Classification de Données Complexes par Globalisation de Mesures de Similarité via les Moyennes Quasi-Arithmétiques

Étienne-Cuvelier*, Marie-Aude-Aufaure**

*ICHEC - Brussels Management School
etienne.cuvelier@ichec.be
**Datarvest
marie-aude.aufaure@datarvest.com

Résumé. La plupart des méthodes de classification sont conçues pour des types particuliers de données: données numériques, textuelles, catégoriques, fonctionnelles, probabilistes ou encore de type graphes. Cependant, les données générées dans notre quotidien sont en général composées de données de types mixtes. Par exemple, si nous considérons la prévention cardiaque dans le domaine de la santé, les applications vont combiner des données issues de capteurs avec d'autres données telles que l'âge, le niveau d'effort, la fréquence cardiaque maximale, des histogrammes de fréquences cardiaques moyennes lors de précédents efforts, etc. Ceci nous amène à la problématique de construire des classes en tenant compte de ces différentes données, et de définir une mesure de similarité à partir des similarités de paires d'objets sur les différents types de variables. Dans cet article nous proposons une méthode de classification basée sur la fusion des matrices de similarité à l'aide des moyennes quasi-arithmétiques qui permet de choisir les différentes "dimensions" des données à considérer, et ce quel que soit le type de données, pour autant qu'une mesure, de similarité ou de dissimilarité existe pour chacun des types de données, ce qui est très souvent le cas.

1 Introduction

A l'ère des Big Data, le volume et la diversité des données disponibles de manière numérique ne cesse de croître. Ces données proviennent de capteurs, de réseaux sociaux, du web, des traces laissées par nos appareils mobiles, de nos achats, des données ouvertes, etc. Cette diversité est d'une grande richesse et est valorisée par les entreprises à travers des applications toujours plus personnalisées. Si les architectures informatiques passent bien à l'échelle, avec notamment du stockage dans de nouveaux outils de type NoSQL, l'utilisation des technologies distribuées avec Hadoop et son écosystème ou encore le traitement en temps contraint avec Spark et les librairies associées, le croisement de toutes ces données restent encore un défi à l'heure actuelle. De nombreuses méthodes de classification ont été conçues pour un type de données particulier. Or, il peut être utile dans de nombreuses applications, de pouvoir résumer et synthétiser un ensemble de données de types divers provenant de sources hétérogènes. Par

exemple, si nous considérons dans le domaine des smart cities une application permettant à un piéton dans une zone urbaine de déterminer la trajectoire optimale pour minimiser l'exposition à la pollution, il sera nécessaire de tenir compte de données issues de capteurs, de la météo, de la vitesse et de la direction du vent, de la tranche horaire de la journée pour estimer la densité de trafic, de l'âge du piéton et d'éventuels éléments de santé pour déterminer le degré de risque d'exposition, etc. Des solutions pour pouvoir gérer ces données mixtes ont été proposées dans la littérature. Liu et al. (2016) catégorisent les méthodes de classification de données mixtes de la manière suivantes : méthodes basées sur la conversion d'attributs, méthodes dites d'ensemble, méthodes basées sur les prototypes et les autres méthodes.

