Détection et Explication des Données Régulières et Irrégulières

Rahul Nath, Grégory Smits, Olivier Pivert

Détection et Explication des Données Régulières et Irrégulières

Rahul Nath[1], Grégory Smits[2], and Olivier Pivert[3]
[1] Department of Informatics, University of Bergen, Bergen Norway
Email : rahul.nath@uib.no
[2] IMT Atlantique, Lab-STICC, UMR CNRS 6285, Brest, France
Email : gregory.smits@imt-atlantique.fr
[3] IRISA - Université de Rennes 1, UMR 6074, Lannion, France
Email : olivier.pivert@irisa.fr

Résumé :

Un vocabulaire flou, composé de partitions floues associées à des variables linguistiques, joue un rôle crucial dans de nombreuses méthodes de description de données. Cependant, la construction d'une partition appropriée pour décrire la structure intrinsèque des données est une tâche difficile. Cet article introduit une nouvelle stratégie visant à inférer des partitions possibles à partir de la distribution des données pouvant être utilisées pour décrire de manière discriminante à la fois les zones denses de points et les zones éparses. Les partitions sont ensuite exploitées pour identifier la structure intrinsèque des points réguliers mais également pour fournir des explications contrastives sur les anomalies trouvées dans les zones éparses.

Mots-clés :

Inférence de vocabulaire flou, résumé linguistique, détection d'anomalies, explication d'anomalies

Abstract:

Fuzzy partitions associated with linguistic variables are particularly useful to provide users with a description of the data. However, designing fuzzy partitions that make it possible to linguistically describe the data distribution and its inner structure is a tedious task. This paper introduces a novel strategy to infer possible fuzzy partitions from the data distribution with the objective to have available modalities to describe both dense and sparse regions. A data inner structure as well as the anomalies are then identified using these partitions whose terms are also used to provide users with contrastive explanations about the found anomalies.

Keywords:

Fuzzy vocabulary inference, linguistic summaries, anomaly detection, anomaly explanation

1 Introduction

Cet article répond au besoin de nombreux utilisateurs de disposer d'outils pragmatiques pour analyser de nouvelles données. L'approche proposée fournit une description interprétable de la structure intrinsèque des données, structure composée des régions denses de points, et également des points qui ne suivent pas cette structure et qui sont considérés comme des anomalies. Une anomalie est un point qui possède une combinaison inhabituelle de valeurs. Une description informative d'un tel point doit mettre en avant les valeurs qui permettent de le distinguer des régularités.

Les points analysés sont initialement définis dans un espace composé d'attributs numériques et catégoriels dont les domaines sont généralement non commensurables. L'utilisation de Sous-Ensembles Flous (SEF) pour former un vocabulaire permet de réécrire les points avant leur analyse afin d'évoluer dans un espace normalisé sur l'intervalle unité, espace constitué d'autant de dimensions qu'il y a de termes dans le vocabulaire. Les SEFs permettent également d'intégrer des connaissances contextuelles sur la façon dont les données doivent être comparées, notamment en introduisant une relation d'indistingabilité au sein d'intervalles de valeurs formant les noyaux des SEFs. Associés à des variables linguistiques, l'utilisation des SEFs permet finalement de générer des descriptions linguistiques interprétables des motifs observables.

Afin d'aider un utilisateur final dans sa démarche d'analyse d'un nouveau jeu de données, l'approche proposée dans cet article vise à fournir une description interprétable, composée de termes issus d'un vocabulaire flou, de la structure intrinsèque des données et également des points anormaux trouvés. Pour atteindre cet objectif, deux problématiques sont abordées dans ce travail :

1. l'inférence de partitions floues à partir de la distribution des données,
2. puis l'identification et la description des zones denses et éparses de points. Les zones denses sont interprétées comme des régularités et les points des zones éparses comme des anomalies.

Les partitions suggérées peuvent ensuite être ajustées et chacune des modalités qualifiée linguistiquement par l'utilisateur pour disposer d'une vocabulaire adapté aux données.

L'article introduit donc une approche coopérative autour d'un vocabulaire flou dont un premier découpage est suggéré automatiquement à partir des données pour décrire la structure intrinsèque des points dits réguliers mais également pour identifier les possibles anomalies. Ces anomalies sont expliquées linguistiquement à l'aide des termes du vocabulaire et de manière contrastive par rapport aux régularités.

