Voir le processus de création d'exemples contrefactuels comme une source de connaissance - Application au classifieur Naïf de Bayes

Vincent Lemaire*, Nathan Le Boudec*,**, Françoise Fessant*, Victor Guyomard*

* Orange Innovation, Lannion, France
** Université de technologie de Compiègne, France

Résumé. Il existe aujourd'hui de nombreux algorithmes de compréhension des décisions d'un algorithme d'apprentissage automatique. Parmi ceux-ci, on trouve ceux basés sur la génération d'exemples contrefactuels. Cet article propose de considérer ce processus de génération comme une source de connaissance qui peut être stockée puis utilisée de différentes manières. Ce processus est illustré dans le cas des modèles additifs et en particulier dans le cas du classifieur naïf de Bayes, dont on montre des propriétés intéressantes pour ce faire.

1 Introduction

L'apprentissage automatique, l'une des branches de l'intelligence artificielle, a connu de nombreux succès ces dernières années. Les décisions prises par ces modèles sont de plus en plus précises, mais aussi de plus en plus complexes. Il apparaît néanmoins que certains de ces modèles sont apparentés à des boîtes noires : leurs décisions sont difficiles, voire impossibles, à expliquer (Bodria et al., 2023). Ce manque d'explicabilité peut entraîner un certain nombre de conséquences indésirables : manque de confiance de l'utilisateur, réduction de l'utilisabilité des modèles, présence de biais, etc. C'est à partir de ces besoins qu'est né le domaine de la XAI (eXplainable AI). Le XAI (Saeed et Omlin, 2023; Allen et al., 2023) est une branche de l'intelligence artificielle qui vise à rendre les décisions prises par les modèles d'apprentissage automatique intelligibles pour les utilisateurs.

Parmi les méthodes XAI, le raisonnement contrefactuel est un concept issu de la psychologie et des sciences sociales (Miller, 2019). Il consiste à examiner les alternatives possibles aux événements passés (Stepin et al., 2021). Les humains utilisent souvent le raisonnement contrefactuel en imaginant ce qui se passerait si un événement ne s'était pas produit, et c'est ce qu'est exactement le raisonnement contrefactuel. Appliquée à l'intelligence artificielle, la question est, par exemple, "Pourquoi le modèle a-t-il pris cette décision plutôt qu'une autre ?" ou "En quoi la décision aurait-elle été différente si une certaine condition avait été modifiée ?

Dans le cadre du raisonnement contrefactuel, cet article propose de considérer ce processus de génération comme une source de connaissances qui peut être stockée puis exploitée de différentes manières. Ce processus est illustré dans le cas des modèles additifs et en particulier dans le cas du classificateur de Bayes naïf, dont on montrera des propriétés intéressantes pour ce faire.

Voir le processus de création d'exemples contrefactuels comme une source de connaissance

Le reste de cet article est organisé comme suit : la section 2 présente les concepts clés utilisés dans le reste de l'article, de sorte qu'il puisse être lu indépendamment. La section 3 présente la première contribution de cet article en montrant qu'il est possible de trouver des "trajectoires additives" de contrefactuels dans le cas du classificateur de Bayes naïf. La section 4 présente la deuxième contribution de cet article en détaillant comment mettre en base ces trajectoires, cette connaissance, et comment l'exploiter. Enfin, avant de conclure, la section 5 illustre, à l'aide d'un problème de désabonnement, comment un clustering, appliqué sur cette base de données, génère de nouvelles connaissances.

Note : Dans la suite de cet article on s'intéressera aux problèmes de classification supervisée où un modèle prédictif f est entraîné à l'aide d'une base de N exemples, chacun décrit par un ensemble de d variables explicatives (un vecteur $X = \{X_1, ..., X_d\}$, issu d'une distribution $\mathcal{X}$) de manière à prédire une variable cible catégorielle notée $Y = \{y_1, ..., y_C\}$ issue d'une distribution $\mathcal{Y}$.

2 Concepts

2.1 Exemple countrefactuel et semi-factuel

En apprentissage automatique, une explication contrefactuelle vise à expliquer pourquoi un résultat particulier a été obtenu en suggérant des modifications hypothétiques des caractéristiques d'entrée, X, qui auraient pu conduire à une prédiction différente (Lemaire et al., 2010; Wachter et al., 2018). En d'autres termes, elle identifie les facteurs qui auraient pu influencer un résultat particulier. Le raisonnement contrefactuel peut être défini (de manière informelle) comme suit.

