Apprentissage de classes déséquilibrées

HAX907X - Apprentissage statistique

EL SAWADOGO Kader
GERMAIN Marine
LABOURAIL Célia
MARIAC Damien

4 octobre 2025

Résumé

Ce document présente le rapport du projet réalisé dans le cadre du cours d'apprentissage statistique. Il détaille les objectifs, la méthodologie, les résultats obtenus et les conclusions tirées.

Table des matières

1 Contexte et Objectifs **2**

2 Méthodologie Damien **3**

3 Forces et Faiblesses Célia **4**

4 Exemples d'applications Kader **5**

5 BRAIN STORM —————— **6**

6 Les différentes méthodes **6**
6.1 Méthodes au niveau des données 6
6.1.1 Sur-échantillonnage (Over-sampling) 6
6.1.2 Sous-échantillonnage (Under-sampling) 6
6.1.3 SMOTE (Synthetic Minority Over-sampling Technique) 6
6.2 Méthodes au niveau de l'algorithme 7

1 Contexte et Objectifs

Dans le cadre de ce travail, nous nous intéressons à l'apprentissage supervisé dans des contextes où les classes sont déséquilibrées. Ce problème est fréquent dans de nombreux domaines, tels que la détection de fraudes, la médecine ou le diagnostic industriel. Un déséquilibre important des classes peut conduire à des modèles biaisés : ils tendent à prédire correctement les classes majoritaires en négligeant les classes minoritaires.

La problématique scientifique que nous étudions est donc la suivante : "Comment atténuer le déséquilibre de classes pour améliorer la performance des modèles ?"

L'objectif de ce travail est d'étudier des méthodes permettant de résoudre ce déséquilibre. Plus précisément, nous nous concentrons sur trois approches issues d'articles scientifiques :

1. Le **Random Over-Sampling** (ROS) cité dans l'article *Survey on Deep Learning with Class Imbalance* [2], une méthode de rééchantillonage qui consiste à augmenter artificiellement la proportion de la classe minoritaire.
2. Le **Random Under-Sampling** (RUS), une autre méthode de rééchantillonage, cité également dans l'article [2] qui consiste à réduire la proportion de la classe majoritaire
3. La **Synthetic Minority Over-sampling Technique** (SMOTE) présenté dans l'article [1] qui génère de nouvelles instances synthétiques pour la classe minoritaire à partir des observations existantes

2 Méthodologie Damien

3 Forces et Faiblesses Célia

4 Exemples d'applications Kader

5 BRAIN STORM ——————

6 Les différentes méthodes

6.1 Méthodes au niveau des données

Les méthodes data-level agissent directement sur les données d'entraînement pour atténuer le déséquilibre entre classes. Le principe est soit d'augmenter le poids de la classe minoritaire en lui fournissant plus d'exemples (réels ou synthétiques), soit au contraire de réduire le poids de la classe majoritaire en éliminant certains de ses exemples. L'objectif est d'obtenir une distribution de classes plus équilibrée, ce qui force l'algorithme d'apprentissage à prêter autant d'attention à la minorité qu'à la majorité. Ces techniques peuvent toutefois introduire de la variance (sur-apprentissage) ou du biais supplémentaire, il faut donc les appliquer judicieusement.

6.1.1 Sur-échantillonnage (Over-sampling)

Le sur-échantillonnage (oversampling) consiste à ajouter des copies ou des variantes des exemples de la classe minoritaire jusqu'à augmenter sa fréquence dans le jeu de données. Dans sa forme la plus simple, le sur-échantillonnage aléatoire (Random Over-Sampling, ROS) duplique aléatoirement des instances minoritaires existantes jusqu'à atteindre un équilibre désiré. Par exemple, si l'on dispose de 100 exemples minoritaires et 1000 majoritaires, le ROS peut répliquer les minoritaires (éventuellement plusieurs fois chacun) jusqu'à en obtenir 1000, rétablissant ainsi un ratio équilibré. On échantillonne avec remise parmi les indices de la classe minoritaire pour générer de nouvelles instances d'entraînement.

