Chapitre 4

Agrégation par classification dans les cubes de données

" Il n'y a pas de conditions, de classes, de rang, dans la nature. Les hommes seuls ont cherché à mettre de l'ordre, là où il y en avait déjà et ils ont établi le désordre ! "

Gilbert Louvain, "*La Catherine de Montréal*"

4.1 Introduction

Avec le développement des supports de stockage, nous assistons aujourd'hui à la collecte de grandes masses de données dans les entreprises. En plus, à l'image de la multiplication des secteurs d'activité de ces dernières, les données collectées ne sont pas seulement volumineuses, mais aussi de plus en plus diverses présentant ainsi des contextes sémantiquement différents. Afin d'assurer la gestion et la structuration de ces données, les entreprises ont compris la nécessité d'intégrer les systèmes d'aide à la décision dans leurs systèmes informatiques.

La technologie des entrepôts de données et de l'analyse en ligne OLAP est une solution qui a été largement adoptée pour répondre à ce genre de besoin. En effet, un entrepôt de données permet de centraliser des données historisées en les structurant selon plusieurs axes d'analyse. On parle souvent de structure multidimensionnelles des données [CD97]. Les opérations OLAP permettent à leur tour une exploration des données d'un cube moyennant des outils visuels dédiés à cette structure multidimensionnelle. Classiquement, les axes d'analyse, ou les dimensions, sont aussi organisés selon une ou plusieurs hiérarchies exprimant plusieurs niveaux d'abstraction. Ces derniers traduisent des niveaux de granularité de l'information véhiculée par les dimensions et les mesures. Par exemple, dans une dimension géographique, le niveau *Continent* présente une granularité supérieure à celle du niveau *Pays*. Par conséquent

chaque modalité du niveau *Pays* appartient naturellement à un groupe de modalités du niveau *Continent*.

Cette organisation des modalités selon l'ordre hiérarchique des dimensions est une organisation *rigide* qui est mise en place lors de la phase conceptuelle de l'entrepôt de données. Elle obéit à un ordre d'appartenance naturel dicté par la hiérarchie des concepts existants et couramment utilisés dans le monde réel. Elle peut aussi obéir à des ordres particuliers liés à des contextes d'application spécifiques dont seuls les experts du domaine en font usage et comprennent leurs sens. Cette organisation des modalités ne permet malheureusement pas de rendre compte des liens de proximité ou de ressemblance des faits en fonction des données du cube. Par exemple, dans le contexte d'une entreprise de vente à distance, avec les outils OLAP, on considère classiquement que la *France*, l'*Italie* et l'*Espagne* appartiennent au même groupe des *Pays* de l'*Europe*. En revanche, ces outils ne permettent pas de considérer, par exemple, que la *France* et le *Canada* appartiennent au même groupe des pays où des niveaux de ventes des ouvrages littéraires francophones sont semblables.

Ce deuxième type d'agrégation traduit des connaissances liées à la structure des données. Classiquement, ce cas de figure correspond à un problème de *classification automatique* dans le domaine de la fouille de données. En effet, il s'agit de faire émerger des groupes d'objets semblables au sens d'une métrique donnée. Cependant, il s'agit d'un *apprentissage non supervisé* car on ne sait pas *a priori* quelles classes ou groupes on va obtenir suite à cette classification. En général, le recours aux techniques de classification automatique sous-entend l'existence de certains regroupements dans les données.

La classification des données dans un contexte non supervisé est couramment rencontré. Ce problème peut conduire à des décisions qui dépendent d'enjeux économiques importants et qui peuvent avoir de lourdes conséquences dans la vie des entreprises. Par exemple, de nos jours, dans le domaine du *marketing*, il est indispensable de concevoir des produits, ou des campagnes publicitaires, spécifiques à des types de clients potentiels. Un service de marketing va chercher à identifier des groupes de clients semblables selon différents critères (l'âge, la classe socio-professionnelle, le pouvoir d'achat, la localisation géographique, etc.). Une telle connaissance sur ses clients permet ainsi à l'entreprise de mieux personnaliser et de cibler d'une manière efficace les différentes classes de consommateurs. Cependant, les systèmes d'aide à la décision se basant sur les entrepôts de données et l'analyse en ligne ne disposent pas de moyens pour découvrir de telles connaissances.

Dans ce chapitre, nous proposons une approche pour la *structuration* et la *classification* des données multidimensionnelles. Nous agrégeons les faits d'un cube de données selon leur ordre de proximité et non plus selon l'ordre d'appartenance hiérarchique de leurs modalités dans les dimensions. Pour cela, nous utilisons la classification ascendante hiérarchique (CAH) en vue de construire des classes correspondant à de nouveaux agrégats dans le cube. Ainsi, la classification est perçue

comme une technique d'agrégation sémantique dans les cubes de données. Dans cette approche, la mise en œuvre de la classification dans les données multidimensionnelles se base sur la deuxième approche de couplage entre l'analyse en ligne et la fouille de données que nous avons présentée dans le chapitre 2. Comme le montre la figure 4.1, des opérations OLAP sont utilisés afin d'extraire les données, notamment les individus et les variables, nécessaires à la classification.

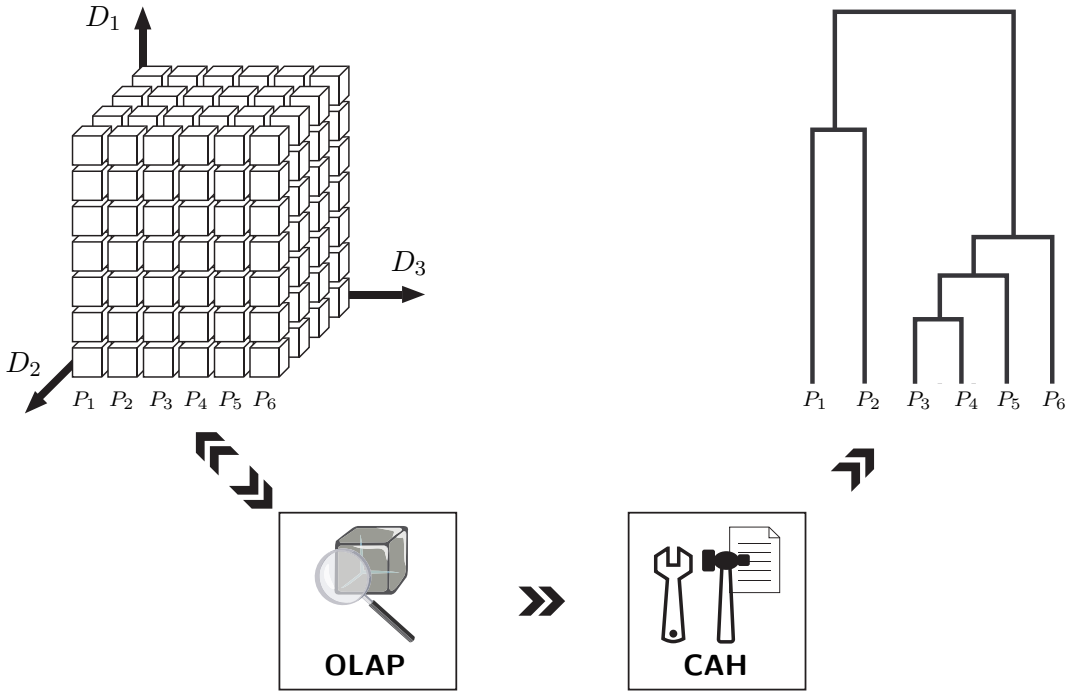


FIG. 4.1 – Étapes de l'agrégation par classification dans les cubes de données

Dans [MRBB04, MBR04], nous avons introduit une première formalisation de notre approche de classification dans les cubes de données. Dans [MBR06a], nous avons amélioré et appliqué notre approche à un cas de données complexes. Ce cas d'application concerne des données de mammographies relatives à des dossiers de patientes atteintes du cancer du sein. Notre étude de cas, que nous présenterons en détails dans le chapitre 6, a montré l'intérêt de notre approche et les enjeux importants dans un système d'aide à la décision. Dans ce chapitre, nous reprenons la formalisation de notre approche et présentons aussi des critères d'évaluation des agrégats fournis par la classification des données multidimensionnelles.

