Qualification du biais de données dans le processus de la science des données

Ginel Dorleon, Nathalie Bricon-Souf, Imen Megdiche, Olivier Teste

Qualification du biais de données dans le processus de la science des données

[1]Gine Dorleon*, [1]Imen Megdiche**
[1]Nathalie Bricon-Souf***, [1]Olivier Teste****

[1]Institut de Recherche en Informatique de Toulouse, Toulouse, France
*ginel.dorleon@irit.fr, **imen.megdiche@irit.fr,
***nathalie.souf@irit.fr, ****olivier.teste@irit.fr

Dans le contexte de l'apprentissage machine, les données constituent la principale ressource pour guider les prises de décisions. Cependant, lorsque des biais existent dans les données, cela affecte de façon significative l'interprétation des décisions. Par biais, nous entendons toute erreur survenue dans la collecte, le traitement ou l'utilisation des données due à des méthodes ou à des algorithmes et qui peut conduire à des préjugés. Dans ce travail, notre étude est basée sur la définition et la qualification du biais dans les données. À l'aide d'un processus d'apprentissage, Fig.1, nous avons identifié les endroits sur le processus où les biais peuvent survenir. Nous avons défini ces biais en appuyant sur des travaux existants puis nous avons qualifié ces biais en utilisant le processus décrit à la Fig.1

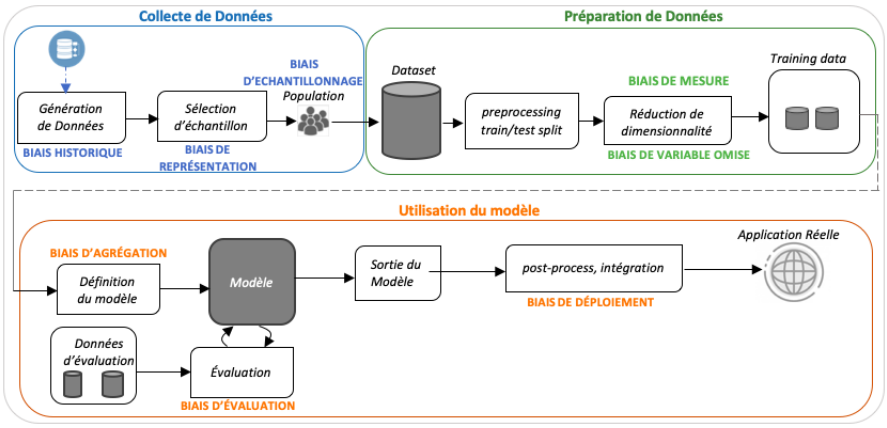

FIG. 1 – *Schéma représentant les trois étapes du processus*

Nous avons représenté ce processus en trois étapes afin de mieux illustrer les biais. Chaque étape est dédiée à un objectif spécifique :

1. Collecte de Données : Cette étape contient les actions consacrées à recueillir des informations du monde réel tout en extrayant un échantillon d'une population. Trois types

de biais sont identifiés à cette étape. Le biais historique qui concerne les problèmes socio-techniques qui existent déjà dans l'environnement où les données sont collectées. Le biais d'échantillonage qui est dû à un échantillonnage où les individus n'ont pas la même probabilité d'être sélectionnés. Le biais de représentation qui provient des critères utilisés pour échantillonner la population.

2. Préparation de Données : Cette étape contient les actions qui aident à mesurer et à sélectionner les caractéristiques pour construire un ensemble de données d'apprentissage. Deux types de biais sont identifiés à cette étape. Le biais de mesure qui provient de la façon dont nous choisissons et mesurons une caractéristique particulière. Le biais de la variable omise qui se produit lorsqu'une ou plusieurs caractéristiques importantes sont omises pendant la phase d'apprentissage du modèle.
3. Utilisation du Modèle : Cette étape regroupe les actions liées à l'entraînement, à l'évaluation et au déploiement du modèle. Trois types de biais sont identifiés. Le biais d'agrégation qui apparaît lorsque le modèle est utilisé sur des sous-groupes avec des distributions conditionnelles différentes. Le biais d'évaluation qui se produit lorsque les données d'évaluation utilisées pour évaluer le modèle ne représentent pas la population cible initiale. Enfin le biais de déploiement qui surgis s'il y a un décalage entre le problème que le modèle est conçu pour résoudre et la façon dont il est réellement utilisé.

Nous avons ensuite souligné les défis liés aux différentes étapes du processus d'apprentissage dans la science des donnés. Nous avons considéré ces défis sous 3 catégories : Imbalanced Data, Feature Selection, Model Deployment. Dans le tableau 1 ci-dessous, nous résumons ces défis et les biais respectifs susceptibles de découler de ces défis.

Défis / Biais	Histor.	Échantil.	Représ.	Mésur.	Var. Omise	Agréga.	Éval.	Déploi.
Imbalanced Data	✓	✓	✓					
Feature Select.				✓	✓			
Model Deploy.						✓	✓	✓

TAB. 1 – *Défis et biais associés*

Bien que nous ne proposions pas de nouvelles méthodes pour atténuer ces biais, nous avons cependant attiré l'attention, dans le tableau 2, sur des solutions antérieures qui cependant doivent être utilisées dans leur contexte respectif. Notre objectif est que cette revue soit un guide utile pour sensibiliser les lecteurs et motiver la recherche sur la notion de biais.

Étape:	**Collecte de données**		**Préparation de données**		**Déploiement du modèle**	
Type de biais	Biais d'échantillonnage	Biais de représentation	Biais de mesure	Biais variable omise	Biais d'agrégation	Biais d'évaluation
Solution exist ante	*Ressampling*	*Reweighing*	*Ensemble feature selection*	*Double cross-validation*	*Multitask learning*	*Subgroup evaluation*

TAB. 2 – *Solution existante relative (3e ligne du tableau) à chaque type de biais.*