Comparaison de résumés linguistiques

Marie-Jeanne Lesot, Grégory Smits

Comparaison de résumés linguistiques

Marie-Jeanne Lesot[1] Grégory Smits[2]
[1] Sorbonne Université CNRS, LIP6, Paris, Marie-Jeanne.Lesot@lip6.fr
[2] IMT Atlantique, Brest, Gregory.Smits@imt-atlantique.fr

Résumé :
Même lorsque des données tabulaires ne peuvent pas être exploitées directement, à cause de leur volumétrie ou du fait de leur caractère privé, il est possible que leurs résumés soient disponibles pour réaliser des analyses. Cet article propose de fournir des descriptions linguistiques des différences majeures entre deux jeux de données compatibles, c'est-à-dire qui traitent le même sujet. Une première stratégie exhaustive est proposée en générant une phrase décrivant les différences dans chaque sous-espace induit par le vocabulaire flou sur lequel sont construits les résumés. Une seconde stratégie est ensuite proposée pour les résumés condensés, qui ne contiennent que des phrases informatives. Les expériences réalisées sur des données artificielles confirment la pertinence de cette seconde stratégie en termes de coûts de calcul et d'informativité des changements identifiés.

Mots-clés :
Résumés linguistiques flous, résumés comparatifs, résumés différentiels.

Abstract:
When tabular data cannot be directly mined, due to their size or for privacy reasons, their summary may still be available for analysis. The approach proposed in this paper provides users with a linguistic description of the data changes between the fuzzy linguistic summaries of two datasets. A first strategy processes exhaustive summaries containing one sentence for each of the subspaces that can be formed using terms from the vocabulary. A second strategy is proposed for condensed summaries, that involve informative sentences only. Experimentation conducted on artificial datasets confirm the relevance of this second strategy in terms of computational cost and informativity of data changes that can be tracked.

Keywords:
Fuzzy linguistic summaries, comparative summaries, differential summaries

1 Introduction

Les résumés linguistiques flous offrent des vues d'ensemble lisibles du contenu de jeux de données, sous la forme d'ensembles de phrases qui suivent des schémas prédéfinis, par exemple $Q\,\mathcal{X}\,sont\,P$ (cf [4]). L'instanciation de ce schéma peut par exemple conduire à *quelques données sont* A_1*.médium et* A_2*.faible*, où A_1 et A_2 sont des attributs descriptifs des données $\mathcal{X}$, *médium* et *faible* des termes linguistiques imprécis qui leur sont associés. De tels résumés fournissent des vues concises et personnalisées de la distribution des données : ils constituent des descriptions linguistiques des clusters qui composent les données, ainsi que de quelques comportements plus rares ; la propriété de personnalisation vient de l'utilisation des termes linguistiques, dont les significations peuvent être définies individuellement pour chaque utilisateur.

Cet article porte sur la tâche de comparaison de résumés extraits de deux jeux de données $\mathcal{X}$ et $\mathcal{X}'$, décrits par les mêmes attributs et vocabulaire. $\mathcal{X}$ et $\mathcal{X}'$ peuvent correspondre à deux sous-populations d'un même ensemble de données, comme des étudiants ayant choisi des options différentes, ou des jeux de données collectés à des dates différentes, comme deux promotions d'étudiants d'une même formation. L'objectif est de fournir un aperçu intelligible des différences des distributions des données décrites par leurs résumés linguistiques.

La question de telles comparaisons est étudiée dans les tâches de détection et de traitement de *data drift* et *concept drift* (cf [2]). Toutefois ces méthodes considèrent le plus souvent que les données sont disponibles. Or il est possible qu'elles ne soient pas accessibles, en raison de contrainte de propriété ou de stockage. Dans de tels cas, les méthodes de fouille de données peuvent être appliquées aux résumés linguistiques des données au lieu des données elles-mêmes [9]. Cet article propose COPILS, *COmParIson of fuzzy Linguistic Summaries*, pour comparer les résumés linguistiques flous : COPILS génère des résumés dits *différentiels* ou *comparatifs*, qui permettent de comparer des jeux de données de façon lisible. Elle génère

ainsi des explications des données, s'insérant par là dans le cadre de l'IA explicable dynamique.

L'article est structuré de la façon suivante : la section 2 décrit formellement le contexte considéré. La section 3 présente l'approche proposée COPILS, qui est expérimentalement étudiée sur des jeux de données synthétiques dans la section 4. La section 5 conclut l'article.