La section 2 positionne l'approche vis-à-vis des travaux existants sur l'explication d'anomalies. Dans la section 3.1, l'approche proposée est détaillée : L'inférence de partitions à partir de la distribution des données (section 4) puis le regroupement des points réguliers et l'identification des anomalies (section 5). Les résultats d'expérimentations sont présentés dans la section 6.

2 Positionnement scientifique

L'approche proposée aborde de manière unifiée deux problématiques majeures de la communauté, l'aide à la construction d'un vocabulaire flou et son utilisation pour l'analyse de données.

2.1 Vocabulaire flou et analyse de données

Le travail présenté s'inscrit pleinement dans le paradigme *Computing with Words* [17] visant à représenter et manipuler les connaissances extraites des données à l'aide de variables linguistiques. Ces variables linguistiques permettent une réécriture contextuelle et personnalisable des données initialement définies dans un espace numérique et catégoriel. Il est d'une part essentiel que l'utilisateur en charge de l'analyse des données ait une bonne compréhension des définitions formelles associées aux termes du vocabulaire, c'est pourquoi plusieurs éditeurs manuels de vocabulaire flou ont été implémentés [15]. Mais d'autre part, la construction manuelle d'un vocabulaire approprié à des données n'est pas une tâche aisée. Des approches dites coopératives visent ainsi à assister l'utilisateur dans cette tache de construction du vocabulaire. Dans [4], une famille de partitions par attribut numérique est générée à l'aide d'un algorithme de regroupement hiérarchique et un critère d'interprétabilité, dépendant du nombre de modalités et de leur forme, est utilisé pour choisir une de ces partitions. Dans [8], une partition est apprise en utilisant des opérations de morphologie mathématique pour modéliser la distribution des points d'entraînement et d'évaluation utilisés pour une tache d'apprentissage automatique. Plus généralement, lorsqu'un vocabulaire doit être utilisé pour décrire une structure de données, comme e.g. dans [12], il est nécessaire de vérifier que les partitions conduisent à une description linguistique interprétable et fidèle vis-à-vis de la structure à décrire. Une mesure d'adéquation entre les espaces numériques/catégoriels et linguistiques est introduite dans [6] et exploitée dans [13] pour ajuster un vocabulaire aux données. Alors que des résumés linguistiques des données ont été utilisés pour identifier les anomalies [14], cet article aborde également la question de la description des anomalies par rapport aux zones de données régulières.

2.2 Description linguistique de données et explication d'anomalies

Une application majeure du paradigme *Computing with Words* consiste à générer un résumé des différentes combinaisons de termes du vocabulaire observables dans les données. Un tel résumé ne décrit cependant pas la struc-

ture intrinsèque des données, structure composée de groupes homogènes représentant des phénomènes fréquents considérés comme réguliers. L'approche introduite dans [12] vise à décrire une telle structure en groupes, mais ne permet cependant pas de distinguer les points réguliers des points dits irréguliers. Dans l'article [11], il a été montré comment un vocabulaire peut influencer un processus de détection d'anomalies. Mais cet article s'intéresse à l'utilisation d'un vocabulaire, inféré à partir des données pour fournir une description contextualisée des anomalies, ce qui selon [7] est la façon la plus interprétable et informative de décrire les anomalies

3 Notions préliminaires

3.1 Notations

$\mathcal{D} = \{x_1; x_2; \ldots; x_m\}$ représente l'ensemble des m points définis sur n attributs $\{A_1; A_2; \ldots; A_n\}$ de domaines $\{D_1; D_2; \ldots; D_n\}$. Ce travail ne considère que des attributs numériques. Le vocabulaire flou est formellement défini par un ensemble de variables linguistiques $\mathcal{V} = \{V_1; \ldots; V_n\}$, où pour $i = 1..n$, V_i est une séquence de q_i modalités $\langle v_{i,1}; \ldots; v_{i,q_i}\rangle$ qui discrétisent le domaine D_i. Pour chaque modalité v, on note μ_v sa fonction caractéristique et l_v son étiquette linguistique. Pour des raisons d'interprétabilité [4], il est imposé que chaque partition soit forte [10], i.e. $\forall y \in D_i$, $\sum_{j=1}^{q_i} \mu_{i,j}(y) = 1$ et que toute valeur y satisfasse au plus deux modalités adjacentes.

