Stratégies coalitionnelles pour une explication efficace des prédictions individuelles.

Elodie Escriva *,**, Gabriel Ferrettini **,
Julien Aligon **, Jean-Baptiste Excoffier *,
Chantal Soulé-Dupuy **

*Kaduceo, 31 Allée Jules Guesde, 31400 Toulouse, France
firstName.lastName@kaduceo.com
**Université de Toulouse-Capitole, IRIT, (CNRS/UMR 5505)
2 rue du Doyen-Gabriel-Marty 31042 Toulouse cedex 9, France
firstName.lastName@irit.fr

Résumé. Ce papier est un résumé des travaux publiés dans le journal Information Systems Frontiers (Ferrettini et al., 2021). Face aux nombreuses applications de l'apprentissage machine (ML) dans de nombreux domaines, la nécessité de comprendre le fonctionnement des modèles en boite noire est devenu croissante, particulièrement chez les non-experts. Plusieurs méthodes fournissant des explications sur les prédictions des modèles existent, avec des temps de calculs longs ou des hypothèses restrictives sur les interactions entre attributs. Ce papier détaille des méthodes basées sur la détection de groupes d'attributs pertinents – appelés coalitions – influençant la prédiction. Nos résultats montrent que les méthodes coalitionnelles sont plus performantes que celles existantes, comme SHAP. Le temps d'exécution est réduit en préservant la précision des explications. Ces méthodes permettent une augmentation des cas d'utilisation afin d'accroître la confiance entre les modèles ML, les utilisateurs et toute personne affectée par une décision impliquant ces modèles.

1 Introduction

Dans de nombreux domaines, l'absence d'explications des prédictions et le manque de transparence des modèles "boite noire" est un frein à l'utilisation de l'apprentissage automatique. La majorité des algorithmes de la littérature ne fournit pas d'informations claires sur la production de prédictions. Cependant, plusieurs techniques agnostiques ont été décrites pour comprendre les prédictions individuelles des modèles (Strumbelj et Kononenko, 2010; Ribeiro et al., 2016; Casalicchio et al., 2018). Les méthodes agnostiques sont applicables à tous types de modèles, indépendamment de leur structure interne, contrairement aux méthodes spécifiques à un seul type de modèles. Pour l'explication d'une prédiction individuelle, les valeurs de Shapley sont une base intéressante pour calculer les influences de chaque attribut sur cette prédiction. Cette technique se base sur la théorie des jeux collaboratifs en considérant les attributs comme des joueurs

participant à une prédiction (Strumbelj et Kononenko, 2010). La méthode explicative se basant sur les valeurs de Shapley est appelée *méthode complète*. Cette méthode est cependant très coûteuses à calculer, avec une complexité exponentielle. Une première approximation de la *méthode complète* est la méthode *k-depth* qui réduit le temps de calcul et la complexité en considérant seulement les groupes d'attributs ayant une taille inférieure à un paramètre k (Ferrettini et al., 2020a). Enfin, une technique populaire, SHapley Additive explanations (SHAP), approxime les valeurs de Shapley en reproduisant localement les comportements du modèle par des modèles linéaires (Lundberg et Lee, 2017). La méthode *KernelSHAP* est utilisable pour tous les modèles et plusieurs autres variantes sont proposées pour des modèles spécifiques, telles que *LinearSHAP* pour les modèles linéaires, *TreeSHAP* pour les modèles basés sur des arbres de décisions et *DeepSHAP* pour les réseaux neuronaux. Ces méthodes restent cependant longues à calculer (Van den Broeck et al., 2021) et se basent sur des hypothèses restrictives comme la linéarité locale qui ne prend pas en compte les dépendances entre attributs.

Même optimisées, toutes ces techniques restent délicates à mettre en place dans un contexte plus large d'analyses prédictives et interactives, où les temps de réponses sont bien essentiels. Ainsi, l'intérêt de nos travaux porte sur l'obtention d'explications calculées de manière efficace, autant en précision qu'en temps de calcul. Nous proposons une technique d'amélioration de la *méthode complète*, en réduisant son temps de calcul, tout en conservant une précision importante, quel que soit le modèle prédictif choisi. A cette fin, nous nous inspirons des techniques de la littérature portant sur la sélection d'attributs et nous permettant de sélectionner les coalitions d'attributs les plus pertinentes. Ces coalitions sont ensuite intégrés dans notre adaptation du calcul des valeurs de Shapley afin d'obtenir les explications correspondant aux prédictions.

