MASTER EN SCIENCES MATHÉMATIQUES

Une première approche de l'analyse des données symboliques

Mathot, V.

Award date:
1997

Facultés Universitaires Notre-Dame de la Paix
Namur
Facultés des sciences - Département de Mathématiques

UNE PREMIÈRE APPROCHE DE L'ANALYSE DES DONNÉES SYMBOLIQUES

Mémoire présenté pour l'obtention du grade
de Licencié en Sciences
Mathématiques
par

MATHOT Valérie

Promoteur : J.-P. Rasson

Année académique 1996-1997

Je remercie tout particulièrement Sandrine LISSOIR Laurent PETIT pour leur aide et leur soutien.

Je remercie aussi mon promoteur, Jean-Paul RASSON, pour la confiance qu'il m'a accordée.

Et je tiens surtout à remercier mes parents et mes frères et soeur pour m'avoir soutenue et encouragée durant ces quatre années.

Résumé

Ce travail comportera cinq parties.

Dans la première, nous décrirons diverses sortes d'objets symboliques. Ces objets permettront de représenter des données complexes mais aussi des classes d'individus, des classes de classes, ... On les appellera encore *atomes de connaissances.* Ils se présenteront sous forme de conjonctions de propriétés. Celles-ci seront définies à l'aide de descripteurs (applications) pouvant prendre des valeurs multiples et pondérées par diverses sémantiques.

Dans la deuxième partie, nous comparerons les objets symboliques aux objets numériques et distinguerons les types d'analyse (en particulier celui que nous utiliserons dans la partie quatre).L'étape suivante serait de parvenir à appliquer certaines méthodes de l'analyse des données classique sur ces objets symboliques. La troisième partie comprendra des rappels concernant des outils de l'analyse classique.

Au cours de la troisième partie, nous tenterons d'utiliser ces techniques dans le cadre symbolique (booléen). Il faudra, pour ce faire, construire une distance entre objets symboliques booléens.

Nous terminerons en essayant de déterminer une distance entre objets symboliques modaux probabilistes. Ce sera l'objectif de notre cinquième et dernière partie.

Abstract

This work is divided into five parts.

In the first part, different kinds of symbolic objects are described. These objects allow to consider complex data, clusters, clusters'groups, ... and are represented by conjunctions of properties. The applications, by which these objects are described, are called descriptors and can take multiple values, weighted by different semantics.

In the second part, symbolic objects are compared to numeric ones and various types of analysis are presented.

Classical analysis tools are described in the third part.

The aim of the fourth part is the application of classical technics to boolean symbolic objects. For that, a metric between these objects is necessary.

In the last part, some trials are made to find a metric between modal probabilist symbolic objects.

Introduction

L'accumulation de données de toutes sortes s'est intensifiée ces dernières années. Extraire de ces données des informations utiles dans un but explicatif ou décisionnel est l'une des préoccupations majeures auxquelles sont confrontées les entreprises. L'analyse de données classique (développée principalement au début des années 60) permet uniquement considérer des informations présentées sous forme de tableaux numériques où p variables en colonnes prennent des valeurs sur n objets (individus) rangés en lignes.

Ce type de représentation ne suffit plus. En effet, la complexité grandissante des données disponibles devient très difficile à exprimer dans le carcan des tableaux de l'analyse de données classique.

C'est dans ce contexte que l'on introduit la notion d'objet symbolique.
Ceux-ci permettront une adéquation plus grande à la réalité que les objets utilisés habituellement en analyse des données.

L'analyse des données aura comme objectifs principaux :

- représenter nos connaissances au moyen d'expressions à la fois symboliques et numériques;
- pouvoir manipuler et employer ces expressions afin d'en extraire le plus d'informations possible afin d'aider à la prise de décisions, l'analyse, la synthèse, l'organisation des ensembles de données dont on dispose.

Première partie

Présentation des objets symboliques (atomes de connaissance)

Introduction

L'approche symbolique a pour objectif principal une meilleure représentation de la réalité multidimensionnelle.
Le but de cette première partie sera une description claire des objets utilisés dans le cadre de l'analyse des données symboliques ainsi que d'établir leur principales propriétés.