On pose $f : \mathcal{X} \mapsto \mathcal{Y}$ un modèle d'apprentissage automatique tel que, pour un individu donné X, $f(X) = \hat{y}_i$. Dans ce cas, un exemple contrefactuel est un nouvel exemple X' tel que $f(X') \neq \hat{y}_i$ et $X \neq X'$. Si X avait pour certaines de ses variables explicatives des valeurs $(X'_1, X'_2, ...)$ différentes tandis que toutes les autres variables restaient à l'identique, la classe $\hat{y}_j \neq \hat{y}_i$ aurait été retournée.

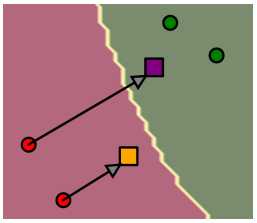

FIG. 1 – *Illustration d'un contrefactuel et d'un semi-factuel. Les point rouge représentent des exemples initiaux (X). Le orange représente un semi-factuel, le point violet représente un contrefactuel et la ligne blanche représente la limite de décision entre la classe rouge et la classe verte.*

Ci-dessus, $\hat{y}_i$ est connue et factuelle, tandis que $\hat{y}_j$ est le résultat non attendu, qui ne s'est pas produit, contrefactuel. On note néanmoins qu'une modification de X en X' n'entraîne pas nécessairement un changement dans la prédiction de classe : c'est ce que l'on appelle un

exemple semi-factuel (Fernández et al., 2022). La connaissance des exemples contrefactuels ou semi-factuels permet d'expliquer comment changer, modifier, les décisions du modèle : "*Votre prêt bancaire n'a pas été accepté MAIS SI vous aviez eu plus d'ancienneté dans notre compagnie la décision aurait été inverse (ou plus proche de l'acceptation)*." Ces deux notions sont illustrées dans la figure 1. La compréhension produite par un une méthode d'explication dite "contrefactuelle" est locale car elle s'applique à un individu en particulier et basée sur l'exemple ("instance based") puisqu'elle est produite sous la forme d'un nouvel exemple.

2.2 Informativité et actionnabilité des variables

Dans le cadre de la prise de décision, et en particulier dans le contexte du raisonnement contrefactuel, l'identification des variables informatives et actionnables est essentielle. On définit une variable informative comme étant est une variable qui a un impact significatif sur la valeur de la sortie du modèle prédictif. Toutefois, il ne suffit pas de savoir quelles sont les variables informatives. Il est également important d'identifier les variables actionnables que l'on définit comme étant une variable sur laquelle il est possible d'agir. Le type de variable le plus précieux est la variable actionnable informative, c'est-à-dire qui non seulement a un impact significatif sur la variable de sortie mais aussi sur laquelle il est possible d'agir pour améliorer, influencer, le résultat.

2.3 Notion de trajectoires

La littérature scientifique propose de nombreuses méthodes qui permettent de générer des exemples contrefactuels soit modèle dépendant (selon le type de classifieur), soit agnostique au modèle prédictif Stepin et al. (2021); Brughmans et al. (2023). Dans cet article on s'intéresse aux méthodes qui induisent la notion de trajectoire. Si on reprend la Figure 1 on s'intéresse aux méthodes qui permettent de s'approcher pas à pas de la frontière de décision (jusqu'à la franchir) par modifications successives de l'exemple initial. On appelle dans ce cas 'trajectoire' la succession de X' résultante de ces opérations successives (voir Figure 2).

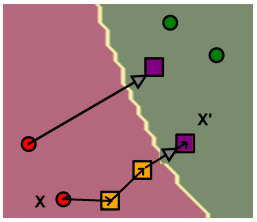

FIG. 2 – *Illustration de deux contrefactuels : l'un atteint en un 1 pas, le second en 3 pas*

On sera d'autant plus intéressé au cours de cet article par les méthodes et/ou modèles prédictifs permettant une notion d'addivité dans cette trajectoire. C'est à dire que les modifications univariées de X s'additionnent et permettent de s'approcher de la frontière de décision à l'aide de modification variable par variable, pas à pas, comme illustré dans la Figure 2. On démontrera plus loin que c'est le cas pour le classifieur Naïf de Bayes.

3 Recherche optimisée de contrefactuels dans le cas du classifieur naïf de Bayes

3.1 Rappels sur le classifieur naïf de Bayes

Le classifieur naïf de Bayes (NB) est un outil largement utilisé dans les problèmes de classification supervisée. Il a pour avantage de se montrer efficace pour de nombreux jeux de données réels (Hand et Yu, 2001). Cependant, l'hypothèse naïve d'indépendance des variables peut, dans certains cas, dégrader les performances du classifieur. Aussi, des méthodes proposant de réaliser de la sélection de variables ont vu le jour (Langley et Sage, 1994). Elles consistent en la mise en place d'heuristiques d'ajout et de suppression de variables afin de sélectionner le meilleur sous-ensemble de variables maximisant un critère de performance du classifieur, selon une approche wrapper (Guyon et Elisseeff, 2003). Il a été montré que moyenner un grand nombre de classifieurs Bayésiens naïfs sélectifs, réalisés avec différents sous-ensembles de variables, revenait à ne considérer qu'un seul modèle avec une pondération sur les variables. La formule de Bayes sous l'hypothèse d'indépendance des variables conditionnellement aux classes devient :