Ce papier est organisé comme suit. La section 2 décrit nos propositions de sélection de coalitions d'attributs basées respectivement sur l'analyse en composante principale (PCA), le facteur d'inflation de variance (VIF) et la corrélation de Spearman. Cette section formalise également notre adaptation du calcul des valeurs de Shapley. La section 3 compare nos propositions à celles de la littérature, à savoir LIME, SHAP et K-depth, appliquées à 243 jeux de données et 2 modèles prédictifs (Random Forest et SVM). Les résultats montrent que notre stratégie coalitionnelle surpasse celles de la littérature, en particulier lors de l'utilisation d'un modèle SVM. La section 4 conclut ce papier et propose de nouvelles perspectives de travail.

2 Méthode d'explication coalitionnelle

L'une des faiblesses de la littérature dans le domaine de l'explicabilité porte sur la non prise en compte des interactions entre attributs. Une stratégie possible est d'identifier les groupes d'attributs corrélés au moyen d'une méthode de coalition. Ne considérer que ces groupes d'attributs pertinents permet alors de réduire la complexité liée au calcul des valeurs de Shapley pour l'obtention d'explications. Nous proposons plusieurs algorithmes de groupement d'attributs se basant sur l'Analyse en Composantes Principales (PCA), le facteur de corrélation de Spearman (Spearman) et le facteur d'inflation de la variance (VIF). Ces méthodes sont inspirées des méthodes de sélection d'attributs utilisées dans le pré-processing des jeux de données. La multiplication des méthodes

permet de proposer un panel divers de groupes d'attributs pouvant s'appliquer à différents types de jeu de données. Pour chaque algorithme, un paramètre *alpha* permet d'être plus ou moins permissif sur l'ajout d'un attribut dans un coalition donnée. Cela permet, in fine, de contrôler la taille des groupes obtenus et de jouer sur la complexité du calcul des explications. Le rôle du seuil *alpha* est détaillé explicitement dans la section 4 de l'article Ferrettini et al. (2021).

2.1 Création de coalitions d'attributs

Coalition basée sur PCA. La PCA consiste à transformer les attributs corrélés entre eux en des nouvelles variables non-corrélées. L'objectif est de réduire le nombre de variables au maximum afin que l'information ne soit pas redondante. En utilisant cette approche, nous considérons les attributs servant à créer une nouvelle variable comme un groupe pertinent d'attribut pour la création d'explications.

Coalition basée sur VIF. Le facteur VIF est une métrique estimant la multicolinéarité des attributs. En calculant le facteur VIF de chaque attribut, il est possible de détecter des groupes en comparant les facteurs VIF obtenus avant et après retrait de l'attribut. Il est alors possible de considérer deux options : les attributs sont regroupés lorsqu'ils ont une forte colinéarité entre eux (lorsque la suppression d'un attribut diminue fortement le facteur VIF) ou regrouper des attributs non-colinéaire. Nous appelons respectivement ces deux options les VIF coalitions et les *Reverse*-VIF coalitions.

Coalition basée sur Spearman. Une limite du facteur VIF est la prise en compte seulement des corrélations linéaires. La coefficient de Spearman résout ce problème en calculant la corrélation entre toutes les paires d'attributs. A partir de la matrice des coefficients de Spearman, il est possible de déterminer des groupes d'attributs. Idem que pour VIF, il y a deux possibilités pour former des groupes : prioriser les groupes avec des attributs ayant une forte corrélation ou des attributs non-corrélés. Ces méthodes sont appelées Spearman coalition et *Reverse*-Spearman coalition.