Dans le premier chapitre, nous définirons les divers types d'objets symboliques booléens en allant du plus simple au plus complexe. Ces objets, comme leur nom l'indique, ne pourront prendre que deux valeurs : $Vrai$ ou $Faux$. Ils apparaîtront sous forme de conjonctions logiques de propriétés. Celles-ci seront décrites au moyen de variables de l'analyse de données classique et de sous-ensembles de valeurs auxquels nous les restreindrons. Nous désignerons encore ces variables par le terme descripteurs. Elles seront de type qualitatif ou quantitatif. Notons que selon leur type, elles auront des 'ensembles images' de formes différentes. Ainsi, en ce qui concerne les variables quantitatives continues, ces 'ensembles images' seront des intervalles de la droite réelle (ou même une union d'intervalles). Par contre, pour les variables quantitatives discrètes ou qualitatives, il s'agira d'ensembles discrets. Il faudra encore distinguer les cas quantitatif discret et qualitatif ordinal du cas qualitatif nominal. En effet, en premier lieu, nous parlerons d'ensembles discrets ordonnés et, en second lieu, d'ensembles discrets quelconques.

Nous commencerons par examiner le cas où les objets sont décrits par une seule variable. Dans ce cas, il s'agira simplement de constater si oui ou non la valeur prise par la variable sur un individu de la population appartient à un ensemble (ou intervalle) prédéterminé. Ce seront les évènements élémentaires.

Ensuite, nous complexifierons le problème en travaillant, non plus avec une seule variable, mais avec une conjonction de plusieurs. Nous parlerons alors d'objets assertions. Par la suite, nous tenterons de prendre en considération plusieurs individus simultanément ce qui nous amènera aux objets hordes. Et, finalement, une conjonction de ces derniers objets conduira à la définition des objets de synthèse. La représentation s'affinera par l'introduction de propriétés liant les individus, les descripteurs, et mêmes les individus et les descripteurs.
Le chapitre 1 se clôturera sur l'étude de quelques propriétés et qualités des objets symboliques booléens, de leurs classes et classifications. En particulier, une propriété très importante sera que tous ces objets pourront se mettre sous une forme syntaxique équivalente.

Le chapitre 2 sera consacré à la présentation des objets modaux. Grâce à ceux-ci, nous espérons obtenir une réponse (renseignement) plus nuancée que simplement $Vrai$ ou $Faux$. Nous allons tenter d'atténuer cet aspect déterministe. Pratiquement, nous attribuerons un mode ou jugement à l'ensemble de valeurs de chacune des variables. En ce cas, les objets symboliques ne seront plus qualifiés de booléens mais bien de modaux. La raison de cette appellation vient de ce que le mode posé fait que l'objet symbolique ne prend plus ses valeurs dans $\{Vrai, Faux\}$ mais bien dans un ensemble ordonné (en général [0, 1]).
Nous présenterons quelques exemples de modes : l'intensité, la probabilité, ...
Ces modes permettront, en fait, de traduire la connaissance (dite aussi sémantique) du domaine étudié. Ils feront intervenir des notions telles que le vague, le doute, l'incertain, ...

Les objets symboliques se rencontrent dans bon nombre de situations :

- lorsque, par exemple, un expert (chef d'atelier, banquier, médecin, biologiste, ...) décide de décrire des connaissances issues de son expérience (comportement d'une chaîne d'usinage, type de clientèle d'une succursale, maladie, plante, ...);

- quand on désire décrire une classe d'objets classiques obtenue par une clas-

sification automatique plus riche et plus explicative qu'en donnant le centre de gravité et la variance;

- si les objets à étudier nécessitent l'utilisation de variables complexes dont la valeur prise peut être soit un arbre, un graphe, d'autres objets, ...

- ...

Chapitre 1

Les objets symboliques booléens

1.1 Définition 1 : Evènement Elémentaire Booléens (EEB)

1.1.1 Description générale

L'élément de base dans la construction des objets symboliques est appelé *évènement élémentaire booléen.*
Motivons-en la définition par un exemple :
Nous désirons exprimer le fait que le *chien Charlie* appartienne à l'une des races suivantes :

$$Cocker.\ Caniche,\ Terrier,\ Berger\ Allemand$$

sans toutefois préciser de laquelle il s'agit exactement. Une idée serait de le faire en employant une variable y (la *race*) d'argument *chien* et en disant que l'événement

$$e(chien\ Charlie) = Vrai$$

si et seulement si la race du *chien Charlie* est soit

$$Cocker,\ Caniche,\ Terrier,\ Berger\ Allemand.$$

Si, ensuite, nous nous donnons un ensemble de *chiens* :

$$\{Poupette,\ Charlie,\ Murphy,\ Loly,\ Brutus\}$$

et que nous regardons dans ce groupe ceux pour lesquels e prend la valeur $Vrai$, ils formeront un ensemble que nous appelerons extension de e. Nous constatons ainsi qu'un évènement permet de désigner, non seulement un individu *chien*, mais aussi l'ensemble des individus qui le vérifient (c'est-à-dire appartiennent à l'une des races proposées).

