélectifs de la recapture de la sérotonine (ISRS), peuvent également augmenter le risque d'hémorragie, en particulier lorsqu'ils sont pris à fortes doses ou en association avec d'autres médicaments [14].

Nous définissons le problème de classification de l'apprentissage automatique comme suit :

Prédiction du risque d'hémorragie
Entrée: *Historique des prescriptions médicales des patients, historique des admissions des patients à l'hôpital.*
Sortie: *Le patient est-il à risque d'hémorragie ou non ?*

Pour étiqueter les données, nous avons d'abord dû définir comment extraire les informations sur les hémorragies induites par les médicaments. Pour ce faire, nous avons examiné le dossier d'admission à l'hôpital du patient, qui comprend la raison de l'admission codée à l'aide du système de classification internationale des maladies (CIM). Ce système est un système de classification médicale standardisé utilisé pour coder et classer les procédures médicales, les symptômes et les diagnostics [25]. En analysant le système de classification internationale des maladies, nous avons pu définir une liste de codes CIM qui représentent les hémorragies induites par les médicaments.

En ce qui concerne les données d'entrée, elles comprennent les informations relatives à l'admission à l'hôpital d'un patient et les détails de sa prescription actuelle. Les médicaments sont codés à l'aide du National Drug Code (NDC), un code unique à 10 chiffres utilisé pour identifier les médicaments aux États-Unis. Cependant, NDC est spécifique aux États-Unis et est trop spécifique, car des codes distincts existent pour les différents dosages, formes et présentations d'un médicament [23]. Nous avons donc utilisé le système de classification *ATC* (Anatomical Therapeutic Chemical), qui organise les médicaments en fonction de leurs propriétés thérapeutiques et de leur site d'action anatomique [24]. Pour remédier à cette différence, nous avons mis en correspondance les codes NDC à leurs codes *ATC* correspondants. Certains médicaments ont plusieurs codes *ATC* ; dans ce cas, nous les avons tous considérés.

Enfin, nous avons codé les médicaments du patient en utilisant one hot encoding. Il s'agit d'un processus utilisé dans l'apprentissage automatique pour convertir des données catégorielles en une représentation numérique pouvant être utilisée par des algorithmes d'apprentissage automatique. Cela implique de créer un vecteur binaire qui a une valeur pour chaque médicament possible, la valeur étant 1 si le médicament est présent et 0 sinon. Par exemple, s'il y a trois médicaments - M_1, M_2 et M_3 - chaque médicament ou ordonnance médicale serait représenté par un vecteur binaire de longueur trois. Le médicament M_1 serait représenté par le vecteur [1,0,0], le médicament M_2 serait représenté par [0,1,0] et le médicament M_3 serait représenté par [0,0,1]. Une ordonnance associant un médicament M_1 et M_2 serait représentée par [1,1,0]. Cela permet aux algorithmes de travailler avec des données catégorielles, ce qui peut être utile dans de nombreuses applications telles que la classification de texte.

Le jeu de données résultant était très déséquilibré avec une classe minoritaire représentant seulement 3,47 % des patients présentant un risque hémorragique. Cette nature déséquilibrée du jeu de données peut poser un défi important au modèle pour prédire avec précision la classe minoritaire. En effet, le modèle peut devenir biaisé en faveur de la classe majoritaire, ce qui a entraîné de mauvaises performances lors de la prédiction de la classe minoritaire. Pour résoudre ce problème, nous avons utilisé le suréchantillonnage comme technique courante pour équilibrer le jeu de données.

4.2 Problème rencontré avec l'oversampling

Après suréchantillonnage, nous avons remarqué que les performances du modèle s'amélioraient considérablement à la fois sur l'apprentissage et la validation qui étaient suréchantillonnées, mais a donné de mauvais résultats sur les données originales en termes de métriques de performance. De plus, prédire le risque d'hémorragie est une

tâche difficile, car il se produit rarement et il est difficile de prédire si une prescription entraînera une hémorragie. Cependant, nous avons obtenu un score f1 de 90% pour prédire l'hémorragie sur les données d'apprentissage ce qui semblait trop optimiste.

Pour étudier ce problème, nous avons procédé à une analyse des prédictions du modèle afin de déterminer s'il répondait toujours le problème initial. Nous formulons l'hypothèse que le modèle apprenait à prédire si une instance était synthétique, au lieu de prédire s'il appartient à la classe minoritaire, ce qui, en effet, revient presque au même, puisqu'une grande majorité des instances appartenant à la classe minoritaire sont synthétiques.