2 Contexte et état de l'art

2.1 Notations et définitions

$\mathcal{X} = \{x_1, \ldots, x_n\}$ désigne un ensemble de n données, décrites par d attributs $\{A_1, \ldots, A_d\}$, qui peuvent être numériques ou catégoriels, respectivement définis sur les domaines D_j, $j = 1 \ldots d$. Chaque donnée est notée $x = (A_1.x, \ldots, A_d.x)$. Un vocabulaire flou $\mathcal{V}$ est défini comme un ensemble de partitions floues $\mathcal{V} = \{V_1, \ldots, V_d\}$, où $V_j = \langle A_j, \{\mu_{j1}, \ldots, \mu_{jq_j}\}, \{l_{j1}, \ldots, l_{jq_j}\}\rangle$ associe l'attribut A_j à des modalités, définies comme des sous-ensembles flous μ_{js} de son domaine D_j, et à des étiquettes linguistiques l_{js}. On fait l'hypothèse que les variables linguistiques forment une partition forte : $\forall j \in \{1..d\}, \forall y \in D_j, \sum_{s=1}^{q_j} \mu_{js}(y) = 1$. Comme illustré sur la fig. 1, une partition $\mathcal{Q}$ est également définie sur l'univers $[0, 1]$ pour décrire les cardinalités relatives.

Phrases. Chaque phrase d'un résumé est une instanciation d'un schéma prédéfini, appelé protoforme, écrit $Q\,\mathcal{X}\,sont\,P$ ou $Q\,R\,\mathcal{X}\,sont\,P$ [4] : Q est l'étiquette linguistique du quantificateur flou, pris dans $\mathcal{Q}$. Le *summarizer* P et le *qualifier* R sont des conjonctions de termes pris dans le vocabulaire $\mathcal{V}$. Cet article considère le premier cas, le second pouvant en être vu comme une instanciation, appliquée à un sous-ensemble flou des données restreint par le filtre R. Une telle protoforme peut être illustrée par la phrase *quelques données sont A_1.élevé et A_2.faible*.

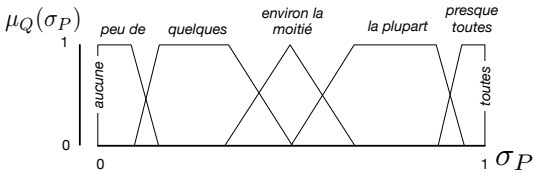

FIGURE 1 – Partition de 7 quantificateurs $\mathcal{Q} = \langle[0, 1], \{\mu_1, \cdots, \mu_7\}, \{\text{peu de}, \cdots, \text{toutes}\}\rangle$ pour décrire des cardinalités relatives

Chaque phrase est associée à un degré de vérité τ qui mesure à quel point elle est adéquate pour représenter les données [4] :

$$\tau(Q\,\mathcal{X}\,sont\,P) = \mu_Q(\sigma_P(\mathcal{X})),$$

où $\sigma_P(\mathcal{X}) = \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \mu_P(x)$. Si $P = A_{(1)}.m_{(1)}$ et $\ldots$ et $A_{(k)}.m_{(k)}$, où $A_{(i)}$ sont des attributs et $m_{(i)}$ des étiquettes linguistiques des modalités de leurs partitions associées, dont les fonctions d'appartenance sont $\mu_{(i)}$, sa fonction d'appartenance μ_P est définie comme $\mu_P(x) = \top_{i=1}^{k}\mu_{(i)}(A_{(i)}.x)$, où $\top$ est une t-norme.

Résumés. Un résumé $\mathcal{S}$ est un ensemble de phrases $Q\,\mathcal{X}\,sont\,P$ décrivant les données. Les *résumés exhaustifs* contiennent une phrase pour chaque P possible, c'est-à-dire pour chaque combinaison possible de $A.m$, y compris l'absence de l'attribut A. Pour chacune d'entre elles, le quantificateur le plus approprié, c'est-à-dire $Q \in \mathcal{Q}$ qui maximise $\mu_Q(\sigma_P(\mathcal{X}))$ est sélectionné. De tels résumés ont une longueur exponentielle car leur nombre de phrases est $\Pi_{j=1}^{d}(1+q_j)$. Un filtre évident consiste à exclure les phrases contenant un *summariser* P pour lequel le quantificateur associé est *aucun*.