1.1.2 Définition

Un évènement élémentaire booléen représenté par l'expression symbolique : $e_i = [y_i = V_i]$ *est défini par la fonction :*

$$e_{y_i V_i} : \Omega \to \{Vrai\ Faux\}\ tq\ e_{y_i V_i} = Vrai \iff y_i \in V_i \subset O_i$$

où $y_i : \Omega \to O_i$ $1 \leq p$ *et* Ω *est l'ensemble des objets élémentaires (dits aussi individus).*
Chacune des variables y_i *sera de type soit quantitatif, soit qualitatif.*

1.1.3 Extension

L'extension de l'événement e est $\mid e \mid_\Omega = \{\omega \in \Omega \dagger y_i(\omega) \in V_i\}$

1.1.4 Exemple

Soit un ensemble d'individus $\Omega = \{\omega_1 \ldots \omega_5\}$
et les variables : y_1, y_2, y_3 désignant respectivement la classe d'âge, la nationalité et la taille.

	y_1	y_2	y_3
ω_1	[1, 15]	*espagnol*	1.65
ω_2	]15, 30]	*belge*	1.80
ω_3	]30, 45]	*belge*	1.29
ω_4	]45, 60]	*suisse*	1.30
ω_5	]60, 75]	*espagnol*	1.58

Considérons la première variable de ce tableau, l'évènement élémentaire noté $e = [y_1 = \{[1, 15],]30, 60]\}]$ est défini par la fonction $e_{y_i V_i}(\omega) = Vrai$ si et seulement si $y_1(\omega) \in \{[1, 15],]30, 60]\} = V_1$.
Son extension est $\mid e \mid_\Omega = \{\omega_1, \omega_3, \omega_4\}$.

1.2 Objet Assertion Booléen (OAB)

1.2.1 Description générale

Supposons que le fait de savoir qu'un *chien* appartient à l'une des races proposées ne nous suffit plus parce que nous avons aussi besoin de savoir si la couleur de son pelage est de l'une des deux teintes : *rousse*, *noire*.
Il nous faudra alors exprimer ces deux conditions dans une seule expression. Pour y parvenir, il suffirait simplement de faire la conjonction des évènements élémentaires : e_1 et e_2, où e_1 serait l'évènement e de l'exemple introductif précédent et e_2 décrirait la couleur du pelage.
Nous dirons qu'un *chien* appartient à l'ensemble décrit par cette conjonction $(e_1 \wedge e_2)$ si et seulement si les deux conditions (celle imposée par e_1 et celle imposée par e_2) sont vérifiées.

Un objet assertion est donc une conjonction d'évènements élémentaires.

1.2.2 Définition

Un objet assertion de représentation symbolique

$$a = [y'_1 = V_1] \wedge \ldots \wedge [y'_q = V_q] \quad avec\, V_i \subset O'_i$$

est défini par la fonction $a_{YV} : \Omega \to \{Vrai, Faux\}$ *telle que*

$$a_{YV}(\omega) = Vrai \iff \forall i = 1 \ldots q \quad y'_i(\omega) \in V_i$$

où $y = (y'_1, \ldots y'_q)$
avec $y'_i \in \{y_1 \ldots y_p\}$

et $y'_i : \Omega \to O'_i \in \{O_1 \dots O_p\}$.
On note $V = \{V_1 \dots V_q\}$ $V_i \in O'_i$.

1.2.3 Extension

L'extension d'un OAB est $| a |_\Omega = \{\omega \in \Omega \ tq \ y'_i(\omega) \in V_i \ \forall i = 1 \dots q\}$
= ensemble des objets élémentaires
$\omega \in \Omega$ tq a est vérifié

1.2.4 Exemples

Soit Ω un ensemble de champignons décrits au moyen de deux variables y_1 *et* y_2 se rapportant, respectivement, à la taille du pied et à la couleur du chapeau : il est alors possible de désigner grâce à une assertion booléenne soit un individu, soit un ensemble d'individus. Voyons cela à l'aide de cas particuliers.