$$P(C_k|X) = \frac{P(C_k)\prod_i P(X_i|C_k)^{W_i}}{\sum_{j=1}^{K}(P(C_j)\prod_i P(X_i|C_j)^{W_i})} \tag{1}$$

où W_i représente le poids de la variable i. La classe prédite est celle qui maximise la probabilité conditionnelle $P(C_k|X)$. Les probabilités $P(X_i|C_i)$ peuvent être estimées par intervalle à l'aide d'une discrétisation pour les variables numériques. Pour les variables catégorielles, cette estimation peut se faire directement si la variable prend peu de modalités différentes ou après un groupage dans le cas contraire.

3.2 Critère à optimiser pour la recherche de contrefactuels

Soit X un exemple et deux classes C_1 et C_2. La recherche d'un contrefactuel consiste à optimiser, augmenter, la probabilité d'appartenir à la classe cible C_1 si initialement X est prédit par le modèle comme appartenant à C_2 (et réciproquement). On peut, pour cela, mettre au point un algorithme glouton coûteux en temps de calcul et ne présentant pas nécessairement les propriétés d'additivité énoncées précédemment section 2.3. On propose ci-dessous de poser autrement le problème en ré-écrivant l'équation 1 et en regardant comment augmenter la probabilité d'appartenance à une classe d'intérêt donnée. Pour atteindre cet objectif, et maximiser $P(C_j|X')$ vis-à-vis de la valeur initiale de $P(C_j|X)$ nous exploiterons la proposition suivante :

Proposition 1. *Si on pose X et X' comme étant deux éléments de l'espace d'entrée $\mathcal{X}$, on montre que pour un problème de classification à deux classes la recherche de contrefactuels de X revient à examiner l'évolution de la valeur de Δ lorsqu'on modifie certaines des valeurs de X à X', tel que :*

$$\begin{aligned}\Delta(X, X') = &\left(\sum_{i=1}^{d} W_i(\log P(X_i|C_1) - \log P(X_i|C_2))\right) \\ &- \left(\sum_{i=1}^{d} W_i(\log P(X'_i|C_1) - \log P(X'_i|C_2))\right)\end{aligned} \tag{2}$$

Preuve : Si on repart de l'équation 1

$$P(C_j|X) = \frac{P(C_j)\prod_{i=1}^{d} P(X_i|C_j)^{W_j}}{\sum_z [P(C_z)\prod_{i=1}^{d} P(X_i|C_z)^{W_j}]}$$

en posant :

$$L_j(X) = \log\left(P(C_j)\prod_{i=1}^{d} P(X_i|C_j)^{W_i}\right) = \log P(C_j) + \sum_{i=1}^{d} W_i \log P(X_i|C_j),$$

on a alors :

$$P(C_j|X) = \frac{e^{L_j(X)}}{\sum_z e^{L_z(X)}} = \frac{1}{\sum_z e^{L_z(X)-L_j(X)}} = \frac{1}{1+\sum_{z\neq j} e^{L_z(X)-L_j(X)}}$$

et donc dans le cas à deux classes :

$$P(C_j|X) = \frac{1}{1+e^{L_{j'}(X)-L_j(X)}} \tag{3}$$

On voit alors que pour se rapprocher de la classe C_j, il suffit de réduire la quantité $L_{j'}(X) - L_j(X)$, et donc de réduire :

$$\log P(C_{j'}) + \sum_{i=1}^{d} W_i \log P(X_i|C_{j'}) - \log P(C_j) - \sum_{i=1}^{d} W_i \log P(X_i|C_j)$$

Comme $P(C_j)$ et $P(C_{j'})$ sont constants, c'est équivalent à faire décroître :

$$\sum_{i=1}^{d} W_i \log P(X_i|C_{j'}) - \sum_{i=1}^{d} W_i \log P(X_i|C_j)$$

et donc à s'intéresser à la distance :

$$\begin{aligned}\Delta(X, X') = &\sum_{i=1}^{d} W_i(\log P(X_i|C_{j'}) - \log P(X_i|C_j)) \\ &- \sum_{i=1}^{d} W_i(\log P(X'_i|C_{j'}) - \log P(X'_i|C_j))\end{aligned}$$

Si Δ est positif alors on se rapproche de la frontière de décision (voir on la franchit) si Δ est négatif on s'éloigne de la frontière de décision donc à l'opposé de l'objectif souhaité. L'algorithme de recherche de contrefactuel est alors simple. Il suffit de calculer, pour un exemple donné X, la valeur de Δ pour chaque variable explicative et pour chaque valeur de cette variable explicative. Puis muni de ces valeurs d'itérer les changements successifs afin d'obtenir un exemple contrefactuel. Ces changements variable par variable ont la propriété d'être additifs.