Plusieurs approches ont été proposées pour élaguer davantage l'ensemble de phrases et ne conserver que les plus pertinentes, conduisant aux *résumés condensés*. Certaines n'évaluent pas les phrases individuellement, mais globalement, en mesurant leur redondance. Elles exploitent par exemple des relations connues entre les valeurs $\tau(Q\,\mathcal{X}\,sont\,P)$ et $\tau(Q\,\mathcal{X}\,sont\,P\,et\,P')$. Ces méthodes diffèrent par le critère d'élagage qu'elles

utilisent [8, 11]. D'autres méthodes évitent la génération de phrases non pertinentes supprimées ultérieurement, en utilisant des approches intégrées : certaines exploitent les relations entre résumés linguistiques et règles d'association [5, 6], d'autres exploitent un principe d'antimonotonie de critères de qualité, par exemple le degré de focus, en plus de degré de vérité [12, 13].

2.2 Distances au niveau linguistique flou

La comparaison de jeux de données à travers le prisme d'un vocabulaire flou repose sur des mesures de distance à différents niveaux. D'abord, pour les valeurs correspondant à un unique attribut, $y, y' \in D$, une distance peut être calculée pour prendre en compte la partition floue associée à D noté V, plus précisément les degrés d'appartenance aux modalités, ainsi que le nombre de modalités qui les séparent [3] :

$$d_V(y, y') = \frac{1}{q_V - 1} \times \\ |\mu_{I(y)}(y) - \mu_{I(y')}(y') + I(y') - I(y)|, \quad (1)$$

où q_V est le nombre de modalités dans V et $I(y) \in \{1, \ldots, q_V\}$ l'indice de la zone, définie par les bornes inférieures des noyaux, à laquelle y appartient.

Pour comparer deux phrases, il a été proposé dans [14] la mesure

$$d(Q\,\mathcal{X}\ sont\ P\,, \tau, Q'\,\mathcal{X}'\ sont\ P'\,, \tau') = \\ 1 - \min(sim_1(Q, Q'), sim_2(P, P'), sim_3(\tau, \tau')), \quad (2)$$

où sim_i sont des mesures de similarité à définir. L'utilisation du minimum implique que les trois termes jouent le même rôle dans la comparaison. Cette distance a été utilisée pour structurer un ensemble de phrases en groupes de phrases similaires [14], mais elle ne semble pas appropriée pour la tâche considérée dans cet article. En effet, des phrases contenant le même *summariser* mais des quantificateurs différents doivent être considérées comme plus proches que des phrases qui appliquent un même quantificateur à des *summarisers* différents. Une mesure présentant cette sémantique est proposée dans la section 3.

La comparaison de résumés linguistiques peut être aussi être considérée comme liée à la question de motifs émergents [1, 10] : ces derniers caractérisent un jeu de données par opposition à une référence à l'aide d'itemsets dont la qualité est définie comme le quotient de leurs supports calculés sur les deux jeux de données.

3 Approche proposée : COPILS

Cette section décrit puis illustre sur des exemples la méthode COPILS qui permet de comparer des résumés linguistiques flous extraits de deux jeux de données.

3.1 Comparaison de résumés exhaustifs

La comparaison de résumés exhaustifs est une tâche aisée, car chaque *summariser* P apparaît dans une phrase exactement de chaque résumé.

COPILS extrait les couples de phrases ayant le même *summariser* mais des quantificateurs différents : $s = Q\,\mathcal{X}\,sont\ P, \tau$ et $s' = Q'\,\mathcal{X}'\,sont\,P, \tau'$. Pour chaque couple (s, s'), il génère une caractérisation différentielle de la forme suivante (d_{quant} est définie dans l'éq. 3 ci-dessous) :
– si $Q =$ *aucun* et $Q' \neq$ *aucun*, l'apparition d'un groupe dans un sous-espace initialement vide est décrit, dans le résumé différentiel, par
AJOUT d = $d_{quant}(s, s')$ *:* $Q'\ \mathcal{X}'$ *sont* P, τ'
– si $Q \neq$ *aucun* et $Q' =$ *aucun*, la disparition d'un groupe est décrite par
SUPP. d = $d_{quant}(s, s')$ *:* $Q\ \mathcal{X}$ *sont* P, τ
– si $Q \neq$ *aucun* et $Q' \neq$ *aucun*, le changement de cardinalité est décrit par
MODIF. d = $d_{quant}(s, s')$ *:* $Q\mathcal{X}$ *sont* $P, \tau \Rightarrow Q'\mathcal{X}'$ *sont* P, τ'.
d_{quant} est une quantification de la dissimilarité entre quantificateurs :