1. Un objet élémentaire (individu) tel que $y_1 = 1$ *et* $y_2 = blanc$ pourra s'exprimer sous la forme de l'objet assertion a où $a = [y_1 = 1] \wedge [y_2 = blanc]$.

2. L'ensemble des objets élémentaires (les champignons) de Ω dont la taille est comprise entre 0 *et* 10 et dont la couleur du chapeau est soit *marron*, soit *noire* s'écrira : $a = [y_1 = [0.10]] \wedge [y_2 = \{marron , noire\}]$.

3. Partons d' un ensemble de champignons $\Omega = \{\omega_1 , \omega_2, \omega_3\}$ tel que les individus ω_i sont décrits
   ω_1 par $y_1 = 3$ *et* $y_2 = noire$;
   ω_2 par $y_1 = 2$ *et* $y_2 = blanche$ et
   ω_3 par $y_1 = 1$ *et* $y_2 = blanche$.
   A chaque ω_i , il est possible d'associer un objet symbolique, que nous noterons a_i avec :
   $a_1 = [y_1 = 3 \wedge y_2 = noire]$;
   $a_2 = [y_1 = 2 \wedge y_2 = blanche]$;
   $a_3 = [y_1 = 1 \wedge y_2 = blanche]$.

Cherchons l'extension de l'objet symbolique

$$a = \underbrace{[y_1 = \{2, 3\}]}_{(1)} \wedge \underbrace{[y_2 = \{noire, blanche\}]}_{(2)}.$$

Nous commencerons par récrire les événements (1) et (2) dont la conjonction donne a sous forme de disjonctions :

(1) devient $[y_1 = 2] \vee [y_1 = 3]$;
(2) devient $[y_2 = noire] \vee [y_2 = blanche]$.

Ainsi, nous reformulons a comme suit :

$$a = [[y_1 = 2] \vee [y_1 = 3]] \wedge [[y_2 = noire] \vee [y_2 = blanche]].$$

Ce qui s'écrit encore

$$a = \omega_1 \vee \omega_2 \vee z_1 \vee z_2$$

avec $z_1 = [y_1 = 3] \wedge [y_2 = blanche]$ et $z_2 = [y_1 = 2] \wedge [y_2 = noire]$.

Et nous aurons finalement :

$| a |_\Omega = \{\omega_1, \omega_2\}$ et
$| a |_{\Omega'} = \{\omega_1, \omega_2, z_1, z_2\}$.

Nous remarquons que l'extension de a dans Ω , $| a |_\Omega$ ne contient pas les individus z_1 et z_2 car ceux-ci n'appartiennent pas à Ω .
Par contre, nous les retrouvons dans Ω' et donc dans $| a |_{\Omega'}$. Ω' est l'ensemble des objets élémentaires $\{\omega_1, \omega_2, z_1, z_2\}$.

1.2.5 Cas particuliers

Etant donné que V_i est une partie de O'_i , deux cas extrêmes peuvent survenir :

1. y'_i prend une valeur quelconque dans O'_i c'est-à-dire $V_i \equiv \{O'_i\}$

2. y'_i n'est pas défini (ce qui est admissible car y_i est une fonction) c'est-à-dire $V_i \equiv \emptyset$.

Exemple

L'extension de

$$a = [\text{existence chapeau} = non] \wedge [\text{couleur chapeau} = \emptyset] \wedge [\text{ lieu de pousse} = \{O_3\}]$$

est formée de l'ensemble des champignons sans chapeau sur lesquels la variable couleur chapeau n'est pas définie (i.e ne prend aucune valeur dans l'ensemble O_2 des couleurs possibles et V_2 est vide) et dont le lieu de pousse est quelconque.

1.2.6 Objets Assertions Simplifiés

Lorsque $V_i = O_i'$, nous pouvons dans certains contextes supprimer le terme $[y_i' = O_i']$ dans la conjonction.
Ainsi, par exemple, si y_2 ='couleur du chapeau' ne peut prendre que deux valeurs (*blanche* ou *noire*) ; cela entraine que l'assertion
$a = [y_1 = 1] \wedge [y_2 = \{blanc, noire\}]$ devient $a = [y_1 = 1]$.

