Mesures d'adéquation entre vocabulaire expert et structure de données

Adequacy between an expert vocabulary and a data structure

Marie-Jeanne Lesot[1] Grégory Smits[2] Olivier Pivert[3]

[1] LIP6, UPMC Univ Paris 06, CNRS, UMR 7606, 4 place Jussieu, 75005 Paris, marie-jeanne.lesot@lip6.fr
[2] IRISA - IUT, 3, rue E. Branly - BP 30219, 22302 Lannion cedex, gregory.smits@irisa.fr
[3] IRISA - ENSSAT, 6, rue de Kerampont - BP 80518, 22305 Lannion cedex, pivert@enssat.fr

Résumé :

L'adéquation entre un vocabulaire expert utilisé pour décrire linguistiquement un ensemble de données et la structure de ces dernières est requise pour garantir la pertinence et la cohérence de l'expression des résultats d'un processus de découverte des données personnalisé. Cet article propose deux mesures pour répondre à cette tâche : la première est basée sur la comparaison des partitionnements respectivement obtenus à partir de la représentation initiale des données et à partir des données réécrites ; la seconde évalue, dans l'espace réécrit, les regroupements identifiés dans l'espace initial. Les expérimentations sur des données artificielles montrent qu'elles permettent d'identifier des vocabulaires pertinents.

Mots-clés :

Variables linguistiques, classification non supervisée, adéquation, structure des données

Abstract:

The adequacy between an expert vocabulary used to linguistically describe a data set and the structure of the latter is required to guarantee the relevance and consistency of the expression of the results obtained in a personalised knowledge discovery task. This paper proposes two measures to address this task : the first one is based on the comparison between the partitions obtained from the initial data and the rewritten data representations ; the second one assesses, in the rewritten space, the clusters obtained from the initial data. Experimental results obtained from artificial data show these measures make it possible to identify relevant vocabularies.

Keywords:

Linguistic variables, clustering, adequacy, data structure

1 Introduction

La prise en compte des préférences et des besoins des utilisateurs permet de personnaliser le processus de découverte de connaissances dans des ensembles de données et d'augmenter leur pertinence [2]. Ces préférences peuvent en particulier être exprimées par le biais d'un vocabulaire expert, modélisé par des variables linguistiques floues, qui permet de décrire des propriétés des données qui ont du sens pour l'analyste. La pertinence de l'utilisation de tels vocabulaires pour la personnalisation a été démontrée dans de nombreux contextes applicatifs, parmi lesquels la recherche par facettes [12], les requêtes flexibles coopératives [1] ou le résumé de données [10, 13].

Il est alors nécessaire que la définition du vocabulaire soit cohérente par rapport à la structure sous-jacente des données : le vocabulaire induit une relation d'indistinguabilité sur les données, puisque deux objets ne peuvent être distingués s'ils satisfont une modalité floue au même degré. Cette relation est légitime, puisqu'elle correspond à des données qui sont également appréciées par l'expert ; toutefois, elle doit être définie de sorte à être compatible avec la structure des données. En effet, d'une part, deux données similaires par leurs descriptions numériques ne doivent pas être séparées par leurs descriptions linguistiques. Réciproquement, deux données indistinguables d'après le vocabulaire ne doivent pas appartenir à des sous-groupes de données distincts. De telles inadéquations devraient conduire à nuancer le vocabulaire de l'analyste, afin de préserver à la fois l'adéquation avec les données et la compatibilité avec ses préférences subjectives exprimées par les modalités floues.

Le problème de l'adéquation entre vocabulaire expert et structure de données a été soulevé [10, 13], mais n'a pas été étudié. Dans cet article, nous proposons de l'interpréter comme un

problème de compatibilité entre deux processus de classification non supervisée de données, plus précisément la compatibilité entre la structure sous-jacente des données, extraite par un processus de clustering appliqué aux données initiales et la structure induite par le vocabulaire, identifiée par un processus de clustering appliqué aux données réécrites comme des vecteurs de leurs degrés d'appartenance aux modalités floues considérées. Nous proposons deux critères pour mesurer cette compatibilité : le premier évalue l'accord des partitions obtenues par les deux processus de clustering cités ; le second mesure à quel point les clusters obtenus par clustering des données initiales sont pertinents en les évaluant dans l'espace de représentation réécrit induit par le vocabulaire considéré.

