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

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

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 sous-espace assez rapidement. Un point régulier quant à lui se retrouvera rarement isolé dans un sous-espace, ou au mieux sera isolé après de nombreuses coupes/séparations aléatoires. Avec ce partitionnement complètement aléatoire, l'information structurelle des données est perdue. En effet, deux points très proches dans leur espace de définition se retrouvent très fréquemment dans des feuilles différentes d'un arbre d'isolation. Récupérer cette information structurelle après la construction d'une forêt d'isolation serait alors une tâche ardue. L'objectif de la méthode proposée dans cet article, nommée RIFIFI (*Revised Isolation Forest to Identify Fraud and the data Inner structure*), est de préserver au mieux l'information structurelle lors de la construction de la forêt d'isolation, afin de reconstituer les clusters de points réguliers. Pour ce faire, le processus de sélection des séparations est revisité et n'est plus complètement aléatoire, mais plutôt guidé par la volonté de préserver au maximum la proximité entre les points appartenant au même groupe de données.

Après une revue de la littérature concernant la détection et l'explication d'anomalies dans la section 2, l'algorithme des forêts d'isolation sera rappelé dans la section 3. Notre méthode sera décrite dans la section 4. Puis, des premières expérimentations montreront la pertinence de l'approche pour construire une partition des données régulières (section 5). Les perspectives de ce travail seront finalement présentées dans la section 6.

2 Etat de l'art

2.1 Détection d'anomalies

La détection d'anomalies en apprentissage automatique peut être un problème supervisé, semi-supervisé ou non supervisé. Le cas non supervisé est le plus attrayant à cause du caractère imprévisible des anomalies et de la difficulté à étiqueter des jeux de données. Local Outlier Factor (LOF) (Breunig et al., 2000), les One-Class Support Vector Machines (Amer et al., 2013) et les forêts d'isolation (Liu et al., 2012) sont parmi les méthodes non supervisées les plus populaires. La dernière méthode citée est particulièrement attirante pour la détection d'anomalies, car elle est rapide, possède peu d'hyperparamètres, ne requiert pas de calcul de distance entre paires de points et est interprétable à l'échelle d'un arbre. Plusieurs variantes des forêts d'isolation ont été proposées dans la littérature. Certaines variantes se focalisent sur le calcul du score d'anomalie, mais maintiennent intact le processus de construction des arbres de la forêt. C'est le cas de Mensi et Bicego (2021) où cinq nouvelles fonctions pour le calcul des scores d'anomalie sont proposées. D'autres en revanche modifient la construction des arbres mais pas le calcul des scores. Dans Liu et al. (2010) et Hariri et al. (2019), des séparations obliques sont utilisées, mais avec des objectifs différents : la détection des clusters d'anomalies pour le premier et l'amélioration de la consistance des scores pour le second. Dans Cortes (2021), les séparations ne sont plus complètement aléatoires et ont pour objectif de minimiser l'écart-type pondéré induit par chaque séparation. La méthode proposée dans cet article produit également des séparations non complètement aléatoires, mais l'objectif recherché est la préservation de la structure des points réguliers.

2.2 Explication d'anomalies

L'explication d'anomalies a reçu moins d'attention dans la littérature que l'explication des classifieurs. Pourtant, à cause de la nature diverse des anomalies, l'explication d'anomalies mérite un traitement particulier. Dans Yepmo et al. (2022), quatre catégories d'explications ont été identifiées : l'explication par importance d'attributs, l'explication par valeurs d'attributs, l'explication par comparaison de points et l'explication par analyse de la structure intrinsèque des données. L'autre travail faisant un état de l'art de l'explication d'anomalies, Panjei et al. (2022), fait une distinction entre les catégories d'explications suivantes : les méthodes proposant un classement des anomalies, celles révélant les relations de cause à effet entre anomalies, et enfin celles identifiant les attributs responsables de l'anormalité des points ou des groupes de points. Dans les deux cas, il est précisé que les techniques trouvant les attributs marginaux sont les plus fréquentes dans la littérature (Gupta et al., 2018; Mokoena et al., 2022). Alors que les explications par comparaison de points se concentrent sur deux points du jeu de données, que les explications révélant des relations de cause à effet se concentrent sur les anomalies détectées, les explications par analyse de la structure intrinsèque offrent une vue globale sur l'anomalie à expliquer par rapport au jeu de données, et sont donc plus détaillées.