4.3 Une méthode d'exploration de la détectabilité des données synthétiques

Pour tester notre hypothèse, nous avons défini un nouveau problème d'apprentissage automatique pour détecter des données synthétiques. Nous avons généré un nombre d'instances synthétiques égal au nombre d'instances de la classe minoritaire en utilisant le suréchantillonnage, nous avons ensuite retiré des instances de la classe majoritaire et labellisé les instances comme étant synthétiques (1) ou originaux (0). Cette approche a été appliquée à différentes méthodes de suréchantillonnage afin de déterminer la facilité avec laquelle les données synthétiques générées par ces méthodes pourraient être détectées. Les données synthétiques de moindre qualité étant plus facilement détectées, cette méthode permet d'évaluer la qualité des différentes techniques de suréchantillonnage. Notre jeu de données ainsi affiné a été utilisé pour résoudre le problème suivant :

Détection de données synthétiques
Entrée : *Classe minoritaire VS données synthétiques produites par suréchantillonnage.*
Sortie : *Instance synthétique ou originale ?*

Nous avons évalué notre méthode en utilisant les métriques suivantes :

1. **Précision, rappel et score F1 :** La précision mesure la proportion d'instances positives correctement prédites parmi toutes les instances positives prédites, tandis que le rappel mesure la proportion d'instances positives correctement prédites parmi toutes les instances positives réelles. Le score F1 est la moyenne harmonique de la précision et du rappel. Ces mesures sont particulièrement utiles lorsqu'il s'agit de données très déséquilibrées, car elles fournissent une mesure de la capacité du modèle à identifier la classe minoritaire [12, 18].

2. **Area under the precision-recall curve (AUPRC) :** AUPRC fournit un score unique qui capture le compromis entre précision et rappel pour différents seuils de décision. Il s'agit d'une mesure utile pour les données très déséquilibrées, car elle se concentre sur la classe positive et peut fournir une évaluation plus informative que la précision ou ROC AUC [7].

3. **Receiver operating characteristic (ROC) et area under the curve (AUC):** ROC trace le taux de vrais positifs par rapport au taux de faux positifs pour différents seuils de décision. L'AUC mesure l'aire sous la courbe ROC et fournit un score unique qui indique la performance globale du modèle. ROC et AUC sont utiles pour comparer des modèles qui ont des seuils de décision différents [12, 18, 9].

4. **Matrice de confusion :** Une matrice de confusion fournit une répartition détaillée des prédictions du modèle, y compris les vrais positifs, les vrais négatifs, les faux positifs et les faux négatifs. Cela peut aider à identifier le nombre d'exemples correctement et incorrectement classés par le modèle pour chaque classe.

5. **Kappa de Cohen :** Le kappa de Cohen mesure l'accord inter-juges entre le jeu de données original et le jeu de données suréchantillonné. Cela peut être utile pour évaluer dans quelle mesure les données synthétiques capturent la vraie nature du problème [19].

En utilisant plusieurs métriques et techniques pour évaluer les performances du modèle avec précision, nous pouvons acquérir une compréhension plus complète des forces et des limites du modèle. Cela peut nous aider à prendre des décisions sur la manière d'améliorer le modèle ou de l'utiliser dans des applications pratiques, car aucune mesure unique ne peut fournir une image complète de l'efficacité du modèle.

Nous avons choisi le réseau de neurones comme approche d'apprentissage en nous basant sur des études précédentes qui ont montré l'efficacité de l'apprentissage profond pour les tâches de classification déséquilibrées [26]. Pour l'implémentation actuelle, nous avons utilisé un réseau de neurones avec deux couches cachées contenant respectivement 30 et 20 neurones. Pour éviter le problème de **"neurones morts"**, nous avons opté pour la fonction d'activation ***LeakyReLU***, qui s'est avérée performante dans des applications similaires [16]. La couche de sortie a été conçue avec une fonction d'activation ***sigmoid***, couramment utilisée dans les problèmes de classification binaire.

Pour garantir l'efficacité du modèle d'apprentissage, nous avons utilisé une technique de réduction du taux d'apprentissage appelée ***ReduceLROnPlateau***. Cette technique nous a permis d'ajuster dynamiquement le taux d'apprentissage de l'optimiseur pendant la phase apprentissage, en fonction d'une métrique surveillée telle que validation loss. Ce faisant, nous avons pu aider le modèle à sortir des plateaux et à continuer de s'améliorer, même à l'approche de la convergence. Notre modèle a été

entrainé sur 100 époques, ce qui était suffisant pour assurer un apprentissage complet et la convergence du modèle.