$$d_{quant}(s, s') = \frac{1}{|\mathcal{Q}| - 1} \times |I(Q) - I(Q')|, \quad (3)$$

où $I(Q)$ est l'indice du quantificateur Q dans

la partition $\mathcal{Q}$ de taille $|\mathcal{Q}|$. Ainsi, dans le cas de la partition illustrée sur la fig. 1, une modification du quantificateur *la plupart* en *quelques*, de degré $|3-5|/(7-1)$ = 1/3, est plus importante qu'une modification en *environ la moitié*, associée à $|5-4|/(7-1)$ = 1/6. COPILS affiche les changements dans l'ordre décroissant de $d_{quant}(s, s')$, afin de prioriser les plus significatifs, augmentant ainsi la lisibilité des résumés générés.

Propriétés Une première spécificité de COPILS est qu'un *summariser* associé à un même quantificateur dans les deux résumés ne conduit pas à une phrase comparative, même si leurs degrés de vérité τ et τ' respectifs diffèrent : une telle différence qui n'induit pas un changement de quantificateur n'est pas assez significative pour être mentionnée. Ainsi, les différences sont mesurées à l'échelle des termes linguistiques utilisés dans la partition $\mathcal{Q}$ associée aux cardinalités relatives. Ceci permet d'intégrer les préférences de l'utilisateur dans l'expression du résultat, en tenant compte de la granularité qu'il indique comme faisant sens pour lui.

Une seconde propriété est que la comparaison est aussi exhaustive que les résumés initiaux : une modification de quantificateur pour un *summariser* complexe P, constitué d'une conjonction de modalités, implique qu'au moins l'une de ses composantes est également associée à une telle modification. Aussi, toutes génèrent une phrase dans le résumé différentiel, ce qui peut conduire à des résultats de grande taille. Il peut être pertinent de sélectionner certaines d'entre elles, en mettant en œuvre une stratégie d'élagage, similaire à celles mentionnées dans la section 2.1. Toutefois, il est difficile de déterminer de façon générique si un utilisateur préfère se concentrer sur les différences de *summariser* maximaux ou, au contraire, sur les minimaux, en termes de nombre de modalités. Aussi, COPILS n'applique pas de stratégie d'élagage par défaut.

Globalement, la mesure de dissimilarité entre phrases utilisée par COPILS peut être formalisée comme suit : si les *summariser*s diffèrent, elle est arbitrairement élevée, sinon elle dépend de la différence entre les quantificateurs :

$$d(s, s') = \begin{cases} 1 & \text{si } P \neq P', \\ d_{quant}(Q, Q') & \text{sinon.} \end{cases} \quad (4)$$

A la différence de la mesure rappelée dans l'équation (2), elle donne un rôle asymétrique au *summariser* et au quantificateur et elle ne dépend pas des degrés de vérité.

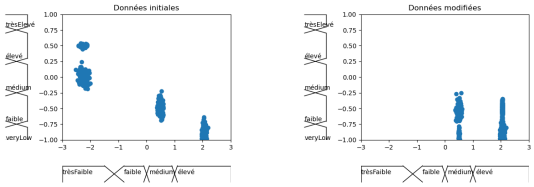

FIGURE 2 – Exemple illustratif : $\mathcal{X}$ et $\mathcal{X}'$

TABLEAU 1 – Résumés de $\mathcal{X}$

Résumé exhaustif : 10 phrases
'peu de données sont A_1.veryLow τ=1
'quelques données sont A_1.low' τ=1
'quelques données sont A_1.medium' τ=1
'quelques données sont A_1.high' τ=1
'la plupart des données sont A_2.low' τ=1
'quelques données sont A_2.medium'τ=1
'peu de données sont A_1.veryLow et A_2.low' τ=1
'quelques données sont A_1.low et A_2.low' τ=1
'quelques données sont A_1.medium et A_2.medium' τ=1
'quelques données sont A_1.high et A_2.low' τ=1

Résumé condensé : 5 phrases
'la plupart des données sont A_2.low' τ=1
'peu de données sont A_1.veryLow et A_2.low' τ=1
'quelques données sont A_1.low et A_2.low' τ=1
'quelques données sont A_1.medium et A_2.medium' τ=1
'quelques données sont A_1.high et A_2.low' τ=1