3 L'algorithme des forêts d'isolation

Dans cette section, l'algorithme des forêts d'isolation est rappelé.

Chaque arbre d'une forêt d'isolation est construit sur un échantillon tiré aléatoirement du jeu de données. À chaque étape de la construction d'un arbre d'isolation (voir algorithme 1), un attribut a puis une valeur v dans l'intervalle de valeurs de a sont sélectionnés aléatoirement. Les points ayant une valeur inférieure à v sur l'attribut a sont transférés vers le fils gauche du noeud courant, et les autres vers le fils droit. Le processus est répété récursivement à partir de la racine de l'arbre qui contient toutes les données de l'échantillon, jusqu'à ce que l'une des deux conditions suivantes soit remplie :

— le noeud n'est plus séparable (il contient un seul point) ;
— la profondeur limite d'un arbre, paramètre prédéfini de la méthode, est atteinte.

L'algorithme possède les hyperparamètres suivants : le nombre d'arbres dans la forêt T, la taille d'un échantillon Ψ et la profondeur limite d'un arbre h_{lim}. Les valeurs par défaut de ces hyperparamètres sont les suivantes : $T = 100$, $\Psi = 256$ et $h_{lim} = 8$. Les autres notations utilisées tout au long de l'article sont :

— $\mathcal{D}$ le jeu de données,
— $\mathcal{A}$ l'ensemble des attributs,
— x un point et $x.a$ sa valeur sur l'attribut $a \in \mathcal{A}$.

Dans l'algorithme 1, la méthode $noeud(fils\_gauche, fils\_droit, D, d, a, v)$ renvoie un nouveau noeud contenant les points appartenant à D, situé à la profondeur d, ayant pour séparation la droite d'équation $a = v$, pour fils gauche *fils_gauche* et pour fils droit *fils_droit*.

La figure 1a montre un exemple de jeu de données en dimension 2 ainsi que les séparations (en noir) d'un arbre. Chaque sous-espace encadré par des séparations est une feuille de l'arbre. Les anomalies se retrouvent assez rapidement isolées dans leurs feuilles respectives, et les clusters sont très souvent traversés par des coupes. Ce dernier constat implique que les points appartenant au même cluster se retrouvent fréquemment dans des feuilles différentes.

Algorithm 1 Forêt d'isolation classique : $construire\_arbre$

Entrées : un échantillon $D \subset \mathcal{D}$, la profondeur d du noeud courant
Sortie : un noeud d'un arbre d'isolation
if $|D| = 1$ ou $d > h_{lim}$ **then**
    Renvoyer $noeud(null, null, D, d, null, null)$ ▷ Feuille (noeud externe)
else
    $a \leftarrow random(\mathcal{A})$ ▷ Sélection aléatoire d'un attribut
    $v \leftarrow random(range(a))$ ▷ Sélection aléatoire d'une valeur
    $D_l \leftarrow \{x \in D / x.a < v\}$
    $D_r \leftarrow \{x \in D / x.a \geq v\}$
    Renvoyer
$noeud(construire\_arbre(D_l, d+1), construire\_arbre(D_r, d+1), D, d, a, v)$ ▷ Noeud interne
end if

4 RIFIFI

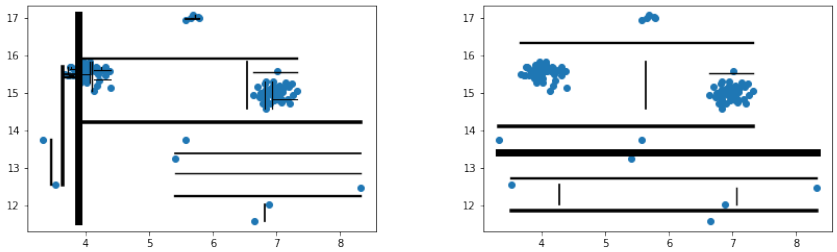
4.1 Principe

RIFIFI diffère des forêts d'isolation classiques sur la génération des séparations. Tandis que les forêts d'isolation classiques utilisent des séparations complètement aléatoires (section 3), RIFIFI possède un critère de conservation des séparations basé sur la densité du sous-espace au voisinage de celles-ci. L'hypothèse est la suivante : si un nombre important de points se retrouve dans le voisinage de la coupe, elle est potentiellement en train de séparer un cluster. Une autre coupe doit donc être générée. Le but recherché est l'encadrement des clusters de points réguliers par les séparations, de manière à ce que certaines feuilles contiennent un cluster, ou une portion importante de cluster. Deux nouveaux hyperparamètres sont introduits en plus des hyperparamètres de la méthode classique : la taille de la marge α autour de la séparation qui représente son voisinage, et le seuil de densité η. Si η points tombent dans la marge autour de la séparatrice, elle est écartée.

