# Vers des explications de type "post hoc" pour les systèmes de recommandation

Willeme Verdeaux*[1] and Nicolas Labroche†[1]

[1]Laboratoire dÍnformatique Fondamentale et Appliquée de Tours – Université de Tours : EA6300, Institut National des Sciences Appliquées - Centre Val de Loire, Institut National des Sciences Appliquées, Centre National de la Recherche Scientifique – France

**Résumé**

Les modèles post-hoc sont des algorithmes de type " boite blanche " qui visent à substituer localement un modèle complexe, généralement de type boîte noire. Ces méthodes simples (modèles post-hoc) sont ensuite, avec un peu de formatage, présentées à l'utilisateur final et sont employés directement en tant qu'explication de la prédiction faite par le modèle complexe. Dans le contexte général de la classification, il a été montré que ces algorithmes de substitutions ne peuvent pas toujours être en mesure de capturer une explication locale, c'est-à-dire être spécifique à une prédiction d'instance donnée. En effet, ils ont plutôt tendance à traduire davantage un comportement général de la boîte noire. Ce problème est encore plus complexe dans un scénario de recommandation où les classes et les frontières de décision ne sont pas explicitement définies et où les données sont par nature très creuses. Nous avons montré dans des travaux que nous avons soumis récemment dans un journal international, qu'il est possible d'aborder ces problèmes avec un échantillonnage efficace autour de l'instance de recommandation à expliquer, ce qui permet d'apprendre un modèle de substitution local approprié. Dans cette optique, nos travaux introduisent aussi de nouvelles approches pour capturer cette localité autour d'une instance à expliquer dans le contexte des systèmes de recommandation. De plus, et contrairement aux travaux précédents, nous montrons également qu'il est possible de réaliser un modèle d'explication simple, mais de meilleure qualité basé sur une fonction de " perte " complexe et expressive telle que celle utilisée dans les RankNets. Nos expériences montrent que nos méthodes sont aussi précises que les méthodes de la littérature en termes de fidélité du modèle de substitution à la boîte noire, mais sont beaucoup plus efficaces pour récupérer localement des caractéristiques explicables significatives.

**Mots-Clés:** Système de recommandation, Explication, Modèle surrogate, Modèle de substitution, XAI

*Intervenant

†Auteur correspondant: nicolas.labroche@univ-tours.fr