Exemple illustratif La fig. 2 montre, à gauche, un jeu de données $\mathcal{X}$ et, à droite, sa version modifiée $\mathcal{X}'$, avec les partitions floues associées à A_1 (abscisses) et A_2 (ordonnées). Après suppression des phrases associées au quantificateur *aucun*, le résumé exhaustif de $\mathcal{X}$, qui pourrait contenir jusqu'à $(q_1+1)\times(q_2+1) = 30$ phrases, contient les 10 phrases indiquées dans le tab. 1 qui donne également le résumé condensé (voir définition ci-dessous).

COPILS conduit alors au résumé différentiel de taille 12 :

– MODIF. $d = 2/3$: *'peu de données sont A_1.veryLow' τ=1 $\Rightarrow$ 'presque toutes les données sont A_1.veryLow' τ=1*,
– AJOUT $d = 0.5$: *'environ la moitié des données sont A_2.veryLow' τ=0.69*,
– AJOUT $d = 0.5$: *'environ la moitié des données sont A_1.veryLow et A_2.veryLow' τ=0.69*,
– SUPP. $d = 1/3$: *'quelques données sont A_1.low' τ=1*,
– SUPP. $d = 1/3$: *'quelques données sont A_1.medium' τ=1*,
– SUPP. $d = 1/3$: *'quelques données sont A_1.low et A_2.low' τ=1*,
– SUPP. $d = 1/3$: *'quelques données sont A_1.medium et A_2.medium' τ=1*,
– MODIF. $d = 1/3$: *'la plupart des données sont A_2.low' τ=1 $\Rightarrow$ 'quelques données sont A_2.low' τ=0.9*,
– AJOUT $d = 1/6$: *'peu de données sont A_1.veryLow et A_2.medium' τ=1*,
– MODIF. $d = 1/6$: *'quelques données sont A_1.high' τ=1 $\Rightarrow$ 'peu de données sont A_1.high' τ=1*,
– MODIF. $d = 1/6$: *'quelques données sont A_2.medium' τ=1 $\Rightarrow$ 'peu de données sont A_2.medium' τ=1*,
– MODIF. $d = 1/6$: *'quelques données sont A_1.high et A_2.low' τ=1 $\Rightarrow$ 'peu de données sont A_1.high et A_2.low' τ=1*.
Il montre toutes les différences et, ainsi, tous les changements attendus. Il reste cependant difficile à lire du fait des redondances qu'il contient.

3.2 Comparaison de résumés condensés

Afin de réduire la taille des résumés différentiels, COPILS comporte une variante permettant de traiter des résumés condensés, obtenus par élagage des résumés exhaustifs.

Plus précisément, parmi les divers élagages possibles (voir la brève discussion dans la section 2.1), nous considérons l'approche qui repose sur les phrases maximales [13]. Elles sont définies comme des instanciations de la protoforme $s = Q\,\mathcal{X}\,sont\,P$ telles qu'aucune instanciation $s' = Q\,\mathcal{X}\,sont\,P et\,P'$ n'est valide. Ainsi, si la phrase *'quelques données sont A_1.medium et A_2.medium'* est sélectionnée, les phrases concernant *A_1.medium et A_2.medium* individuellement ne sont conservées que si elles couvrent une proportion différente des données. Un résumé condensé ne contient alors que des phrases maximales, ce qui conduit à une taille nettement inférieure à celle d'un résumé exhaustif (cf tab. 1 et fig. 3).

La difficulté du traitement des résumés condensés vient du fait qu'une phrase de *summariser* P dans l'un des résumés peut ne pas avoir d'équivalent dans l'autre, suite à une suppression éventuelle. La comparaison de résumés condensés nécessite donc une phase de mise en correspondance des phrases.

Mise en correspondance optimale. COPILS repose sur une variante de l'algorithme des mariages stables : deux phrases sont mises en correspondance s'il n'existe pas de phrases plus proches encore non associées avec lesquelles elles puissent être mises en correspondance respectivement. La variante permet de traiter des nombres de phrases qui diffèrent entre les résumés, conduisant éventuellement à des phrases qui restent non associées.