Ce chapitre est organisé de la façon suivante. La section 4.2 expose les objectifs et les motivations de notre proposition. Nous développons, dans la section 4.3, une formalisation selon laquelle nous définissons les individus et les variables de notre

agrégation par classification. La section 4.4 est consacrée pour décrire les principes et les propriétés de l'algorithme de la CAH. Afin d'aider l'utilisateur à choisir les meilleures classes induites par la CAH, nous proposons dans la section suivante des critères d'évaluation de la qualité de ces dernières. Dans la section 4.6, nous concluons ce chapitre et proposons de nouvelles pistes de recherche pour notre approche.

4.2 Objectifs et motivations

La construction d'un cube de données cible un problème d'analyse précis. Le choix des dimensions et des mesures dépend des besoins de l'analyse. D'une manière générale, une dimension est organisée sur plusieurs hiérarchies correspondant à des niveaux d'observation différents. Chaque hiérarchie comporte un ensemble de modalités, et chaque modalité d'une hiérarchie regroupe des modalités de la hiérarchie immédiatement inférieure selon un ordre d'appartenance logique. Par exemple, il est possible de structurer une dimension géographique selon quatre niveaux hiérarchiques : *Ville* → *Région* → *Pays* → *Continent*.

Toutefois, la granularité d'une dimension est fortement dépendante du niveau de précision exigé par l'analyse. Par exemple, si l'analyse exclut les mesures du niveau régional, on peut limiter la dimension géographique aux niveaux : *Ville* → *Pays* → *Continent*. En revanche, l'organisation des modalités d'une dimension est toujours régie par un ordre d'appartenance logique dicté par l'usage naturel des objets ou des concepts du monde réel. Par exemple, il est naturel de dire que la modalité du continent *Europe* d'une dimension géographique contient les modalités des pays *France*, *Italie* et *Espagne*.

Le cube des ventes de la figure 4.2 (a) est constitué de trois dimensions : une dimension géographique D_1, une dimension de produits D_2 et une dimension temporelle D_3. Le cube contient également une mesure M_1 désignant le niveau des ventes. La dimension géographique est organisée selon deux niveaux hiérarchiques : le niveau *Continent* (H_1^1) et le niveau *Pays* (H_2^1). L'idée de notre proposition consiste à exploiter les mesures contenues dans un cube de données afin de regrouper les modalités d'une de ses dimensions. Ainsi, si on veut agir sur la dimension géographique, les modalités du niveau *Pays* sont vues comme des individus qu'on peut décrire par des mesures significatives provenant du cube. Comme le montre la figure 4.2 (a), on peut considérer "*les niveaux des ventes en 2004*" et "*les niveaux des ventes du produit EN-700*" comme descripteurs des individus. D'après cet exemple, le *Canada* est caractérisé par 26 unités de ventes du produit *EN-700* et 17 unités de ventes en *2004*. En adoptant une technique de classification, il est possible de regrouper les pays les plus semblables au sens de ces deux descripteurs.

Contrairement à l'agrégation OLAP classique basée sur le sens de l'appartenance logique, notre approche constitue une nouvelle forme d'agrégation sémantique qui

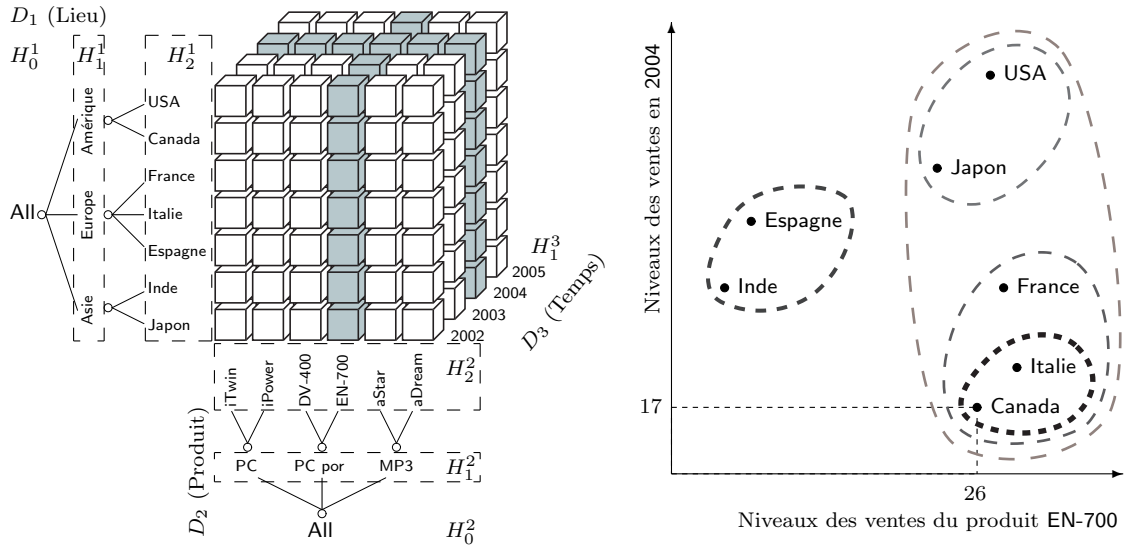


FIG. 4.2 – Agrégation (a) classique dans le contexte OLAP et (b) agrégation par classification

tient compte des faits réels contenus dans le cube de données. Dans le contexte OLAP (Figure 4.2 (a)), on dit que les pays *France*, *Italie* et *Espagne* forment un agrégat puisqu'ils appartiennent tous au continent européen. Alors que dans le contexte de la classification automatique (Figure 4.2 (b)), l'agrégation sémantique nous permet de constater que le *Canada* et l'*Italie* forment un agrégat plus significatif du point de vue de l'utilisateur puisqu'ils représentent des pays où les niveaux de ventes sont semblables.