2.2 Explications coalitionnelles

Dans la *méthode complète*, la complexité provient principalement du nombre exponentiel de combinaisons d'attributs. La méthode coalitionnelle est une adaptation des valeurs de Shapley qui considère uniquement les groupes d'attributs pertinents, réduisant ainsi la complexité tout en minimisant la perte de précision par rapport à la méthode complète Ferrettini et al. (2020b). L'influence d'un attribut sur la prédiction du modèle, pour une instance spécifique, est ainsi calculée de la manière suivante :

$$coal\mathcal{I}_{a_i}^{C}(x) = \sum_{g' \subseteq g \setminus a_i,\ g \in G_{a_i}} \frac{|g'|! * (|g| - |g'| - 1)!}{\sum_{g \in G_{a_i}} |g|!} * (\Delta_{C,(g' \cup a_i)}(x) - \Delta_{C,g'}(x)) \quad (1)$$

$$\Delta_{C,S}(x) = f^*_{C,S}(x) - f^*_{C,\varnothing}(x)$$

où a_i est un attribut du jeu de données, g une coalition d'attributs, G l'ensemble des coalitions, G_{a_i} l'ensemble des groupes de G contenant a_i et Δ la différence entre la

prédiction attendue pour S et la prédiction attendue en l'absence de données avec $f_{C,S}(x)$ la fonction de classification sur le groupe d'attribut S pour l'instance x et la classe C prédite par le modèle. La pénalisation liant les cardinalités de g et g' est une adaptation de la pénalisation des valeurs de Shapley permettant de prendre en compte la redondance possible des attributs dans plusieurs coalitions.

Afin de contrôler la complexité du calcul des méthodes de coalition, un algorithme de bissection permet de déduire le seuil *alpha* itérativement en fonction d'un pourcentage de complexité. Ce pourcentage correspond au ratio entre le nombre de combinaisons d'attributs présentes dans les coalitions et le nombre total de combinaisons avec tous les attributs. Le zéro de la bissection est défini comme le moment où le nombre de combinaisons d'attribut pour un *alpha* est égal au nombre de combinaisons voulu par l'utilisateur via la complexité. Si l'on souhaite un temps de calcul court, il est possible de fixer une proportion faible (par exemple 10%), tandis qu'une valeur plus élevée (par exemple 50%) peut être utilisée lorsque des résultats plus précis sont nécessaires sans contrainte de temps. Ce pourcentage permet également de simplifier le choix du paramètre *alpha*, qui est très dépendant de la méthode de coalition et du jeu de données.

3 Expériences

Dans cette section, nous souhaitons évaluer l'approche coalitionnelle par rapport aux autres approches de la littérature, en prenant comme référence la *méthode complète*. Les méthodes sont départagées par rapport à leur temps d'exécution et leur précision par rapport à la base de référence. Les méthodes coalitionnelles sont d'abord comparées entre elles avant de comparer les plus prometteuses aux méthodes de la littérature. En raison du manque de place, nous ne présentons qu'une partie des résultats obtenus dans (Ferrettini et al., 2021). Des résultats supplémentaires sont disponibles à la section 5 de l'article complet, comme une comparaison des modèles utilisés, une caractérisation des coalitions d'attributs et une étude de cas.

3.1 Protocole expérimental

Les tests ont été réalisés sur une collection de 243 jeux de données de classification [1] accessibles sur la plateforme OpenML (Vanschoren et al., 2013). Le tableau 1 décrit, par nombre d'attributs, le nombre de jeux de données ainsi que le nombre moyen d'instances. Le nombre d'instances est très variable en fonction du nombre d'attributs et influe directement sur le temps d'exécution. Afin d'avoir des temps comparables, les temps d'exécution pour chaque jeu de données sont normalisés en divisant par le nombre d'instances pour obtenir un temps par instance. Le temps de calcul des influences via la *méthode complète* étant très long, nous avons sélectionné uniquement les jeux de données ayant au maximum 9 attributs afin de conserver un temps de calcul raisonnable, sans sélection d'attributs pour réduire la dimension des jeux de données.