La mise en correspondance nécessite une mesure de dissimilarité entre *summariser*s (éq. 5) : seules les phrases pour lesquelles elle est inférieure à un seuil η fixé par l'utilisateur sont conservées. Nous proposons de la définir comme

$$d_{sum}(P, P') = \frac{1}{\max(|P|, |P'|)} \sum_{j=1}^{d} d^j_{mod}(P, P'), \quad (5)$$

où $d^j_{mod}(P, P')$ compare les composantes, combinées conjonctivement dans P et P', qui concernent l'attribut j. Si P ou P' n'en contient pas, la distance vaut 1, sinon elle est égale à la différence, en valeur absolue, entre les indices de ces modalités dans V_j, suivant le principe de l'éq. (3), en remplaçant la partition des quantificateurs par V_j.

La différence entre phrases de résumés condensés est définie comme la différence entre *summariser*s éventuellement combinée à celle entre quantificateurs :
si $d_{sum}(P, P') > \eta$, $d(s, s') = 1$

sinon, $d(s, s') = \max(d_{sum}(s, s'), d_{quant}(s, s'))$.

Par rapport à la distance rappelée dans l'éq. (2), celle-ci présente les mêmes caractéristiques que l'éq. (4) : elle ne dépend pas des degrés de vérité et donne un rôle asymétrique aux comparaisons entre *summariser*s et quantificateurs, en les combinant dans une procédure hiérarchique. En effet, la comparaison entre *summariser*s joue un rôle prépondérant, la distance entre quantificateurs n'intervenant que dans un second temps.

Forme du résumé différentiel généré. Les résumés différentiels construits à partir de résumés condensés ajoutent une caractérisation aux 3 cas considérés pour les résumés exhaustifs : les situations d'ajout (resp. suppression) de données correspondent aux cas où une phrase du résumé de $\mathcal{X}'$ (resp. $\mathcal{X}$) n'est pas mise en correspondance. Les phrases associées conduisent à deux types de caractérisations, selon que leurs *summariser*s sont identiques (ce sont alors des modifications, comme précédemment) ou non. Dans ce dernier cas, elles sont interprétées comme de possibles déplacements d'un sous-espace à un autre, conduisant à
DEPL. POSS. $d = d(s, s')$: *$Q\mathcal{X}$ sont $P, \tau \Rightarrow Q'\mathcal{X}'$ sont P', τ'.*
Ici aussi, les caractérisations différentielles sont présentées par COPILS dans l'ordre décroissant de leur distance associée. Cette dernière est définie comme la distance entre les phrases mises en correspondance dans le cas de DEPL. POSS. et comme la distance entre les quantificateurs dans les trois autres cas.

Exemple illustratif. Pour les données représentées sur la fig. 2, le résumé condensé de $\mathcal{X}$ contient seulement 5 phrases (cf tab. 1). Ainsi, le groupes de données avec les valeurs les plus élevées pour l'attribut A_2 est décrit par l'unique phrase *'quelques données sont A_1.médium et A_2.médium'* : la propriété de phrase maximale garantit que les phrases *'quelques données sont A_1.medium'* et *'quelques données sont A_2.medium'* sont également valides et ne sont pas incluses. COPILS génère alors le résumé différentiel de taille 6 suivant :
– DEPL. POSS $d = \max(1/6, 7/12)$: *'quelques données sont A_1.low et A_2.low', τ=1 $\Rightarrow$ 'environ la moitié des données sont A_1.veryLow et A_2.veryLow', τ=0.69*,
– DEPL. POSS. $d = \max(1/6, 1/3)$: *'quelques données sont A_1.medium et A_2.medium', τ=1 $\Rightarrow$ 'peu de données sont A_1.veryLow et A_2.medium', τ=1*,
– MODIF. $d = 1/3$: *'la plupart des données sont A_2.low', τ=1 $\Rightarrow$ 'quelques données sont A_2.low', τ=0.9*.
– AJOUT $d = 1/3$: *'quelques données sont A_1.veryLow et A_2.veryHigh', τ=1*,
– AJOUT $d = 1/6$: *'presque toutes les données sont A_1.veryLow', τ=1*,
– MODIF. $d = 1/6$: *'quelques données sont A_1.high et A_2.low', τ=1 $\Rightarrow$ 'peu de données sont A_1.high et A_2.low', τ=1*.

Le résultat couvre bien les changements attendus entre $\mathcal{X}$ et $\mathcal{X}'$, sous une forme plus lisible que dans le cas exhaustif. De plus, le fait d'autoriser des mises en correspondances partielles permet de suggérer des déplacements possibles.

