Intégration de connaissances en XAI avec les intégrales de Gödel

Adulam Jeyasothy, Agnès Rico, Marie-Jeanne Lesot, Christophe Marsala, Thibault Laugel

Intégration de connaissances en XAI avec les intégrales de Gödel

Adulam Jeyasothy[1] Agnès Rico[2] Marie-Jeanne Lesot[1] Christophe Marsala[1] Thibault Laugel[3]
[1] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[2] Univ. Lyon 1, Lyon, France
[3] AXA, Paris, France

Résumé :

Les exemples contre-factuels constituent une forme populaire d'explications souvent générées par optimisation d'une fonction de coût qui combine différentes composantes de la qualité de l'explication. Cet article se concentre sur l'agrégation finale du terme objectif, qui dépend de la tâche d'apprentissage considérée, et du terme subjectif, qui dépend de l'utilisateur considéré et plus précisément de ses connaissances. Il examine les propriétés souhaitées de cette agrégation et propose d'utiliser deux formes de l'intégrale de Gödel, en soulignant l'expressivité et la pertinence qu'elles offrent.

Mots-clés :

XAI, exemples contre-factuels, fonctions d'agrégation, connaissances, intégrale de Gödel, intégrale de Sugeno.

Abstract:

Counterfactual examples constitute a popular form of explanations that are most often generated through the optimisation of a cost function that combines different components of the explanation quality. This paper focuses on the final aggregation of the objective term, that depends on the considered machine learning task, and the subjective term, that depends on the targeted user and more precisely on their knowledge. It discusses the desired properties of this aggregation operator and proposes to use two forms of the Gödel integral operator, highlighting the expressiveness and appropriateness they offer.

Keywords:

XAI, counterfactual examples, aggregation functions, prior knowledge, Gödel integral, Sugeno integral.

1 Introduction

Parmi les méthodes d'explication post-hoc locales [7] qui expliquent la prédiction d'un classifieur donné pour une instance donnée, les exemples contre-factuels [20] identifient des modifications à appliquer à l'instance considérée pour obtenir une prédiction différente : ils répondent à la question de l'utilisateur "Que modifier pour obtenir la prédiction souhaitée ?". Une grande variété d'approches a été proposée pour y répondre (par exemple [1, 8, 10, 19]), reposant sur la définition d'une fonction de coût à optimiser. Cette dernière comprend différents termes qui constituent différentes composantes de la qualité de l'explication. Une question cruciale porte alors sur leur agrégation.

L'article se concentre sur l'étape finale qui combine un terme objectif avec un terme subjectif. Le premier, appelé pénalité, fait référence à la combinaison de critères numériques qui dépendent uniquement du classifieur considéré, de l'instance étudiée et d'informations supplémentaires telles que la densité des données. Le terme subjectif, appelé incompatibilité, permet de personnaliser une explication et dépend de l'utilisateur considéré, en particulier de ses connaissances. Cet article examine les propriétés spécifiques que cette agrégation requiert afin de proposer une explication adaptée aux attentes de l'utilisateur. Il se concentre sur le cas d'explications contre-factuelles pour la définition de la pénalité et de l'incompatibilité, mais son principe peut être appliqué chaque fois que ces deux quantités peuvent être définies. Il propose ensuite d'utiliser les intégrales de Gödel [3], en soulignant leur expressivité et leur adéquation. À notre connaissance, il constitue la première application des intégrales de Gödel dans le domaine de l'intelligence artificielle explicable (XAI).

L'article est organisé comme suit : la section 2 résume les principes des explications contre-factuelles, en présentant le problème d'agrégation soulevé. La section 3 examine les caractéristiques désirées pour l'opérateur d'agrégation. La section 4 présente l'opérateur choisi : l'intégrale de Gödel et son utilisation. La section 5 illustre la richesse de ce choix sur des données classiques en 2D. La section 6 conclut l'article, en discutant des perspectives.

2 Contexte : explications contre-factuelles enrichies

Cette section présente les principes des explications contre-factuelles et leur personnalisation par intégration les connaissances utilisateur. Elle aborde ensuite l'étape cruciale de l'agrégation.