Sur un domaine D, la distribution marginale des points notée P est lissée à l'aide d'une opération de convolution de noyau K. La probabilité, après lissage, d'une valeur $a \in D$ est notée $\rho(a)$ et calculée comme suit :

$$\rho(a) = (P * K)(a) = \sum_{m=\inf(D)}^{\sup(D)} P(a-m)K(a). \quad (1)$$

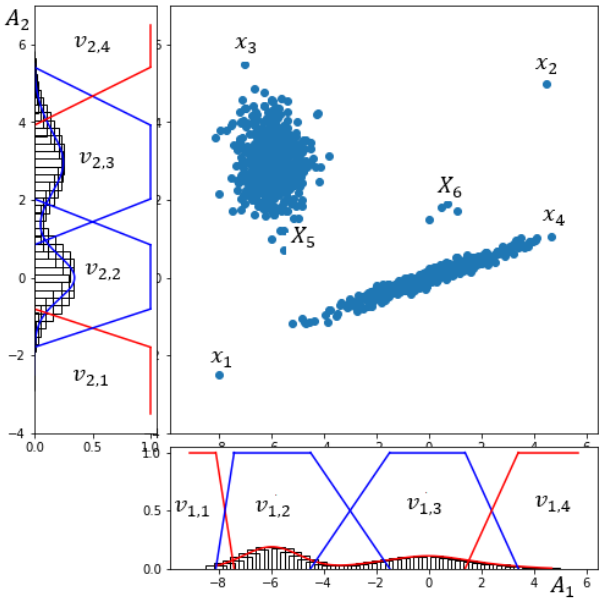


FIGURE 1 – Exemple illustratif de partitions représentant la distribution des données

3.2 Exemple illustratif

La Figure 1 illustre des données définies sur deux attributs A_1 en abscisse et A_2 en ordonnée. Deux motifs réguliers peuvent être observés, un sphérique autour de $A_1 \approx -6.5$ et $A_2 \approx 3$, et un second elliptique autour de $A_2 \approx 0$ et pour A_1 allant de -4 à 4. Quelques points éparses considérés comme des irrégularités sont labellisés x_1 à X_6, cette dernière annotation décrivant un groupe d'anomalies. Deux exemples de partitions floues sont suggérés sur chaque axe avec en bleu des modalités couvrant des zones denses et en rouge des zones éparses des domaines.

4 Inférence de partitions floues pour décrire les données

Cette section décrit comment des partitions floues sont suggérées à partir de l'analyse de la distribution des données puis utilisées pour décrire les motifs réguliers observables ainsi que les irrégularités.

4.1 Inférence de partitions floues

Soit A un attribut numérique de domaine D, une partition $\langle A, \{v_1, \ldots, v_q\}, \{l_1, \ldots, l_q\}\rangle$ est construite à partir de la distribution lissée ρ (Eq. 1) à l'aide de deux seuils de densité. Le premier β sert à identifier les intervalles de D de suffisamment haute densité et le second γ ($1 \geq \beta > \gamma \geq 0$) à l'opposé sert à délimiter les intervalles de faible densité.

Definition 1 *Une modalité v est dite de type* **Zone Dense (ZD)** *si sa fonction caractéristique μ_v couvre une zone dense :*

— $\mu_v(a) = 1$ *ssi.* $\rho(a) \geq \beta$.

À l'opposé, une modalité v' est qualifiée de type **Zone Eparse (ZE)** *si elle couvre une zone de faible densité :*

— $\mu_{v'}(a) = 1$ *ssi.* $\rho(a) \leq \gamma$.

Pour une variable linguistique V, on note par $\hat{V}$ et $\hat{V}_x$ ses modalités ZD et par $\hat{V}_x$ les modalités de type ZD satisfaites par x.

◇

L'algorithme de construction d'une partition à partir de la distribution marginale lissée des données est simple et efficace dans la mesure où il traite en temps linéaire les valeurs distinctes observées dans $\mathcal{D}$ sur chaque attribut. Paramétrée par β et γ, la première étape consiste à identifier les intervalles de largeur maximale décrivant des zones denses et éparses, ces zones constitueront les noyaux des modalités de type ZD et ZE. Des transitions graduelles sont ensuite ajoutées entre modalités adjacentes pour obtenir une partition forte. Un éditeur de vocabulaire, tel que celui introduit dans [9], peut ensuite être utilisé pour ajuster les modalités et les étiqueter linguistiquement.

