Vers un partitionnement des données à partir d'une forêt d'isolation

Véronne Yepmo*, Grégory Smits**, Marie-Jeanne Lesot***, Olivier Pivert*

* Université de Rennes 1 - IRISA - UMR 6074 - Lannion, France
{veronne.yepmo-tchaghe, olivier.pivert}@irisa.fr,
** IMT Atlantique - Lab STICC - UMR 6285 - Brest, France
gregory.smits@imt-atlantique.fr,
*** Sorbonne Université - LIP6 - Paris, France
marie-jeanne.lesot@lip6.fr

Résumé. Cet article effectue un pas vers une extraction d'explications contrastives entre anomalies et structure intrinsèque des points réguliers. Il propose une variante de l'algorithme des forêts d'isolation ayant pour objectif principal la préservation de la structure des données régulières en vue de sa reconstitution plus aisée. Les expérimentations menées sur des jeux de données synthétiques montrent que cette variante des forêts d'isolation détériore moins la structure des données régulières que la méthode classique. Par conséquent, la première citée peut servir de base pour une approche unifiée de détection et d'explication d'anomalies.

1 Introduction

Contrairement à la détection d'anomalies qui a été intensivement explorée dans la littérature, l'explication d'anomalies reste un sujet ouvert. Même si des travaux récents ont essayé de combler le vide (Kopp et al., 2020; Mokoena et al., 2022), il a été précisé dans Yepmo et al. (2022) que les explications d'anomalies les plus détaillées, c'est-à-dire celles prenant en compte la structure des données régulières, manquent de références. Celles-ci expliquent les anomalies détectées par rapport à un/des groupe(s) de données régulières, et non comme des points isolés du reste des données. Il est possible d'extraire ce type d'explications à l'aide d'un pipeline. Ce dernier consisterait alors premièrement en la détection d'anomalies à l'aide d'un algorithme dédié, suivie d'un partitionnement des données régulières à l'aide d'un algorithme de clustering, puis d'une identification des anomalies relativement à chaque cluster de données régulières et finalement en la génération d'explications contextuelles. Le travail proposé dans cet article suggère l'encapsulation des différentes étapes du pipeline sous une méthode unifiée ayant pour base la forêt d'isolation ou FI (Liu et al., 2012) qui est un algorithme de détection d'anomalies.

Une forêt d'isolation est un ensemble d'arbres binaires construits chacun en partitionnant récursivement et aléatoirement l'espace de données. Chaque arbre d'isolation est construit sur un échantillon différent du jeu de données, avec l'hypothèse qu'une anomalie, qui est par définition un point rare et distant des autres points dits réguliers, se trouvera isolée dans un