2.1 Principes des exemples contre-factuels

Les exemples contre-factuels [20] expliquent la prédiction donnée par un modèle d'apprentissage automatique $f : \mathcal{X} \rightarrow \mathcal{Y}$ (où $\mathcal{X}$ l'espace des données et $\mathcal{Y}$ l'espace de sortie, par exemple $\mathcal{Y} = \{0, 1\}$ pour la classification binaire) pour une instance $x_0 \in \mathcal{X}$. Ils répondent à la question : "Quelles modifications, de x_0 à x'_0, permettent $f(x'_0) \neq f(x_0)$?". L'exemple contre-factuel de base x'_0 est défini comme l'instance la plus proche associée à une prédiction souhaitée. La contrainte de proximité, définie par les normes l_2 [11, 12] ou l_1 [20], vise à minimiser l'effort fourni par l'utilisateur pour obtenir le résultat souhaité. Des critères comme la parcimonie [2, 12] ou la plausibilité [16] peuvent être ajoutés à la proximité.

On note $P_{f,x_0}(e)$ le terme de pénalité qui définit la qualité d'un exemple contre-factuel candidat e en fonction de ces différentes composantes. L'exemple contre-factuel est formellement défini comme la solution du problème :

$$e^* = \arg\min_{e \in \mathcal{X}} P_{f,x_0}(e) \text{ avec } f(e) \neq f(x_0) \quad (1)$$

2.2 Intégration de connaissances utilisateur

Au-delà des critères objectifs qui définissent la fonction de pénalité uniquement en fonction de la tâche d'apprentissage considérée, une deuxième classe de critères adopte un point de vue plus subjectif sur les candidats contre-factuels et fait dépendre leur qualité de l'utilisateur. Ces critères permettent de personnaliser les explications, en augmentant leur pertinence et leurs avantages pour l'utilisateur considéré.

Les approches existantes diffèrent à la fois sur le type de connaissances de l'utilisateur qu'elles considèrent et sur la manière dont ces dernières sont intégrées dans la génération d'explications. Ces connaissances peuvent par exemple prendre la forme d'un ensemble d'attributs [9, 18], un ensemble d'invervalles associés à chaque attribut [15], la monotonie des modifications [13] ou encore des liens entre les attributs [4], notamment des liens de causalité [5, 13].

Dans cet article, la connaissance de l'utilisateur est notée E, indépendamment de sa forme. Les exemples contre-factuels candidats incompatibles avec cette connaissance doivent alors être pénalisés. Nous notons $I_{E,x_0}(e)$ la valeur de l'incompatibilité.

2.3 Problème d'agrégation

La génération d'un exemple contre-factuel peut être formulée comme un problème d'optimisation, dont la fonction de coût à minimiser combine $P_{f,x_0}(e)$ et $I_{E,x_0}(e)$. Une question cruciale est celle de l'agrégation de ces deux termes.

La définition de la pénalité soulève déjà un problème d'agrégation, car elle combine plusieurs éléments. Néanmoins, on peut considérer que ces composantes sont de même nature, puisqu'elles constituent des critères objectifs qui ne dépendent que de la tâche d'apprentissage automatique considérée.

La combinaison de la pénalité et de l'incompatibilité quant à elle s'applique à des composants de nature différente si l'on considère le premier comme objectif et le second comme subjectif. Aussi, la discussion sur l'agrégation peut être considérée comme plus riche, nécessitant des opérateurs plus expressifs. Une première discussion de ce type est proposée dans [9] qui soutient que les opérateurs conjonctifs sont trop stricts, les opérateurs disjonctifs trop laxistes et qui propose d'utiliser des opérateurs de compromis comme des moyennes pondérées. Nous proposons ici d'étendre cette discussion, en examinant de façon plus détaillée les propriétés

que l'opérateur d'agrégation devrait posséder (Section 3), puis d'appliquer les intégrales de Gödel (Section 4).

Formellement, la question considérée est la sélection d'un opérateur d'agrégation agg pour définir la fonction de coût :

$$cost_{f,E,x_0}(e) = agg(P_{f,x_0}(e), I_{E,x_0}(e)) \quad (2)$$

à minimiser sous la contrainte $f(e) \neq f(x_0)$.

Pour alléger les notations, nous omettons les indices fixes f, x_0 et E et l'exemple contre-factuel candidat e lorsqu'il n'y a pas d'ambiguïté. Ainsi, nous étudions le choix d'un opérateur pour calculer $agg(P, I)$. Une étape de normalisation classique dans $[0, 1]$ est appliquée aux critères P et I, l'étude de leur commensurabilité est laissée pour des travaux futurs.

3 Caractéristiques désirées

Il existe une littérature très riche sur les opérateurs d'agrégation [6, 14], aussi bien sur leurs définitions que sur leurs propriétés. Cette section examine certaines des propriétés qu'un opérateur d'agrégation doit vérifier pour répondre aux exigences du contexte XAI décrit dans la section précédente.