Example 1 *En utilisant $\beta = \overline{\rho}$, où $\overline{\rho}$ est la densité moyenne, et $\gamma = \frac{1}{4}\beta$, les partitions illustrées sur la Figure 1 sont obtenues. Elles comportent chacune deux modalités ZD (en bleu) couvrant les pics de densité, et deux modalités ZE (en rouge).*

4.2 Partitionnement des données

La prochaine étape est de combiner des modalités de type ZD pour identifier des sous-espaces denses, comme pourrait le faire une approche de partitionnement par grille [5]. Cependant, l'approche proposée dans cet article se "contente" d'identifier des sous-espaces denses convexes de forme délimitée par un combinaison conjonctive de modalités de type ZD. Ces sous-espaces seront ensuite exploités pour identifier les irrégularités (Sec. 5).

Definition 2 *Une* **Région Dense guidée par le Vocabulaire (RDV)** *est un hyper rectangle flou délimité par l'intersection de modalités de type ZD. Formellement, soit une famille d'ensembles de modalités ZD trouvées pour chaque attribut $\{\hat{V}^1, \ldots, \hat{V}^m\}$. Un RVD ψ est un sous-ensemble de modalités ZD, $\psi \in \hat{V}^1 \otimes \ldots \otimes \hat{V}^m$, tel que :*

$$\frac{1}{|\mathcal{D}|} \times \Sigma^{\mathcal{D}}_{\bigwedge_\psi} \geq \zeta, \qquad (2)$$

où $\bigwedge_\psi$ est la conjonction formée par les modalités de ψ et $\Sigma^{\mathcal{D}}_{\bigwedge_\psi}$ sa cardinalité scalaire [2]. ζ est un seuil de taille minimum pour qu'un groupe de points constitue un motif régulier. Un RDV est une caractérisation guidée par le vocabulaire d'un groupe de points suffisamment nombreux pour constituer une régularité.

◇

Definition 3 Partition des Données guidée par un Vocabulaire (PDV) *Une PDV Ψ est un ensemble de RDV tel que :*

— $\forall \psi \in \Psi$, *ψ est une RDV,*

— $\forall \psi \in \Psi, \nexists \psi' \in \hat{V}^1 \otimes \ldots \otimes \hat{V}^m$ *tel que ψ' est un RDV et $\psi' \supset \psi$.*

◇

Par définition (Def. 3), une PDV correspond à la bordure positive maximale dans le treillis conjonctif formé des combinaisons possibles de modalités ZD contenant au plus une modalité par attribut. Le critère d'arrêt de l'exploration est que la combinaison de modalités ZD

doit être une RDV. Il est évident de montrer la monotonicité de la propriété d'une RDV par rapport à l'inclusion entre combinaisons. Soit $\psi, \psi' \in \mathcal{V}_1^d \otimes \ldots \otimes \mathcal{V}_m^d$, si $\psi \subseteq \psi'$ alors $\Sigma_{\bigwedge_{\psi_i \in \psi}} \geq \Sigma_{\bigwedge_{\psi'_i \in \psi'}}$. Une implémentation efficace d'un algorithme à la *Apriori* [1] est utilisée pour identifier cette bordure et construire une PDV. Cette étape de la procédure d'analyse est évidemment la plus coûteuse en temps car l'espace d'exploration augmente de manière exponentielle par rapport au nombre d'attributs et linéairement par rapport au nombre de modalités par attribut.

5 Détection et description d'anomalies

Les RDV identifiées à l'aide du vocabulaire représentent les motifs réguliers observables dans les données. Les points ne suivant pas ces motifs sont considérés comme des anomalies qu'il est possible de décrire également à l'aide du vocabulaire.