L'article présente successivement les différentes étapes de la méthode proposée : la section 2 décrit l'étape de réécriture des données selon le vocabulaire à évaluer ; la section 3 décrit l'étape de clustering, justifiant l'algorithme ainsi que les mesures de comparaison de données choisis ; la section 4 présente les critères d'adéquation proposés. La section 5 décrit les résultats obtenus sur une base de données artificielles.

2 Etape de réécriture

2.1 Données et vocabulaire

Les données, notées $\mathcal{D}$, sont constituées d'objets $\{x_1, x_2, ..., x_n\}$ décrits par m attributs $\mathcal{A} = \{A_1, A_2, ..., A_m\}$, catégoriels ou numériques, respectivement définis sur les domaines $\mathcal{D}_j$.

Le vocabulaire $\mathcal{V}$ dont l'adéquation est à mesurer est défini comme un ensemble de variables linguistiques, qui associent chaque attribut à une étiquette linguistique et une partition floue de type Ruspini [11] : formellement, pour l'attribut A_j, $j = 1..m$, on note a_j le nombre de modalités associées et $\mathcal{V}_j = \{v_{j1}, \dots v_{ja_j}\}$ les sous-ensembles flous associés. La propriété de partition de Ruspini impose $\forall j = 1..m$, $\forall x \in \mathcal{D}_j$, $\sum_{k=1}^{a_j} \mu_{v_{jk}}(x) = 1$.

2.2 Données réécrites

Chaque donnée peut alors être réécrite, en calculant les degrés d'appartenance à chaque modalité de chaque variable linguistique puis en concaténant ceux-ci : la donnée x est représentée par le vecteur à $\sum_{j=1}^{m} a_j$ composantes $\langle \mu_{v_{11}}(x), \dots, \mu_{v_{1a_1}}(x), \dots, \mu_{v_{m1}}(x), \dots, \mu_{v_{ma_m}}(x) \rangle$.

Il faut noter que chaque point ne peut satisfaire partiellement que deux modalités pour chaque attribut. Aussi le vecteur ci-dessus comporte au plus $2m$ composantes non nulles.

2.3 Indistinguabilité

Les représentations numériques ou catégorielles offrent des descriptions précises mais peu interprétables d'un objet. Au contraire, les variables linguistiques floues améliorent l'interprétabilité, en particulier lorsqu'elles sont définies par l'expert en charge de l'analyse des données. Elles induisent cependant une imprécision qui ne permet pas de distinguer des objets qui satisfont au même degré une modalité floue. Cette relation d'indistinguabilité est légitime puisqu'elle correspond à des objets qui ne sont pas différenciés par l'expert, mais également préférés.

Néanmoins, pour garantir la pertinence et la cohérence des résultats exprimés avec ce vocabulaire, elle ne doit pas être décorrélée de la structure sous-jacente des données, mais être compatible avec celle-ci. La Figure 1 par exemple illustre deux types d'incompatibilité non souhaitables : les données du cluster C'_3, qui appartiennent à un même cluster d'après la structure sous-jacente, ont des descriptions linguistiques distinctes qui les affectent à 3 clusters différents, C_2, C_3 et C_4. Réciproquement, les objets des clusters C'_1 et C'_2 ont la même réécriture, mais devraient être distingués d'après la structure sous-jacente.

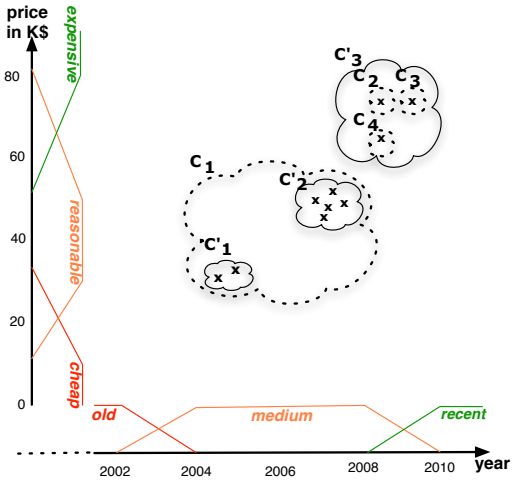

Figure 1 – Comparaison des partitions basées sur le vocabulaire expert (clusters C_1 à C_4) et sur la représentation initiale des données (clusters C'_1 à C'_3).