Pour les tests, nous avons utilisé les modèles Random Forest (RF) et Support Vector Machine (SVM) avec un noyau non-linéaire Radial-Basis-Function (RBF) fournis par la librairie Scikit-Learn. Pour chaque modèle, paramètre *alpha*, pourcentage de complexité

1. Disponible sur `https://www.openml.org/s/107/tasks`

Nb d'attributs	1	2	3	4	5	6	7	8	9
Nb de jeux de données	3	21	44	25	38	26	34	28	24
Nb moyen d'instances	724	736	1688	560	843	600	456	750	479

TAB. 1 – *Statistiques sur les jeux de données utilisés par nombre d'attributs*

et jeux de données, nous avons calculé les influences des attributs en utilisant notre méthode coalitionnelle ainsi que les méthodes de la littérature : complète, *k-depth*, Kernel SHAP et Tree SHAP [2]. Les tests ont été effectués sur un processeur AMD Ryzen 3700 avec 8 x 3.6 GHz cores et 32 GB of RAM.

Concernant les algorithmes de coalitions, les coalitions sont construites avec les méthodes de la section 2 selon trois pourcentages de complexité pour la bissection : $\{10\%, 25\%, 50\%\}$. Pour comparer les méthodes, la méthode complète est utilisée comme base de référence. L'erreur entre une méthode d'intérêt et la référence est calculée à partir de la distance euclidienne normalisée :

$$err(\mathcal{I}, x) = d(\mathcal{I}^C(x), complete\mathcal{I}^C(x)) \tag{2}$$

$$d(i, j) = \frac{1}{2\sqrt{n}} \sum_{k=1}^{n} \sqrt{(i_k - j_k)^2} \tag{3}$$

où n est le nombre d'attributs, x une instance du jeu de données, $complete\mathcal{I}$ la méthode complète, $\mathcal{I}^C(x) = [i_1, ..., i_n]$ le vecteur des influences de l'instance x pour la classe C prédite par le modèle via la méthode $\mathcal{I}$. L'erreur est calculée pour chaque méthode comme la somme des erreurs sur toutes les instances d'un jeu de données, puis moyennée sur tous les jeux de données. Ainsi, une faible erreur indique une estimation plus précise par rapport à la *méthode complète*. Comme indiqué précédemment, le temps de calcul est aussi pris en compte lors des expérimentations. Ce temps comprend le calcul des influences pour toutes les méthodes, additionné du temps de calcul des coalitions pour la méthode coalitionnelle. Le temps moyen par instance est alors obtenu en moyennant le temps par instance sur tous les jeux de données.

3.2 Résultats

Afin de visualiser les résultats d'erreurs par rapport à la *méthode complète* et de rapidité, nous utilisons les résultats moyens sur les deux modèles - RF et SVM - indépendamment du nombres d'attributs afin de construire une carte de performance. Le temps de calcul est normalisé par rapport à la *méthode complète* sur l'axe horizontal et l'erreur est présentée sur l'axe vertical. La figure 1 présente une comparaison de l'erreur et du temps de calcul uniquement pour les méthodes coalitionnelles. Toutes les méthodes coalitionnelles sont placées au dessus de la *méthode complète*, signifiant une perte de précision pour toutes ces méthodes. Elles sont aussi à gauche de la méthode complète, signifiant que leurs temps de calcul moyen par instance sont inférieurs à celui

2. Les codes utilisés sont disponibles aux liens suivants : `https://github.com/slundberg/shap` et `https://github.com/kaduceo/coalitional_explanation_methods`

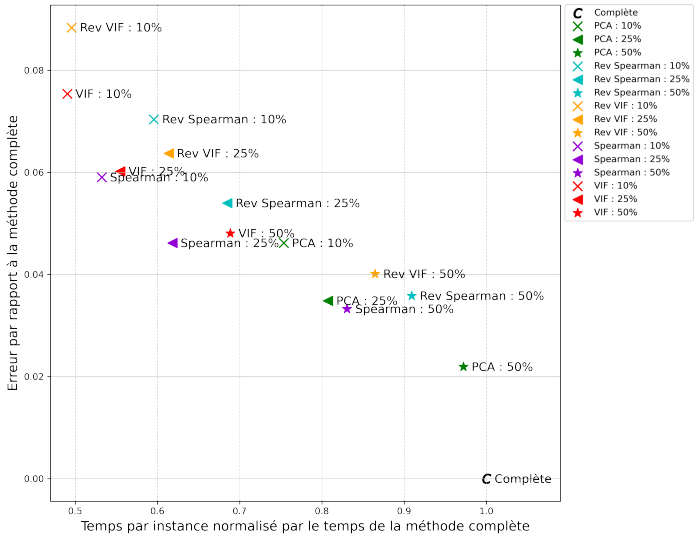

FIG. 1 – *Carte de performance moyenne des méthodes coalitionnelles pour RF et SVM.*