5.1 Détection d'anomalies

Definition 4 Anomalie guidée par le Vocabulaire *Soit une PDV Ψ. Un point x est une anomalie s'il ne correspond à aucun des motifs (i.e. RDV) identifiés, plus formellement si* $\max_{\psi \in \Psi} \mu_\psi(x) < 1$.
$\diamond$

Afin de rendre la définition d'une anomalie plus graduelle, et ainsi de prioriser leur gestion, un score d'anomalie est calculé pour chaque point. Le score d'anomalie d'un point x, noté $As(x, \Psi)$, est inversement proportionnel à sa proximité vis-à-vis des motifs formant la PDV Ψ. $As(x, \Psi)$ est défini dans l'intervalle unité et est maximal si x ne correspond à aucun RDV de Ψ et est minimal si x s'intègre complètement dans un de ces motifs.

$$As(x, \Psi) = 1 - \max_{\psi \in \Psi} \mu_{\bigwedge_\psi}(x), \qquad (3)$$

où $\mu_{\bigwedge_\psi}(x)$ est le degré de satisfaction de x vis-à-vis ψ.

Example 2 *Sur les données de la Figure 1, les RDV trouvées sont* $\{\{v_{1,2}; v_{2,3}\}, \{v_{1,3}; v_{2,2}\}\}$ *avec* $\zeta = 0.3$. *Le point* x_2 *satisfait la modalité ZD* $v_{2,3}$ *uniquement avec* $\mu_{v_{2,3}}(x_2) = 0.1$. *Son score d'anomalie est donc* $As(x_2) = 1 - \max(\min(\mu_{v_{1,2}}(x_2), \mu_{v_{2,3}}(x_2)), \min(\mu_{v_{1,3}}(x_2), \mu_{v_{2,2}}(x_2))) = 1$. *Soit* x *un point de l'ensemble* X_6 *où* $\mu_{v_{1,3}}(x) = 1$, $\mu_{v_{2,2}}(x) = 0.2$ *et* $\mu_{v_{2,3}}(x) = 0.8$. *Son score d'anomalie est alors* $As(x) = 1 - \max(\min(0, 0.8), \min(1, 0.2)) = 0.8$.

5.2 Description linguistique des anomalies

Afin d'aider l'utilisateur final à mieux comprendre les raisons de l'anormalité d'un point, une explication linguistique, s'appuyant sur les termes du vocabulaire, est associée à chaque point dont le score d'anomalie est positif.

Soit x un point tel que $As(x, \Psi) > 0$. Si x partage au moins, même partiellement, une modalité de type ZD avec au moins un des motifs (RDV) observés, alors une explication constrastive est produite. Pour chaque RDV $\psi \in \Psi$ telle que $\hat{V}x \bigcap \psi \neq \emptyset$, une explication de la forme suivante est produite :
Contrairement au motif régulier ψ :

- *x satisfait partiellement :*
  $\forall v_{i,j} \in \hat{V}x \cap \psi$ st. $0 < \mu_{v_{i,j}}(x) < 1$
  - *A_i is $l_{i,j}$ à un degré de $\mu_{v_{i,j}}(x)$.*
- *(De plus,) x possède des valeurs rarement observées :* $\forall v_{i,j} \in \mathcal{V} \setminus \hat{V}x$ st. $\mu_{v_{i,j}}(x) > 0$:
  - *A_i is $l_{i,j}$ à un degré de $\mu_{v_{i,j}}(x)$.*

Dans le cas où x ne satisfait pas de modalité de type ZD, seules les explications sur les valeurs rares sont générées.

Example 3 *Le point x_3 est une anomalie possible ($As(x_3, \Psi) = 0.9$) expliquée comme suit : Contrairement au motif régulier* $\{l_{1,2}; l_{2,3}\}$ *:*

- *x_3 possède des valeurs rarement observées :*

— A_2 *is* $l_{2,4}$ *à un degré de* 1.

6 Expérimentations

Des premières expérimentations ont été conduites sur différents jeux de données artificielles pour pour mettre en avant la pertinence du vocabulaire généré à partir des données et le rôle central qu'il peut jouer pour identifier et caractériser les motifs réguliers et les anomalies.

6.1 Données et hyper-paramètres

Trois jeux de données, D_1, D_2 et D_3 décrits dans la Table 1, ont été utilisés. D_1 and D_2 ont deux dimensions comme illustré par les Figure 1 et 2 respectivement. D_3 est un jeu en trois dimensions classiquement utilisé pour tester des approches de partitionnement en sous-espaces, chaque motif régulier de D_3 n'existant que dans deux des trois dimensions. Des points dont les valeurs sont générées aléatoirement sont ajoutés à ces données et labellisés comme anomalies.