Nous tentons par notre proposition de fournir des agrégats mettant en évidence les liens réels entre les faits contenus dans les données. Cette forme d'agrégation permet de véhiculer des informations sémantiquement plus riches que celles fournies par l'agrégation classique d'OLAP. Prenant en compte ces objectifs, notre choix s'est porté sur la CAH. Nous justifions ce choix par les points suivants :

1. nous constatons l'existence d'une forte analogie entre les résultats de la CAH et la structuration d'une dimension d'un cube de données. En plus, les objectifs énoncés pour notre approche correspondent bien à la stratégie de la CAH. Cette analogie est due, en grande partie, à l'aspect hiérarchique qui constitue un lien pertinent entre la CAH et les hiérarchies des dimensions d'un cube ;
2. le choix de la CAH n'est pas dû uniquement à l'aspect hiérarchique. Contrairement à la Classification Descendante Hiérarchique (CDH), la CAH adopte une stratégie agglomérative partant de la partition la plus fine où chaque individu est

vu comme une classe. Cette propriété permet à notre démarche d'inclure, dans ses résultats, les modalités les plus fines de la classification. De plus, la stratégie ascendante est plus rapide que la stratégie descendante. La complexité de la CAH est généralement polynomiale, tandis que celle de la CDH est exponentielle [CGLT]. En effet, lors de la première étape d'une méthode ascendante, il faut évaluer toutes les agrégations possibles de deux individus parmi n, soit $n(n-1)/2$ possibilités, tandis qu'un algorithme descendant basé sur l'énumération complète évalue toutes les divisions des n individus en deux sous-ensembles non vides, soit $2^{n-1} - 1$ possibilités ;

3. les résultats de la CAH sont compatibles avec l'esprit de l'analyse en ligne et peuvent être réutilisés par les opérateurs de navigation classiques de l'OLAP. La CAH fournit plusieurs partitions d'individus où chaque partition correspond à un niveau hiérarchique. En passant d'un niveau de partition à celui qui lui est immédiatement supérieur, deux classes sont agrégées ensemble pour former un nouvel agrégat. Inversement, en passant d'un niveau de partition à celui qui lui est immédiatement inférieur, un agrégat est divisé en deux classes. Ce comportement donne à notre proposition un aspect exploratoire comparable à celui des opérateurs OLAP de forage vers le haut (*roll-up*) et vers le bas (*drill-down*).

Cependant, comme toutes les techniques de fouille de données, la CAH requiert la définition des individus à classifier et les variables qui vont intervenir dans le calcul des proximités entre ces individus. Dans la suite, nous fournissons une formalisation détaillée selon laquelle l'utilisateur peut extraire, à partir du cube de données, les individus et les variables de la classification.

4.3 Définitions et formalisation

Nous proposons, dans cette section, un cadre formel pour définir les individus et les variables du problème de classification à partir d'un cube de données. Soit un cube de données $\mathcal{C}$ ayant les propriétés définies selon les notations de la section 3.4. Notons Ω l'ensemble des individus et Σ l'ensemble des variables de la classification à définir.

4.3.1 Individus de la classification

Supposons que nous cherchons à agir sur les modalités du niveau hiérarchique H_j^i de la dimension D_i du cube $\mathcal{C}$. Il est à noter que le choix de H_j^i dépend entièrement des besoins de l'analyse et des objectifs fixés par l'utilisateur. C'est donc à l'utilisateur de fixer la dimension D_i, le niveau hiérarchique H_j^i et les modalités qu'il souhaite classifier dans l'ensemble $\mathcal{A}_{ij}$.

Ainsi, statistiquement parlant, l'ensemble des individus Ω à agréger par la CAH correspond à l'ensemble des modalités choisies par l'utilisateur dans $\mathcal{A}_{ij}$. Soit :

$$\Omega \subset \mathcal{A}_{ij} = \{a_1^{ij}, \ldots, a_t^{ij}, \ldots, a_{l_{ij}}^{ij}\} \tag{4.3.1}$$

Par exemple, dans le cube de la figure 4.2 (a), un utilisateur peut choisir de classifier des modalités appartenant au niveau *Pays* (H_2^1) de la dimension *Lieu* (D_1). Dans ce cas, l'ensemble Ω des individus à classifier appartiennent à l'ensemble $\mathcal{A}_{12}$=$\{$*USA*, *Canada*, *France*, *Italie*, *Espagne*, *Inde*, *Japon*$\}$.

4.3.2 Variables de la classification

Soit $\mathcal{A}$ l'ensemble des n-uplets des modalités des hiérarchies du cube $\mathcal{C}$ y compris les agrégats totaux des dimensions :

$$\mathcal{A} = \prod_{i=1}^{d} \underbrace{\mathcal{A}_{ij}}_{j\in[0,n_i]} = \underbrace{\mathcal{A}_{1j}}_{j\in[0,n_1]} \times \ldots \times \underbrace{\mathcal{A}_{ij}}_{j\in[0,n_i]} \times \ldots \times \underbrace{\mathcal{A}_{dj}}_{j\in[0,n_d]} \tag{4.3.2}$$

On considère qu'une mesure M_q du cube de données $\mathcal{C}$ peut s'écrire selon une fonction de l'ensemble $\mathcal{A}$ dans l'ensemble des réels $\mathbb{R}$.

$$M_q : \mathcal{A} \longrightarrow \mathbb{R} \tag{4.3.3}$$

En d'autres termes, cette fonction fait correspondre à un agrégat d'un cube une valeur scalaire dans $\mathbb{R}$. Par exemple, dans le cube de données de la figure 4.2 (a), en utilisant les notations précédentes, on peut écrire les expressions suivantes :

- M_1(*Canada*, *DV-400*, *All*) désigne la mesure du niveau des ventes du produit *DV-400* au *Canada* durant toutes les années ;
- M_1(*Amérique*, *All* , *2004*) désigne la mesure du niveau des ventes de tous les produits dans le continent américain en *2004*.

Rappelons que l'objectif de notre approche consiste à établir une agrégation sémantique qui tient compte de l'information contenue dans les données d'un cube. Pour cela, nous considérons les mesures du cube comme des variables quantitatives décrivant la population des individus Ω. Cependant, il faut tout de même respecter certaines contraintes logiques et statistiques fondamentales dans le choix de ces variables :

- **Première contrainte :** Aucun niveau hiérarchique de la dimension retenue pour les individus ne doit être générateur des variables de la classification. En effet, décrire un individu par une variable exprimant une propriété qui le contient, ou qui l'agrège, n'aura aucun sens logique. Il serait, par exemple,

insensé de vouloir décrire le niveau des ventes en *Europe* par celui de la *France*. Inversement, une variable qui spécifie des propriétés d'appartenance à un individu ne peut servir que pour la description de cet individu particulier. Par exemple, le niveau des ventes en *France* ne peut servir de descripteur que pour le continent européen et sera inutilisé pour la description des niveaux de ventes des autres continents.

– **Seconde contrainte :** Par dimension, on ne peut choisir qu'un seul niveau hiérarchique pour générer les variables. Cette contrainte est essentielle pour assurer l'indépendance des variables de la classification. En effet, la valeur d'une modalité peut s'obtenir par combinaison linéaire des valeurs des modalités qui lui appartiennent dans la hiérarchie inférieure. Par exemple, la somme des valeurs des ventes pour chaque mois d'une année correspond bien à la valeur totale des ventes de l'année en question.