Cette démarche n'est pas toujours possible sans risquer de perdre de l'information car l'emploi de la variable y_i permet parfois de distinguer les objets symboliques qui sont concernés par elle de ceux qui ne le sont pas.
Ainsi, par exemple, la variable rayon n'intervient pas dans la description d'un losange mais bien la variable surface.

1.3 Objets Hordes

1.3.1 Description générale

Si nous reprenons le tableau de données,

	y_1	y_2	y_3
ω_1	[1, 15]	*espagnol*	1.65
ω_2	]15, 30]	*belge*	1.80
ω_3	]30, 45]	*belge*	1.29
ω_4	]45, 60]	*suisse*	1.30
ω_5	]60, 75]	*espagnol*	1.58

nous savons que pour les événements $e_1 = [y_1 = [1, 15]]$ *et* $e_2 = [y_2 = suisse]$, nous arriverons à trouver des éléments $\omega \in \Omega$ tels que ceux-ci soient vérifiés; nous avons en effet que ω_1 *et* ω_4 conviennent c'est-à-dire que $e_1(\omega_1) = Vrai$ *et* $e_2(\omega_4) = Vrai$.

Par contre, nous pouvons constater qu'il n'existe aucun individu satisfaisant simultanément e_1 et e_2.

Nous nous proposons alors de travailler avec un type d'objet plus complexe qui permettra de décrire plusieurs individus dans une seule expression.

C'est sur cette idée que repose la notion d'objet horde.

En voici un exemple :

$h = [y_1(u_1) = [1, 15]] \wedge [y_2(u_2) = suisse]$.

Dans cet exemple, nous considèrons les deux individus u_1 et u_2 en même temps.

Notons que, pour le cas où nous traitions un seul individu à la fois, nous avions volontairement omis d'écrire les arguments des descripteurs y_i.

Cela ne s'avérait d'ailleurs pas nécessaire puisque l'argument était identique pour tous les y_i.

Avant d'en donner la définition formelle, signalons que : y_i , y_i' , O_i' , V_i sont définis comme pour les objets assertions.

1.3.2 Définition d'un objet horde

Un objet horde représenté par l'expression symbolique :

$$h = [y'_1(u_1) = V_1] \wedge \ldots \wedge [y'_p(u_p) = V_q]$$

est défini par la fonction $h_{YV} : \Omega^q \rightarrow \{Vrai\, Faux\}$ *telle que* $\forall\ W = (\omega'_1, \ldots, \omega'_q) \in \Omega^q$, *on ait :* $h_{YV}(W) = V \iff \forall i\ y'_i(\omega'_i) \in V_i$

Lorsque tous les u_i sont égaux, nous nous trouvons dans le cas particulier où l'objet horde est un OAB.

1.3.3 Extension d'un objet horde

Soit h un objet défini sur Ω^q.
L'extension de h est l'ensemble des éléments $W \in \Omega^q\ tq\ h_{YV}(W) = Vrai$.
$h^{-1}_{YV}(Vrai) = \mid h \mid_\Omega = \{W = (\omega'_1, \ldots, \omega'_q) \in \Omega^q\ tq\ y'_i(\omega'_i) \in V_i\}$.

1.3.4 Exemple

Nous utilisons le tableau de données

	y_1	y_2	y_3
ω_1	[1, 15]	*espagnol*	1.65
ω_2	]15, 30]	*belge*	1.80
ω_3	]30, 45]	*belge*	1.29
ω_4	]45, 60]	*suisse*	1.30
ω_5	]60, 75]	*espagnol*	1.58

et nous nous donnons l'objet horde $h = [y_1(u_1) = [1, 15]] \wedge [y_2(u_2) = espagnol]$.
L'extension de h dans Ω est :
$\mid h \mid_\Omega = \{(\omega_1, \omega_1), (\omega_1, \omega_5)\}$.
Si nous considèrons l'objet assertion $a = [y_1 = [1, 15]] \wedge [y_2 = espagnol]$, son extension $\mid a \mid_\Omega$ est réduite à l'objet élémentaire ω_1.
Ce qui illustre bien qu'un OAB est un cas particulier d'objet horde.

1.4 Objets de synthèse booléens

1.4.1 Description générale

Les objets de synthèse permettront de travailler non seulement avec des individus différents simultanément mais aussi sur plusieurs populations comprenant elles-mêmes différentes classes d'individus.
Soient $\Omega_1, \ldots, \Omega_k$, k ensembles d'objets élémentaires avec

Ω_1 défini sur l'ensemble des variables J_1 ;
... ...
... ...
Ω_k défini sur l'ensemble des variables J_k ;

avec $\#(J_i) = P_i$.