En effet si on pose quatre exemples X^0, $X^{'1}$, $X^{'2}$ et $X^{'3} \in \mathcal{X}$ étant respectivement (i) un exemple initial X^0, puis le même exemple pour lequel on a modifié une seule variable explicative l pour $X^{'1}$, m pour $X^{'2}$ et enfin un exemple cumulant les deux modifications univariées l et m pour $X^{'3}$ tel que :

Voir le processus de création d'exemples contrefactuels comme une source de connaissance

$$\exists!\ l \text{ tel que } X_l^{'1} \neq X_l^0$$

$$\exists!\ m \text{ tel que } X_m^{'2} \neq X_m^0 \text{ et } m \neq l$$

et

$$X_k^{'3} = \begin{cases} X_l^{'1}, & \text{if } k = l \\ X_m^{'2}, & \text{if } k = m \\ X_k^0, & \text{sinon} \end{cases}$$

alors il est évident, d'après l'additivité sur l'ensemble des variables de l'équation 2, que l'on a : $\Delta(X_0, X'_3) == \Delta(X_0, X'_1) + \Delta(X_0, X'_2)$. Modifier une variable puis l'autre est équivalent à les modifier simultanément dans le calcul de Δ. On note aussi que cette additivité est démontré à partir de l'équation 3 on a donc l'assurance de faire progresser la valeur **normalisée** de la probabilité de la classe d'intérêt, $P(C|X)$, ce qui est un plus.

Note : La liste de valeurs de Δ peut être potentiellement très grande si le nombre de valeurs distinctes des variables explicatives est grande. Néanmoins il est fréquent pour certains classifieurs [1] naïf de Bayes (Yang et Webb, 2003, 2009) de discrétiser les variables numériques et grouper les modalités des variables catégorielles, de manière supervisée, afin d'avoir une estimation des densités conditionnelles ($P(X_i|C)$) exploitées ensuite dans le calcul de $P(C|X)$. C'est ce qui a été réalisé ici dans cet article à l'aide des méthodes (Boullé, 2006) et (Boullé, 2005) respectivement pour les variables numériques et catégorielles. Ces opérations, supervisées de discrétisation et de groupage, produisent souvent un nombre limité d'intervalles ou de groupe de modalités. Ceci permet alors d'obtenir un nombre raisonnable de valeurs à tester.

4 Création et exploitation d'une base de connaissance

4.1 Création d'une base de connaissance

Dans ce qui précède nous avons montré comment faire progresser la probabilité d'appartenance à une classe d'intérêt et quantifier cette progression à l'aide de l'équation 2. Nous avons aussi montré que cette quantité est additive à mesure que l'on change les valeurs des variables explicatives une à une. Nous proposons à présent de stocker ces valeurs de Δ dans une table qui à la forme de celle présentée dans la Table 1.

On rappelle qu'on part du postula que les variables numériques ont été préalablement discrétisées et qu'un groupement de modalités a été réalisé pour les variables catégorielles. Chaque variable est donc représentée par un nombre limité de valeurs (correspondant aux valeurs de $P(X|C)$). La table 1 donne, à titre illustratif, les valeurs stockées dans le cas où le modèle prédictif utilise deux variables explicatives X_1 et X_2 respectivement discrétisées (ou groupées) en 3 et 2 intervalles (groupes) de valeurs (I). Pour chaque individu, l, chaque ligne de la table, on mémorise les valeurs de l'équation 2 où $\Delta(X^l_{i,*\to m}, X^l)$ est la valeur de Δ lorsque l'on modifie la valeur de la variable i de sa valeur initiale '$*$' pour la valeur de l'intervalle (groupe) m (que l'on simplifie comme étant $\Delta(X^l_{i,*\to m})$ dans le tableau).

1. Hormis la version Gaussienne.

	X_1		X_2		
	I_1	I_2	I_1	I_2	I_3
X^1	$\Delta(X^1_{1,*\to 1})$	$\Delta(X^1_{1,*\to 2})$	$\Delta(X^1_{2,*\to 1})$	$\Delta(X^1_{2,*\to 2})$	$\Delta(X^1_{2,*\to 3}$
X^2	$\Delta(X^2_{2,*\to 1})$	$\Delta(X^2_{2,*\to 2})$	$\Delta(X^2_{2,*\to 1})$	$\Delta(X^2_{2,*\to 2})$	$\Delta(X^2_{2,*\to 3})$

TAB. 1 – *Illustration de la base de connaissance crée sous la forme des Δ, ici dans le cas à deux variables et deux exemples.*

Nous détaillons dans les sections suivantes comment exploiter la connaissance ainsi stockée. Dans cet article seul le classifieur naïf de Bayes est considéré mais toute autre classifieur et / ou méthode de création de contrefactuels permettant de produire une table de données similaire pourrait être utilisés.