6.1.2 Sous-échantillonnage (Under-sampling)

À l'inverse, le sous-échantillonnage (undersampling) vise à réduire la proportion de la classe majoritaire en retirant certains de ses exemples du jeu de données. Le sous-échantillonnage aléatoire (Random Under-Sampling, RUS) élimine au hasard des instances de la classe majoritaire jusqu'à atteindre un ratio plus équilibré avec la minorité.le RUS est utile lorsque l'on dispose d'une grande abondance de données majoritaires, potentiellement redondantes. Plutôt que de tout utiliser, ce qui peut être coûteux et inutile, on peut se permettre d'en élaguer une partie. En réduisant drastiquement le nombre d'exemples majoritaires, on élimine le biais numérique et on accélère l'entraînement (moins de données à parcourir).

6.1.3 SMOTE (Synthetic Minority Over-sampling Technique)

Plutôt que de copier des instances existantes, SMOTE crée de nouvelles instances minoritaires artificielles en interpolant entre des exemples réels.Concrètement, pour chaque exemple minoritaire original, SMOTE sélectionne aléatoirement l'un de ses k plus proches voisins (minoritaire également), puis génère un nouvel exemple situé aléatoirement le long du segment joignant les deux points dans l'espace des *features*. En répétant ce procédé, on peut synthétiser autant d'exemples minoritaires que souhaité.

6.2 Méthodes au niveau de l'algorithme

???????

Annexe

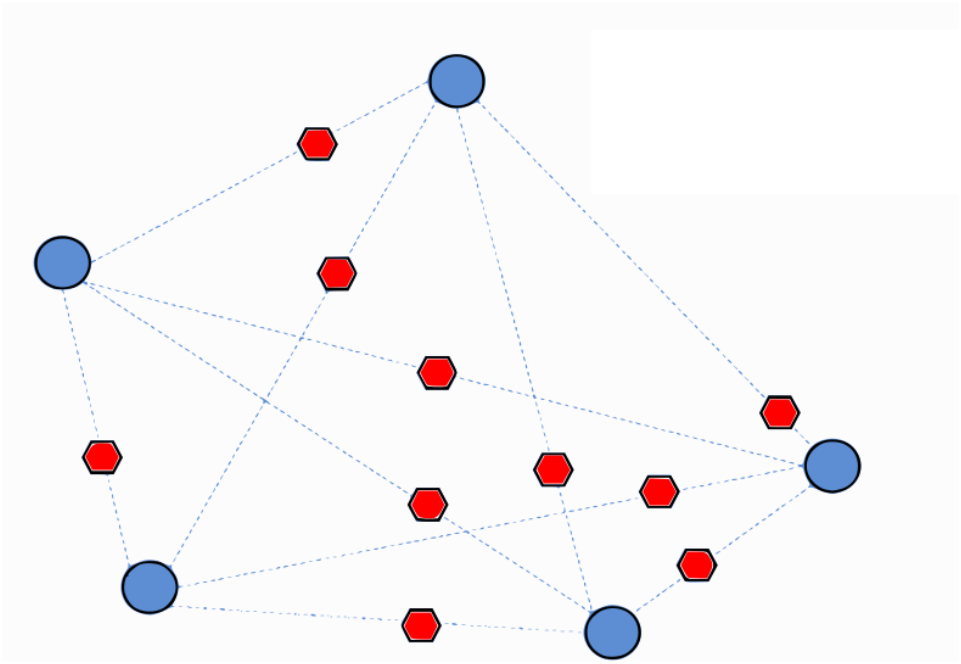

FIGURE 1 – Smote illustration

Les points bleu correspondent à la classe minoritaire dont on souhaite générer de nouveaux exemples, les points rouge sont les points générés par SMOTE.

Références

[1] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In : *Journal of artificial intelligence research* 16 (2002), p. 321-357.

[2] Justin M Johnson et Taghi M Khoshgoftaar. "Survey on deep learning with class imbalance". In : *Journal of big data* 6.1 (2019), p. 1-54.