TABLEAU 1 – Statistiques sur les données

	Taille	# d'anomalies
D1	1009	9
D2	530	10
D3	420	20

Les valeurs des hyper-paramètres de la méthode proposée ont été fixées empiriquement puis l'impact de leur variation sur les résultats a été analysé. L'impact de la fonction noyau utilisé pour lisser la distribution des données a peu d'influence sur la construction des partitions qui forment le vocabulaire. Par contre, le nombre, le type et la forme des modalités de types ZD dépendent fortement des valeurs des hyper-paramètres β and γ. Une valeur par défaut de $\beta = \overline{p}$, i.e. la densité moyenne observée, semble pertinente car elle garanti l'existence d'au moins une modalité de type ZD, modalité qui couvrirait l'ensemble du domaine en cas de distribution uniforme des points. γ détermine

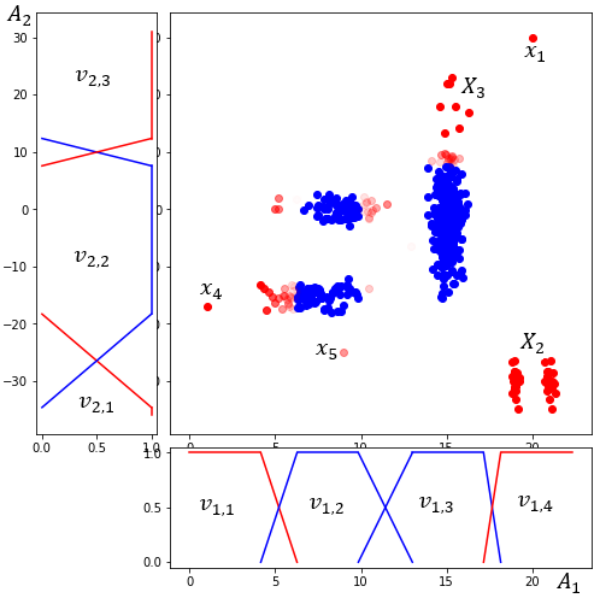


FIGURE 2 – Jeu de données D_2

à la fois la forme des modalités de type ZE et la gradualité des transitions entre modalités adjacentes, sa valeur par défaut est $\gamma = \frac{1}{4}\beta$. Les valeurs suivantes ont ainsi été utilisées pour β : $\langle 0.077, 0.012\rangle$ pour D_1, $\langle 0.048, 0.015\rangle$ pour D_2, et $\langle 0.17, 0.34, 0.17\rangle$ pour D_3. La valeur par défaut pour le seuil ζ est de 0.3

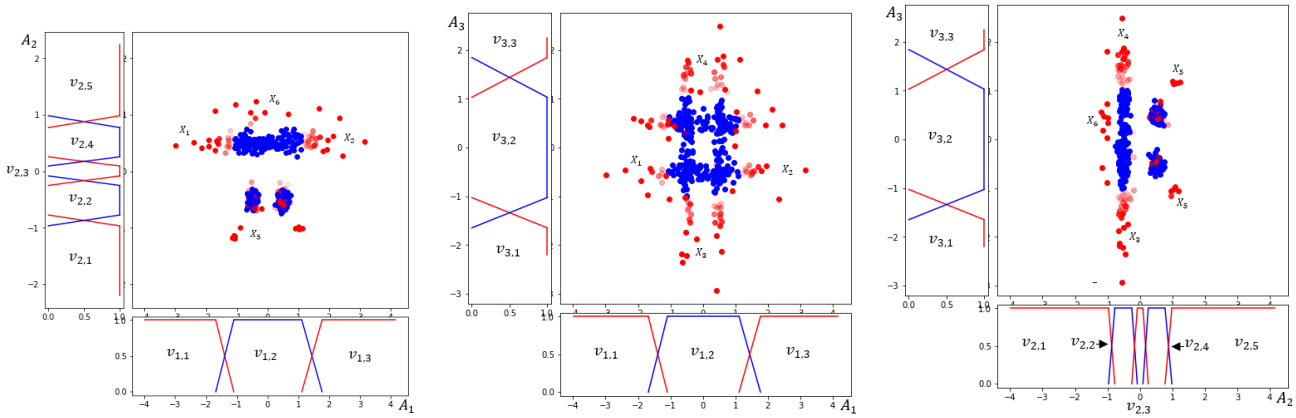
6.2 Résultats et discussion

En utilisant les valeurs par défaut des hyper-paramètres, les vocabulaires générés (cf. Fig. 1, 2 et 3) par la méthode ont permis d'identifier les motifs réguliers indiqués dans la Table 2. Chaque RDV décrit un sous-espace dense interprété comme un motif régulier et caractérisable à l'aide des termes du vocabulaire.