4.2 Générations de contrefactuel ayant certaines propriétés

4.2.1 Minimisation du nombre de changements

Nous pouvons nous fixer comme critère de trouver le contrefactuel ayant la propriété d'impliquer le plus petit nombre de variables modifiées. Pour ce faire, nous allons exploiter la base de connaissance. Pour un individu donné, X, il suffit d'aller lire la valeur du plus grand Δ puis comme on a la propriété d'additivité de lire la seconde valeur du plus grand Δ pour une deuxième variable et ainsi de suite. A chaque étape on vérifie si ($\hat{f}(\mathbf{X}') > 0.5$).Si tel est le cas le contrefactuel a été trouvé [2].

4.2.2 Prise en compte de contraintes ou critères métier

Dans d'autre cas l'optique peut être de trouver le contrefactuel le plus proche mais sous des "contraintes" métier définies par l'utilisateur. Par exemple, on pourrait contraindre la recherche de contrefactuels à ne réaliser des changements que dans des intervalles adjacents (par exemple des intervalles de valeurs proches). Muni de la table 1 on serait autorisé à faire passer des individus de l'intervalle 1 à 2 pour la seconde variable mais pas de 1 à 3. L'utilisateur peut aussi contraindre la recherche par le fait qu'une des variables doit toujours être modifiée à une valeur donnée, etc. Ce type de contraintes est aisé à prendre en compte et à insérer dans un algorithme de recherche de contrefactuel muni de la table de connaissance proposée.

La littérature sur les contrefactuels pose d'ailleurs à ce sujet des propriétés intéressantes qui sont par exemple (i) la notion de minimalité : avoir un exemple contrefactuel qui diffère le moins de l'exemple initial ; (ii) le réalisme : le contrefactuel généré ne doit pas impliquer des changements qui n'ont pas de sens du point de vue des données (par exemple la diminution de "l'âge" d'un individu) , appelé aussi plausibilité (Nemirovsky et al., 2022) ; (iii) créer des contrefactuels similaires à des exemples réels ou dans des régions denses de la classe d'intérêt pour avoir des contrefactuels robustes (Guyomard et al., 2023) : dans ce cas on peut utiliser, dans le cas du classifieur naïf de Bayes, une distance Bayésienne tel que proposée dans (Lemaire et al., 2020) ; ...

2. On pourrait à l'opposer maximiser le nombre de changements mais ce qui revêt souvent peu d'intérêt en pratique .

Voir le processus de création d'exemples contrefactuels comme une source de connaissance

Toutes ces propriétés sont facilement atteignables muni de la Table 1 puisque l'utilisateur peut choisir la liste et l'ordre des variables sur lesquels ils souhaitent intervenir ainsi qu'une distance de son choix entre X et X'.

4.3 Connaissance additionnelle exploitable

4.3.1 Action préventives et réactives

Jusqu'à présent nous avons essentiellement parlé de la création d'exemples contrefactuels ceci afin d'expliquer les décisions du modèle (comme évoqué en introduction de cet article) mais potentiellement aussi dans le but de pouvoir réaliser des actions réactives. En effet si, par exemple, un client d'une banque est prédit comme "partant" (churneur) l'exemple contrefactuel nous désigne une ou plusieurs actions à mener afin d'essayer de le garder comme client : on parle alors d'**actions "réactives"**.

Réciproquement, l'étude, a posteriori, des trajectoires contrefactuelles est d'un intérêt majeur, car elle nous permet aussi de détecter quand une trajectoire s'approche de la frontière. Dans de telles situations, des mesures réactives peuvent être prises pour inverser la tendance et éviter des résultats indésirables. Cette approche est particulièrement pertinente pour prédire, par exemple, le désabonnement, car elle permet d'identifier les clients qui s'engagent dans un processus de désabonnement. En agissant de manière proactive, il est possible de mettre en place des stratégies ciblées pour conserver ces clients et les ramener à un service de qualité.

Enfin notre base de connaissance peut aussi nous permettre de réaliser des **actions "préventives"**. En reprenant la Figure 1 on peut chercher à créer un semi-factuel qui s'éloignerait de la frontière de décision. "Le client n'est pas prédit comme partant mais est néanmoins proche de la frontière de décision". Dans ce cas il suffit alors de s'intéresser aux valeurs négatives de Δ et de faire des pas d'éloignement selon le désir de l'utilisateur. Par exemple tous les individus qui sont à un pas de franchir la frontière de décision, ici facilement identifiables, pourraient être concernés.