L'impact de ce critère sur la procédure d'isolation est illustré sur la figure 1. Avec ce nouveau critère, les séparations traversent plus rarement les clusters de points, et les anomalies se retrouvent isolées. Toutefois, étant donné qu'un échantillonnage est effectué lors de la construction des arbres, des séparations peuvent tout de même traverser des clusters de points. Dans ce cas, les feuilles d'un arbre RIFIFI contiennent des portions de clusters qui sont des groupes de points inséparables. C'est pourquoi les informations de chaque arbre de la forêt doivent être combinées.

4.2 Algorithme

L'algorithme 2 présente les détails de RIFIFI. Pour éviter de générer des séparations dans des intervalles qui ont déjà été écartées parce que beaucoup de points s'y trouvaient, l'ensemble des intervalles testés est stocké. Si la méthode n'a pas pu trouver de séparation valide dans tout l'intervalle de valeurs d'un attribut, cet attribut est écarté. Les attributs écartés sont donc également stockés. Si la méthode ne parvient pas à trouver de séparation valide peu importe



(a) Séparations d'un arbre d'une forêt d'isolation classique

(b) Séparations d'un arbre d'une forêt RIFIFI

FIG. 1 – Exemples de séparations (en noir) d'un arbre : FI VS RIFIFI. L'épaisseur du trait décroît avec la profondeur de la séparation.

l'attribut, alors une feuille est retournée, car le groupe de points en question est considéré comme inséparable. Ce groupe de points est un cluster ou une portion de cluster.

En comparaison avec une forêt d'isolation classique, une forêt RIFIFI induit un surcoût relatif au stockage des intervalles de valeurs exclues. Ce surcoût est dans le cas le plus défavorable une constante qui vaut $|\mathcal{A}|*(100/\alpha+1)$. La complexité temporelle quant à elle diffère de celle d'une forêt d'isolation classique par la sélection des séparations. Cette différence est dans le cas le plus défavorable linéaire en fonction du nombre de dimensions/attributs : $\mathcal{O}(|\mathcal{A}|)$.

4.3 Indice d'inséparabilité

Les trois conditions d'arrêt du processus de construction d'un arbre sont les suivantes :

— le noeud contient un point isolé ;
— le noeud contient un ensemble de points qui n'ont pas pu être séparés peu importe l'attribut ;
— la profondeur limite est atteinte.

Ce deuxième type de feuilles est le plus intéressant car il contient idéalement une portion de cluster. Par conséquent, si des points se retrouvent fréquemment dans la même feuille, ils appartiennent vraisemblablement au même cluster. Nous définissons ainsi l'*indice d'inséparabilité* (équation 1) entre deux points comme étant le nombre moyen de fois où ils se retrouvent ensemble dans la même feuille. Les points peuvent donc être combinés progressivement sur la base de leur indice d'inséparabilité à l'aide d'un clustering ascendant hiérarchique pour reconstituer une partition du jeu de données.

$$sim(x_1, x_2) = \frac{1}{T} \sum_{f \in \mathcal{F}} \mathbb{1}_f(x_1, x_2) \tag{1}$$

avec $\mathcal{F}$ l'ensemble des feuilles de la forêt, et $\mathbb{1}_f(x_1, x_2) = 1$ si $x_1, x_2 \in f$ et 0 sinon.