Discussion sur la monotonie. Tout d'abord, l'opérateur d'agrégation considéré doit être croissant en chacun de ses deux arguments. Ainsi, pour le premier argument P, étant donné deux candidats contre-factuels e_1 et e_2 tels que $P(e_1) \leq P(e_2)$ et $I(e_1) = I(e_2)$, on souhaite que $cost(e_1) \leq cost(e_2)$.

Discussion sur la commutativité. Nous défendons l'idée que l'opérateur considéré ne doit pas être commutatif. Comme évoqué dans la section 2, les deux critères considérés, P et I, ont des sémantiques différentes, étant respectivement de nature objective et subjective. Ils ne sont donc pas équivalents et il se peut que $agg(x, y) \neq agg(y, x)$ parce que $y = P(e)$ n'a pas la même signification que $y = I(e)$.

Discussion sur le comportement des variables. Nous défendons que l'opérateur considéré doit avoir des comportements différents en fonction des valeurs des variables : selon les régions, il doit être conjonctif, disjonctif ou de compromis. En effet en XAI, l'une des difficultés du choix de l'agrégation est qu'elle doit être adaptée à tous types d'utilisateurs, qui ont des motivations et des besoins différents. Nous proposons que les utilisateurs expriment leurs besoins sous forme de contraintes sur les critères, par exemple comme des limites sur les valeurs minimales de la pénalité et de l'incompatibilité : ils peuvent fixer des seuils d'acceptabilité $\delta_P, \delta_I \in \mathbb{R}$ et imposer $P(e) < \delta_P$ et $I(e) < \delta_I$. Pour la pénalité, δ_P peut dépendre de la valeur de référence obtenue par l'exemple contre-factuel e_P^* qui minimise la pénalité, comme défini dans l'équation 1 : la contrainte peut être exprimée en termes de perte de qualité par rapport à e_P^*, en l'absence de seuil dépendant de x_0. Pour l'incompatibilité, il est difficile de définir une telle valeur de référence. Nous proposons donc de considérer deux contraintes :

$$P(e) - P(e_P^*) < \delta_P \quad \text{et} \quad I(e) < \delta_I \quad (3)$$

Ces contraintes divisent l'espace des critères, décrit par les couples $(P(e), I(e))$, en quatre zones différentes, selon que deux, une seule ou aucune contrainte est satisfaite. Une propriété souhaitable est la fonction d'agrégation offre des comportements différents dans ces zones, dont l'interprétation n'est pas la même.

Discussion sur la priorité. La différence de sémantique des deux critères peut impliquer une préférence, induisant une relation d'ordre entre eux, qui peut être interprétée comme un comportement prioritaire souhaité. Cette préférence n'est évidemment pas la même pour tous les utilisateurs et participe à l'étape de personnalisation de l'explication. Si un utilisateur exprime par exemple une préférence pour la pénalité par rapport à l'incompatibilité, alors parmi deux candidats contre-factuels ayant la même norme dans l'espace (P, I), celui qui a le P le plus faible doit être favorisé.

Une deuxième possibilité consiste à intégrer la notion de priorité par le choix des seuils dans l'équation (3). Dans le cas où la pénalité est préférée à l'incompatibilité, une condition plus forte sur la pénalité que sur l'incompatibilité est attendue : δ_P devrait être inférieur à δ_I.

4 Opérateur choisi : intégrales de Gödel

Cette section propose de répondre aux exigences décrites dans la section précédente, en utilisant les intégrales de Gödel : elle rappelle d'abord leur définition générale, puis discute leur instanciation dans le cadre de l'XAI, c'est-à-dire leur application à P et I, avant de commenter et d'illustrer leur sémantique.

4.1 Rappel de la définition

Les intégrales de Gödel [3] sont une variante de l'intégrale classique de Sugeno [17] utilisée en prise de décision multicritère. Cette dernière a deux expressions équivalentes, une forme min-max et une forme max-min [17]. En les généralisant avec la conjonction ou l'implication de Gödel, on obtient deux opérateurs différents, qui constituent la famille des intégrales de Gödel.

Notations. L'ensemble des critères d'évaluation est noté $\mathcal{C} = \{1, \cdots, n\}$, ils sont évalués numériquement, par des valeurs dans $L = [0, 1]$.