Les méthodes de conversion d'attributs convertissent les différents types d'attributs en un type unifié, puis utilisent une méthode de classification appropriée à ce type. Dans les méthodes de type ensemble, développées en classification supervisée, chaque classifieur tente de résoudre la même tâche pour améliorer la précision et la robustesse, et les résultats sur l'application à des jeux de données variés donne de bons résultats. Cette approche a également été appliquée à la classification non supervisée pour combiner de multiples partitionnements d'un ensemble d'objets dans une seule classification consolidée. Les difficultés sont liées à la mise en correspondance des labels des clusters, et le fait que le nombre et la forme des clusters peuvent varier. La classification de données mixtes peut être réalisé en deux étapes : une étape consistant à générer un ensemble de clusters sur le même jeu de données, et une étape pour les combiner et former le résultat final. Durant la première étape, différents algorithmes peuvent être utilisés, ou différentes initialisations des paramètres ou encore différents sous-ensembles d'objets. La seconde étape est la plus critique et consiste à trouver une fonction de consensus pour générer le résultat final. Strehl et al. (2002) proposent trois algorithmes différents pour la fonction de consensus, basés sur la transformation de l'ensemble des labels des clusters en une représentation sous forme d'hypergraphe. Plusieurs scénarios ont été testés sur des jeux de données réels et synthétiques (classification sur des ensembles d'objets différents, sur le même jeu de données mais avec des ensembles de caractéristiques différents, sur le même jeu de données avec pour objectif d'améliorer la qualité et la robustesse). Les méthodes basées sur les prototypes comme les k-moyennes consistent à utiliser un prototype pour représenter la classe. Ces méthodes ont été appliquées aux données mixtes : Cheung et Jia (2013) a défini un algorithme itératif basé sur la notion de similarité objet-classe et fourni une métrique de similarité unifiée pouvant s'appliquer à des données catégoriques, numériques et mixtes. Ahmad et Dey (2007) ont proposé une nouvelle fonction de coût et une mesure de distance pour les données mixtes basée sur la co-occurrence des valeurs. Les expérimentations ont été réalisées sur des jeux de données numériques, catégoriques et mixtes. Enfin, la dernière catégorie de méthodes regroupe les méthodes hiérarchiques ou basées sur la densité. L'algorithme proposé dans Rodriguez et Laio (2014) est basé sur le fait que le centre de la classe est caractérisé par une densité plus forte que ses voisins et est à une distance assez grande de points avec une densité plus forte. Pour chaque point, on calcule sa densité locale et sa distance des points ayant une densité plus forte. Une fois les centres des clusters identifiés, chaque point restant est associé à la même classe que son plus proche voisin de densité supérieure en une seule passe. Plusieurs extensions de cet algorithme aux données mixtes (Liu et al. (2017) et Jinyin et al. (2017)) ont défini une mesure de distance unifiée et une extension de l'algorithme proposé dans Rodriguez et Laio (2014). Li et Biswas (2002) ont proposé un algorithme hierarchique basé sur une approche agglomérative et la mesure de similarité définie par Goodall (1966) pour construire

les partitions. La plupart de ces travaux s'appliquent à des données mixtes de type numérique et/ou catégorique.

Nous souhaitons proposer un cadre théorique pour traiter conjointement tous types de données pour lesquels une mesure de similarité ou de dissimilarité existe. Dans la section 2 nous décrivons ce cadre de globalisation de mesures de similarité/dissimilarité fondée sur les moyennes quasi-arithmétiques. La section 3 illustre un cas d'usage mixant des données numériques et probabilistes. Et enfin nous terminons avec les conclusions et perspectives dans la section 4 .

2 Globalisation de Mesures de Similarité/Dissimilarité

Beaucoup de techniques de classification, supervisées ou non, sont basées sur les notions de mesures de similarité ou de dissimilarité. Nous commencerons donc par rappeler brièvement ces deux concepts et la façon dont ils sont liés. Nous détaillerons ensuite comment nous nous proposons de combiner des mesures de ces deux types en une mesure résultante à l'aide des moyennes quasi-arithmétiques.

2.1 Mesures de Similarité et de Dissimilarité

Sur un domaine de données, noté V, les notions de mesures de similarité et de dissimilarité peuvent être définies comme suit.

Définition 1 : *Une mesure de similarité entre deux éléments de $u, v \in V$ est toute fonction $s : V \times V \rightarrow \mathbb{R}^+$ qui satisfait les propriétés suivantes :*

Séparation : $s(u, u) = k$*, où* k *est une constante,*

Symétrie : $s(u, v) = s(v, u)$*,*

Maximalité : $s(u, v) \leq s(u, u) = k$*.*

Définition 2 : *Une mesure de dissimilarité entre deux éléments de $u, v \in V$ est toute fonction $d : V \times V \rightarrow \mathbb{R}^+$ qui satisfait les deux propriétés suivantes :*

Séparation : $d(u, u) = 0$*,*

Symétrie : $d(u, v) = d(v, u)$*.*

Toute mesure de dissimilarité d peut être transformée en mesure de similarité s, et vice et versa, au travers d'une fonction ϕ strictement décroissante :

$$s(u, v) = \phi(d(u, v)) \text{ et } d(u, v) = \phi^{-1}(s(u, v)) \quad (1)$$

avec comme condition que $\phi(0) = k$. La fonction de densité gaussienne est un exemple de fonction utilisée dans ce cas.

Si un phénomène est décrit, non seulement par plusieurs variables, mais surtout par plusieurs types de variables (numériques, catégorielles, probabilistes, intervalles, arborescentes, fonctionnelles, ...) alors la plupart du temps il est possible de calculer la similarité (dissimilarité) entre deux objets sur base d'un type de variable choisi, mais il n'est en général pas possible de

calculer de telles mesures en prenant en compte plus de deux types de variables, voire tous en même temps. Une solution pour contourner ce problème serait donc de disposer d'un moyen pour calculer une mesure de similarité/dissimilarité résultante sur base de différentes mesures existant entre deux objets.