4 Résultats expérimentaux

Cette section étudie expérimentalement les deux variantes de COPILS, en termes de nombre de phrases et de temps de calcul, sur des données artificielles.

4.1 Génération de données artificielles

L'utilisation de données synthétiques permet de contrôler la complexité de la tâche de résumé, en choisissant les paramètres de génération. Afin d'obtenir des résumés pertinents, et ainsi des comparaisons pertinentes, il est important de garantir l'adéquation entre la discrétisation des domaines d'attributs opérée par les partitions floues et la distribution des données [7]. Dans ce but, nous proposons un processus de génération qui prend en entrée le vocabulaire.

Plus précisément, nous générons des données ayant entre 2 et 8 attributs numériques, as-

sociées à des partitions floues définies manuellement pour chacun d'eux, comportant entre 3 à 6 modalités. Les données sont ensuite générées par des distributions gaussiennes, dont les moyennes et écarts-types sont choisis de façon à ce que chaque cluster soit principalement couvert par une modalité sur chaque attribut, garantissant l'adéquation du vocabulaire.

Etant donné un jeu de données $\mathcal{X}$ généré selon cette procédure, une version modifiée $\mathcal{X}'$ est ensuite produite de façon à ce que les modifications attendues dans le résumé comparatif soient connues à l'avance : chaque cluster de $\mathcal{X}$ est modifié avec une probabilité uniforme, selon deux types de modifications. La première supprime aléatoirement certaines données du cluster ou en ajoute, selon sa distribution gaussienne ; la seconde supprime le cluster considéré et en crée un nouveau, dans un sous-espace choisi aléatoirement. Il peut différer du sous-espace initial par un nombre variable d'attributs.

Les hyper-paramètres du processus de génération sont le nombre d'attributs, le vocabulaire flou et le nombre de clusters. Les autres paramètres sont fixés aléatoirement. Le nombre total de données est constant : il influence seulement l'étape préliminaire de génération des résumés de $\mathcal{X}$ et $\mathcal{X}'$. COPILS prend en entrée ces résumés et ne dépend pas des jeux de données eux-mêmes. Les données de la fig. 2 correspondent à ce processus avec 2 attributs, associés à des partitions de 4 et 5 modalités illustrées sur la figure, et 4 clusters.

Taille des résumés initiaux A titre d'obervation préliminaire, le graphe de gauche de la fig. 3 montre la moyenne et l'écart-type de la taille des résumés linguistiques de $\mathcal{X}$, exhaustifs et condensés, calculés sur 10 jeux de données dans chaque configuration. Il montre que, comme attendu, pour plus de 4 attributs, les résumés exhaustifs contiennent plus de 100 phrases, ce qui les rend illisibles pour les utilisateurs. Les résumés condensés augmentent également exponentiellement en fonction du nombre d'attributs, mais avec une facteur plus faible, conduisant à des résultats plus lisibles.

TABLEAU 2 – Taille des résumés différentiels : #C : nombre de clusters, #Ch : nombre moyen de changements, #Texh (#Tcond) taille des résumés exhaustifs (condensés)

#C	2	3	4	5
#Ch	1.4	2.2	2.9	3.9
#Texh	41.4	51.4	66.7	75.6
#Tcond	6.7	8.5	11.6	12

Taille des résumés différentiels Le graphe central de la fig. 3 montre la moyenne et l'écart-type de la taille des résumés différentiels. Comme attendu, dans le cas exhaustif, ils sont identiques aux longueurs des résumés de $\mathcal{X}$; dans le cas condensé, ils peuvent être légèrement supérieurs. Ils sont également trop élevés pour un utilisateur dans le cas exhaustif et acceptables dans le cas condensé, montrant que l'approche COPILS est prometteuse.

Temps de calcul Le graphe de droite de la fig. 3 montre que, comme attendu, le temps de génération des résumés différentiels dépend exponentiellement du nombre d'attributs, avec un degré plus élevé pour les résumés exhaustifs que pour leur variante condensée. Ceci est dû à la taille des entrées traitées par COPILS.

On observe que, pour 3 attributs ou plus, l'augmentation du temps de calcul dû à la mise en correspondance est négligeable par rapport au gain de traitement d'entrées de taille bien inférieure : le nombre nettement plus faible de phrases à traiter compense le surcoût de calcul.