4.4 Résultats et analyse

Les résultats du tableau 1 indiquent que, pour les quatre techniques de suréchantillonnage, le réseau de neurones a obtenu de bons résultats en termes de métriques d'évaluation et a donc été en mesure de prédire des données synthétiques avec un degré élevé de précision. Parmi toutes les techniques de suréchantillonnage, *SMOTEN* a obtenu le score le plus élevé en termes de score f1, de rappel, de précision, d'accuracy, de cohen kappa et d'AUC. L'algorithme Borderline *SMOTE* conduit également à des scores élevés dans toutes les mesures d'évaluation, à l'exception de l'AUC.

Par conséquent, nous pouvons facilement prédire si une instance est synthétique ou non. Cette prédiction est beaucoup plus facile que celle du risque d'hémorragie. Cela confirme donc notre hypothèse : le modèle initial prédisait en fait la nature synthétique des données au lieu du risque d'hémorragie.

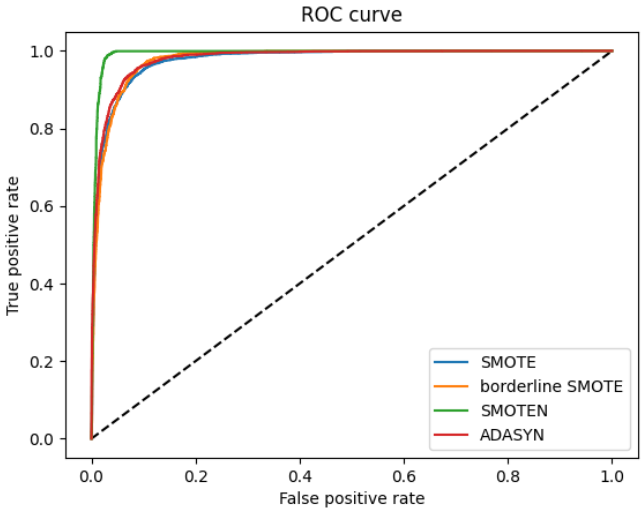

Figure 1: Courbe ROC pour la détection de la nature synthétique des données.

Comme expliqué ci-dessus, la courbe ROC et la courbe Precision-Recall fournissent des informations importantes sur les performances d'un modèle de classification binaire. Par conséquent, nous avons tracé les deux courbes pour obtenir une évaluation plus complète des performances du modèle. La figure 1 résume la courbe ROC pour les quatre algorithmes de suréchantillonnage. Elle indique que le modèle est très précis dans la distinction entre les instances positives et négatives. En fait, une AUC de 0,5 suggère une classification aléatoire, tandis qu'une AUC de 1 suggère une classification parfaite. Les valeurs AUC pour *SMOTE*, borderlineSMOTE et *ADASYN* sont

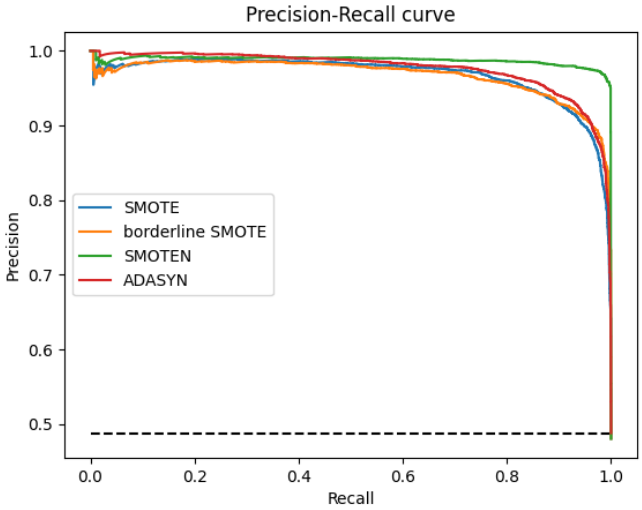

Figure 2: Courbe de précision et de rappel pour la détection de la nature synthétique des données.

de 0,97, ce qui indique que les performances du modèle sont très proches de la perfection, avec seulement un petit nombre de faux positifs et de faux négatifs. De plus, nous avons observé que les données suréchantillonnées générées par *SMOTEN* sur nos données étaient les plus faciles à détecter, comme le confirme la figure 2, qui résume la courbe Rappel-Précision.