Définition 3 : *Si $s_i (i \in \{1 \cdots p\})$ sont différentes mesures de similarité calculées entre deux objets $u, v \in V$, alors nous appellerons mesure de similarité jointe (ou résultante) toute mesure de similarité σ calculée sur base des s_i à l'aide d'un opérateur $\mathcal{M} : \mathcal{I}m(s_1) \times \cdots \times \mathcal{I}m(s_p) \rightarrow [0, K]$ (avec $K \in \mathbb{R}_0^+$) :*

$$\sigma(u, v) = \mathcal{M}\left(s_1(u, v), \cdots, s_p(u, v)\right). \tag{2}$$

Si $\forall i \in \{1 \cdots, n\}, \mathcal{I}m(s_i) = \mathcal{I}m(\sigma) = [0, K]$ alors σ est appelée mesure de similarité jointe normalisée.
On définira aisément de façon similaire une mesure de dissimilarité jointe.

Comme on ne peut raisonnablement comparer que des choses comparables, dans la suite nous ne considérerons que des mesures de similarité jointes normalisées avec, de façon assez classique, $K = 1$.

Toute la difficulté étant de trouver un opérateur $\mathcal{M}$ satisfaisant. Nous proposons d'utiliser les moyennes quasi-arithmétiques.

2.2 Les Moyennes Quasi-Arithmétiques

Définition 4 : *Soit $[a, b]$ un intervalle réel fermé et $p \in \mathbb{N}_0$. Une moyenne quasi-arithmétique est une fonction $\mathcal{M}_\phi^{(p)} : [a, b]^p \rightarrow [a, b]$ définie comme suit :*

$$\mathcal{M}_\phi^{(p)}(\overline{u}) = \mathcal{M}_\phi^{(p)}(u_1, \ldots, u_p) = \phi^{-1}\left(\sum_{i=1}^{p} \alpha_i \phi\left(u_i\right)\right) \tag{3}$$

avec ϕ fonction continue strictement monotone définie sur $[a, b]$, $\forall i \in 1, \cdots, p : \alpha_i \in [0, 1]$ et $\sum_{i=1}^{p} \alpha_i = 1$.
Les moyennes quasi-arithmétiques forment une extension des moyennes classiques (Fodor et Roubens (1994)). Ainsi, si $\phi(x)$ est respectivement égale à x, x^2, $\log(x)$ et x^{-1}, l'expression (3) génère respectivement les moyennes classiques suivantes : arithmétique, quadratique, géométrique et harmonique. Dans la suite nous noterons $\mathcal{M}_{id}^{(p)}$ la moyenne arithmétique.
Une propriété de base des moyennes quasi-arithmétiques est que (Bullen et al. (1988)) :

$$\min(u_1, \ldots, u_p) \leq \mathcal{M}_\phi^{(p)}(u_1, \ldots, u_p) \leq \max(u_1, \ldots, u_p). \tag{4}$$

Kolmogorov (1930) a montré que les moyennes quasi-arithmétiques possèdent aussi les propriétés suivantes :

Idempotence : $\mathcal{M}_\phi^{(p)}(u, \ldots, u) = u,\ \forall u \in [a, b]$,

Continuité : pour tout $p \in \mathbb{N}_0, \mathcal{M}_\phi^{(p)}$ est une fonction continue sur $[a, b]^p$,

ϕ	Croissant	Décroissant
Convexe	$\mathcal{M}_{id}^{(p)}(\overline{u}) \leq \mathcal{M}_{\phi}^{(p)}(\overline{u})$	$\mathcal{M}_{id}^{(p)}(\overline{u}) \geq \mathcal{M}_{\phi}^{(p)}(\overline{u})$
Concave	$\mathcal{M}_{id}^{(p)}(\overline{u}) \geq \mathcal{M}_{\phi}^{(p)}(\overline{u})$	$\mathcal{M}_{id}^{(p)}(\overline{u}) \leq \mathcal{M}_{\phi}^{(p)}(\overline{u})$

TAB. 1 – *Comparaison du résultat de la Moyenne Arithmétique $\mathcal{M}_{id}^{(p)}(\overline{u})$ comparée à une Moyenne Quasi-Arithmétique $\mathcal{M}_{\phi}^{(p)}(\overline{u})$ selon les propriétés du générateur ϕ.*