En conclusion, et en supposons que l'ensemble des individus Ω est sélectionné par l'utilisateur selon l'équation (4.3.1), l'ensemble Σ des variables de la classification de notre agrégation appartient donc à l'ensemble suivant :

$$\Sigma \subset \left\{ \begin{array}{l} V \ / \ \forall t \in \{1, \ldots, l_{ij}\} \\ \\ \underbrace{V(a_t^{ij})}_{j \in \{1,\ldots,n_j\}} = M_q(\ All\ , \ldots,\ All\ ,\ \underbrace{a_t^{ij}}_{j \in \{1,\ldots,n_j\}}\ ,\ All\ , \ldots,\ All\ ,\ \underbrace{a_v^{sr}}_{r \in \{1,\ldots,n_s\}}\ ,\ All\ , \ldots, All\ ) \\ \\ \text{avec } s \neq i,\ r \text{ est unique pour chaque } s,\ v \in \{1, \ldots, l_{sr}\} \text{ et } q \in \{1, \ldots, m\} \end{array} \right\} \quad (4.3.4)$$

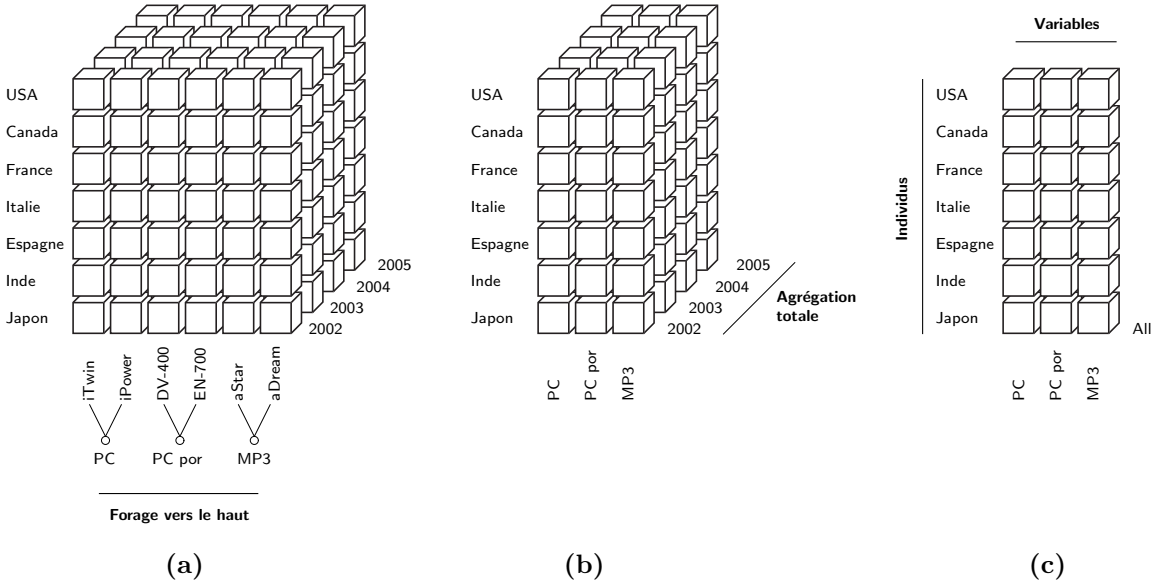
Selon cette équation, un utilisateur peut définir un ensemble de variables en sélectionnant une mesure M_q, une dimension D_s et un niveau hiérarchique H_r^s. Afin de permettre une analyse ciblée, l'utilisateur peut aussi sélectionner dans l'ensemble $\mathcal{A}_{sr}$ des modalités a_v^{sr} particulières qui répondent au mieux à la nature et aux objectifs de l'analyse qu'il souhaite mener.

Pour mieux comprendre cette formalisation, revenons à l'exemple du cube de la figure 4.2 (a). Supposons qu'un utilisateur souhaite classer les pays selon les niveaux des ventes par famille de produits et/ou par année. Dans ce cas, on retient les modalités du niveau *Pays* de la dimension *Lieu* (D_1) comme individus statistiques. On aura donc :

$$\Omega = \{\mathit{USA}, \mathit{Canada}, \mathit{France}, \mathit{Italie}, \mathit{Espagne}, \mathit{Inde}, \mathit{Japon}\}$$

En respectant la première contrainte suscitée, on ne peut plus réutiliser la dimension D_1 pour la génération des variables. De plus, et en respectant la seconde contrainte, on ne peut choisir qu'un seul niveau hiérarchique de D_2 et/ou de D_3 comme générateur de variables. Dans l'exemple de la figure 4.3, on choisit le niveau *Famille de produits* de la dimension D_2 pour générer les variables. Ainsi, dans un

premier temps, nous regroupons les modalités de la dimension D_2 en passant du niveau *Produit* (figure 4.3 (a)) au niveau *Famille de produits* (figure 4.3 (b)) par une opération de forage vers le haut (*roll-up*). Ensuite, nous effectuons des forages totaux vers le haut (agrégations totales) dans toutes les autres dimensions du cube à l'exception de la dimension D_1, retenue pour les individus, et la dimension D_2, retenue pour les variables. Dans notre exemple, cette opération porte sur la dimension D_3 (figure 4.3 (c)). On obtient donc un cube à deux dimensions exprimant les valeurs des ventes pour les modalités de D_1 croisées avec celles de D_2, c'est-à-dire les valeurs des ventes par ville pour chaque produit. De la même manière, on peut générer des variables à partir de D_3 en faisant un forage total vers le haut dans la dimension D_2.



(a) **(b)** **(c)**

FIG. 4.3 – Exemple de domaines d'individus et de variables d'une agrégation par classification dans le cube des ventes

4.4 Classification ascendante hiérarchique

Les données extraites sont mises sous la forme d'un tableau individus-variables comme le montre la figure 4.3 (c). Notons X, de rang (n, p), le tableau individus-variables obtenu à partir du cube de données $\mathcal{C}$. Les n lignes de X représentent les individus de Ω et les p colonnes de X représentent les variables de Σ.

La technique de la CAH – comme toutes les techniques de classification automatique – est destinée à produire des groupements d'individus de Ω du tableau X selon leurs p caractéristiques décrites par les variables de Σ. D'une manière générale,

les techniques de classification ont tendance à fournir des groupes où les individus de chaque groupe sont les plus semblables possible et les groupes sont les plus dissemblables possible. Ces techniques font appel à des mesures de distance pour évaluer les degrés de dissemblance ou de ressemblance entre deux individus ou deux groupes d'individus. Nous exposons dans la suite les mesures que nous prenons en compte dans le cadre de notre agrégation pour le calcul de la distance entre deux individus et la distance entre deux groupes d'individus.

4.4.1 Mesure de distance entre individus

La dissemblance entre deux d'individus est évaluée par la notion de dissimilarité dont le sens mathématique peut se traduire par divers critères de mesure quantitative. D'une manière générale, une distance d entre deux individus est une application à valeurs positives ou nulles obéissant aux propriétés suivantes :

1. $d(A, B) = 0$ si et seulement si $A = B$
2. $d(A, B) = d(B, A)$ (symétrie)
3. $d(A, B) \leq d(A, C) + d(B, C)$ (inégalité triangulaire)

Dans le cadre de notre approche, nous prenons en compte les distances les plus usuelles pour les données quantitatives. L'utilisateur peut choisir une de ces distances pour le calcul des dissimilarités entre les individus à classifier. Formellement, soient A et B deux individus de la matrice des données X décrits par les p variables quantitatives. On note, $A = (a_1, \ldots, a_p)$ et $B = (b_1, \ldots, b_p)$. La distance entre A et B peut se calculer selon :

- *la distance euclidienne* qui calcule la racine carrée de la somme des différences carrées entre les coordonnées de deux individus :

$$d(A, B) = \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}$$

- *la distance euclidienne carrée* qui calcule la somme des différences carrées entre les coordonnées de deux individus :

$$d(A, B) = \sum_{i=1}^{p} (a_i - b_i)^2$$

- *la distance de Chebychev* qui calcule la valeur absolue maximale des différences entre les coordonnées de deux individus :

$$d(A, B) = \max_{i=\{1,\ldots,p\}} |a_i - b_i|$$

– *la distance de Manhattan* qui calcule la somme des valeurs absolues des différences entre les coordonnées de deux individus :

$$d(A,B) = \sum_{i=1}^{p} |a_i - b_i|$$

– *la distance de Minkowski* qui est une métrique de distance générale. Elle s'écrit :

$$d(A,B) = \sqrt[\lambda]{\sum_{i=1}^{p} |a_i - b_i|^{\lambda}}$$

où λ est un entier naturel non nul. On note que le cas $\lambda = 1$ correspond à la distance de Manhattan. La distance euclidienne correspond aussi à la distance de Minkowski quand $\lambda = 2$. La distance de Chebychev est aussi un cas particulier de la distance de Minkowski pour $\lambda = \infty$.