Par conséquent, le tableau 1 et les figures 1 et 2 suggèrent que les techniques de suréchantillonnage peuvent être facilement détectées dans une large mesure. Cependant, le choix de l'algorithme de suréchantillonnage doit dépendre des caractéristiques spécifiques du jeu de données et des mesures d'évaluation qui nous intéressent.

Bien qu'il soit connu que l'algorithme de suréchantillonnage ne se comporte pas de la même manière sur différents jeux de données, les résultats des tests sur les données médicales, comprenant des prescriptions de médicaments, et constituant un jeu de données fortement déséquilibré, indique fortement que le suréchantillonnage ne sera pas une technique considérable pour équilibrer nos données. Une analyse et une expérimentation plus approfondies peuvent être nécessaires pour déterminer l'approche la plus efficace pour équilibrer le jeu de données des prescriptions médicales en question.

4.5 Comprendre pourquoi les données synthétiques sont facilement détectées

Les prescriptions de médicaments suréchantillonnées peuvent ne pas refléter fidèlement les données du monde réel, car elles sont facilement détectables par les algorithmes d'apprentissage automatique. Pour mieux comprendre cette problématique, nous avons formulé les hypothèses suivantes :

	F1 Score	Recall	Precision	Accuracy	Cohen Kappa	AUC
SMOTE	0.92	0.94	0.90	0.92	0.84	0.97
Borderline *SMOTE*	0.93	0.96	0.91	0.93	0.86	0.97
SMOTEN	0.97	0.99	0.96	0.97	0.95	0.99
ADASYN	0.92	0.93	0.91	0.92	0.84	0.97

Table 1: Évaluation comparative des métriques de performance de divers algorithmes d'oversampling pour la classification de données synthétiques.

Hypothèse 1 : Sur ou sous-représentation des drogues. (Problème #4 dans la section 3)
Les prescriptions médicales contiennent généralement un nombre limité de médicaments. Cependant, les données synthétiques générées peuvent contenir un nombre plus petit ou plus grand de médicaments, ce qui entraîne une sous-représentation ou une surreprésentation des médicaments, ce qui pourrait entraîner des écarts entre les données synthétiques et réelles.

Le tableau suivant 2 montre que les quatre méthodes de suréchantillonnage (*SMOTE*, *Borderline SMOTE* , *SMOTEN* et *ADASYN*) ont entraîné une diminution du nombre moyen de codes *ATC* pour les médicaments dans les données suréchantillonnées par rapport aux données originales. Cela indique une sous-représentation des médicaments dans les instances suréchantillonnés.

	Nombre Moyen de Codes *ATC*
Données Originales	34.78
SMOTE	20.11
Borderline SMOTE	21.13
SMOTEN	20.76
ADASYN	19.49

Table 2: Distribution des médicaments dans les données originales et synthétiques.

Hypothèse 2 : Changement de la nature des données. (Problème #2)
SMOTE peut introduire de petites perturbations dans les valeurs des attributs afin de créer des instances synthétiques, ce qui peut entraîner des valeurs non entières ou à virgule flottante pour les attributs discrètes [21]. Par exemple, les médicaments sont représentés par des valeurs discrètes de 0 ou 1, indiquant la présence ou l'absence du médicament dans une ordonnance. Cependant, les données synthétiques générées à des fins d'analyse peuvent contenir des médicaments avec des valeurs continues, ce qui peut entraîner des inexactitudes dans les résultats.

Après une enquête plus approfondie, nous avons constaté que l'application de *SMOTE*, Borderline *SMOTE*, *SMOTEN* et *ADASYN* n'a entraîné aucun changement significatif dans la nature des données suréchantillonnées. Les quatre méthodes de suréchantillonnage appliquées à nos données n'ont pas modifié la nature des données.

Hypothèse 3 : Incohérences dans les codes* ATC *. (Problème #3)
Certains médicaments comme l'aspirine ont plusieurs codes *ATC* , et nous les avons associés à tous leurs codes correspondants dans les données originales. Toutefois, dans les instances synthétiques, un tel médicament peut être associé à un seul de ses codes. Par exemple, une prescription d'aspirine pourrait être codée comme un inhibiteur de l'agrégation plaquettaire mais pas comme un analgésique dans les instances synthétiques.

Hypothèse 4 : Incohérences dans les associations médicamenteuses. (Problème #3)
Les ordonnances synthétiques générées peuvent inclure des associations médicamenteuses incohérentes. Par exemple, des médicaments comme le ramipril et l'énalapril, qui sont tous deux des inhibiteurs de l'enzyme de conversion de l'angiotensine et qui ont les mêmes effets, ne sont donc jamais associés. Cependant, de telles incohérences peuvent se produire dans les instances synthétiques.