Croissance Stricte : pour chaque argument

$$u_i < u'_i \Rightarrow \mathcal{M}_{\phi}^{(p)}(u_1, \ldots, u_i, \ldots, u_p) < \mathcal{M}_{\phi}^{(p)}(u_1, \ldots, u'_i, \ldots, u_p),$$

Symétrie : si π est une permutation de $\{1, \ldots, p\}$, alors

$$\mathcal{M}_{\phi}^{(p)}(u_1, \ldots, u_p) = \mathcal{M}_{\phi}^{(p)}(u_{\pi(1)}, \ldots, u_{\pi(p)}),$$

Décomposable : si $\mathcal{M}_k = \mathcal{M}_{\phi}^{(k)}(u_1, \ldots, u_k)$, alors

$$\mathcal{M}_{\phi}^{(p)}(u_1, \ldots, u_k, u_{k+1}, \ldots, u_p) = \mathcal{M}_{\phi}^{(p)}(\mathcal{M}_k, \ldots, \mathcal{M}_k, u_{k+1}, \ldots, u_p).$$

Certaines de ces propriétés vont permettre de répondre à la question fondamentale suivante : *la moyenne quasi-arithmétique de plusieurs mesures de similarité est-elle une mesure de similarité ?* La conservation de la symétrie est assez évidente. La séparation est conservée via la propriété d'idempotence des moyennes quasi-arithmétiques. Et enfin la propriété de maximalité l'est aussi via la croissance stricte en chaque argument. Par l'inégalité de Jensens, qui établit que si ϕ est convexe, alors

$$\phi\left(\frac{\sum \alpha_i u_i}{\sum \alpha_i}\right) \leq \frac{\sum \alpha_i \phi(u_i)}{\sum \alpha_i} \tag{5}$$

et comme la croissance de ϕ implique la croissance de son inverse, on en conclut donc que dans le cas d'un générateur convexe et croissant la moyenne arithmétique sera inférieure à la moyenne quasi-arithmétique. Bien entendu si la fonction est décroissante, l'ordre entre les deux moyennes sera inversé. Enfin si on tient compte du fait que l'inégalité (5) s'inverse si ϕ est concave, alors on obtient le tableau 1.

Si pour un vecteur de similarités $\overline{s} = (s_i, \cdots, s_p)$ mesurées entre deux objets u et v on a deux moyennes quasi-arithmétiques $\mathcal{M}_{\phi'}^{(p)}$ et $\mathcal{M}_{\phi''}^{(p)}$ telles que (en tenant compte de (4))

$$\min(\overline{s}) \leq \mathcal{M}_{\phi'}^{(p)}(\overline{s}) \leq \mathcal{M}_{x}^{(p)}(\overline{s}) \leq \mathcal{M}_{\phi''}^{(p)}(\overline{s}) \leq \max(\overline{s}) \tag{6}$$

alors on peut interpréter cela comme le fait que $\mathcal{M}_{\phi'}^{(p)}$ "favorise" la dissimilarité entre u et v dans le résultat final alors que $\mathcal{M}_{\phi''}^{(p)}$, favorise la similarité. Reste la question du choix du générateur ϕ. Le tableau 1 montre, sur base de la convexité ou concavité, et en fonction de

la croissance ou décroissance du générateur choisi, dans lequel des cas figures montrés dans l'inégalité (6) on se situera. Pour éclairer le choix du générateur, rappelons que les mesures de similarité et de dissimilarité sont liées via une fonction strictement décroissante (1). Supposons que nous disposions pour nos deux objets à comparer u et v à la fois d'une mesure de similarité s_1 et d'une mesure de dissimilarité d_2. Si le générateur choisi pour calculer la moyenne a aussi les propriétés requises pour être utilisé dans l'expression (1) alors $s_2 = \phi^{-1}(d_2)$ est une similarité induite à partir de d_2 et nous pouvons alors écrire que :

$$\sigma(u,v) \quad = \quad \phi^{-1}\left(\alpha\phi\left(s_1(u,v)\right) + (1-\alpha)d_2(u,v)\right) \tag{7}$$

avec $\alpha \in [0,1]$. Ce qui signifie que si le choix se porte sur un générateur strictement décroissant et utilisable dans (1), non seulement nous pourrions joindre des similarités mais aussi des dissimilarités dans le même calcul.

2.3 Familles de Générateurs

Il existe un ensemble de fonctions qui satisfont aux différentes conditions souhaitables, à savoir, être définies sur $[0,1]$ et être décroissantes : les générateurs de copules archimédiennes.