3 Etape de clustering

3.1 Choix de l'algorithme de clustering

De nombreux algorithmes de clustering peuvent être envisagés a priori [6], toutefois le contexte applicatif impose des contraintes qui guident la sélection : l'algorithme doit être capable de traiter de grands volumes de données, de déterminer automatiquement le nombre de clusters, que l'utilisateur ne peut indiquer a priori, et de traiter des données hétérogènes, décrites par des attributs numériques et catégoriels.

Nous proposons d'utiliser l'algorithme *l-fcmed-select* [8] qui constitue une extension de l'algorithme des c-médoïdes linéarisé *l-fcmed* [7]. Ce dernier présente 3 caractéristiques motivant son utilisation : il définit les centres des clusters comme des médoïdes, c'est-à-dire les données qui minimisent la distance aux membres du cluster, et non comme des points fictifs calculés comme des moyennes par exemple, ceci permet une application à des données hétérogènes ; il effectue une affectation floue, ce qui lui apporte robustesse et indépendance à l'initialisation aléatoire ; il met en œuvre une approximation pour la mise à jour des médoïdes, en recherchant la nouvelle position dans le voisinage de la position précédente, ce qui diminue son temps de calcul.

L'extension *l-fcmed-select* [8] constitue une variante incrémentale de *l-fcmed*, qui est appliqué à des échantillons de données, conduisant à des résultats locaux fusionnés ultérieurement par clustering hiérarchique. De plus, lorsqu'un échantillon a été traité, une étape d'augmentation des clusters teste si les données non encore traitées peuvent être affectées aux clusters identifiés. Cette étape permet d'une part de diminuer la quantité de données restant à traiter et d'autre part d'éviter l'identification de clusters trop similaires aux clusters déjà identifiés. Elle réduit donc le temps de calcul, à la fois pour les étapes de *l-fcmed* et pour leur fusion finale.

En outre, afin de sélectionner le nombre de clusters, *l-fcmed-select* ajoute une étape de sélection de médoïdes : *l-fcmed* identifie c clusters, que cette valeur soit pertinente ou non pour les données. Aussi, *l-fcmed-select* applique à chaque échantillon *l-fcmed* avec une valeur élevée de c et sélectionne ensuite uniquement les clusters pertinents, selon des critères de taille et de compacité [8].

3.2 Choix des mesures de distance

Pour l'espace de représentation initial, nous définissons la distance comme la moyenne des distances calculées pour chaque attribut, c'est-à-dire $d(x,y) = 1/q \sum_{j=1}^{m} d_i(x_j, y_j)$, où d_j est la distance associée à l'attribut A_j, définie comme d_{cat} pour les attributs catégoriels et d_{num} pour les attributs numériques :

$$d_{cat}(x,y) = \begin{cases} 1 & \text{si } x \neq y \\ 0 & \text{sinon} \end{cases} \quad (1)$$

$$d_{num}(x,y) = \frac{|x-y|}{\max(x,y)} \quad (2)$$

La distance pour attribut catégoriel est binaire et vaut 0 si les deux valeurs comparées sont identiques, 1 sinon. La distance pour attributs numériques calcule un écart relatif : ainsi, une différence de prix de 2 000€ n'a pas le même

effet suivant que les valeurs comparées sont de l'ordre de $100\,000$€ ou de $10\,000$€.