4.4.2 Mesure de distance entre groupes d'individus

Les groupes d'individus étant constitués, la CAH a besoin de déterminer ensuite la dissemblance entre ces groupes deux à deux. Il convient donc de se demander sur quelle base peut-on calculer une dissemblance entre un individu et un groupe et par la suite une distance entre deux groupes. Ceci revient à définir une stratégie de regroupement des éléments, c'est-à-dire, se fixer des règles de calcul des distances entre groupements disjoints d'individus, appelées *critères d'agrégation*. Cette distance entre groupements pourra en général se calculer directement à partir des distances des différents éléments impliqués dans le regroupement.

Par exemple, si A, B, C sont trois objets, et si les objets A et B sont regroupés en un seul élément noté H, on peut définir la distance de ce groupement à C par la plus petite distance des divers éléments de H à C :

$$d(H,C) = \min\{d(A,C), d(B,C)\}$$

Cette distance s'appelle le *saut minimal* (*single linkage*) et constitue un critère d'agrégation.

On peut également définir la distance du *saut maximal* (ou diamètre) en prenant la plus grande distance des divers éléments de H à C :

$$d(H,C) = \max\{d(A,C), d(B,C)\}$$

Une autre règle simple et fréquemment employée est celle de la distance moyenne ; pour deux objets A et B regroupés en H :

$$d(H,C) = \frac{\{d(A,C) + d(B,C)\}}{2}$$

Plus généralement, si A et B désignent des sous-ensembles disjoints de l'ensemble des objets, ayant respectivement n_A et n_B éléments, H est alors un sous-ensemble formé de $n_A + n_B$ éléments et on définit :

$$d(H,C) = \frac{\{n_A d(A,C) + n_B d(B,C)\}}{n_A + n_B}$$

On peut aussi utiliser la distance séparant les centres de gravité des deux groupes comme mesure représentant la ressemblance de ces derniers.

$$d(H,C) = d(G(H),C)$$

où $G(H)$ est le centre de gravité de h.

Le *critère de Ward* est une autre mesure de calcul de distance entre deux groupes qui est plutôt basée sur la décomposition inter et intra-groupe de la variance. Le critère consiste à maximiser la variance inter-groupes et à minimiser la variance intra-groupe. La fonction d'agrégation définit la distance entre deux groupes A et B comme la croissance de la variance intra-groupe ESS (*error sum of squares*) après fusion des deux groupes. Par exemple, en supposant que le groupe A contient n éléments i, sa variance intra-groupe est de : $ESS(A) = \sum_{i=1}^{n} d(i - G(A))^2$. La distance entre A et B selon le critère de *Ward* est :

$$d(A,B) = ESS(AB) - [ESS(A) + ESS(B)]$$

où AB est le groupe résultant de la fusion de A et B.

4.4.3 Algorithme de la classification ascendante hiérarchique

Une étape préliminaire dans l'algorithme de la CAH consiste à construire la matrice de dissimilarités. Cette matrice, notée S, est une matrice symétrique carrée (n,n) dont le terme général s_{ij} correspond à la distance entre les individus i et j. Ainsi, la matrice S contient les distances qui séparent toutes les paires d'individus possibles dans les n individus à classifier. Une dissimilarité s_{ij} est calculée selon une mesure de distance d quantitative choisie par l'utilisateur. Nous résumons l'algorithme de la classification ascendante hiérarchique par les étapes suivantes :

- **Étape 1** : les n individus de la matrice X sont affectés chacun à des classes distinctes $\{A_1, A_2, \ldots, A_n\}$;
- **Étape 2** : l'algorithme cherche dans la matrice S la plus petite distance s_{ij}, c'est-à-dire, l'algorithme détermine les classes A_i et A_j les plus proches ;

- **Étape 3** : les deux classes A_i et A_j sont regroupées pour former une nouvelle classe. À chaque étape de l'algorithme, deux classes sont agrégées ensemble, par conséquent, le nombre de classes est réduit de un ;
- **Étape 4** : l'algorithme construit une nouvelle matrice S des distances qui résultent de l'agrégation. Il calcule les distances entre la nouvelle classe et les classes restantes. Les autres distances restent inchangées. Ainsi, à la première itération, on se trouve dans les mêmes conditions de l'étape 1, mais avec seulement $(n-1)$ classes ;
- **Étape 5** : l'étape 2, 3 et 4 sont répétées jusqu'à atteindre une condition d'arrêt. En général, la condition d'arrêt la plus employée par la CAH correspond au cas où il n'y a plus qu'une seule classe regroupant tous les individus et qui constitue la dernière partition. Dans le cadre de notre approche, la condition d'arrêt des agrégations de la CAH correspond à l'itération où le nombre de classes est égale à un nombre n_c $(1 \leq n_c \leq n)$ défini par l'utilisateur. Il est à noter que, par défaut, n_c est égale à 1, ce qui rejoint la condition d'arrêt classique.

4.5 Évaluation des agrégats des modalités

Rappelons que l'objectif de notre proposition est d'utiliser la CAH pour établir une nouvelle agrégation des faits selon la classification des modalités d'une dimension dans un cube de données. Cependant, d'une manière générale, à partir de n individus à classifier, la CAH génère n partitions hiérarchiques. Comme toutes les techniques de classification, le défaut majeur de la CAH réside dans le choix du nombre de classes adéquat qui répond au mieux aux objectifs de l'analyse menée. De plus, la CAH ne donne aucune indication sur la qualité et la pertinence des classes fournies. Par conséquent, il est souvent difficile pour un utilisateur de choisir la meilleure partition au sens de son analyse. Le choix de cette partition est encore difficile quand l'utilisateur fait face à un grand nombre d'individus à classifier. Généralement, pour valoriser les résultats d'un apprentissage non supervisé – comme c'est le cas de la CAH – on fait appel à des connaissances supplémentaires fournies par les experts du domaine étudié.

Dans la littérature, plusieurs travaux ont été proposés pour l'évaluation de la qualité des résultats de la classification [War63, LMJ82, LFSH04]. Cependant, il est à noter qu'on ne peut pas parler d'un critère universel pour l'évaluation des classes. Chaque mesure de qualité dépend fortement de sa propre construction, du sens de la qualité qu'elle exprime et des orientations d'analyse que sous-entend l'utilisateur [LFSH04]. Par conséquent, dans notre approche, nous proposons d'employer plus d'un critère d'évaluation des classes fournies par la CAH. Nous pensons qu'il est judicieux de fournir à l'utilisateur plusieurs critères fournissant plusieurs points de vue de la qualité des classes des modalités d'une dimension.

Ceci permet de mieux aider l'utilisateur dans son choix de la meilleure partition en adéquation avec les objectifs de son analyse.