Les copules sont des fonctions de distributions multivariées utilisées pour joindre des marginales (Nelsen (1999)). La définition des copules archimédiennes est particulièrement proche de la définition des moyennes quasi-arithmétiques :

Définition 5 : *Une copule archimédienne est une fonction* $C[0,1]^p \rightarrow [0,1]$ *définie par l'expression*

$$C(u_1,...,u_p) = \phi^{-1}\left[\sum_{i=1}^{p}\phi(u_i)\right] \tag{8}$$

où $\phi : [0,1] \rightarrow [0,\infty]$ *est une fonction continue strictement décroissante, telle que* $\phi(1) = 0$. *Si* $p = 2$, *alors* ϕ *doit aussi être convexe, et pour* $p > 2$, *alors* ϕ^{-1} *doit de plus être complètement monotone.*

Définition 6 : *Widder (1941) Une fonction continue* ϕ *définie sur un intervalle* $[a,b]$ *est complètement monotone ssi*

$$\forall a < t < b \text{ et } \forall k \geq 1 : (-1)^k \frac{d^k}{dt^k} f(t) \geq 0. \tag{9}$$

Si on se souvient qu'une fonction deux fois différentiable est convexe si et seulement si sa dérivée seconde est non-négative, on constate donc qu'une fonction complètement monotone est aussi convexe, et il en est de même pour sa réciproque. On trouvera dans Nelsen (1999) une liste de familles de générateurs de copules archimédiennes. Dans notre cas nous utiliserons le générateur 4.2.2 extrait de cette liste car son paramètre θ ($\theta \geq 1$) permet d'agir assez souplement sur sa courbure :

$$\phi_\theta(t) = (1-t)^\theta. \tag{10}$$

3 Un Cas d'Usage en Classification

3.1 Les Stations Climatiques Chinoises

Pour illustrer l'utilisation de cette globalisation des mesures de similarité sur différents types de variables, nous aurions pu prendre un exemple classique mixant données catégorielles et données numériques, comme dans les solutions évoquées dans l'introduction. Néanmoins pour montrer l'étendue de la méthode proposée, nous avons choisi d'illustrer notre propos en mixant des données numériques et des données de nature probabiliste. Les données climatiques chinoises (http://cdiac.ornl.gov/ndps/tr055.html) regroupent 14 variables climatiques pour chacune des 4 saisons, et ce pour 60 stations dont les coordonnées et l'élévation sont données. On trouvera la représentation des 60 stations sur la carte de Chine dans les figures 2 et 3. Ces données ont été collectées mensuellement de 1978 à 1988, ce qui donne donc 132 enregistrements pour chacune des variables. Les 132 valeurs de chaque variable peuvent être résumées sans trop de pertes d'information en un histogramme des valeurs pour chaque station, ce qui a été fait dans le package R HistDAWass (Histogram-Valued Data Analysis, Irpino (2016)). Notre choix s'est porté sur les variables concernant les températures moyennes. Quatre variables de type histogramme (ou distributionnel) existent dans le package à ce propos : une pour chaque saison. Conjointement à ces variables probabilistes, nous considérerons les trois variables classiques suivantes : l'altitude, la longitude et la latitude, ces deux dernières variables étant considérées ensembles puisqu'elles donnent la localisation de la station.

Pour des raisons de facilité de mise en oeuvre nous avons choisi d'appliquer une classification hiérarchique ascendante, car un seul calcul des distances ou similarités suffit contrairement à ce qui se passe pour les k-moyennes par exemple. Une classification spectrale aurait pu être appliquée directement pour les mêmes raisons.

3.2 Classification Hiérarchique

Nous avons pratiqué la classification hiérarchique en utilisant le lien moyen. Les distances entre altitudes ont été calculées simplement en utilisant la distance euclidienne unidimensionnelle. La distance "à vol d'oiseau" à été utilisée pour calculer les dissimilarités entre les coordonnées de localisations. Pour les variables distributionnelles la distance utilisée est la L2 de Wasserstein (Irpino et Romano (2007)), implémentée dans le package HistDAWass.