Algorithm 2 RIFIFI : $construire\_arbre$

Entrées : un échantillon $D \subset \mathcal{D}$, la profondeur d du noeud courant, la largeur de la marge α, le seuil de densité η, l'ensemble des intervalles testés I_t, l'ensemble des attributs testés A_t
Sortie : un noeud d'un arbre d'isolation
if $A_t = \mathcal{A}$ ou $|D| = 1$ ou $d > h_{lim}$ **then**
  renvoyer $noeud(null, null, D, d, null, null)$ ▷ Feuille
else
  $a \leftarrow random(\mathcal{A} \setminus A_t)$ ▷ Sélection d'un attribut parmi les attributs non testés
  $v \leftarrow random(domain(a) \setminus \cup_J \{J \in I_t^a\})$ ▷ Sélection d'une valeur parmi les valeurs non testées pour cet attribut
  $marg \leftarrow \frac{\alpha * (\max_{x \in D} x.a - \min_{x \in D} x.a)}{2}$
  $I_t^a \leftarrow I_t^a \cup [v - marg, v + marg]$
  if $\cup_J \{J \in I_t^a\} \supseteq [\min_{x \in D} x.a, \max_{x \in D} x.a]$ **then** ▷ Tout l'intervalle de valeurs a été parcouru et exclu
    $A_t \leftarrow A_t \cup \{a\}$ ▷ L'attribut est rajouté à la liste des attributs exclus
  end if
  $D_m \leftarrow \{x \in D / x.a \in [v - marg, v + marg]\}$ ▷ Points contenus dans la marge
  if $|D_m| \leq \eta$ **then**
    $D_l \leftarrow \{x \in D / x.a < v\}$
    $D_r \leftarrow \{x \in D / x.a \geq v\}$
    Renvoyer $noeud(construire\_arbre(D_l, d+1, \alpha, \eta, I_t, A_t),$
     $construire\_arbre(D_r, d+1, \alpha, \eta, I_t, A_t), D, d, a, v)$ ▷ Noeud interne
  end if
  Renvoyer $construire\_arbre(D, d, \alpha, \eta, I_t, A_t)$ ▷ Sélection d'une autre séparation
end if

5 Expérimentations

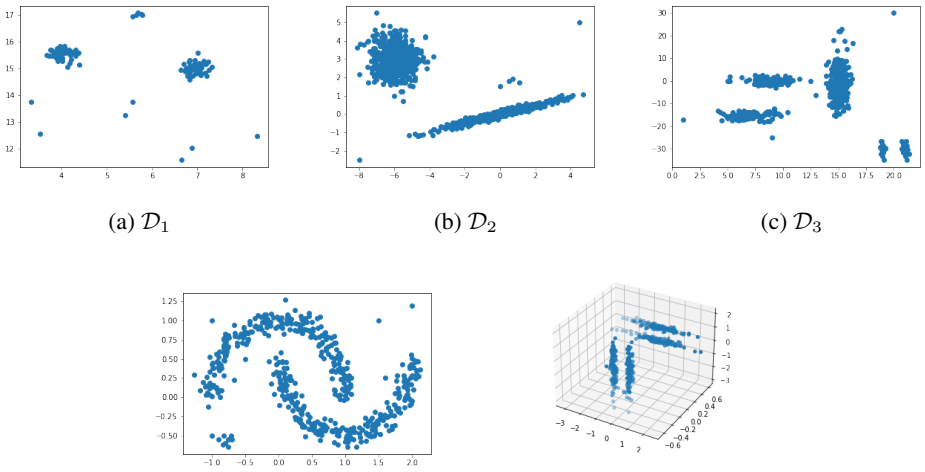
L'objectif de cette section est de vérifier si RIFIFI préserve la structure des données régulières ; ou en d'autres termes, de vérifier si chaque feuille est une portion de cluster, et si les informations contenues dans les feuilles peuvent servir à reconstituer cette structure. À cet effet, il faut vérifier que :

— les feuilles de RIFIFI contiennent plus de points que celles des FI ;
— les points appartenant à la même feuille proviennent du même cluster ;
— l'indice d'inséparabilité peut servir à reconstituer une partition du jeu de données.

Dans cette section expérimentale, les paramètres de RIFIFI sont les suivants : $T = 100$, $\Psi = 256$, $h_{lim} = 8$, $\alpha = 5\%$ de l'intervalle initial de valeurs pour chaque attribut et $\eta = \alpha * n/2$ où n est le nombre de points dans le noeud courant. L'intuition derrière la valeur de α est la suivante : si deux points sont séparés par moins de $\alpha * range(a)$ sur un attribut a, alors ces deux points ne devraient pas être séparés par une coupe. En supposant une distribution uniforme des points du noeud, $\alpha * n$ points devraient se retrouver dans la marge. Par conséquent, si moins de $\alpha * n/2$ points s'y trouvent, elle peut être considérée comme contenant relativement peu de points.