de la *méthode complète*. Globalement, et logiquement, un faible pourcentage de complexité donne des résultats plus rapides mais avec une erreur plus importante. Parmi les quatre méthodes les plus rapides, Spearman 10% est la méthode ayant l'erreur la plus faible, alors que PCA 50% a une erreur faible avec un temps de calcul proche de la *méthode complète*. Spearman 25% et 50% et PCA 25% semblent les méthodes les plus équilibrées entre précision et rapidité. Globalement, les algorithmes PCA et Spearman offrent de meilleurs résultats que les algorithmes *Reverse*-Spearman, *Reverse*-VIF et VIF.

La figure 2 compare les méthodes coalitionnelles Spearman et PCA avec les méthodes de la littérature *k-depth*, Kernel SHAP, Tree SHAP et la *méthode complète*. Pour Tree SHAP, les résultats sont calculés uniquement avec le modèle Random Forest. La méthode Kernel SHAP est très longue à calculer avec une erreur élevée, écrasant les résultats de toutes les autres méthodes à gauche du graphique. En enlevant cette méthode du graphique, la figure montre que les algorithmes PCA 10% et Spearman 10% et 25% ont des résultats similaires à la *3-depth* et *4-depth*, Spearman étant à chaque fois plus précis. PCA 25% et Spearman 50% ont des résultats proches de la *4-depth*. Tree SHAP a les moins bons résultats parmi les méthodes de la littérature, avec un temps de calcul plus de deux fois celui de la *méthode complète* et une erreur élevée.

Globalement, les méthodes coalitionnelles sont plus rapides et précises que les méthodes de la littérature et semblent être des alternatives prometteuses à la méthode complète, notamment pour les jeux de données complexes où l'interdépendance des attributs est plus présente. La réduction du temps de calcul semble également permettre d'utiliser la méthode coalitionelle sur des jeux de données de plus de 9 attributs, ce qui est difficilement possible avec la *méthode complète* ou *SHAP*.

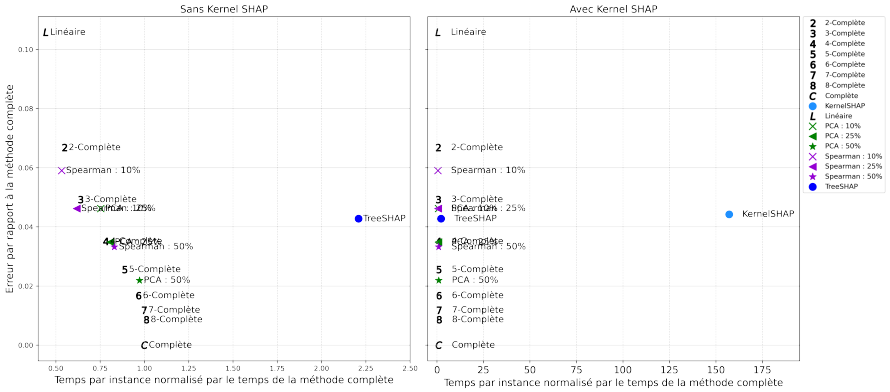

FIG. 2 – *Carte de performance des méthodes coalitionnelles Spearman et PCA contre les méthodes de la littérature pour les modèles RF et SVM.*