Comme d'autres opérateurs, par exemple les intégrales de Sugeno, les intégrales de Gödel permettent de modéliser et de prendre en compte le fait que les critères, mais aussi des sous-ensembles de critères, ont des poids différents : elles permettent de représenter l'importance des critères individuellement ainsi que leurs interactions. Cette importance est modélisée par une capacité ou mesure floue, $\mu : 2^{\mathcal{C}} \to [0, 1]$ qui associe à chaque sous-ensemble de critères $A \subset \mathcal{C}$ son poids $\mu(A)$. Par définition, cette fonction est croissante par rapport à l'inclusion et satisfait les conditions aux limites $\mu(\emptyset) = 0$ et $\mu(\mathcal{C}) = 1$.

Intégrale de Gödel basée sur la conjonction. La conjonction de Gödel est l'opérateur non commutatif défini sur $[0, 1]^2$ par :

$$\alpha \otimes_G \beta = \begin{cases} 0 & \text{si } \beta \leq 1 - \alpha \\ \beta & \text{sinon.} \end{cases}$$

Elle est croissante en ses deux arguments et satisfait les conditions aux limites : $1 \otimes_G \beta = \beta$, $\alpha \otimes_G 1 = 0$ si $\alpha = 0$ et 1 sinon, et $0 \otimes_G \beta = \alpha \otimes_G 0 = 0$.

L'intégrale de Gödel applique cet opérateur à chaque critère i, associé à son poids α_i : chaque évaluation locale x_i est modifiée en utilisant la conjonction de Gödel comme $\alpha_i \otimes_G x_i$. Ainsi, x_i n'est pas modifié s'il est supérieur au seuil $1 - \alpha_i$, il est fixé à 0 dans le cas contraire. Ce seuil est décroissant par rapport à α_i : il est faible lorsque α_i est élevé, c'est-à-dire lorsque le critère i est important. Ainsi, une petite évaluation x_i sur un critère important est conservée, tandis qu'une petite évaluation sur un critère non important est modifiée en 0.

Enfin, l'intégrale de Gödel étend ces modifications locales individuelles :

$$G^{\otimes}_{\mu}(x) = \max_{A \subseteq \mathcal{C}} \left(\mu(A) \otimes_G \min_{i \in A} x_i \right) \quad (4)$$

Intégrale de Gödel basée sur l'implication. L'intégrale de Gödel qui repose sur l'implication suit le même principe, en remplaçant la conjonction de Gödel par l'implication de Gödel, le maximum par un minimum et l'utilisation de μ par son conjugué. L'implication de Gödel est définie sur $[0, 1]^2$ par :

$$\alpha \to_G \beta = \begin{cases} 1 & \text{si } \alpha \leq \beta \\ \beta & \text{sinon.} \end{cases}$$

Elle satisfait les conditions aux limites : $0 \to_G \beta = 1$ et $\alpha \to_G 1 = 1$.

Comme pour la conjonction, chaque évaluation locale x_i est transformée à l'aide de cet

opérateur : $\alpha_i \rightarrow_G x_i$. Ainsi, dans ce cas, une valeur x_i n'est pas modifiée si elle est inférieure à α_i, sinon elle est transformée en 1, où α_i mesure toujours l'importance du critère i. Par conséquent, lorsqu'on utilise l'implication de Gödel, une petite évaluation sur un critère non important est modifiée en 1, ce qui offre une sémantique qui diffère de celle obtenue avec la conjonction de Gödel.

L'intégrale de Gödel basée sur l'implication étend l'action locale sur les critères avec une capacité :

$$G^{\rightarrow}_{\mu}(x) = \min_{A \subseteq \mathcal{C}} \left(\mu^c(A) \rightarrow_G \max_{i \in A} x_i \right) \quad (5)$$

où μ^c est la capacité conjuguée de μ définie par $\mu^c(A) = 1 - \mu(\overline{A})$ où $\overline{A}$ désigne le complémentaire de A.

4.2 Application aux exemples contrefactuels en XAI

Cette section traite de l'application de $G^{\otimes}_{\mu}$ et $G^{\rightarrow}_{\mu}$ à la pénalité P et à l'incompatibilité I. L'expression formelle des valeurs agrégées $G^{\otimes}_{\mu}(P, I)$ et $G^{\rightarrow}_{\mu}(P, I)$, est donnée ci-dessous, leurs lignes de niveaux sont illustrées sur la figure 1 et leur interprétation est détaillée dans la section suivante.

L'ensemble des critères est $\mathcal{C} = \{P, I\}$, qui sont normalisés et évalués sur $L = [0, 1]$. La capacité μ est définie sur l'univers $2^{\mathcal{C}}$ dont la taille est égale à 4. Deux valeurs sont fixées d'après les conditions aux limites ($\mu(\emptyset) = 0$ et $\mu(\{P, I\}) = 1$), et les autres sont notées $\mu(\{P\}) = \alpha_P$ et $\mu(\{I\}) = \beta_I$.