5.1 Les jeux de données

Les jeux de données sont illustrés sur la figure 2. Chacun d'entre eux contient des clusters et des anomalies : 2 clusters de données régulières pour $\mathcal{D}_1$, $\mathcal{D}_2$ et $\mathcal{D}_4$, 3 clusters de données régulières pour $\mathcal{D}_3$ et 4 clusters de données régulières pour $\mathcal{D}_5$. $\mathcal{D}_5$ est un jeu de données en trois dimensions, et chaque cluster est situé dans un sous-espace de dimension 2. La génération de ce jeu de données est expliquée dans Parsons et al. (2004). $\mathcal{D}_4$ est le jeu de données *moons* composé de deux arcs de cercle entrelacés, dans lequel des anomalies ont été rajoutées manuellement.



(a) $\mathcal{D}_1$ (b) $\mathcal{D}_2$ (c) $\mathcal{D}_3$

(d) $\mathcal{D}_4$ (e) $\mathcal{D}_5$

FIG. 2 – Les jeux de données

5.2 Cardinalité des feuilles et profondeur des arbres

Cette partie des expérimentations sert à évaluer l'impact du choix des séparations dans RIFIFI sur la taille des feuilles. Avec les forêts d'isolation classiques, les séparations sont complètement aléatoires jusqu'à l'isolation d'un point ou l'atteinte de la profondeur limite. On s'attend donc à avoir d'une part des feuilles contenant des points isolés, et d'autre part des feuilles contenant un certain nombre de points, mais plus profondes. Avec RIFIFI, on s'attend à avoir des feuilles contenant des points isolés, des feuilles contenant des points qui n'ont pas pu être séparés et des feuilles ayant atteint la profondeur limite. Idéalement, il devrait y avoir plus de feuilles du second type car l'objectif est de préserver les clusters. RIFIFI devrait donc avoir des arbres moins profonds (la profondeur limite étant plus difficilement atteinte que dans la version classique) et des feuilles contenant plus de points.

Une forêt d'isolation classique et une forêt RIFIFI sont construites sur chacun des jeux de données. Les feuilles contenant des points isolés sont écartées. Puis, la cardinalité moyenne des

feuilles ainsi que les profondeurs moyennes des arbres de chaque type de forêt sont calculées. Les résultats obtenus sont reportés dans le tableau 1.

Jeu de données	Cardinalités moyennes des feuilles		Profondeurs moyennes des arbres	
	FI	RIFIFI	FI	RIFIFI
$\mathcal{D}_1$	3.33	9.59	5.99	4.17
$\mathcal{D}_2$	3.83	19.82	7.11	4.67
$\mathcal{D}_3$	3.87	13.57	7.06	5.49
$\mathcal{D}_4$	2.81	25.11	7.27	4.10
$\mathcal{D}_5$	3.53	7.02	7.14	6.47

TAB. 1 – Tailles des feuilles et profondeurs des arbres

Les feuilles de RIFIFI contiennent plus de points que les feuilles d'une forêt d'isolation classique, et ce sur tous les jeux de données. Les arbres RIFIFI quant à eux sont moins profonds que les arbres d'isolation classiques. Il faudrait maintenant vérifier si le critère de sélection des séparations de RIFIFI permet d'avoir suffisamment de feuilles contenant des points qui n'ont pas pu être séparés.

5.3 Types de feuilles

Les points regroupés dans une feuille correspondent-ils à des portions de clusters ? Le tableau 2 montre le pourcentage de feuilles ayant atteint la profondeur limite dans l'arbre (*feuilles de type 1*) et le pourcentage de feuilles dont les points regroupés ne sont plus séparables (*feuilles de type 2*).