Pour les données réécrites, nous proposons de définir la distance comme la somme des distances obtenues pour chaque variable linguistique, en utilisant pour celle-ci la distance proposée par [3]. Pour un attribut A associé à une variable linguistique à a modalités définissant une partition forte, ordonnées de telle sorte que leurs noyaux $[\underline{K}_i, \bar{K}_i]$ vérifient $\bar{K}_i \leq \underline{K}_{i+1}$, en notant $x = (x_i)_{i=1..a}$ et $y = (y_i)_{i=1..a}$ deux vecteurs de degrés d'appartenance, I la fonction telle que $I(x) = i \Leftrightarrow x \in [\underline{K}_i, \underline{K}_{i+1}]$, et $\eta(x) = I(x) - \mu_{I(x)}(x)$, la distance est définie comme

$$d_A(x, y) = \frac{1}{a-1}|\eta(x) - \eta(y))| \qquad (3)$$

4 Mesures d'adéquation proposées

Deux critères d'évaluation de l'adéquation du vocabulaire considéré à l'ensemble de données sont proposés, basés sur les résultats de l'étape de clustering décrite dans la section précédente.

4.1 Comparaison des partitions

Une première approche consiste à comparer les partitions obtenues lorsque les données $\mathcal{D}$ sont respectivement décrites dans l'espace initial et réécrites selon le vocabulaire $\mathcal{V}$. Ces partitions sont respectivement notées $\mathcal{C}(\mathcal{D})$ et $\mathcal{C}(\mathcal{RD}_\mathcal{V})$.

La comparaison de partitions a donné lieu à de multiples critères [9], beaucoup s'expriment en fonction des 4 quantités suivantes : a est le nombre de paires de données affectées au même cluster dans $\mathcal{C}(\mathcal{D})$ et également au même cluster dans $\mathcal{C}(\mathcal{RD}_\mathcal{V})$, b le nombre de paires de données affectées au même cluster dans $\mathcal{C}(\mathcal{R})$, mais à des clusters différents dans $\mathcal{C}(\mathcal{RD}_\mathcal{V})$, c, symétriquement, le nombre de paires affectées à des clusters différents dans $\mathcal{C}(\mathcal{R})$ mais un même cluster dans $\mathcal{C}(\mathcal{RD}_\mathcal{V})$, d est le nombre de paires affectées à des clusters différents dans $\mathcal{C}(\mathcal{R})$ et dans $\mathcal{C}(\mathcal{RD}_\mathcal{V})$.

L'un des critères les plus utilisés est l'indice de Rand ajusté [5], noté ira dans la suite :

$$ira = \frac{\frac{n(n-1)}{2}(a+d) - Z}{\left(\frac{n(n-1)}{2}\right)^2 - Z}$$

où $Z = (a+b)(a+c) + (c+d)(b+d)$. Il normalise l'indice de Rand par rapport à une distribution de référence des données.

La mesure d'adéquation du vocabulaire à la base de données s'écrit alors

$$ad_{IRA}(\mathcal{V}, \mathcal{D}) = ira(\mathcal{C}(\mathcal{D}), \mathcal{C}(\mathcal{RD}_\mathcal{V})) \qquad (4)$$

4.2 Evaluation croisée

La mesure précédente requiert d'appliquer l'algorithme de clustering deux fois, pour chacun des espaces de représentation considérés, ce qui peut être coûteux. Le second critère mesure la qualité du résultat du clustering effectué dans l'un des espaces de représentation en utilisant la seconde représentation.

Plus précisément, le critère évalue si les clusters identifiés en utilisant la représentation initiale des données sont compacts et séparables au sens des données réécrites. En effet, si deux médoïdes apparaissent indistinguables dans l'espace réécrit, cela signifie que ce dernier n'est pas approprié pour représenter la structure des données.

Il existe de nombreuses définitions de compacité et de séparabilité, ainsi que de nombreuses combinaisons de ces critères [4]. Nous considérons l'indice de Xie-Beni [14]

$$xieBeni(U, \mathcal{W}, \mathcal{D}) = \frac{C(U, \mathcal{W}, \mathcal{D})}{S(\mathcal{W})}$$

$$\text{où} \quad C(U, \mathcal{W}, \mathcal{D}) = \frac{1}{n} \sum_{x_i \in \mathcal{D}} \sum_{w_r \in \mathcal{W}} u_{ri}\, d(x_i, w_r)$$

$$\text{et} \quad S(\mathcal{W}) = \min_{w_r, w_s \in \mathcal{W}} d(w_r, w_s)$$

où U désigne la matrice d'affectation des données, telle que $u_{ri} = 1$ si la donnée x_i est

affectée au cluster r, et 0 sinon. $\mathcal{W} = \{w_r\}$ représente l'ensemble des centres des clusters, et d la mesure de distance. L'indice de Xie-Beni doit être minimisé.