Dans la suite, nous exposons les *critères de l'inertie intra et inter-classes* [LMJ82] et le critère de la *méthode de Ward* [War63]. Nous proposons également un nouveau critère pour la mesure de la qualité des partitions de la CAH. Notre critère se base sur la notion de la *séparabilité des classes* [MBR06a]. Pour la formulation de ces critères, nous adoptons les notations générales suivantes :

- $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ est l'ensemble des individus à classer ;
- chaque individu ω est caractérisé par un poids $P(\omega)$ et est décrit par p variables quantitatives $V_1, V_2, \ldots, V_p$;
- soit $k \in \{0, \ldots, n-1\}$ l'indice des itérations, ou des partitions, de la CAH. L'indice $k = 0$ correspond à la partition initiale où chaque individu représente une classe. En général, une itération k correspond à une partition de $(n-k)$ classes ;
- à l'itération k, les classes A_i et A_j sont regroupées et on passe ainsi de la partition $(k-1)$ à la partition k. $A_1, A_2, \ldots, A_{n-k}$ représentent les classes de la partition courante de Ω ;
- n_i est le nombre d'individus de la classe A_i ;
- $\forall i \in \{1, \ldots, n-k\}$, la classe A_i est caractérisée par le poids $P(A_i) = \sum_{\omega \in A_i} P(\omega)$;
- $G(A_i) = \frac{1}{P(A_i)} \sum_{\omega \in A_i} P(\omega)V(\omega)$ est le centre de gravité de la classe A_i ;
- $G = \sum_{\omega \in \Omega} P(\omega)V(\omega)$ est le centre de gravité de tous les individus de l'ensemble Ω ;
- d représente la distance euclidienne et d^2 représente la distance euclidienne carrée.

4.5.1 Critère des inerties intra et inter-classes

Ce critère se base sur la définition fondamentale de toutes les techniques de classification qui préconise la minimisation des dissemblances des individus d'une même classe et la maximisation de celle des groupes d'individus [LMJ82]. L'inertie intra-classe consiste à minimiser les distances séparant les individus d'une même classe. L'inertie inter-classes consiste plutôt à maximiser les distances entre les centres de gravité des différentes classes.

Pour un groupe d'individus A_i, l'inertie intra-classe est définie selon :

$$I(A_i) = \sum_{\omega \in A_i} P(\omega) d(V(\omega), G(A_i)) \tag{4.5.1}$$

L'inertie intra-classe est la somme pondérée des écarts des individus par rapport au centre de gravité de la classe à laquelle ils appartiennent. Cette inertie permet de

mesurer le degré d'homogénéité de la classe. Plus elle est petite, plus la classe est homogène. L'inertie intra-classe totale d'une partition k, notée $I_{intra}(k)$, est égale à la somme des inerties de ses $(n-k)$ classes :

$$I_{intra}(k) = \sum_{i=1}^{n-k} I(A_i) \tag{4.5.2}$$

Pour une partition k, son inertie inter-classe, notée $I_{inter}(k)$ est définie par la somme pondérée des distances séparant les centres de gravité $G(A_i)$ des classes A_i du centre de gravité G de Ω. Cette inertie permet de mesurer la dissemblance des classes, plus elle est grande, plus les classes sont éloignées.

$$I_{inter}(k) = \sum_{i=1}^{n-k} P(A_i) d(G(A_i), G) \tag{4.5.3}$$

D'après le théorème de *Huygens*, quelque soit la partition d'un ensemble d'individus Ω, la somme de ses deux inerties est une constante égale à l'inertie du nuage entier des individus de Ω.

$$\forall k \in \{0, \ldots, n-1\},\ I_{intra}(k) + I_{inter}(k) = I(\Omega) \tag{4.5.4}$$

On montre aussi que l'inertie intra-classe (respectivement, inter-classes) est globalement croissante (respectivement, décroissante) en fonction de l'indice des partitions k. Rappelons que l'itération k correspond à une partition à $(n-k)$ classes. Ainsi, l'inertie intra-classe (respectivement, inter-classes) est décroissante (respectivement, croissante) en fonction du nombre de classes.

- Pour $k = 0$, chacune des n classes est constituée d'un seul individu. L'inertie intra-classe est nulle puisqu'on évalue l'écart de chaque individu à lui-même. En revanche, l'inertie inter-classe sera maximale puisqu'on évalue la somme des écarts de tous les individus par rapport à leur centre de gravité.

  $$I_{intra}(0) = 0 \text{ et } I_{inter}(0) = I(\Omega)$$

- Pour $k = (n-1)$, tous les individus forment une seule classe, l'inertie intra-classe est maximale, puisqu'on se ramène à l'évaluation de la somme des écarts de tous les individus par rapport à leur centre de gravité. Alors que l'inertie inter-classes est nulle puisqu'on ne dispose que d'une seule classe. Dans ce cas, on se ramène au calcul de l'écart du centre de gravité de tous les individus par rapport à lui-même.

$$I_{intra}(n-1) = I(\Omega) \text{ et } I_{inter}(n-1) = 0$$

Ce critère consiste à calculer, pour chaque partition de la CAH, les inerties intra et inter-classes. La détection d'un changement remarquable de l'inertie intra ou inter-classes en passant d'une partition à une autre est un indicateur pertinent pour l'utilisateur dans le choix du nombre de classes à prendre en compte. Il s'agit d'un compromis entre le nombre de classes, la minimisation de l'inertie intra-classe, la maximisation de l'inertie inter-classes et les objectifs de l'analyse.

Le critère des inerties est cependant un critère globalement monotone qui, dans certaines situations, n'offre pas de comparaisons claires sur la qualité des différentes partitions de la CAH. Ainsi, nous proposons à l'utilisateur le critère de la méthode de *Ward* qui évalue différemment la qualité des classes en mesurant le *coup d'agrégation* en passant d'une partition à une autre dans le processus de construction de la CAH.

4.5.2 Critère de la méthode de *Ward*

La méthode de *Ward*, proposée dans [War63], construit un critère qui permet de détecter un saut remarquable dans les niveaux d'agrégation des classes d'objets. Ce critère consiste à évaluer la variation de l'inertie interne quand deux classes A_i et A_j sont agrégées. À chaque itération de la CAH, ce *coup d'agrégation* est calculé selon la distance euclidienne carrée entre les centres de gravité des classes à agréger pondérées par leur poids respectifs :

$$W(A_i, A_j) = \frac{n_i n_j}{n_i + n_j} d^2(G(A_i), G(A_j)) \tag{4.5.5}$$

Rappelons que l'objectif est de trouver la partition des individus dont les classes sont les plus homogènes possible. Ceci revient à minimiser les inerties internes des classes. Par conséquent, à une itération k, une grande valeur du critère de la méthode de *Ward* indique une grande variation de l'inertie interne quand on passe de la partition $k-1$ à la partition k. Cette variation est un indicateur qui apporte une aide à l'utilisateur en lui apprenant qu'il faut plutôt préférer la partition $k-1$ qui correspond à $(n-k+1)$ classes. D'une manière générale, la méthode de *Ward* fournit plus qu'une variation remarquable de l'inertie interne des agrégats d'une CAH. Une fois encore, c'est à l'utilisateur de choisir la meilleure partition non seulement en fonction des résultats du critère, mais aussi en adéquation avec ses objectifs d'analyse.

Il est à noter que les deux critères précédents sont complètement liés au principe de l'inertie et se basent principalement sur l'idée de maximiser l'homogénéité des individus dans les classes. Afin de fournir à l'utilisateur un point de vue complémentaire à celui de l'inertie, nous proposons aussi un nouveau critère pour

mesurer la qualité des agrégats de la CAH en se basant sur le principe de la séparabilité des classes [ZLM02].