Soit H_i l'ensemble des objets hordes que nous pouvons définir sur Ω_i .
(exemple : si Ω_i est un ensemble de *chiens*, il y aura dans H_i les objets hordes suivants:
h_m = [couleur($chien1$) = $beige$] $\wedge$ [race($chien2$) = $\{Lassie, Teckel\}$]
h_l = [poids($chien1$) = $[2, 4]$] $\wedge$ [âge($chien2$) = $\{5, 8\}$]
etc ....)

1.4.2 Définition

Un objet de synthèse est la conjonction de k objets hordes respectivement définis sur chacun des ensembles $\Omega_1, \ldots, \Omega_k$ (et appartenant respectivement à $H_1, \ldots, H_k$).

Il s'écrira sous la forme générale : $s = h_1 \wedge \ldots \wedge h_k$ avec $h_i \in H_i$.

Les objets hordes et assertions sont des cas particuliers des objets de synthèse lorsque k=1.

1.4.3 Extension d'un objet de synthèse

L'ensemble des éléments de $\hat{\Omega} = \Omega_1 \times \ldots \times \Omega_k$ qui satisfont s constituent l'extension de s notée $\mid s \mid_\Omega$.

1.4.4 Exemple

Nous choisissons 3 ensembles Ω_1 , Ω_2 , Ω_3 avec Ω_1 qui est l'ensemble des fenêtres, Ω_2 celui des portes et Ω_3 celui des toits.
Parmi les fenêtres de Ω_1, nous prenons les couples de fenêtres (u, v) qui répondent aux conditions exprimées par l'objet horde suivant :

$$h_f = [\text{forme fenêtre}(u) = \textit{ronde}] \wedge [\text{type de vitre fenêtre}(v) = \textit{épais}]$$

Nous prenons ensuite les couples (u, v) de portes de Ω_2 satisfaisant à l'objet horde :

$$h_p = [\text{ blindage porte}(u) = \textit{acier}] \wedge [\text{ matière porte}(v) = \textit{chêne}]$$
$$\wedge [\text{ numéro porte}(u) = 12].$$

Et enfin pour les toits (nous ne parlons pas de couple car il n'y a, dans ce cas, qu'un seul individu étudié, il s'agit donc d'un objet assertion), nous prendrons :

$$a = [\text{ type de toit} = \textit{ardoise}] \wedge [\text{surface de toit} = \textit{très grande}]$$

Nous définissons alors notre objet de synthèse *type de maison* par $m = h_f \wedge h_p \wedge a$.
Un objet de l'extension de m dans $\Omega_1^2 \times \Omega_2^2 \times \Omega_3$ est une maison vérifiant les trois objets hordes h_p , h_f *et* a.
Il s'agira de maisons possédant une fenêtre *ronde* et une fenêtre à vitre *épaisse*, une porte en *acier* portant le numéro 12, une porte en *chêne* et un toit en *ardoise* de *très grande* surface.

1.5 Objets munis de méthodes et de propriétés

1.5.1 Introduction

Nous allons généraliser les objets qui viennent d'être définis à des objets munis de méthodes et de propriétés.

Pour ce faire, nous ajoutons par conjonction des événements décrivant, par exemple des méthodes pour calculer une variable ou une fonction (des variables ou des objets élémentaires) ou des propriétés exprimant des liens entre les variables ou entre les objets élémentaires ou encore entre variables et objets élémentaires.
Ces liens dépendent du type d'objets et seront de plus en plus complexes en allant des objets assertions jusqu'aux objets de synthèse.
Ils peuvent tous s'exprimer sous forme de fonction de plusieurs éléments qui peuvent être des variables et des objets élémentaires.

1.5.2 Objets assertions munis de méthodes et de propriétés

1.5.2.1 Introduction

Dans ce cas, les propriétés ne peuvent concerner que les méthodes de calcul de chaque variable ou les propriétés qui les relient.
En effet, comme nous l'avons déjà mentionné dans la description des objets hordes, les descripteurs intervenant dans un OAB portent tous sur un unique et même individu $\omega \in \Omega$. C'est pour cette raison que nous nous étions d'ailleurs permis d'omettre de noter les arguments de ces variables.