L'expression formelle de l'agrégation P et I par les intégrales de Gödel est alors :

$$G^{\otimes}_{\mu}(P, I) = \max(\alpha_P \otimes_G P, \beta_I \otimes_G I, 1 \otimes_G \min(P, I))$$

$$= \begin{cases} \min(P, I) & \text{si } P \leq 1 - \alpha_P \text{ et } I \leq 1 - \beta_I \\ \max(P, I) & \text{si } P > 1 - \alpha_P \text{ et } I > 1 - \beta_I \\ P & \text{si } P > 1 - \alpha_P \text{ et } I \leq 1 - \beta_I \\ I & \text{si } P \leq 1 - \alpha_P \text{ et } I > 1 - \beta_I \end{cases}$$

$$G^{\rightarrow}_{\mu}(P, I) = \min\big((1 - \beta_I) \rightarrow_G P, (1 - \alpha_P) \rightarrow_G I, 1 \rightarrow_G \max(P, I)\big)$$

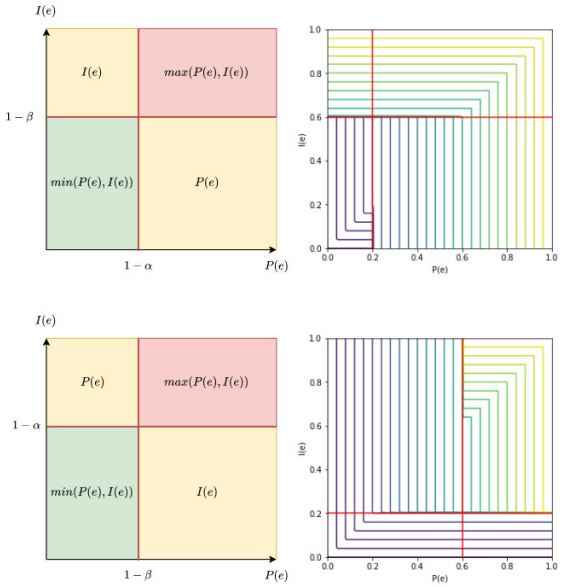


FIGURE 1 – Lignes de niveaux de $G^{\otimes}_{\mu}(P, I)$ (en haut) et $G^{\rightarrow}_{\mu}(P, I)$ (en bas) avec $\alpha_P = 0.8$ et $\beta_I = 0.4$

$$= \begin{cases} \min(P, I) & \text{si } P < 1 - \beta_I \text{ et } I < 1 - \alpha_P \\ \max(P, I) & \text{si } P \geq 1 - \beta_I \text{ et } I \geq 1 - \alpha_P \\ I & \text{si } P \geq 1 - \beta_I \text{ et } I < 1 - \alpha_P \\ P & \text{si } P < 1 - \beta_I \text{ et } I \geq 1 - \alpha_P \end{cases}$$

Propriétés Il est facile de montrer que $G^{\otimes}_{\mu}(P, I)$ et $G^{\rightarrow}_{\mu}(P, I)$ satisfont toutes les propriétés souhaitées présentées dans la section 3 : elles sont monotones en chaque argument, non commutatives, offrent un comportement variable et permettent d'exprimer une hiérarchie de critères.

4.3 Interprétation des intégrales de Gödel dans le domaine de l'XAI

Interprétation des seuils. La correspondance entre les paramètres de Gödel α_P et β_I et les seuils associés aux contraintes discutées dans la section 3 peut être établie en comparant les régions qu'ils définissent respectivement. Par exemple, pour $G^{\otimes}_{\mu}(P, I)$ et le critère P, la contrainte est satisfaite lorsque $P \leq 1 - \alpha_P$, alors que pour $G^{\rightarrow}_{\mu}(P, I)$, la condition est $P \leq 1 - \beta_I$. En les comparant aux contraintes exprimées dans l'équation 3, on obtient pour $G^{\otimes}_{\mu}(P, I)$ à $\delta_P + P(e^*_P) = 1 - \alpha_P$ et $\delta_I = 1 - \beta_I$. Pour $G^{\rightarrow}_{\mu}(P, I)$, on obtient des

contraintes similaires en inversant les variables α_P et β_I. Ces différences sont commentées ci-dessous, en examinant la différence entre les régions induites.