4.5.3 Critère de la séparabilité des classes

Ce nouveau critère se base sur le principe de la séparabilité des classes introduit par Zighed *et al.* [ZLM02]. Afin de mettre en valeur cette notion de séparabilité entre classes, nous utilisons les *graphes de voisinage*. Un graphe de voisinage, ou *graphe de proximité*, est un outil visuel qui permet d'analyser la topologie des objets en exprimant les ressemblances, selon une métrique de distance, entre ces derniers dans leur espace de représentation. Concrètement, un graphe de voisinage permet de représenter l'existence ou l'absence de liaisons entre les individus de Ω. Un graphe est formé d'un ensemble de *sommets*, qui représentent les individus, reliés entre eux par un ensemble d'*arêtes* (arcs non orientés) [BG88]. Un sommet est relié à un autre par une arête s'il est voisin de ce dernier selon une structure de voisinage telles que la *structure des k-plus proches voisins*, la *structure de Gabriel* [GS69], la *structure des voisins relatifs* [Tou80] ou la *structure des polyhèdres de Delaunay*.

Pour notre critère de séparabilité des classes, nous utilisons particulièrement le *graphe de Gabriel* proposé dans [GS69]. Dans un graphe de *Gabriel*, deux individus représentés par les sommets A et B sont reliés par une arête si la hypersphère de diamètre AB ne contient aucun sommet de Ω. La figure 4.4 (a) montre un exemple d'une représentation plane d'un graphe de *Gabriel* construit sur des individus décrits par deux variables X_1 et X_2. D'une manière générale, notons g_Ω le graphe de *Gabriel* construit sur l'ensemble Ω des individus à classifier selon les variables de la classification Σ.

La construction de notre critère évolue d'une manière parallèle à celle de l'algorithme de la CAH. En effet, à une itération $k \in \{0, \ldots, n-1\}$ de la CAH, notre critère consiste à construire les graphes de voisinage dans les classes d'individus de la partition en cours. Dans chaque classe A_i de l'itération k on construit le graphe de *Gabriel*, noté g_{A_i}, des individus appartenant à cette classe. Il est à remarquer que pour une partition donnée, l'union des sous-graphes engendrés par les classes A_i ($i \in \{1, \ldots, n-k\}$) de cette partition ne correspond pas forcément au graphe complet de l'ensemble des individus de Ω :

$$\cup_{i=1}^{n-k}\{g_{A_i}\} \neq g_\Omega \tag{4.5.6}$$

Soit e_{ij}, notée aussi $\{\omega_i \leftrightarrow \omega_j\}$, l'arête reliant le sommet i (représentant l'individus ω_i) au sommet j (à représentant l'individu ω_j) dans un graphe de voisinage. Chaque arête e_{ij} est associée à un poids $P(e_{ij})$ égale à l'inverse de la distance euclidienne qui sépare les deux sommets ω_i et ω_j de cette arête.

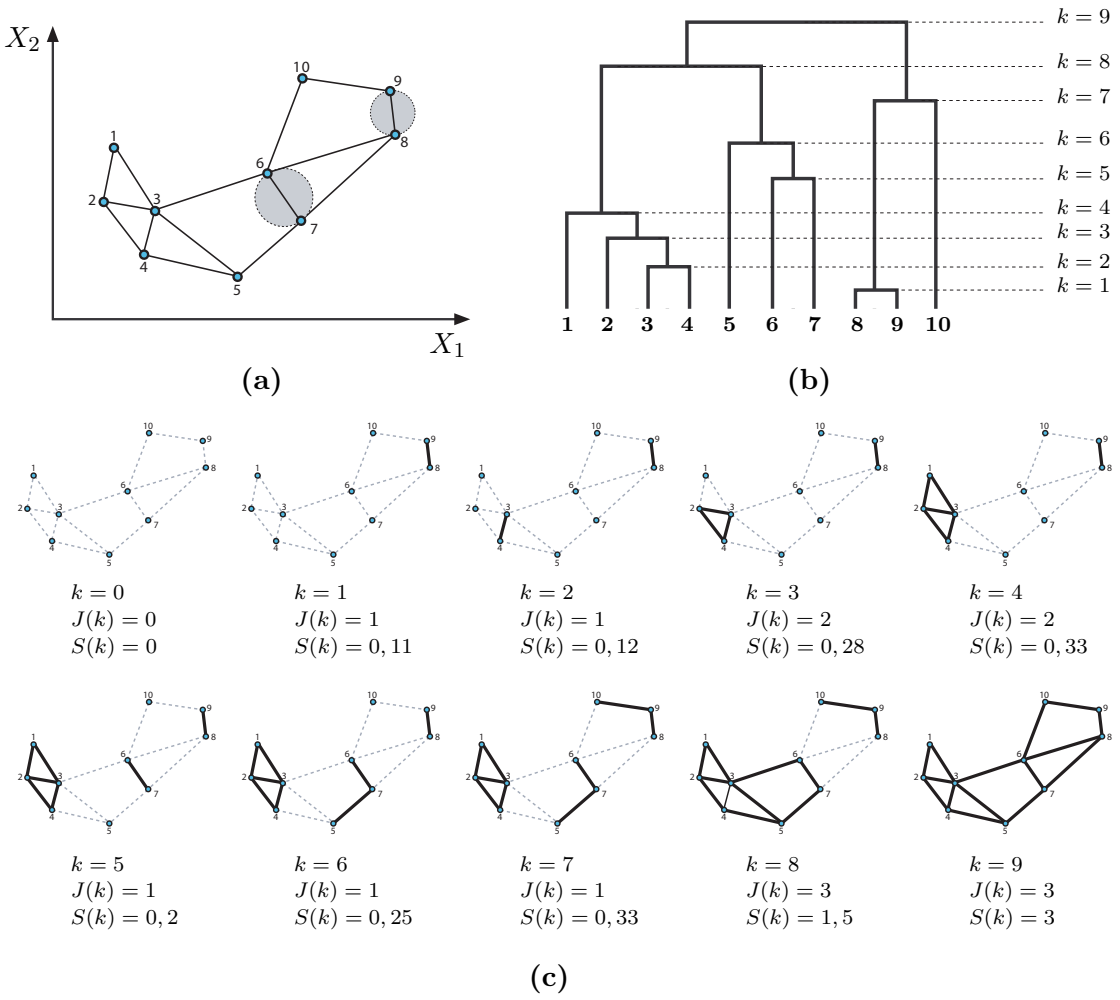


FIG. 4.4 – Principe du critère de la séparabilité des classes

$$P(e_{ij}) = P(\{\omega_i \leftrightarrow \omega_j\}) = \frac{1}{d(\omega_i, \omega_j)} \tag{4.5.7}$$

Le poids associé aux arêtes permet de mesurer l'importance du *degré de séparabilité* entre les sommets d'un graphe de voisinage. En effet, deux sommets séparés par une grande distance correspondent à deux individus peu semblables et qui sont donc facilement séparables dans un processus de classification. Par conséquent, l'arête reliant les deux sommet possède un faible poids dans le graphe de voisinage. En revanche, deux sommets séparés par une petite distance caractérisent deux individus semblables et donc moins facilement séparables. Ainsi, l'arête reliant ces deux sommets possèdent un poids fort dans le graphe de voisinage.

À chaque itération de la CAH, notre critère calcule la somme des poids des nouvelles arêtes construites par les graphes de voisinage dans les classes A_i ($i \in \{1, \ldots, n-k\}$) de la partition en cours. Soit $\mathcal{E}^k$ l'ensemble des nouvelles arêtes construites à l'itération k de la CAH. Dans l'exemple de la figure 4.4, à l'itération $k = 3$, la classe $\{2\}$ est regroupée avec la classe $\{3, 4\}$. Dans ce cas, les arêtes nouvellement construites sont $\{2 \leftrightarrow 3\}$ et $\{2 \leftrightarrow 4\}$. Ainsi, on note $\mathcal{E}^3 = \{\{2 \leftrightarrow 3\}, \{2 \leftrightarrow 4\}\}$.