1.5.2.2 Forme générale

La forme la plus générale d'un objet assertion muni de méthodes et de propriétés sera :

$$a_p = [y_1 = V_1] \wedge \ldots \wedge [y_p = V_p] \wedge \overbrace{y_i = meth[y_l \ldots y_k]}^{(1)} \wedge \ldots \wedge \overbrace{[meth[y_k \ldots y_1]]}^{(2)}$$
$$\wedge \overbrace{[y_i \; R_k \; y_j]}^{(3)} \wedge \ldots\ldots \wedge [y_1 \; R_m \; y_n]$$

- (1) nous donne la méthode pour calculer la variable y_i en fontion de $y_l \ldots y_k$;
- (2) est une information concernant, par exemple, la manière de procéder pour atteindre un objectif ;

- (3) donne la relation liant y_i *et* y_j.
  Nous avons, par exemple, comme cela est décrit dans l'article [1] des dépendances logiques (DL) entre variables.

1.5.2.3 Description des différentes sortes de dépendances logiques

Parmi les dépendances logiques, on distinguera :

- les dépendances conditionnelles (DC);
- les dépendances de corrélation logique (DCL).

Pour représenter des connaissances réelles, la description d'une classe d'individus par un OAB doit tenir compte des différentes DL entre les variables. Ces DL sont, en fait, décrites par des règles entre les variables.

1. les DC :
   - une variable y_i peut devenir inapplicable si une autre variable y_j prend ses valeurs dans un sous-ensemble S_j :

     $$\forall\, \omega \in \Omega \;\; y_j(\omega) \in S_j \subseteq O_j \Longrightarrow y_i(\omega) \text{ est Non Applicable.}$$

   - Exemple :
     Il n'est pas possible de décrire la couleur du collier d'un chien qui n'en possède pas.
2. Les DCL :
   - Un sous-ensemble $S_j \subseteq O_j$ d'une variable y_j peut être en relation avec un sous-ensemble $S_i \subseteq O_i$ d'une autre variable y_i de telle manière que :

     $$y_j(\omega) \in S_j \subseteq O_j \Longrightarrow y_i(\omega) \in S_i \subseteq O_i.$$

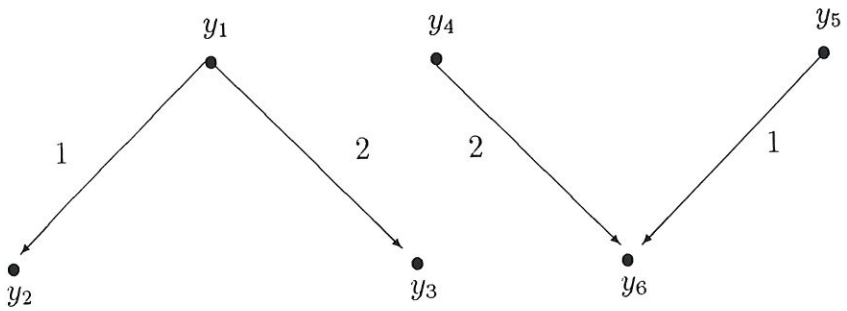
   - Exemple :
     $\forall \omega \in \omega$ si couleur$(\omega) = \{brune\} \Rightarrow$ race$(\omega) = \{Epagneul, Berger\ Allemand.\}$.

1.5.2.4 Représentation graphique

Nous pouvons représenter ces DL graphiquement par un ensemble de graphes connexes[1] Dans chaque graphe, les noeuds sont les variables (ou encore leur ensemble de valeurs).
Ces graphes sont directs (leurs chemins sont de longueur unité: l'arc issu de la variable prémisse de la règle conduit à la variable conclusion).

Ces arcs sont étiquettés:

- ils ont le label 1 s'ils représentent une DC;
- ils ont le label 2 s'ils représentent une DCL.



Représentation de DC et DCL.

1.5.2.5 Extension

Finalement, l'extension d'un objet assertion muni de méthodes et de propriétés sera:

$| a_p |_\Omega = \{\omega \in \Omega$ tels que $y_i(\omega) \in V_i$

1. **Rappel**

1. Un graphe $G(X,U)$ *connexe* (où X est l'ensemble des noeuds et U celui des arcs) est tel que pour tout x, y appartenant à X (x différent de y), il existe une chaîne dont les extrémités sont x et y.
2. Une *chaîne* est une séquence d'arcs $(u_1, u_2, \ldots, u_i, \ldots, u_q)$ telle que pour chaque arc u_i est attaché à u_{i-1} par une de ses extrémités et à u_{i+1} par l'autre de ses extrémités.