5 Conclusions et perspectives

Afin de fournir une vue lisible et intelligible des changements entre jeux de données décrits sur le même espace, dans le cas où seuls leurs résumés linguistiques sont disponibles, COPILS identifie les additions, suppressions et déplacements possibles. Pour les résumés condensés, COPILS exploite des dissimilarités

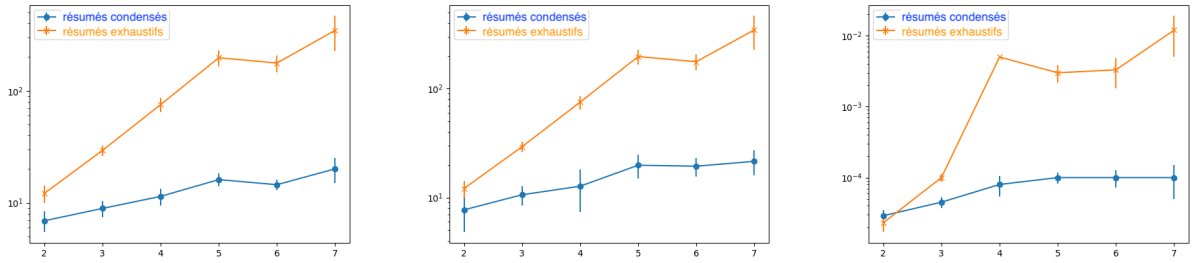

FIGURE 3 – (Gauche) Taille des résumés de $\mathcal{X}$, (milieu) taille des résumés différentiels de $\mathcal{X}$ et $\mathcal{X}'$, (droite) temps de calcul (en secondes) dans les cas exhaustif et condensé, en fonction du nombre d'attributs (échelles log).

appropriées et une mise en correspondance pour générer efficacement des résumés comparatifs de taille acceptable. Les perspectives incluent une évaluation qualitative de l'interprétabilité des résumés différentiels générés. Afin de mesurer leur qualité par le biais de leur utilité, une piste consiste à examiner la capacité des utilisateurs à reconstruire le jeu de données initial à partir des comparaisons produites par COPILS.

Références

[1] G. Dong and J. Li. Mining border descriptions of emerging patterns from dataset pairs. *Knowledge and Information Systems*, 5 :178–202, 2005.

[2] R. Gemaque, A. Costa, R. Giusti, and E. Dos Santos. An overview of unsupervised drift detection methods. *Data Mining and Knowledge Discovery*, 10(6), 2020.

[3] S. Guillaume, B. Charnomordic, and P. Loisel. Fuzzy partitions : a way to integrate expert knowledge into distance calculations. *Information sciences*, 245 :76–95, 2013.

[4] J. Kacprzyk and S. Zadrożny. Protoforms of linguistic data summaries : Towards more general natural-language-based data mining tools. In *Soft computing systems*, pages 417–425. 2002.

[5] J. Kacprzyk and S. Zadrożny. Linguistic summarization of data sets using association rules. In *Fuzz-IEEE*, 2003.

[6] J. Kacprzyk and S. Zadrożny. Derivation of linguistic summaries is inherently difficult : Can association rule mining help ? In *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, pages 291–303. 2013.

[7] M.-J. Lesot, G. Smits, and O. Pivert. Adequacy of a user-defined vocabulary to the data structure. In *Fuzz-IEEE*, 2013.

[8] D. Pilarski. Linguistic summarization of databases with quantirius : a reduction algorithm for generated summaries. *IJUFKS*, 18(3), 2010.

[9] G. Smits, R. R. Yager, and O. Pivert. Interactive data exploration on top of linguistic summaries. In *Fuzz-IEEE*, 2017.

[10] A. Soulet, B. Crémilleux, and F. Rioult. Condensed representation of emerging patterns. In *PAKDD*. 2004.

[11] A. Wilbik and R. M. Dijkman. On the generation of useful linguistic summaries of sequences. In *Fuzz-IEEE*, 2016.

[12] A. Wilbik and J. Kacprzyk. Towards an efficient generation of linguistic summaries of time series using a degree of focus. In *NAFIPS*, 2009.

[13] A. Wilbik, U. Kaymak, and R. M. Dijkman. A method for improving the generation of linguistic summaries. In *Fuzz-IEEE*, 2017.

[14] A. Wilbik and J. M. Keller. A distance metric for a space of linguistic summaries. *Fuzzy Sets and Systems*, 208 :79–94, 2012.