Interprétations des régions. La représentation graphique donnée dans la figure 1 montre que $G_\mu^\otimes(P, I)$ et $G_\mu^\rightarrow(P, I)$ partagent deux régions similaires, l'une en bas à gauche et l'autre en haut à droite. La première correspond aux candidats contre-factuels qui satisfont les deux contraintes et peuvent donc être considérés comme satisfaisants. Leur évaluation ne dépend alors que du meilleur critère, le minimum de P et I (rappelons que le coût global doit être minimisé). Au contraire, dans la région en haut à droite, les candidats ne satisfont aucune contrainte. Afin de les pénaliser, leur score est défini comme le maximum de P et I.

Pour les deux zones restantes, les deux intégrales n'offrent pas la même agrégation, car elles utilisent des principes différents. Pour faciliter la discussion, considérons le cas où seulement la contrainte de pénalité est satisfaite, ce qui correspond à la région supérieure gauche. $G_\mu^\otimes(P, I)$ adopte un comportement de sanction, pénalisant les candidats de cette région selon le critère non satisfait, I, indépendamment de leur valeur de pénalité. Au contraire, $G_\mu^\rightarrow(P, I)$ considère qu'ils sont tous aussi mauvais en ce qui concerne l'incompatibilité et ne les distingue pas par rapport à ce critère, ils sont considérés comme équivalents selon I. $G_\mu^\rightarrow(P, I)$ favorise alors ces candidats selon leur valeur de pénalité. Ceci constitue une différence sémantique majeure qui souligne la richesse et la pertinence des intégrales de Gödel.

Nous commentons enfin l'impact des paramètres de Gödel sur les tailles relatives des quatre régions, en montrant qu'ils jouent le même rôle pour $G_\mu^\otimes(P, I)$ et $G_\mu^\rightarrow(P, I)$ malgré la différence d'interprétation de leurs régions : ils sont basés sur le même principe selon lequel si la capacité associée à un critère est élevée, alors la zone qui minimise uniquement ce critère, en ignorant l'autre critère, est grande, ce qui lui donne en effet plus d'importance. Par exemple, si α_p est élevé, le seuil $1 - \alpha_P$ est faible. Les deux intégrales augmentent la zone qui minimise la pénalité : la zone inférieure droite pour $G_\mu^\otimes(P, I)$ et supérieure gauche pour $G_\mu^\rightarrow(P, I)$. Dans les deux cas, l'aire de cette région est égale à $\alpha_P(1 - \beta_I)$. La même importance est accordée à la pénalité pour des valeurs données des paramètres mais dans des zones différentes.

5 Exemples illustratifs

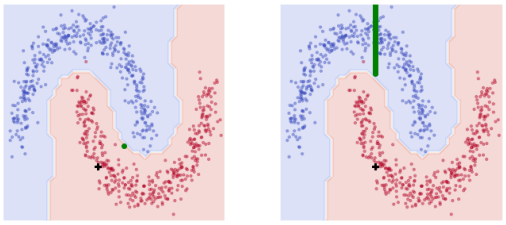
Cette section présente des exemples contre-factuels obtenus avec l'agrégation proposée basée sur Gödel, en les visualisant pour un jeu de données en 2D.

5.1 Données considérées

Les expérimentions sont menées avec l'ensemble de données 2D Half-Moons dont les dimensions sont X_0 (abscisses) et X_1 (ordonnées). Sur les figures 2 et 3, les régions bleues et rouges représentent les classes prédites, les points plus foncés les exemples d'apprentissage. La frontière de décision d'un classifieur SVM entraîné est représentée en blanc (précision : 0.99). La connaissance de l'utilisateur considérée est le singleton $E = \{X_1\}$. Pour permettre des comparaisons visuelles, toutes les expérimentations utilisent la même instance x_0, représentée par une croix noire. La pénalité est définie comme la distance euclidienne normalisée $P = \|x_0 - e\|^2$, l'incompatibilité comme la distance euclidienne normalisée sur l'attribut extérieur à E, $I = \|x_0 - e\|^2_{X_0}$. Le problème d'optimisation n'a pas de solution unique, les solutions sont représentées par des points verts.

5.2 Cas de référence

Nous examinons d'abord quatre fonctions d'agrégation de référence, qui correspondent également à des cas extrêmes des intégrales de Gödel : lorsque le couple (α_P, β_I) est



(a) $P(e)$ (b) $I(e)$

(c) $\min(P(e), I(e))$ (d) $\max(P(e), I(e))$

FIGURE 2 – Explications contre-factuelles générées pour des cas de base

égal à $(1,0), (0,1), (0,0)$ et $(1,1)$, $G_\mu^\otimes(P,I)$ et $G_\mu^\rightarrow(P,I)$ sont équivalentes à $P, I, \min(P,I)$ et $\max(P,I)$ respectivement.