Nous sommes encore en train d'expérimenter et d'essayer de valider ces hypothèses.

5 Discussion

Dans cet article, nous avons décrit un problème que nous avons rencontré lors de l'utilisation du suréchantillonnage : le modèle d'apprentissage automatique apprenait à détecter la nature synthétique des données suréchantillonnées plutôt que les informations pertinentes initialement dans les données originales. Nous avons proposé une méthode pour identifier ce problème et évaluer les méthodes de suréchantillonnage, consistant à essayer d'apprendre dans quelle mesure des données synthétiques peuvent être détectées.

Dans la littérature, de nombreuses études ont exploré les problèmes associés au suréchantillonnage et au *SMOTE*, cependant, à notre connaissance, aucune d'entre elles n'a mentionné l'apprentissage de la nature synthétique des données ni proposé une méthode pour la quantifier.

Les travaux de Tarawneh et al. [22] sont un article de synthèse complet sur le problème de la résolution du déséquilibre des classes dans l'apprentissage automatique et met en évidence la technique de suréchantillonnage

couramment utilisée pour y remédier. Les auteurs affirment que le suréchantillonnage peut entraîner un surapprentissage, une augmentation des coûts de calcul et une réduction des performances de généralisation.

L'article souligne également que le suréchantillonnage peut augmenter le risque de biais du modèle et peut entraîner une diminution des performances de généralisation, en particulier lorsque les données suréchantillonnées sont utilisées pour les tests. En outre, le suréchantillonnage peut augmenter le coût de calcul des modèles de formation, particulièrement dans le cas des grands jeux de données, car il nécessite de générer et de stocker un grand nombre d'instances synthétiques. Les auteurs discutent des limites de l'approche de suréchantillonnage et suggèrent des méthodes alternatives, telles que l'apprentissage sensible au coût et la détection d'anomalies, qui peuvent fournir des solutions plus efficaces au problème de déséquilibre des classes.

Les travaux de R. Buda et al. [3] étudient l'impact du déséquilibre des classes sur les performances des CNN pour les tâches de classification d'images et évalue l'efficacité de différentes stratégies, y compris le suréchantillonnage. Cependant, l'article suggère que le suréchantillonnage seul peut ne pas être suffisant pour résoudre le déséquilibre de classe dans les CNN. En effet, le suréchantillonnage peut entraîner un surapprentissage dans les CNN, où le modèle mémorise les données d'apprentissage et fonctionne mal sur les nouvelles données. De plus, le suréchantillonnage peut créer des instances irréalistes et redondants, entraînant une utilisation inefficace des ressources de calcul.

Plusieurs études ont proposé des modifications à la technique de suréchantillonnage pour résoudre ces problèmes. Rodríguez-Torres et al. [20] ont proposé le suréchantillonnage aléatoire à grande échelle (LRO) pour résoudre les problèmes de déséquilibre des classes sur les grands jeux de données.

Dans cette étude, les performances de LRO ont été comparées à celles de plusieurs autres méthodes de suréchantillonnage, telles que *SMOTE* et *Borderline-SMOTE*. Les résultats ont montré que LRO atteignait une précision et un score F1 plus élevés que les autres méthodes, et était également plus efficace en termes de calcul. Les auteurs ont identifié certaines limites et difficultés de la méthode *SMOTE*, telles que son incapacité à générer des instances diversifiées et sa sensibilité au bruit. Dans l'ensemble, les auteurs suggèrent que LRO pourrait fournir une solution plus efficace et plus évolutive pour les problèmes de déséquilibre des classes sur de grands jeux de données.

En résumé, la littérature met en évidence les limites et les défis potentiels du suréchantillonnage et du *SMOTE* pour traiter les données déséquilibrées dans l'apprentissage automatique, et suggère des approches alternatives et des modifications pour résoudre ces problèmes. Les articles présentés couvrent divers aspects du suréchantillonnage et des problèmes *SMOTE*, notamment le surapprentissage, l'évaluation des performances, la gestion de grands jeux de données, le déséquilibre multi-classes, la gestion du bruit et le suréchantillonnage synthétique.