Comme nous l'avons évoqué dans la section 2.1, la conversion d'une dissimilarité en similarité peut se faire à l'aide d'une fonction strictement décroissante (cf. équation (1)). Le choix de la fonction, et notamment la "vitesse" de sa décroissance aura un impact sur les similarités résultantes. Ainsi, par exemple, une même fonction ϕ appliquée à des dissimilarités $d_1(u, v)$ et $d_2(u, v)$ calculées à partir variables numériques différentes et ayant des échelles non comparables, aura des impacts différents sur le calcul des deux similarités résultantes s_1 et s_2. Quand on travaille avec des données quantitatives classiques, le problème ne se pose pas si l'on veille à réduire chacune des variables en la divisant par son écart-type. Dans le cas de données complexes, de par l'hétérogénéité des types de données possibles (quantitatives, qualitatives, intervalles, distributionnelles, fonctionnelles, ...) cette approche n'est pas applicable pour tous les types de variables. C'est pourquoi nous avons décidé d'appliquer la réduction aux dissimilarités, et ensuite d'appliquer la fonction gaussienne, ce qui revient à appliquer la

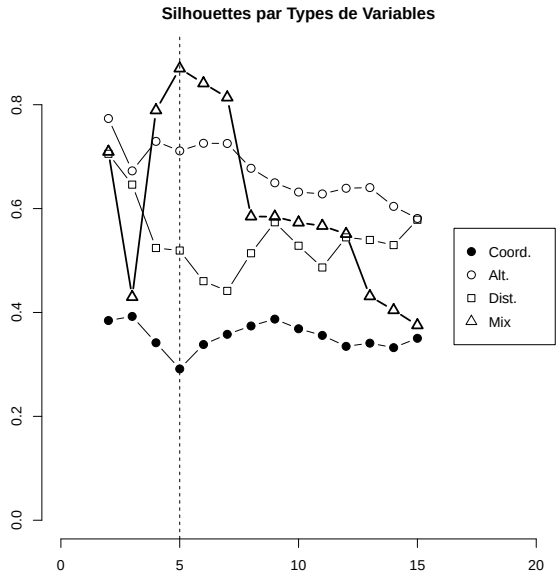

FIG. 1 – *Les silhouettes calculées pour les différentes variables : les coordonnées (Coord.), les altitudes (Alt.), les distributions de températures (Dist.) et la composition des précédentes via les moyennes quasi-arithmétiques (Mix.).*

conversion suivante

$$s(u_i, v_j) = \exp -\frac{d(u_i, v_j)^2}{2s^2}$$

où s est l'écart-type de l'ensemble des valeur $d(u_i, v_j)$ pour la dissimilarité considérée d. La fonction gaussienne a pour avantage que l'effet de la réduction est bien connu dans son cas, et de plus si $d = 0$ alors $s = 1$ ce qui permet de générer des similarités déjà normalisées.

Le choix des poids dépend bien entendu des données et du contexte de la classification effectuée. Dans notre cas la solution des poids égaux pour les trois types de variable été utilisée. Une autre piste possible à explorer pour un choix "automatique" de ces poids serait d'utiliser la mesure d'entropie de chaque variable (Shannon (1948)) quand elle existe.

La détermination du choix du générateur de la moyenne quasi-arithmétique et la valeur optimale de son paramètre sont encore à explorer en profondeur, mais comme nous l'avons écrit précédemment, pour ce premier test nous avons choisi le générateur 4.2.2 présenté en (10) pour de simples raisons pratiques : son paramètre θ permet une maîtrise fine de la courbure, et pour $\theta = 1$ ce générateur donne la moyenne arithmétique classique. Dans notre cas nous avons choisi "arbitrairement" $\theta = 2$.

La similarité résultante devant être convertie en dissimilarité pour l'application de la classification hiérarchique, le choix d'une fonction de conversion (1) est nécessaire. Notre choix, empirique, s'est porté sur la densité d'une normale de moyenne nulle et d'écart-type $\frac{1}{4}$, car sa décroissance est modérée et telle que sa valeur en 1 est assez proche de zéro. Bien entendu d'autres choix sont envisageables.

Sur base des trois matrices de dissimilarités calculées à partir, des altitudes, des coordonnées, des distributions de températures et enfin avec la matrice des dissimilarités calculée à partir des moyennes quasi-arithmétiques des similarités, nous sommes en mesure d'utiliser la classification hiérarchique ascendante.

Nous sommes aussi en mesure d'appliquer la méthode des silhouettes (Rousseeuw (1987)) pour déterminer le nombre optimal de classes. Succinctement expliquée, la silhouette mesure l'adéquation de chaque élément à sa propre classe : plus la silhouette moyenne d'une classification est élevée, plus on peut considérer que les éléments ont été correctement classifiés. La figure 1 montre les résultats pour chaque type de variable. On peut y lire le nombre optimal de classes (selon ce critère) pour chaque type de variable : $k_{opt} = 3$ pour les coordonnées, $k_{opt} = 2$ pour les altitudes et les distributions de températures et enfin, $k_{opt} = 5$ sur base de l'ensemble des variables. On y constate aussi que c'est sur base de l'ensemble des variables que la silhouette moyenne est la plus importante. On peut donc conclure que le calcul d'une similarité par les moyennes quasi-arithmétiques ne détériore pas automatiquement la qualité de la classification résultante. L'amélioration constatée, elle, s'est répétée en utilisant aussi les liens complet et de Ward, alors que pour le lien simple la similarité résultante ne donnait pas une qualité meilleure que pour les autres variables, mais sans faire moins bien [1].