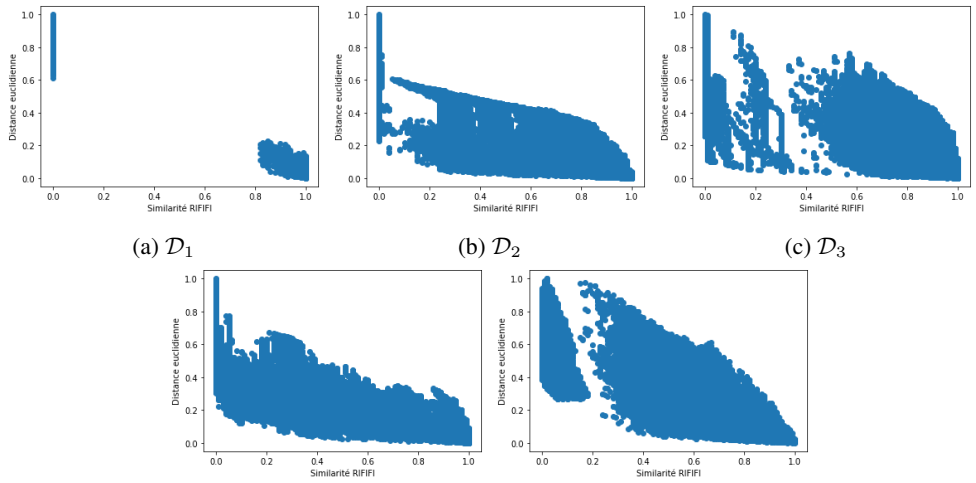
Jeu de données	Feuilles de type 1	Feuilles de type 2
$\mathcal{D}_1$	26.9%	73.1%
$\mathcal{D}_2$	33.8%	66.2%
$\mathcal{D}_3$	34.7%	65.3%
$\mathcal{D}_4$	8.4%	91.6%
$\mathcal{D}_5$	72.5%	27.5%

TAB. 2 – Pourcentages des différents types de feuilles

Il apparaît qu'une proportion non négligeable de feuilles sont de type 2. Ce phénomène est vérifié sur les jeux de données $\mathcal{D}_1$ à $\mathcal{D}_4$, mais pas sur le jeu de données $\mathcal{D}_5$. Ce dernier contient aussi moins de points dans les feuilles, en comparaison avec les autres jeux de données et, les arbres de la forêt RIFIFI, bien que moins profonds que les arbres d'isolation classiques, restent tout de même plus profonds que ceux des forêts construites sur les autres jeux de données (tableau 1). Ceci s'explique par le fait que dans $\mathcal{D}_5$, chaque cluster "n'existe" que dans deux des trois dimensions. Or, le processus d'isolation se poursuit en séparant les points sur la troisième dimension, où ils sont distribués de manière quasi-uniforme.

5.4 Proximité entre les points

Cette partie sert à vérifier la pertinence des regroupements des points au sein des feuilles. Pour chaque paire de points dans le jeu de données, la distance euclidienne entre les points constituant la paire est calculée, de même que l'indice d'inséparabilité. Ces deux valeurs sont normalisées. Pour éliminer l'impact de l'échantillonnage, tous les points du jeu de données sont propagés dans chaque arbre d'isolation afin que les feuillent couvrent la totalité des points. Les résultats sont reportés pour chaque jeu de données sur la figure 3 : pour chaque paire de points, en abscisse l'indice d'inséparabilité RIFIFI et en ordonnée la distance euclidienne.



(a) $\mathcal{D}_1$ (b) $\mathcal{D}_2$ (c) $\mathcal{D}_3$

(d) $\mathcal{D}_4$ (e) $\mathcal{D}_5$

FIG. 3 – Distance euclidienne VS indice d'inséparabilité RIFIFI

Des phénomènes *a priori* contre-intuitifs sont observés en analysant des résultats :

— certains points proches dans l'espace euclidien (distance euclidienne faible), se retrouvent rarement dans la même feuille (indice d'inséparabilité proche de 0). C'est le cas lorsque les deux points, bien que proches dans l'espace euclidien, sont séparables et font donc partie de clusters différents, par exemple les points $(14.22; -0.70)$ et $(8.59; -0.89)$ dans $\mathcal{D}_3$, surtout lorsqu'une dimension est commune. Une séparation entre ces deux points pourrait tout à fait être conservée, car il suffit que son voisinage ne soit pas dense. C'est également le cas lorsqu'un des deux points est très proche du cluster contenant l'autre point, sans pour autant en faire partie, ou encore lorsqu'un des deux points est localisé en bordure du cluster et est donc souvent séparé des autres (les points $(-7.10; 4.57)$ et $(-6.03; 3.16)$ dans $\mathcal{D}_2$) ;

— certains points éloignés dans l'espace euclidien, se retrouvent parfois ensemble dans la même feuille. C'est le cas lorsque les deux points, bien qu'éloignés dans l'espace euclidien font partie du même cluster, par exemple lorsque le cluster est étiré.