La figure 2 montre les résultats obtenus dans chaque cas, illustrant leur diversité. La figure 2a constitue l'explication de référence. La figure 2b montre les explications qui minimisent l'incompatibilité. Pour x_0 des exemples contre-factuels totalement compatibles avec les connaissances existent : les explications générées sont donc des points à la verticale de x_0 qui appartiennent à la classe bleue, avec une incompatibilité égale à 0. La figure 2c est similaire à la figure 2b : dans le cas considéré, I peut être égal à 0, alors que P ne le peut pas ; le minimum conduit donc aux mêmes résultats que l'incompatibilité. Enfin, la figure 2d est associée à la fonction maximum ; les explications générées sont situées à des positions où l'incompatibilité l'emporte sur la pénalité.

5.3 Cas général

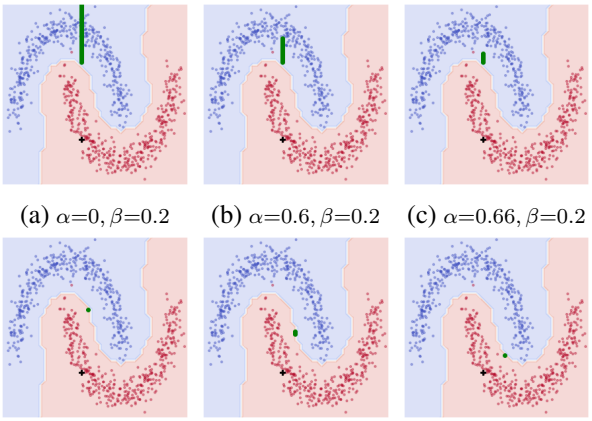
La figure 3 montre des explications générées avec $G_\mu^\otimes$ pour des valeurs moins extrêmes de α_P et β_I, choisies pour illustrer la diversité des résultats. Six cas peuvent être distingués, illustrant la richesse de cet opérateur d'agrégation.

Sur la fig. 3a, identique aux fig. 2b et 2c, les explications générées sont l'ensemble des instances totalement compatibles de l'autre classe, c'est-à-dire $I(e) = 0$. Elles sont obtenues pour $\alpha_P < 0.45$ pour l'instance x_0. Lorsque α_P augmente au-delà de ce seuil, le nombre d'explications générées diminue, comme l'illustrent les fig. 3b et 3c ($\alpha_P = 0.6$ et 0.66 respectivement). Ceci montre l'impact de la prise en compte du seuil α_P dans les intégrales de Gödel : même si les explications sont complètement compatibles, si elles ne satisfont pas la contrainte imposée par la pénalité, elles sont rejetées.

Sur la fig. 3f, identique à la fig. 2a, un seul exemple contre-factuel, qui correspond au point le plus proche de l'autre classe est généré, il a la pénalité la plus faible. Ce cas est obtenu lorsque la contrainte imposée par la pénalité est trop forte, lorsque α_P est trop élevé par rapport à l'incompatibilité. Dans ce cas, il est impossible de trouver une explication compatible, le processus d'optimisation se concentre donc sur la minimisation de la pénalité.

Les figures 3d et 3e représentent un compromis entre les cas extrêmes des figures 3c et 3f, c'est-à-dire des compromis entre la pénalité et l'incompatibilité. Nous illustrons ces cas avec un seuil de pénalité élevé, les exemples contre-factuels générés sont les instances les plus compatibles qui satisfont la contrainte de pénalité. Dans les figures représentées ici, au moins une des contraintes est satisfaite. La figure 2d représente la fonction maximale si aucune des contraintes n'est vérifiée ($\alpha_P > 0,9$ et $\beta_I > 0,95$). Ces valeurs sont associées à des contraintes très fortes. La figure 3e en est une variante, avec une plus grande tolérance sur la valeur de la pénalité.

Les résultats obtenus par $G_\mu^\rightarrow(P,I)$, omis par contrainte de place, montrent des comportements similaires, pour d'autres valeurs des paramètres, en raison de leur différence sémantique (cf section 4.3).