Sur base des silhouettes nous avons effectués la classification avec $k = 5$ pour chaque matrice de dissimilarité et pour les matrices concernant les coordonnées, les altitudes et les distributions nous avons aussi utilisé les valeurs optimales de k données par la méthode des silhouettes, à savoir, respectivement $k = 3$, $k = 2$ et $k = 2$.

Par économie de place nous n'illustrons pas ici les résultats des classifications basée sur les altitudes et sur les coordonnées car ils sont assez prévisibles. Dans le premier cas, avec deux classes, la station himalayenne, culminant à plus de 3700 mètres, forme une classe, et les autres stations forment la seconde. Avec cinq classes, les stations se regroupent en "strates" : de 0 à 300 mètres, de 400 à 1100 mètres, de 1500 à 1900 mètres et enfin les deux stations les plus élevées, 2300 mètres et 3700 mètres, forment à chaque fois un singleton. Dans le cas de la classification sur base des coordonnées, les classes formées sont homogènes et, pour $k = 3$, les classes sont formées des stations du Nord-Est d'une part, des stations de l'Ouest d'autre part, les stations du Sud-Est formant la dernière classe. Pour cinq groupes, c'est la classe de l'Est, la plus clairsemée qui se subdivise, la station himalayenne s'isolant de nouveau.

La classification sur bases des distributions de températures, pour $k = 5$, est illustrée dans la figure 2. On peut y constater que les groupes sont approximativement séparés suivant des axes allant du sud-ouest au nord-est, avec une tendance à la décroissance de l'altitude quand on se dirige dans cette dernière direction.

Enfin si nous examinons maintenant les résultats de la classification sur l'ensemble des variables considérées, nous devons pouvoir espérer que les classes formées tiennent compte des trois groupes de variables pris en comptes. Et c'est ce que nous pouvons constater dans la figure 3 et dans la table 2. Les classes sont réparties en zones géographiques assez homogènes : au nord-est pour la classe 1, à l'ouest pour la classe 2, formant une mince bande centrale pour le 3ème, se regroupant au sud-est pour l'avant dernière et avec la singularité himalayenne pour dernière classe. En ce qui concerne les altitudes, on retrouve la stratification évoquée précédemment avec un peu de recouvrement entre les groupes 1 et 4. Enfin l'ensemble des températures moyennes caractérise clairement les classes 1 (stations les plus froides) et 4

1. Cas non illustrés par manque de place.

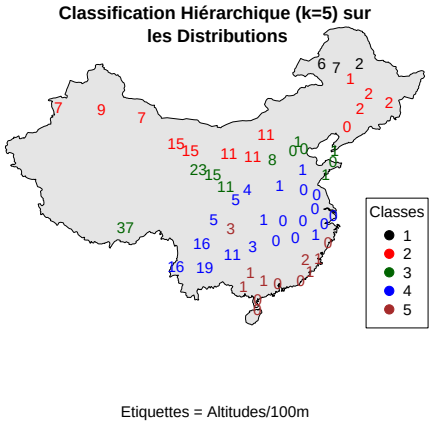

FIG. 2 – *La classification sur base des variables distributionnelles.*

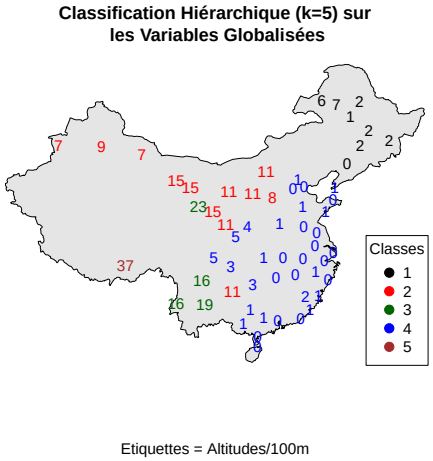

FIG. 3 – *La classification sur base de l'ensemble des variables considérées.*

Classe	Alt	Lat	Long	mTW	mTSp	mTSu	mTF
1	304	46	124	-14	12	18	-7
2	1084	38	102	-3	17	21	2
3	1848	29	101	8	17	19	9
4	103	30	115	7	21	26	13
5	3658	30	91	2	13	14	3

TAB. 2 – *Centres des classes sur l'ensemble des variables considérées : altitudes (Alt), latitudes (Lat), longitudes (Long), températures moyennes en hiver (mTW), au printemps (mTSp), en été (mTSu) et en automne (mTF).*

(stations les plus chaudes), alors que les classes 2 et 3 se distinguent essentiellement par leurs différences de températures hivernales et automnales.