Cette analyse laisse penser que RIFIFI apporte une information de proximité contextualisée, contexte défini par la séparabilité des sous-espaces. x_1 et x_2 puis x_1 et x_3 peuvent être situés à la même distance euclidienne, mais x_1 et x_2 font partie du même cluster, et pas x_3.

5.5 Distances moyennes intra et inter-feuilles

La distance euclidienne moyenne entre les points de chaque feuille et le centre de la feuille sont calculées pour les deux types de forêt. La distance euclidienne moyenne entre les centres des feuilles est également calculée. Les résultats obtenus sont reportés dans le tableau 3.

Jeu de données	Distances moyennes intra-feuilles		Distances moyennes inter-feuilles	
	FI	RIFIFI	FI	RIFIFI
$\mathcal{D}_1$	**0.12**	0.16	1.73	**2.54**
$\mathcal{D}_2$	**0.23**	0.48	4.13	**5.38**
$\mathcal{D}_3$	**1.14**	1.46	13.18	**19.57**
$\mathcal{D}_4$	0.81	**0.69**	1.26	**1.47**
$\mathcal{D}_5$	**0.11**	0.24	1.40	**1.65**

TAB. 3 – Distances moyennes intra-feuilles et inter-feuilles

La distance moyenne intra-feuilles est plus grande chez RIFIFI sur presque tous les jeux de données (excepté $\mathcal{D}_4$ à cause de sa forme), ce qui est compréhensible car les feuilles d'isolation classique contiennent beaucoup moins de points et ces derniers sont proches. Par contre, la distance inter-feuilles est systématiquement plus grande chez RIFIFI, ce qui traduit le fait qu'avec les forêts d'isolation classiques, les points appartenant au même cluster se retrouvent généralement séparés dans des feuilles différentes.

5.6 Indice d'inséparabilité et clustering

Cette partie a pour but de vérifier que l'indice d'inséparabilité défini peut permettre de reconstituer une partition du jeu de données. Pour chaque jeu de données, une forêt RIFIFI est construite. Les anomalies sont identifiées et écartées. Étant donné que la détection d'anomalies est hors de la portée de cet article qui est plutôt focalisé sur la capacité de RIFIFI à préserver les clusters, le processus d'identification des anomalies est volontairement omis. Ensuite, tous les points du jeu de données sont propagés dans chaque arbre d'isolation. Puis, l'indice d'inséparabilité entre chaque paire de points est calculé et un clustering ascendant hiérarchique est appliqué sur la matrice de similarité obtenue. Le nombre de clusters k du jeu de données étant supposé connu, le clustering ascendant hiérarchique est stoppé lorsque k groupes sont construits. L'*Adjusted Rand Index* (ARI) qui permet d'évaluer les résultats du clustering en présence des véritables étiquettes est utilisé comme mesure d'évaluation. Un ARI de 1 entre deux partitions signifie que les deux sont identiques. L'ARI du clustering de chaque jeu de données est reporté dans le tableau 4. Dans le calcul de l'ARI, les anomalies sont considérées comme faisant partie d'un cluster isolé.

L'ARI tend vers 1 pour la plupart des jeux de données. Il est inférieur à 0.95 pour le jeu de données $\mathcal{D}_5$. Cela est dû au fait que beaucoup d'anomalies sont identifiées par RIFIFI sur ce jeu de données. Le seuil d'anomalie aurait dû être plus élevé. L'ARI est assez faible pour $\mathcal{D}_4$, en comparaison avec les autres jeux de données. La cause en est que beaucoup de coupes regroupent une partie de la demi-lune inférieure et la portion de demi-lune supérieure située dans le creux de la première citée. Par conséquent, ces points de la demi-lune supérieure sont affectés au même cluster que les points de la demi-lune inférieure. Un agrégation à l'échelle

Jeu de données	ARI
$\mathcal{D}_1$	1.0
$\mathcal{D}_2$	0.95
$\mathcal{D}_3$	0.97
$\mathcal{D}_4$	0.64
$\mathcal{D}_5$	0.87

TAB. 4 – ARI du clustering de chaque jeu de données

des feuilles (c'est-à-dire un clustering ascendant hiérarchique avec comme point de départ les feuilles des arbres, et non plus les points, et une mesure de similarité du type *Jaccard*) résoudrait ce problème, car les feuilles ayant beaucoup de points en commun et une cardinalité similaire seraient combinées en premier. Les coupes parallèles aux axes ne permettent en outre d'identifier que les groupes elliptiques en utilisant d'indice d'inséparabilité sur les paires de points. C'est l'objet des travaux actuels.