Le critère proposé consiste alors à calculer dans l'espace réécrit la compacité et la séparabilité des clusters identifiés dans l'espace initial :

$$q_{XB}(\mathcal{V}, \mathcal{D}) = xieBeni(U(\mathcal{D}), \mathcal{RW}_\mathcal{V}, \mathcal{RD}_\mathcal{V}) \quad (5)$$

où $U(\mathcal{D})$ représente l'affectation obtenu en partitionnant les données $\mathcal{D}$ dans l'espace initial, $\mathcal{W}$ les centres de clusters associés et $\mathcal{RD}_\mathcal{V}$ leur réécriture par le vocabulaire $\mathcal{V}$.

5 Résultats expérimentaux

Cette section décrit l'étude des critères proposés réalisée sur une petite base de données artificielle représentative, en deux dimensions.

5.1 Données considérées

Les données, représentées sur la Figure 2, ont été générées par un mélange de trois gaussiennes non sphériques générant chacune 150 points. Plusieurs vocabulaires sont comparés, comme représenté à gauche et dans la partie inférieure de la figure : la variable linguistique associée à l'attribut y est définie par une partition appropriée, qui correspond à la distribution des données : les deux modalités isolent le cluster d'en haut des clusters inférieurs et induit des zones indistinguables au sein de chaque cluster.

Pour l'attribut x, 7 partitions sont considérées, classées par ordre croissant de nombre de modalités : $P1$ contient une unique modalité et peut donc être considérée comme trop générale. En effet, toutes les données appartiennent à son noyau et sont donc indistinguables. La partition $P2$ est la partition appropriée, qui correspond à la structure des données. La partition $P2a$ présente également 2 modalités, mais peut être décrite comme absurde : elle est en double contradiction avec la structure des données : d'abord elle conduit à une

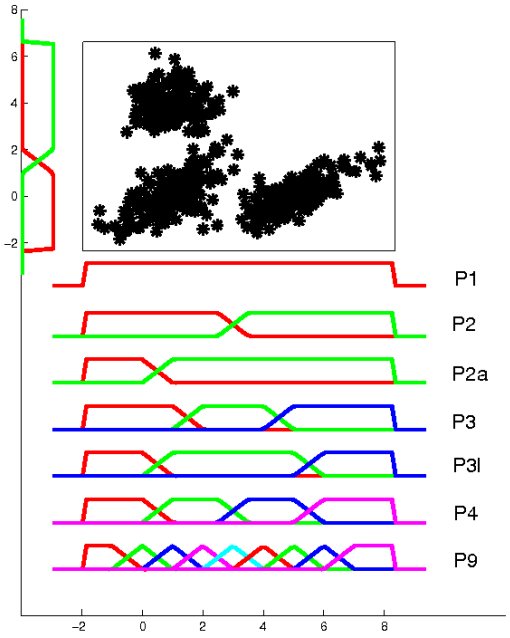

Figure 2 – Données et vocabulaire considérés

distinction artificielle des données du cluster inférieur de gauche ; de plus, elle ne fait pas de différence entre les données du cluster inférieur de droite et certaines des données du cluster inférieur de gauche. La partition $P3$ définit trois modalités qui rendent la partie centrale des données indistinguable, regroupant les deux clusters inférieurs et ignorant leur structure. La partition $P3l$ est une variante de $P3$ où la région de chevauchement entre les deux clusters est plus large. Les partitions $P4$ et $P9$ sont d'autres exemples de partitions trop détaillées, possédant trop de modalités.