Notre classification sur l'ensemble des variables a donc bien pris en compte des variables de natures différentes, et dont les dissimilarités et/ou similarités ont été calculées en fonction du type de chaque variable mais aussi en fonction de leur signification.

4 Conclusions et Perspectives

Nous proposons dans cet article une façon de calculer une mesure de similarité entre deux objets qui globalise, via les moyennes quasi-arithmétiques, les mesures de similarité et/ou de dissimilarité calculées sur des variables de types différents. Son utilisation pour la classification hiérarchique ascendante d'un ensemble de données ayant des variables numériques et distributionnelles montre des résultats encourageants pour la classification future de situations décrites par des variables de différents types. La méthode devra encore être éprouvée avec plus de types de variables et avec d'autres algorithmes de classification, supervisée ou non. L'impact du choix du générateur et des poids sur le résultat sont une piste de développements futurs à étudier prioritairement.

Références

Ahmad, A. et L. Dey (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering 63*(2), 503–527.

Bullen, P., D. Mitrinovic, et P. Vasic (1988). *Means and their inequalities*. D.Reidel Publishing Company.

Cheung, Y.-M. et H. Jia (2013). Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognition 46*(8), 2228–2238.

Fodor, J. et M. Roubens (1994). *Fuzzy Preference Modelling and Multicriteria Decision Support*. Kluwer Academic Publishers.

Goodall, D. W. (1966). A new similarity index based on probability. *Biometrics*, 882–907.

Irpino, A. (2016). *HistDAWass : Histogram-Valued Data Analysis*. R package version 0.1.4.

Irpino, A. et E. Romano (2007). Optimal histogram representation of large data sets : Fisher vs piecewise linear approximations. *Revue des nouvelles technologies de l'information 1*, 99–110.

Jinyin, C., H. Huihao, C. Jungan, Y. Shanqing, et S. Zhaoxia (2017). Fast density clustering algorithm for numerical data and categorical data. *Mathematical Problems in Engineering 2017*.

Kolmogorov, A. N. (1930). Sur la notion de moyenne. *Rendiconti Accademia dei Lincei 12*(6), 388–391.

Li, C. et G. Biswas (2002). Unsupervised learning with mixed numeric and nominal data. *IEEE Transactions on Knowledge and Data Engineering 14*(4), 673–690.

Liu, S., B. Zhou, D. Huang, et L. Shen (2017). Clustering mixed data by fast search and find of density peaks. *Mathematical Problems in Engineering 2017*.

Liu, S.-H., L.-Z. Shen, et D.-C. Huang (2016). A three-stage framework for clustering mixed data. *WSEAS TRANSACTIONS on SYSTEMS*.

Nelsen, R. (1999). *An introduction to copulas*. London : Springer.

Rodriguez, A. et A. Laio (2014). Clustering by fast search and find of density peaks. *Science 344*(6191), 1492–1496.

Rousseeuw, P. J. (1987). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*(Supplement C), 53 – 65.

Shannon, C. E. (1948). A mathematical theory of communication. (27), 623–655.

Strehl, A., E. Strehl, et J. Ghosh (2002). Cluster ensembles-a knowledge reuse framework for combining partitionings. In *Journal of Machine Learning Research*. Citeseer.

Widder, D. V. (1941). *The Laplace Transform*. Princeton University Press.

Summary

Most clustering methods have been designed for specific data types i.e. numerical, textual, categorical, functional, probabilistic or graph. However, datasets generated in our daily life are made of mixed data. Let's consider the health domain, in particular for cardiac disease prevention. The apps developed in this domain will combine data from sensors with many data types like the age of the patient, the effort level, the maximum cardiac frequency, histograms of average cardiac frequency, etc. For summarizing all these data, it would be useful to be able to build clusters on these different data types and to define a global similarity measure from similarities of pairs of objects based on different data types. In this paper, we propose a clustering method based on merging similarity matrices using quasi-arithmetic means, adapted for choosing the different dimensions of data with different types, based on the assumption that a similarity measure exists for each data type.