6 Conclusion

Cet article propose une variante de l'algorithme des forêts d'isolation appelée RIFIFI ayant pour objectif de préserver les clusters présents dans le jeu de données. À cet effet, un nouveau critère de sélection des séparations a été introduit, critère basé sur l'analyse du voisinage des séparations. À travers les premières expérimentations menées, il a été constaté que RIFIFI permet de conserver la proximité entre les points appartenant au même groupe de données, et que la reconstitution d'une partition du jeu de données est donc possible en effectuant un clustering ascendant hiérarchique sur une matrice de similarité basée sur le nombre de fois où les paires de points se retrouvent dans la même feuille. Ce travail constitue un premier pas vers une approche unifiée pour l'extraction d'explications contextuelles d'anomalies. En effet, l'idéal serait d'utiliser directement les informations contenues dans les feuilles, et non de calculer les distances (euclidienne ou pas) entre des paires de points, étant donné que les forêts d'isolation ne requièrent pas ces calculs. Pour ce faire, une agrégation des feuilles similaire à des méthodes de clustering de type *grid-based* pourrait être explorée : chaque feuille de cardinalité importante délimite un sous-espace, et les différents sous-espaces peuvent être combinés pour reconstituer une partition du jeu de données. Ayant donc les anomalies d'un côté, et cette partition de l'autre, il deviendrait possible d'extraire des explications contrastives d'anomalies à l'aide d'une méthode unifiée, sans recourir à un pipeline.

Références

Amer, M., M. Goldstein, et S. Abdennadher (2013). Enhancing one-class support vector machines for unsupervised anomaly detection. In *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, pp. 8–15.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, et J. Sander (2000). Lof : identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104.

Cortes, D. (2021). Revisiting randomized choices in isolation forests. *arXiv preprint arXiv :2110.13402*.

Gupta, N., D. Eswaran, N. Shah, L. Akoglu, et C. Faloutsos (2018). Beyond outlier detection : Lookout for pictorial explanation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 122–138. Springer.

Hariri, S., M. C. Kind, et R. J. Brunner (2019). Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering 33*(4), 1479–1489.

Kopp, M., T. Pevnỳ, et M. Holeňa (2020). Anomaly explanation with random forests. *Expert Systems with Applications 149*, 113187.

Liu, F. T., K. M. Ting, et Z.-H. Zhou (2010). On detecting clustered anomalies using sciforest. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 274–290. Springer.

Liu, F. T., K. M. Ting, et Z.-H. Zhou (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD) 6*(1), 1–39.

Mensi, A. et M. Bicego (2021). Enhanced anomaly scores for isolation forests. *Pattern Recognition 120*, 108115.

Mokoena, T., T. Celik, et V. Marivate (2022). Why is this an anomaly ? explaining anomalies using sequential explanations. *Pattern Recognition 121*, 108227.

Panjei, E., L. Gruenwald, E. Leal, C. Nguyen, et S. Silvia (2022). A survey on outlier explanations. *The VLDB Journal*, 1–32.

Parsons, L., E. Haque, et H. Liu (2004). Subspace clustering for high dimensional data : a review. *Acm sigkdd explorations newsletter 6*(1), 90–105.

Yepmo, V., G. Smits, et O. Pivert (2022). Anomaly explanation : A review. *Data & Knowledge Engineering 137*.

Summary

This paper takes a step towards the extraction of contrastive explanations between anomalies and the intrinsic structure of regular points. It proposes a variant of the isolation forest algorithm whose main objective is to preserve the structure of regular data in order to facilitate its reconstruction. Experiments conducted on synthetic datasets show that this variant of isolation forest deteriorates less the structure of regular data than the classical method. Therefore, the former can serve as a basis for a unified approach to anomaly detection and explanation.