5.2 Protocole expérimental

Le Tableau 1 montre, pour chacune des partitions considérées, les valeurs des deux critères proposés, ad_{IRA} et q_{XB}, en donnant la moyenne et l'écart-type pour 100 initialisations. Chaque initialisation tire aléatoirement un premier médoïde puis sélectionne les autres médoïdes de façon à maximiser leurs distances deux à deux. Des analyses non détaillées ici montrent que, quelle que soit la représentation de données considérée, les résultats du clustering sont stables et peu dépendants de l'initialisation, ce

Tableau 1 – Moyenne et écart-type des critères proposés pour 100 initialisations aléatoires.

$\mathcal{V}$	c	ad_{IRA}	q_{XB}
$P1$	2.0 ± 0	0.54 ± 0.05	Inf
$P2$	3.0 ± 0	0.95 ± 0.06	0.04 ± 0.03
$P2a$	5.0 ± 0.32	0.55 ± 0.03	Inf
$P3$	2.1 ± 0.37	0.56 ± 0.05	0.20 ± 0.03
$P3l$	2.0 ± 0	0.56 ± 0.05	Inf
$P4$	3.0 ± 0.10	0.94 ± 0.06	0.33 ± 0.05
$P9$	2.0 ± 0.10	0.56 ± 0.07	0.23 ± 0.02

qu'indiquent également les faibles écarts-types des valeurs du Tableau 1. La Figure 3 montre les affectations les plus fréquentes obtenues pour chaque représentation.

On peut observer qu'avec la représentation initiale des données, l'algorithme de clustering obtient le résultat attendu. L'indice de Rand ajusté, qui compare les affectations obtenues et attendues d'après le processus de génération des données, vaut 0.96 ± 0.05 : quelques affectations sont erronées, aux frontières des clusters, mais ces erreurs sont rares.

5.3 Résultats de l'adéquation IRA

La mesure ad_{IRA} indique clairement deux types de vocabulaires : le premier, associé à des valeurs proches de 0.94, contient les partitions $P2$ et $P4$; le second, associé à des valeurs autour de 0.56, groupe les autres partitions $P1$, $P2a$, $P3$, $P3l$ et $P9$. On ne note pas de différence significative au sein de ces deux types.

Les partitions $P2$ et $P4$ identifient les trois clusters attendus, correspondant à la structure identifiée par la représentation initiale des données. L'adéquation élevée obtenue correspond donc à un résultat attendu. Comme l'illustre la Figure 3, les affectations sont les mêmes pour $P2$ et $P4$ et ne diffèrent de la partition obtenue à partir de la représentation initiale que pour la partie supérieure du cluster situé en bas à gauche : pour l'attribut x, ces données sont indistinguables des clusters situés en haut et en bas à gauche, pour l'attribut y, elles se situent dans la zone de chevauchement des deux modalités et sont plutôt dans la partie haute. Il est aussi intéressant de noter que le point qui maximise y pour le cluster en bas à droite est affecté au cluster du haut et non au cluster attendu, pour $P2$ comme $P4$: après réécriture, ce point est à égale distance de certains des points du cluster de droite, en raison d'une distance nulle pour l'attribut x mais non nulle pour y, et de certains des points du cluster du haut, en raison de la configuration inverse (distance non nulle pour l'attribut x mais nulle pour y).

Les partitions $P1$, $P2a$, $P3$, $P3l$ et $P9$ échouent à identifier le nombre de clusters obtenu par la représentation initiale : à part $P2a$, elles ne distinguent que 2 clusters, fusionnant les deux clusters du bas en un seul groupe. Le plus souvent, $P2a$ décompose les données en 5 clusters, bien qu'elle soit légèrement plus instable que les autres partitions. Plus précisément, elle identifie 3 clusters correspondant aux combinaisons des modalités définies sur x et y, ainsi que 2 clusters correspondant aux données situées dans les régions de chevauchement des modalités, pour les attributs x comme y.

Il est intéressant de noter que le nombre de clusters identifiés n'est donc pas corrélé au nombre de modalités : la partition trop précise $P9$ n'identifie pas plus de clusters que la partition $P3$. En effet, la différence en termes de distance qu'un nombre trop important de modalités pourrait induire est d'une part atténuée par la normalisation par le nombre de modalités (cf. Eq. 3) et d'autre part dominée par la distance induite par les partitions définies sur les autres attributs.