4.3.2 Création de profils

La dernière possibilité d'exploitation de la base de connaissance, que nous décrirons ici [3], est de réaliser une analyse exploratoire à l'aide d'une technique de clustering. En utilisant la base de connaissance, il est possible de regrouper les individus en fonction des effets de chaque changement possible, donc des effets résultant de chacun de ces changements. L'analyse des clusters élaborés peut être source d'enseignement. Ceci est illustré dans la section suivante.

5 Illustration du clustering sur un cas de désabonnement

1) Jeu de donnée et classifieur utilisé : On utilise dans cette section le jeu de données "Telco Customer Churn" (très utilisé dans l'analyse des résultats de méthodes XAI) offert par une société de télécommunications fictive qui a fourni des services de téléphonie et d'internet à domicile à 7043 clients en Californie. Il s'agit de classifier des personnes qui risquent ou non

3. Le lecteur pourra en imaginer d'autres : statistiques descriptives de la table, nombre d'individus à 1, 2, 3 ... pas de la frontière de décision, visualisation des trajectoires, ...

de quitter cette entreprise. Chaque client est décrit par 20 variables descriptives (3 numériques et 17 catégorielles) plus la variable de classe ('désabonnement' (oui/non) qui à deux modalités (75% de non-churn). Ce jeu de données peut être téléchargé auprès de Kaagle (Kaagle, 2023). Nous utilisons 80% des données en apprentissage et 20% en test. Le classifieur naïf de Bayes est produit à l'aide la librairie Khiops disponible à présent de manière gratuite et en open source sur Github (Khiops, 2023).

Au cours du processus d'apprentissage seules 10 des variables ont été retenues au sein du modèle. On donne ci-dessous l'ensemble des intervalles de valeurs ou de groupes de modalités obtenues lors du processus de pré-traitement réalisé (la valeur entre parenthèse donne le poids de la variable dans le modèle, équation 1, valeurs de 0 à 1) :

- 1 * - Tenure (W_1=0.67) : [0-0.5],]0.5-1.5],]1.5-5.5],]5.5-17.5],]17.5-42.5],]42.5-58.5],]58.5-71.5],]71.5-72]
- 2 - InternetService (W_2=0.78) : [Fiberoptic], [DSL], [No]
- 3 - Contract (W_3=0.37) : [Month-to-month], [Twoyear], [Oneyear]
- 4 - PaymentMethod (W_4=0.29) : [Mailedcheck], [Creditcard(automatic), Electroniccheck, Banktransfer(automatic)]
- 5 - OnlineSecurity (W_5=0.15) : [No], [Yes], [No internet service]
- 6 -TotalCharges (W_6=0.29) : [18.8 ;69.225],]69.225 ;91.2],]91.2 ;347.9],]347.9 ;1182.8],]1182.8 ;2065.7],]2065.7 ;3086.8],]3086.8 ;7859],]7859 ;8684.8]
- 7 * - PaperlessBilling (W_7=0.40) : [Yes], [No]
- 8 - TechSupport (W_8=0.04) : [No], [Yes], [No internet service]
- 9 * - SeniorCitizen ($W_9 = 0.28$) : [0], [1]
- 10 * - Dependents ($W_9 = 0.10$) : [Yes], [No]

Pour l'ensemble des 10 variables, il y a un total de 36 intervalles/groupements et donc 26 valeurs Δ à calculer dans notre base de connaissances. En fait, pour chaque individu et chaque variable, il existe une valeur Δ qui a une valeur nulle, celle qui lui correspond factuellement et qui n'a donc pas besoin d'être calculée.

2) Etape d'analyse du classifieur : Avant de réaliser l'étape de clustering il est important de s'intéresser aux variables conservées lors de l'étape d'entraînement du modèle de classification. Par exemple, même si l'analyse de la variable 'Tenure' pourrait être intéressante, elle n'est pas à l'évidence une variable actionnable. En effet il n'est pas possible de modifier l'ancienneté d'un client pour le rendre potentiellement moins infidèle. Il va de même pour les variables 'SeniorCitiezn, 'Dependents' . Nous retirons aussi la variable 'PaperlessBilling' qui a un très faible impact sur les résultats du clustering décrit ci-dessous.

De ce fait ces 4 variables ne sont pas conservées lors de l'étape de clustering ci-après ; on ne conserve que les variables informatives, influentes et actionnables [4](voir Section 2.2).