TABLEAU 2 – PDV proposées pour D_1, D_2 et D_3

Jeu de données	PDV
D1	$\{\{v_{1,3}; v_{2,2}\}, \{v_{1,2}; v_{2,3}\}\}$
D2	$\{\{v_{1,2}; v_{2,2}\}, \{v_{1,3}; v_{2,2}\}\}$
D3	$\{\{v_{1,2}; v_{2,2}\}, \{v_{2,4}; v_{3,2}\}, \{v_{1,2}; v_{3,2}\}\}$

Pour quantifier la qualité des PDV trouvées et leur capacité à discriminer les points réguliers



(a) Attributs A_1 et A_2 (b) Attributs A_1 et A_3 (c) Attributs A_2 et A_3

FIGURE 3 – Jeu de données $D3$

TABLEAU 3 – Score de précision de la distinction régularité vs. anomalie

Jeu de données	Précision
D1	0.8938
D2	0.9554
D3	0.8484

des anomalies, la mesure suivante a été utilisée :

$$acc(\Psi, R) = \frac{1}{|R|} \times \sum_{x \in R} \max_{\psi \in \Psi} \mu_\psi(x), \qquad (4)$$

où $R \subseteq D$ est le sous-ensemble des points réguliers, et ψ est la conjonction de modalités de type ZD, le calcul de μ_ψ repose sur une t-norm, le minimum dans notre cas. La Table 3 indique les scores de précision obtenus.

L'évaluation selon la mesure acc et complétée par une analyse du score d'anomalie illustrée sur la Figure 4 qui indique l'AUC du score d'anomalie. La variation de ce score est donnée selon différentes valeurs de l'hyper-paramètre β ($\pm 5\%$ de la valeur par défaut), et en fixant $\gamma = \frac{1}{4}\beta$). La Figure 4 compare également l'AUC obtenu par l'approche proposée avec des méthodes classiques dédiées à la détection d'anomalies : LOF et les forêts d'isolation (IF).

La Figure 4 montre l'efficacité de l'approche avec des scores comparables à ceux de LOF et IF. Sur D_1 on constate que le résultat est sensible au choix de la valeur de l'hyper-paramètre

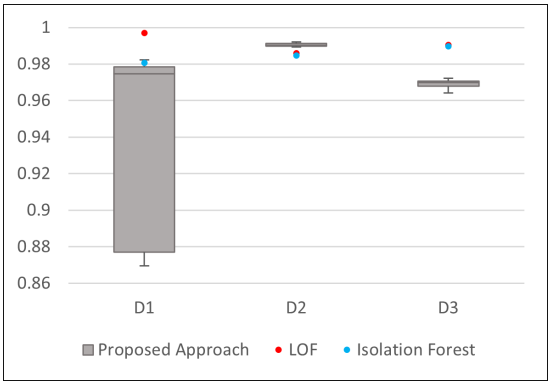


FIGURE 4 – AUC du score d'anomalie, sensibilité par rapport à β

γ. Pour une faible valeur de γ pour D_1, une seule modalité de type ZD est construite par partition, rendant ainsi impossible la distinction entre régularités et anomalies.

Cependant, l'approche apporte globalement deux avantages majeurs. Tout d'abord, le fait de calculer un score d'anomalie selon une distance vis-à-vis des motifs réguliers (i.e. les RDV) rend l'approche relativement stable comparativement à LOF e.g. qui utilise le voisinage local des points et aux IF qui peinent à identifier des anomalies locales. Le second avantage concerne le rôle central donné au vocabulaire qui permet de manière intrinsèque de fournir des explications linguistiques facilement interprétables sur les données.

