Découverte de labels dupliqués par l'exploration du treillis des classifieurs binaires

Quentin Labernia*, Victor Codocedo*, Mehdi Kaytoue*, Céline Robardet*

*Université de Lyon, CNRS, INSA-Lyon, LIRIS UMR5205, F-69621, France
prenom.nom@insa-lyon.fr

Résumé. L'analyse des données comportementales représente aujourd'hui un grand enjeu. Tout individu génère des traces d'activité et de mobilité. Lorsqu'elles sont associées aux individus, ou labels, qui les ont créées, il est possible de construire un modèle qui prédit avec précision l'appartenance d'une nouvelle trace. Sur internet, il est cependant fréquent qu'un utilisateur possède différentes identités virtuelles, ou labels doublons. Les ignorer provoque une grande réduction de la précision de l'identification. Il est ainsi question dans cet article du *problème de déduplication de labels*, et l'on présente une méthode originale basée sur l'exploration du treillis des classifieurs binaires. Chaque sous-ensemble de labels est classifié face à son complémentaire et des contraintes rendent possible l'identification des labels doublons en élaguant l'espace de recherche. Des expérimentations sont menées sur des données issues du jeu vidéo STARCRAFT 2. Les résultats sont de bonne qualité et encourageants.

1 Introduction

Les capteurs sont ancrés dans la vie quotidienne. Cachés dans les voitures, les smartphones, les objets connectés, ils enregistrent une multitude de mesures. Ces capteurs, qu'ils soient autonomes ou intégrés à un système plus complexe, génèrent des données comportementales riches. Correctement analysées, elles participent à la résolution de divers défis industriels et à la création de services et applications pour le grand public.

On s'intéresse ici à une technique d'identification d'individus se basant sur de données comportementales. De telles méthodes sont utiles d'un point de vue sécuritaire (détection de fraudes ou d'usurpations) ou de celui de la gestion des données privées (évaluation de techniques d'anonymisation de données). Il est vérifié dans de nombreux domaines qu'un individu peut être reconnu via les traces qu'il a générées : il est par exemple possible d'identifier de manière unique un individu à l'aide de quelques points d'intérêt dans l'espace et le temps (De Montjoye et al. (2013)). Encore, la manière d'interagir avec le clavier permet la reconnaissance d'un individu écrivant son mot de passe (Peacock et al. (2004)) ou encore jouant à un jeu vidéo (Yan et al. (2015)).

De nombreux utilisateurs possèdent de multiples identités virtuelles, appelées par la suite *labels doublons*, dont les relations sont à priori inconnues. Certaines problématiques de ciblage marketing concernent l'association de cookies provenant de périphériques différents à un