4 Conclusion

Ce papier présente une nouvelle méthode d'explication des modèles prédictifs, la méthode coalitionnelle, qui se base sur les valeurs de Shapley. Cette méthode résout les limitations des méthodes de la littérature quant à l'interdépendance des attributs grâce à la recherche de groupes pertinents d'attributs. Les expérimentations comparant la méthode coalitionnelle aux méthodes de la littérature, *SHAP* et *k-depth*, montrent une baisse de la complexité et du temps de calcul grâce à la méthode coalitionnelle, tout en conservant une précision acceptable. Les méthodes de coalition *PCA*, *Spearman* et *Reverse VIF* montrent les résultats les plus prometteurs. La méthode coalitionnelle est notamment plus performante pour les jeu de données complexes, où l'interdépendance des attributs est plus susceptible d'être présente. Cependant, un des avantages de *SHAP* par rapport aux méthodes coalitionelles est l'absence de ré-entrainement des modèles, même si la création de perturbations pour simuler l'absence des attributs peut être coûteux. Un point restant à évaluer serait de comparer notre méthode et la littérature sur des jeux de données avec une dimension supérieur à 9 attributs.

Aussi, un important axe de travail futur est de confronter les explications générées par ces techniques, notamment la méthode coalitionnelle, à des utilisateurs dans des cadres réels. Ceci afin de mesurer le niveau d'intérêt et l'apport des ces explications dans un processus d'aide à la décision.

Une perspective à plus long terme est également de prendre en compte le contexte dans lequel l'analyse prédictive est menée. En effet, les explications fournies pour une instance particulière ne peuvent être totalement satisfaisantes pour tout utilisateur dans toutes les situations. Le degré d'expertise de l'utilisateur semble très importante à considérer : un utilisateur expert sera certainement plus intéressé par des explications très précises par rapport à un novice. De plus, le processus d'analyse peut avoir un impact sur le type d'explications à considérer. Par exemple, les explications ne doivent pas être analysées de la même manière si l'analyse est réalisée de manière exploratoire ou confirmatoire.

Stratégies coalitionnelles pour une explication efficace des prédictions individuelles.

Références

Casalicchio, G., C. Molnar, et B. Bischl (2018). Visualizing the feature importance for black box models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 655–670. Springer.

Ferrettini, G., J. Aligon, et C. Soulé-Dupuy (2020a). Explaining single predictions : A faster method. In *SOFSEM 2020 : Theory and Practice of Computer Science*, Lecture Notes in Computer Science, pp. 313–324. Springer International Publishing.

Ferrettini, G., J. Aligon, et C. Soulé-Dupuy (2020b). Improving on coalitional prediction explanation. In *Advances in Databases and Information Systems*, ADBIS 2020. Lecture Notes in Computer Science, vol. 12245, pp. 122–135. Springer, Cham.

Ferrettini, G., E. Escriva, J. Aligon, J.-B. Excoffier, et C. Soulé-Dupuy (2021). Coalitional strategies for efficient individual prediction explanation. *Information Systems Frontiers*.

Lundberg, S. M. et S.-I. Lee (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, Volume 30, pp. 4765–4774. Curran Associates, Inc.

Ribeiro, M. T., S. Singh, et C. Guestrin (2016). "Why should I trust you ?" : Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Strumbelj, E. et I. Kononenko (2010). An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research 11*, 1–18.

Van den Broeck, G., A. Lykov, M. Schleich, et D. Suciu (2021). On the tractability of SHAP explanations. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

Vanschoren, J., J. N. van Rijn, B. Bischl, et L. Torgo (2013). Openml : Networked science in machine learning. *SIGKDD Explorations 15*(2), 49–60.

Summary

As Machine Learning (ML) is now widely applied in many domains, in both research and industry, the need to understand how black-box algorithms work has grown, especially among non-experts. Several approaches had thus been developed to provide clear insights of a model prediction for a particular observation but at the cost of long computation time or restrictive hypothesis that does not fully take into account interaction between attributes. This paper provides methods based on the detection of relevant groups of attributes -named coalitions- influencing a prediction. Compared to the literature, our results show that coalitional methods outperform existing ones such as SHapley Additive exPlanation (SHAP). Computation time is shortened while preserving an acceptable accuracy of individual prediction explanations. Therefore, this enables wider practical use of explanation methods to increase trust between developed ML models, end-users, and whoever impacted by any decision where these models played a role.