7 Conclusion

Dans cet article, une approche pragmatique assistant un utilisateur lors de l'analyse de données est proposée. Une première étape crucial concerne l'inférence de partitions floues à partir de la distribution des données. Les modalités et la couverture de leurs combinaisons conjonctives permettent d'identifier des sous-espaces denses qui modélisent des motifs réguliers. Les points ne suivant pas ces régularités sont considérés comme des anomalies, dont les raisons du caractère suspicieux peuvent être expliquées en utilisant les termes du vocabulaire. Les premières expérimentations menées sur des données artificielles montrent la pertinence des connaissances extraites des données qu'il s'agisse de l'identification des motifs réguliers ou des anomalies. Cependant, les partitions suggérées ne sont pertinentes qu'en cas de distributions multimodales des données et sont utilisables uniquement pour caractériser des sous-espaces convexes denses et séparés. Il apparait donc important de combiner cette stratégie coopérative avec des outils de visualisation des données pour permettre aux utilisateurs d'identifier les sous-espaces à séparer pour mieux décrire les différents motifs fréquents observables. L'inférence de variables linguistiques à affecter aux différentes modalités des partitions est un problème intéressant à aborder et qui complèterait l'apport de l'approche pour l'utilisateur final.

Ce travail s'inscrit dans le cadre du projet Sea Defender financé par la Direction Générale de l'Armement.

Références

[1] Christian Borgelt. An implementation of the fp-growth algorithm. In *Proceedings of the 1st international workshop on open source data mining : frequent pattern mining implementations*, pages 1–5, 2005.

[2] Didier Dubois and Henri Prade. Fuzzy cardinality and the modeling of imprecise quantification. *Fuzzy sets and Systems*, 16(3) :199–230, 1985.

[3] Sanjay Goil, Harsha Nagesh, and Alok Choudhary. Mafia : E±cient and scalable subspace clustering for very large data sets. In *Proc. 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Citeseer*, pages 443–452. Citeseer, 1999.

[4] Serge Guillaume and Brigitte Charnomordic. Generating an interpretable family of fuzzy partitions from data. *IEEE transactions on fuzzy systems*, 12(3) :324–335, 2004.

[5] Douglas M Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

[6] Marie-Jeanne Lesot, Grégory Smits, and Olivier Pivert. Adequacy of a user-defined vocabulary to the data structure. In *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE, 2013.

[7] Ninghao Liu, Donghwa Shin, and Xia Hu. Contextual outlier interpretation, 2018.

[8] Christophe Marsala. Fuzzy partition inference over a set of numerical values. In *Proc. of the IEEE Int. Conf. on Fuzzy Systems*, pages 1512–1517. Citeseer, 1995.

[9] Pierre Nerzic, Grégory Smits, Olivier Pivert, and Marie-Jeanne Lesot. Massive data exploration using estimated cardinalities. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE, 2022.

[10] Enrique H Ruspini. A new approach to clustering. *Information and control*, 15(1) :22–32, 1969.

[11] Grégory Smits, Marie-Jeanne Lesot, Véronne Yepmo Tchaghe, and Olivier Pivert. Panda : Human-in-the-loop anomaly detection and explanation. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems : 19th International Conference, IPMU 2022, Milan, Italy, July 11–15, 2022, Proceedings, Part II*, pages 720–732. Springer, 2022.

[12] Grégory Smits and Olivier Pivert. Linguistic and graphical explanation of a cluster-based data structure. In *International Conference on Scalable Uncertainty Management*, pages 186–200. Springer, 2015.

[13] Grégory Smits, Olivier Pivert, and Marie-Jeanne Lesot. Vocabulary elicitation for informative descriptions of classes. In *2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS)*, pages 1–8. IEEE, 2017.

[14] Anna Wilbik and James M Keller. Anomaly detection from linguistic summaries. In *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7. IEEE, 2013.

[15] Ronald R Yager, Marek Z Reformat, and Nhuan D To. Drawing on the ipad to input fuzzy sets with an application to linguistic data science. *Information Sciences*, 479 :277–291, 2019.

[16] Véronne Yepmo, Grégory Smits, and Olivier Pivert. Anomaly explanation : A review. *Data & Knowledge Engineering*, 137 :101946, 2022.

[17] Lotfi A Zadeh. Fuzzy logic= computing with words. In *Computing with words in information/intelligent systems 1*, pages 3–23. Springer, 1999.