Soit maintenant $J(k)$ la somme des poids des nouvelles arrêtes construites par les graphes de voisinage à l'itération k :

$$J(k) = \sum_{e \in \mathcal{E}^k} P(e) \tag{4.5.8}$$

Le but de notre critère est d'évaluer la séparabilité des classes pour chaque partition fournie par la CAH. Deux classes ont plus de chance d'être facilement séparables quand elles sont reliées par des arêtes dont la somme des poids est relativement faible. Cependant, les nouvelles arêtes construites, même si leurs poids rendent compte du degré de séparabilité des classes de la partition en cours, ne tiennent pas compte du nombre des classes de cette partition. Dans notre critère nous voulons tenir compte du nombre de classes de la partition considérée afin d'exprimer la séparabilité moyenne par classe dans une partition. Enfin, pour une partition k, notre critère de séparabilité des classe $S(k)$ s'écrit comme suit :

$$S(k) = \frac{J(k)}{n-k} = \frac{\sum_{e \in \mathcal{E}^k} P(e)}{n-k} \tag{4.5.9}$$

$S(k)$ calcule la moyenne des poids des arêtes nouvellement construites au moment où l'algorithme de la CAH passe de l'itération $(k-1)$ à l'itération k. En effet, dans la formule de notre critère on divise $J(k)$ par le nombre de classes $(n-k)$ de l'itération k afin d'obtenir un indice sur le degré de séparabilité en fonction du nombre de classes concernées.

Ce critère est relativement simple à interpréter. Une faible valeur de $J(k)$ indique qu'en passant de la partition $(k-1)$ à la partition k, des arêtes de faibles poids ont été construites. Par conséquent, les classes regroupées à cette étape ont plus de chance d'être séparables. Ainsi, dans ce cas, on peut préférer la partition $(k-1)$ – dont les classes présentent une meilleure propriété de séparabilité – à la partition k.

Par exemple, la figure 4.4 (c) montre l'évolution de la construction des arêtes des graphes de voisinage dans les classes fournies à chaque itération de la CAH selon la figure 4.4 (b). Pour simplifier l'exemple, supposons que toutes les arêtes ont le même poids égal à l'unité ($P(e) = 1$). La figure 4.4 (c) fournit la somme des poids des arêtes construites $J(k)$ et la valeur du critère de séparabilité $S(k)$ à chaque itération k de la CAH. Nous remarquons, dans cet exemple, que $S(k)$ atteint une valeur relativement

faible à la partition $k = 5$. Ce constat peut représenter un indice pour l'utilisateur l'aidant ainsi à préférer la partition précédente $k = 4$ qui se compose de six classes ayant une bonne propriété de séparabilité.

4.6 Conclusion et perspectives

Nous avons proposé, dans ce chapitre, une deuxième approche dans le cadre du couplage entre l'analyse en ligne et la fouille de données. Notre proposition a pour but d'apporter une nouvelle structuration dans les données multidimensionnelles en se basant sur une technique de classification automatique. En effet, la classification permet de créer des groupes d'objets sémantiquement semblables en fonction d'un certain nombre de critères. Par conséquent, ces groupes d'objets traduisent des connaissances fortement intéressantes dans un processus d'aide à la décision.

Compte tenu de ses propriétés particulièrement adaptées au contexte des données multidimensionnelles, nous avons opté pour l'utilisation de la CAH. Notre approche consiste à classifier les modalités d'une dimension d'un cube de données, non pas selon leur ordre hiérarchique classique des dimensions, mais en tenant plutôt compte d'un ordre de proximité calculé en fonction de variables extraites à partir du cube. Par conséquent, notre approche est capable de créer des agrégats sémantiques de faits selon les nouveaux groupements des modalités fournis par la CAH. Alors que les opérateurs OLAP classiques agrègent les faits selon des groupes de modalités traduisant des liens hiérarchiques pré-définis, notre agrégation par classification permet de créer de nouveaux agrégats qui reflètent plutôt des connaissances contenus dans les données du cube.

Afin d'évaluer la qualité des agrégats fournis par notre approche et surtout pour aider l'utilisateur à choisir la meilleure partition d'agrégats parmi celles obtenues par la CAH, nous proposons aussi d'utiliser deux critères classiques, à savoir le critère de l'inertie intra et inter-classes et le critère de la méthode de *Ward*. De plus, dans un souci de fournir à l'utilisateur un point de vue complémentaire à la qualité des agrégats exprimée selon l'inertie, nous introduisons également un nouveau critère basé sur le principe de la séparabilité des classes.

Des améliorations possibles sont à prévoir pour notre présente approche. Nous pensons que, en dehors de sa vocation de structuration et de classification, il est possible aussi d'exploiter notre méthode d'agrégation en vue d'améliorer l'organisation des faits OLAP selon leur ordre de ressemblance dans l'espace de représentation d'un cube de données. En effet, en classifiant les modalités de chaque dimension d'un cube, on réorganise implicitement les faits dans l'espace de représentation du cube. Ceci permet potentiellement de faire émerger des régions intéressantes dans le cube de données, où les faits OLAP sont décrits par des modalités qui sont les plus semblables possible au sens de la classification.

Dans le cadre d'une plateforme générale pour l'analyse et la fouille dans les cubes de données, nous avons prévu une implémentation qui concrétise notre approche d'agrégation par classification. Dans cette implémentation, les outils d'analyse en ligne OLAP sont exploités afin d'interagir avec l'algorithme de la CAH et d'extraire, à partir du cube de données étudié, les données nécessaires pour la construction des agrégats. Une extension de notre agrégation par classification aux données complexes est aussi envisagée. Cette perspective sous-entend la définition au préalable d'une méthodologie d'entreposage et de construction de cubes de données complexes. Elle sous-entend également, sur un plan technique, l'adaptation de notre implémentation à ce nouveau modèle de données. Ces derniers points, en plus d'un cas d'application concernant des données complexes de mammographies, feront l'objet du chapitre 6.

Explication dans les cubes de données par règles d'association

Résumé

Ce chapitre expose notre troisième proposition de couplage entre l'analyse en ligne et la fouille de données. Différemment à nos deux premières propositions, celle-ci adapte un algorithme de fouille afin d'extraire des connaissances directement à partir de la structure multidimensionnelle des données.

Notre proposition s'inscrit dans une démarche explicative dans les cubes de données en se basant sur les règles d'association. Nous mettons en place un nouvel algorithme, de type Apriori, pour une recherche guidée des règles d'association dans les cubes de données. Une visualisation graphique des règles d'association extraites est également proposée afin de mieux valoriser les connaissances qu'elles véhiculent.

Sommaire

Publications

[MRBM06] Messaoud R.B., Rabaséda S.L., Boussaid O., Missaoui R., « Enhanced Mining of Association Rules from Data Cubes », in *Proceedings of the* 9^{th} *ACM International Workshop on Data Warehousing and OLAP (DOLAP'2006)*, pp. 11–18, Arlington, VA, USA : ACM Press. November 2006.

[MBR06c] Messaoud R.B., Boussaid O., Rabaséda S.L., « Mining Association Rules in OLAP Cubes », in *Proceedings of the* 1^{st} *International Conference on Innovations in Information Technology (IIT'2006)*, Dubai, UAE : IEEE Communications Society. November 2006.