Ces résultats montrent que la méthode proposée identifie le vocabulaire approprié et pénalise les vocabulaires incorrects. Toutefois, elle ne distingue pas parmi ces derniers, ce qui peut être un inconvénient dans le cas où l'objectif est d'aider un expert à adapter un vocabulaire.

On peut observer que les deux types de vocabulaires ne sont pas justifiés par le nombre

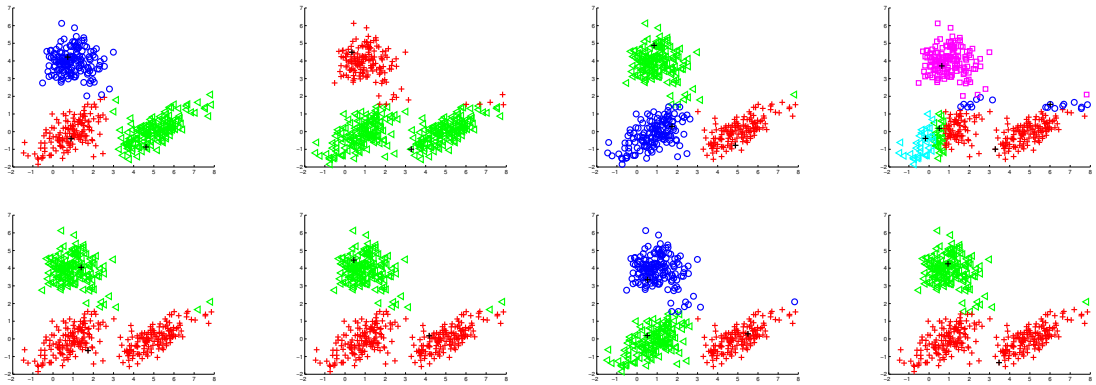

Figure 3 – Affectations obtenues : (première ligne) représentation initiale et après réécriture avec les partitions $P1$, $P2$, $P2a$; (seconde ligne) après réécriture avec les partitions $P3$, $P3l$, $P4$, $P9$.

de modalités, par exemple selon que ce dernier est trop élevé ou trop faible, mais par la présence ou l'absence d'une modalité reliant les deux clusters du bas : pour les deux partitions obtenant une valeur élevée de ad_{IRA}, $P2$ et $P4$, la frontière entre les clusters correspond à une transition entre des modalités. Au contraire, pour les vocabulaires obtenant une valeur faible, la région frontalière est rendue indistinguable par une modalité à cheval entre les deux clusters. Pour ces données, la taille du chevauchement n'a pas d'influence : qu'elle soit élevée ($P1$, $P2a$ ou $P3l$), moyenne ($P3$) ou faible ($P9$), la mesure prend la même valeur.

5.4 Résultats de l'adéquation XB

La dernière colonne du Tableau 1 donne les valeurs de q_{XB}, qui doit être minimisé. A titre de référence, l'évaluation dans l'espace de représentation initiale, c'est-à-dire la mesure de la compacité et de la séparabilité dans l'espace où le clustering a lieu, conduit à 0.10 ± 0.02.

On peut observer que $P2$ est significativement la meilleure partition : représenter les données par ce vocabulaire rend les clusters encore plus compacts et séparables que dans l'espace initial. En effet, toutes les données situées dans les noyaux des modalités sont bien affectées à un même cluster, et donc à distance nulle : les diamètres des clusters sont donc très faibles. De plus, les médoïdes sont à distance maximale les uns des autres, car ils sont décrits par les modalités extrêmes de chaque modalité.

Les partitions $P3$ et $P9$ sont classées deuxième ex aequo : elles ont également une séparabilité maximale, mais une compacité plus faible. Il faut souligner que ces comparaisons de compacité et séparabilité sont basées sur l'interprétation, et restent conceptuelles et non numériques : les échelles de distance sont différentes, et une normalisation est difficile.