3) Clustering réalisé : L'opération de clustering réalisée est usuelle : (i) on utilise la table des valeurs des Δ calculée sur l'ensemble de test, (ii) l'apprentissage d'un k-means avec la distance L2 (Hartigan et Wong, 1979) pour différentes valeurs de k est réalisé ($k \in \{2, 12\}$), (iii) et finalement on conserve le k-means dont la valeur de k corresponds au 'point elbow', ici $k = 4$, (Thorndike, 1953) de la courbe représentant la distance de reconstruction globale versus la valeur de k.

4. Toutes les variables auraient pu être conservées mais le clustering aurait été biaisé par des variables inintéressantes d'un point de vue de la création d'exemples contrefactuels.

Voir le processus de création d'exemples contrefactuels comme une source de connaissance

4) Les Clusters obtenus sont présentés dans la Figure 3

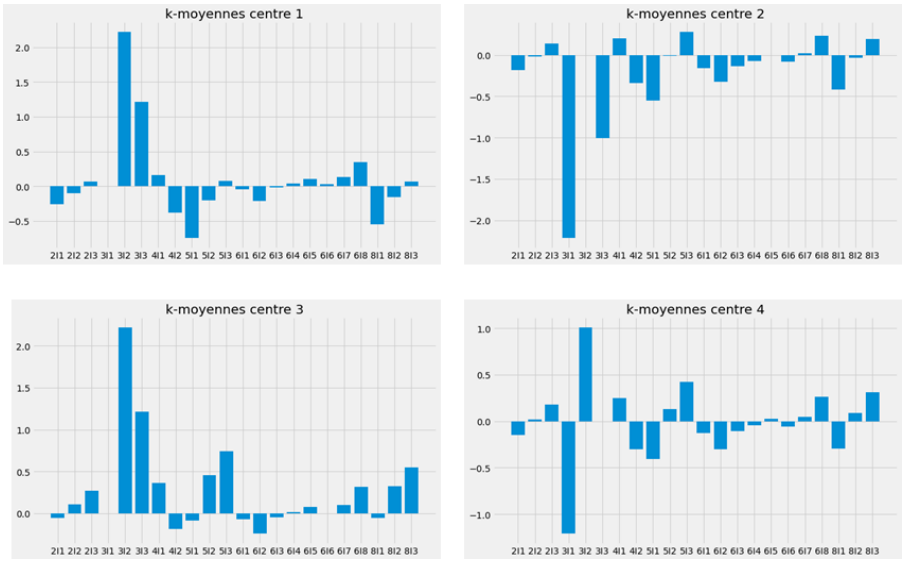

FIG. 3 – *Profil moyen des individus des clusters représentés sous la forme d'histogrammes. Les noms des valeurs en abscisses réfèrent au numéro des variables et numéro des intervalles (groupes) décrits ci-dessus. Par exemple '3I2' réfère la troisième variable ('Contract') et son deuxième intervalle / groupe de valeurs ('Twoyear'). Les valeurs en ordonnées sont les valeurs moyennes du cluster (*Δ*).*

Une analyse de ces 4 clusters, couplée aux prédictions du classifieur, indique :

— Cluster 1 (10% de la population et contenant 2% de clients prédits comme infidèles) : individus que l'on peut rendre moins infidèles principalement au moyen de la variable 3 ('Contrat') -c'est-à-dire en essayant de leur faire souscrire un contrat annuel ('Twoyear' ou 'OneYear') ; NB - cette action de marketing est assez difficile à réaliser.

— Cluster 2 (24% de la population et ne contenant aucun client prédit comme infidèle) : des individus très peu sensibles au fait de devenir moins infidèles (des valeurs moyennes de Δ surtout négatives). Potentiellement à ne pas adresser dans une campagne marketing 'réactive' (ce qui est conforme aux prédictions du classifieur) mais plutôt avec une campagne préventive utilisant la variable 'Contract' ou la variable 'PaymentMethod' (paiement par carte ou par prélèvement automatique).

— Cluster 3 (45% de la population et contenant 47% de clients prédits comme infidèles (la quasi totalité des individus prédits infidèles)) : certaines analogies avec les individus du cluster 1 pour la variable 'Contract'. D'autre part, on observe par contre que la 5è ('OnlineSecurity') et 8è variable ('TechSupport') sont ici des variables "à effet de levier" pour réduire le taux de désabonnement. Proposer une option de sécurité ou ou d'assistance est très intéressant pour ces individus.

— Cluster 4 (21% de la population et ne contenant aucun client prédit comme infidèle) : des individus opposés en partie à ceux du premier cluster, par exemple pour la variable

'Contract' à qui ici il ne faut pas proposer un contrat 'TwoYear'.

L'analyse des clusters obtenus ici est non exhaustive par manque de place. Elle relève de l'analyse exploratoire, où le data scientist et l'expert métier consacreront le temps nécessaire pour affiner leurs analyses conjointes. Cependant, l'analyse réalisée ici permet néanmoins d'identifier des actions 'réactives' intéressantes à mener auprès des individus du cluster 3 ou ou des actions préventives auprès des individus du cluster 2.