(a) α=0, β=0.2 (b) α=0.6, β=0.2 (c) α=0.66, β=0.2

(d) α=0.75, β=0.9 (e) α=0.8, β=0.95 (f) α=0.9, β=0.2

FIGURE 3 – Exemples contre-factuels minimisant $G_{\mu}^{\otimes}(P, I)$ pour différents α_P et β_I

6 Conclusion

Cet article propose d'appliquer les intégrales de Gödel pour combiner une évaluation subjective liée aux connaissances de l'utilisateur, avec l'évaluation objective, liée uniquement à la tâche d'apprentissage considérée. Il examine les propriétés requises d'une fonction d'agrégation dans ce contexte, présente les avantages des intégrales de Gödel et en propose une application innovante dans le domaine de l'XAI.

Des perspectives expérimentales incluent l'évaluation, par des utilisateurs réels, des explications générées, cruciale dans les études sur l'XAI. Cette évaluation pourrait également porter sur les propriétés retenues et examiner la pertinence de variantes, par exemple en imposant une monotonie stricte. Les perspectives théoriques incluent l'élicitation des paramètres de l'intégrale de Gödel, ainsi que l'étude d'autres intégrales, comme celle de Sugeno.

Références

[1] A. Artelt, B. Hammer. On the computation of counterfactual explanations - A survey. *arXiv preprint arXiv :1911.07749*, 2019.

[2] S. Dandl, C. Molnar, M. Binder, B. Bischl. Multi-Objective Counterfactual Explanations. *Parallel Problem Solving from Nature – PPSN XVI*, Springer, 2020.

[3] D. Dubois, H. Prade, A. Rico, B. Teheux. Generalized qualitative Sugeno integrals. *Inf. Sci.*. vol. 415, pp. 429–445, 2017.

[4] M. Drescher, A. Perera, C. Johnson, L. Buse, A. Drew, M. Burgman. Toward rigorous use of expert knowledge in ecological research. *Ecosphere*, vol. 4, no. 7, 2013.

[5] C. Frye, C. Rowat, I. Feige. Asymmetric Shapley values : incorporating causal knowledge into model-agnostic explainability. *Proc. of NeurIPS*, vol. 33, 2020.

[6] M. Grabisch, J.L. Marichal, R. Mesiar, E. Pap. Aggregation Functions. *Encyclopedia of Mathematics and its Applications*. Cambridge Univ. Press, no. 127, 2009

[7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi. A Survey of Methods for Explaining Black Box Models *ACM Comput. Surv., 2018*. vol. 51, no. 5.

[8] R. Guidotti. Counterfactual explanations and how to find them : literature review and benchmarking. *Data Mining and Knowledge Discovery, 2022*. pp. 1-55.

[9] A. Jeyasothy, T. Laugel, M.-J. Lesot, C. Marsala, M. Detyniecki. A general framework for personalising post hoc explanations through user knowledge integration. *Int. J. of Approximate Reasoning, IJAR*, vol. 160, 2023.

[10] A.-H. Karimi, G. Barthe, B. Schölkopf, I. Valera. A survey of algorithmic recourse : contrastive explanations and consequential recommendations. *ACM Computing Surveys (CSUR)*, ACM New York, 2022.

[11] M. T. Lash, Q. Lin, N. Street, J. G. Robinson, J. Ohlmann. Generalized Inverse Classification. *Proc. of the SIAM Int. Conf. on Data Mining*, 162–170, 2017.

[12] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, M. Detyniecki. Comparison-based Inverse Classification for Interpretability in Machine Learning. *Proc. of Int. Conf. on IPMU, 2018*. pp. 100–111. Springer

[13] D. Mahajan, C. Tan, A. Sharma. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. *NeurIPS workshop, 2019*.

[14] T. Calvo and G. Mayor and R. Mesiar. Aggregation Operators : New Trends and Applications. *Springer*, vol. 97, 2022.

[15] G. Navas-Palencia. Optimal counterfactual explanations for scorecard modelling. *arXiv preprint arXiv :2104.08619*, 2021

[16] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, P. Flach. FACE : Feasible and Actionable Counterfactual Explanations. *Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society, AIES*, 2020

[17] M. Sugeno. Theory of fuzzy integrals and its applications. *Tokyo Institute of Technology*, 1974.

[18] B. Ustun, A. Spangher, Y. Liu. Actionable Recourse in Linear Classification. *Proc. of the ACM Conf. on FAccT*, ACM, pp. 10-19, 2019.

[19] S. Verma, J. Dickerson, K. Hines. Counterfactual explanations for machine learning : A review. *arXiv preprint arXiv :2010.10596, 2020*.

[20] S. Wachter, B. Mittelstadt, C. Russell. Counterfactual Explanations without Opening the Black Box : Automated Decisions and the GDPR. *Harvard journal of law & technology, 2018*. vol. 31., pp. 841–887