La partition $P4$, classée quatrième, a une compacité comparable à celles de $P3$ et $P9$, mais n'obtient pas une séparabilité maximale : pour l'attribut x, les médoïdes sont respectivement réécrits comme $(0, 1, 0, 0)$ and $(0, 0, 1, 0)$ et non plus affectés aux modalités extrêmes.

Enfin, pour les partitions $P1$, $P2a$ et $P3l$, q_{XB} est infini en raison d'un dénominateur nul : au moins 2 médoïdes ne sont pas distinguables selon le vocabulaire, ce qui en indique le manque d'adéquation. Cette valeur infinie, qui ne permet pas de faire de différence entre ces trois vocabulaires, pourrait être raffinée par le nombre de paires de médoïdes indistinguables.

Le classement des partitions pertinentes est donc plus fin et différent de celui induit par la mesure ad_{IRA}, ce qui indique la complémentarité de ces mesures de sémantiques différentes.

6 Conclusion

Dans le but de personnaliser le processus de découverte, nous avons considéré le problème de l'adéquation entre un vocabulaire expert et la structure des données, en proposant deux mesures. La première repose sur la comparaison de deux partitionnements des données, obtenus à partir de la représentation initiale des données et d'une réécriture des données. La seconde mesure, moins coûteuse en temps de calcul, est basée sur l'évaluation dans l'espace réécrit des regroupements identifiés dans l'espace initial. Les expérimentations préliminaires ont montré que les deux mesures permettent d'identifier les vocabulaires pertinents.

Les travaux en cours visent à étendre les expérimentations à des bases de données plus complexes, plus bruitées ou en dimensions supérieures, et en particulier pour des données réelles, qui posent également le problème des valeurs manquantes. La comparaison de vocabulaire différant par le degré de flou des partitions est également envisagée. Les perspectives incluent aussi l'interprétation plus détaillée des mesures proposées, en particulier pour proposer des adaptations du vocabulaire, pour améliorer l'adéquation à la structure des données tout en conservant la subjectivité du vocabulaire expert.

Références

[1] P. Bosc, A. Hadjali, O. Pivert, and G. Smits. An approach based on predicate correlation to the reduction of plethoric answer sets. In *Advances in Knowledge Discovery and Management*, volume 398, pages 213–233. Springer, 2012.

[2] J. Chomicki. Preference formulas in relational queries. *ACM Transactions on Database Systems*, 28 :1–40, 2003.

[3] S. Guillaume, B. Charnomordic, and P. Loisel. Fuzzy partitions : a way to integrate expert knowledge into distance calculations. *Information sciences*, pages 76–95, 2012.

[4] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17 :107–145, 2001.

[5] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1) :193–218, 1985.

[6] A. Jain, M. Murty, and P. Flynn. Data clustering : a review. *ACM Computing survey*, 31(3) :264–323, 1999.

[7] R. Krishnapuram, A. Joshi, O. Nasraoui, and L. Yi. Low complexity fuzzy relational clustering algorithms for web mining. *IEEE Trans. on Fuzzy Systems*, 9(4) :595–607, 2001.

[8] M.-J. Lesot and A. Revault d'Allonnes. Credit-card fraud profiling using a hybrid incremental clustering methodology. In *Scalable Uncertainty Management*, pages 325–336. Springer, 2012.

[9] M. Meila. Comparing clustering, an axiomatic view. In *Proc. of the Int. Conf. on Machine Learning*, pages 577–584, 2005.

[10] G. Raschia and N. Mouaddib. Saintetiq : a fuzzy set-based approach to database summarization. *Fuzzy Sets and Systems*, 129(2) :137 – 162, 2002.

[11] E. Ruspini. A new approach to clustering. *Information and Control*, 15(1) :22 – 32, 1969.

[12] G. Smits and O. Pivert. A fuzzy-summary-based approach to faceted search in relational databases. In *Advances in Databases and Information Systems*, volume 7503, pages 357–370. Springer, 2012.

[13] L. Ughetto, W. Voglozin, and N. Mouaddib. Database querying with personalized vocabulary using data summaries. *Fuzzy Sets and Systems*, 159(15) :2030–2046, 2008.

[14] X. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4) :841–846, 1991.