6 Conclusion et perspectives

Le suréchantillonnage, en conclusion, peut être un outil précieux pour améliorer les performances des modèles d'apprentissage automatique sur des jeux de données déséquilibrés. Cependant, nos résultats suggèrent que les algorithmes de suréchantillonnage peuvent introduire un certain nombre de problèmes. La mauvaise qualité des données synthétiques dans la classe minoritaire peut amener le modèle d'apprentissage automatique à apprendre à prédire les données synthétiques plutôt que les informations pertinentes, ce qui entraîne de mauvaises performances sur les données du monde réel.

Par conséquent, il est essentiel d'analyser en profondeur les méthodes de suréchantillonnage pour s'assurer que de tels problèmes sont évités. Nous avons proposé une méthode d'évaluation des algorithmes de suréchantillonnage qui tient compte à la fois de leur efficacité et de leur potentiel à introduire du bruit détectable. En évaluant la capacité du modèle à distinguer les données synthétiques des données réelles, nous pouvons identifier les techniques de suréchantillonnage qui introduisent trop de bruit et ne sont pas efficaces. Cette approche permet aux chercheurs de sélectionner les meilleures techniques de suréchantillonnage pour leurs jeux de données spécifiques et d'améliorer la précision et la généralisation de leurs modèles. En plus, cela peut aider à déterminer si le suréchantillonnage est une option viable pour équilibrer le jeu de données.

Pour les recherches futures, notre objectif principal est de développer un algorithme de suréchantillonnage spécialement conçu pour répondre aux particularités uniques des données médicamenteuses. En outre, nous souhaitons étudier les avantages potentiels de différentes techniques existantes, telles que l'apprentissage par transfert, l'apprentissage par ensemble, d'augmentation des données et l'apprentissage sensible aux coûts, afin d'améliorer les performances des modèles d'apprentissage automatique sur des jeux de données déséquilibrés. Ainsi, une direction future intéressante de ce travail serait (1) d'essayer différentes approches individuellement et de comparer leurs performances ; et (2) d'essayer de combiner les différentes approches ensemble.

Remerciements

Ce travail a été financé par l'agence nationale de la recherche (ANR) via le projet ABiMed [grant number ANR-20-CE19-0017-02].

References

[1] Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–30.

[2] Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 475-482, Springer.

[3] Buda, R., Maki, A., Mazurowski, M. A. (2018). A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks. *Neural Networks*, 106, 249-259.

[4] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3), 15.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1), 321–357.

[6] Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, 110, 24–31.

[7] Davis, J. and Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning*, 233-240.

[8] Drummond, C., & Holte, R. (2003). C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets*

[9] Fawcett, T. (2006). An Introduction to ROC Analysis. In *Pattern Recognition Letters*, 27(8), 861-874.

[10] Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, 878-887, Springer.

[11] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *IEEE International Joint Conference on Neural Networks*, 1322–1328.

[12] He, H. and Garcia, E.A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.

[13] Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3), 489–501.

[14] Johnathan W. Hamrick, Diane Nykamp (2015). Drug-Induced Bleeding. *US Pharmacist*, 40(12), HS17-HS21.

[15] Johnson, A., Bulgarelli, L., Pollard, T., Celi, L. A., Mark, R., & Horng, S. (2021). MIMIC-IV (version 1.0). *PhysioNet*.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[17] Longadge, Rushi and Dongre, Snehalata (2013). Class imbalance problem in data mining: Review. In *International Journal of Computer Science and Network*

[18] Powers, D.M.W. (2011). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation. In *Journal of Machine Learning Technologies*, 2(1), 37-63.

[19] McHugh, M.L. (2012). Interrater Reliability: The Kappa Statistic. In *Biochemia Medica*, 22(3), 276-282.

[20] Rodríguez-Torres F, Martínez-Trinidad JF, Carrasco-Ochoa JA (2022). An Oversampling Method for Class Imbalance Problems on Large Datasets. *Applied Sciences*, 12(7), 3424.

[21] Rok Blagus and Lara Lusa (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14,106 .

[22] Tarawneh, S., Al-Betar, M. A., & Mirjalili, S. (2022). Stop oversampling for class imbalance learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2), 340-354.

[23] U.S. Food and Drug Administration. (n.d.). National Drug Code (NDC) Directory.

[24] WHO Collaborating Centre for Drug Statistics Methodology. (2013). Guidelines for ATC classification and DDD assignment 2013. Oslo, Norway: WHO Collaborating Centre for Drug Statistics Methodology.

[25] World Health Organization. (2016). International classification of diseases, 11th revision (ICD-11). Geneva: World Health Organization.

[26] Xu, Youjun & al. Deep Learning for Drug-Induced Liver Injury. *Journal of chemical information and modeling* vol. 55,10 (2015): 2085-93.