6 Conclusion

Dans le cadre des méthodes d'explication des résultats d'un modèle d'apprentissage automatique, cet article a proposé de considérer le processus de génération d'exemples contrefactuels comme une source de connaissances qui peut être stockée puis exploitée de différentes manières. Ce processus a été illustré dans le cas des modèles additifs et en particulier dans le cas du classifieur naïf de Bayes, dont les propriétés intéressantes à cette fin ont été montrées. Nous avons aussi suggéré les quantités qui peuvent être stockées et les différentes manières de les exploiter. Certains des résultats ont été illustrés sur un problème de désabonnement, mais l'approche est tout aussi exploitable dans d'autres domaines d'application.

Références

Allen, G. I., L. Gan, et L. Zheng (2023). Interpretable machine learning for discovery : Statistical challenges & opportunities. *Arxiv preprint :2308.01475*.

Bodria, F., F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, et S. Rinzivillo (2023). Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery 37*(5), 1719–1778.

Boullé, M. (2005). A Bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research 6*, 1431–1452.

Boullé, M. (2006). MODL : a Bayes optimal discretization method for continuous attributes. *Machine Learning 65*(1), 131–165.

Brughmans, D., P. Leyman, et D. AU Martens (2023). Nice : an algorithm for nearest instance counterfactual explanations. *Data Mining and Knowledge Discovery*.

Fernández, R. R., I. M. De Diego, J. M. Moguerza, et F. Herrera (2022). Explanation sets : A general framework for machine learning explainability. *Information Sciences 617*, 464–481.

Guyomard, V., F. Fessant, T. Guyet, T. Bouadi, et A. Termier (2023). Generating robust counterfactual explanations. In *European Conference on Macine Learning*.

Guyon, I. et A. Elisseeff (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res. 3*, 1157–1182.

Hand, D. J. et K. Yu (2001). Idiot's bayes-not so stupid after all ? *International Statistical Review 69*(3), 385–398.

Hartigan, J. A. et M. A. Wong (1979). A k-means clustering algorithm. *JSTOR : Applied Statistics 28*(1), 100–108.

Kaagle (2023). Telco customer churn dataset. [https://www.kaggle.com/datasets/blastchar/telco-customer-churn], last visited 08/22/2023.

Khiops (2023). Github khiops. [https://github.com/KhiopsML/khiops], last visited 08/22/2023.

Langley, P. et S. Sage (1994). Induction of selective bayesian classifiers. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, USA, pp. 399–406. Morgan Kaufmann Publishers Inc.

Lemaire, V., C. Hue, et O. Bernier (2010). *Data Mining in Public and Private Sectors : Organizational and Government Applications*, Chapter Correlation Analysis in Classifiers, pp. 204–218. IGI Global.

Lemaire, V., O. A. Ismaili, A. Cornuéjols, et D. Gay (2020). Predictive k-means with local models. In W. Lu et K. Q. Zhu (Eds.), *Trends and Applications in Knowledge Discovery and Data Mining*, Cham, pp. 91–103. Springer International Publishing.

Miller, T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence 267*, 1–38.

Nemirovsky, D., N. Thiebaut, Y. Xu, et A. Gupta (2022). Countergan : Generating counterfactuals for real-time recourse and interpretability using residual gans. In *Conference on Uncertainty in Artificial Intelligence*, Machine Learning Research, pp. 1488–1497. PMLR.

Saeed, W. et C. Omlin (2023). Explainable ai (xai) : A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems 263*, 110273.

Stepin, I., J. M. Alonso, A. Catala, et M. Pereira-Fariña (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access 9*, 11781–11803.

Thorndike, R. L. (1953). Who belongs in the family ? *Psychometrika 18*, 267–276. The method can be traced to speculation by Robert L. Thorndike.

Wachter, S., B. Mittelstadt, et C. Russell (2018). Counterfactual explanations without opening the black box : Automated decisions and the gdpr. *Harvard Journal of Law and Technology 31*(2), 841–887.

Yang, Y. et G. Webb (2003). A comparative study of discretization methods for naive-bayes classifiers. In *Proceedings of PKAW, vol. 2002*.

Yang, Y. et G. Webb (2009). Discretization for naive-bayes learning : Managing discretization bias and variance. *Machine Learning 74*, 39–74.

Summary

There are now many comprehension algorithms for understanding the decisions of a machine learning algorithm. Among these are those based on the generation of counterfactual examples. This article proposes to view this generation process as a source of creating a certain amount of knowledge that can be stored to be used in different ways. This process is illustrated in the additive model and, more specifically, in the case of the naive Bayes classifier, whose interesting properties for this purpose are shown.