Cette définition est tirée de [14]

et pour lequels les méthodes et propriétés décrites ci-dessus sont vérifiées.}

L'indice p du a signifie propriétés.

1.5.2.6 Exemple

$$\text{Soit } a_p = \underbrace{[\text{ couleur chapeau } = \{blanc, jaune\}]}_{(1)} \wedge \underbrace{[\text{ taille du pied } = [0, 10]]}_{(1)}$$
$$\wedge \underbrace{[\text{ taille du chapeau } = [0, 15]]}_{(1)} \wedge \underbrace{[\text{ taille du chapeau } = Meth(c)]}_{(2)}$$
$$\wedge \underbrace{[\text{ meth(calcong) } = f(l,c)]}_{(3)} \wedge \underbrace{[\text{ taille chapeau } > \text{ taille pied }]}_{(4)}$$

avec $Meth(c)$ qui donne la méthode de calcul de la taille du chapeau : on prend la longueur de la plus grande lamelle; et $f(l,c)$ est la méthode qui permet de calculer la longueur du champignon (qui n'est pas une variable) en fonction de celle du pied et du chapeau.
Si nous comparons avec la présentation théorique : (1) correspond à $[y_i = V_i]$; (2) à $Meth(y_p \ldots$; (3) à $meth(y_k \ldots y_1)$ et (4) à $[y_i\, R_k\, y_j]$.

1.5.2.7 Remarque

Une propriété peut aussi s'exprimer sous forme de règle :
Ainsi, par exemple, si nous savons qu'une anémone ne pousse qu'à partir de $mars$ ($[y_1 = mars]$) au $nord$ ($[y_2 = nord]$) et à partir d'$avril$ ($[y_1 = avril]$) au sud ($[y_2 = sud]$), alors :

$$\begin{aligned} a = &[y_1 = (mars, avril)] \wedge [y_2 = \{nord, sud\}] \\ &\wedge [[y_2 = nord] \Rightarrow [y_1 = mars]] \\ &\wedge [\,[y_2 = sud] \Rightarrow [y_1 = avril]]. \end{aligned}$$

1.5.3 Objets Hordes munis de méthodes et de propriétés.

1.5.3.1 Introduction.

En plus des propriétés définies pour les objets assertions, il existe des méthodes qui peuvent dépendre de plusieurs variables et objets élémentaires ainsi que des

propriétés liant ces objets ou les valeurs prises par les variables sur ces objets. Ceci était impossible dans le cadre des objets assertions qui ne disposaient que d'un seul argument.

1.5.3.2 Forme générale.

La forme la plus générale que puisse prendre un objet horde muni de méthodes et de propriétés est donc :

$$h_p = \underbrace{\omega_{ap}}_{(i)} \wedge \underbrace{[y_1(u_1) = V_1] \wedge [y_q(u_q) = V_q]}_{(ii)}$$

$$\wedge \underbrace{[y_l(u_i) = F(u_j \dots u_k)]}_{(1)} \wedge \dots \wedge \underbrace{[meth[y_l\, y_i \dots u_j\, u_k]}_{(2)}$$

$$\wedge \underbrace{[y_1(u)\, R_k\, y_m(v)[}_{(3)} \wedge \dots \wedge \underbrace{]y_m(w)\, R_1\, y_k(t)]}_{(4)}$$

- (i) représente la forme générale d'un OAB muni de méthodes et de propriétés;

- (ii)est la forme habituelle d'un objet horde;

- (1),(3)et(4) désignent les propriétés reliant les objets élémentaires ou encore les valeurs prises par des variables sur ces objets;

- (2) exprime la méthode et donne les objets élémentaires et variables dont elle dépend.

1.5.3.3 Extension

Finalement, l'extension d'un objet horde muni de méthodes et de propriétés sera :

$| h_p |_\Omega = \{(u_1 \dots u_p) \in \Omega^p$
tels que les méthodes et propriétés ci-dessus sont vérifiées $\}$.

1.5.3.4 Exemple

$h_p = [\text{couleur}(u_1) = \{gris, noir\}] \wedge [F(u_1\ u_2)] \wedge [\ \text{taille}(u_1) < \text{taille}(u_3)]$
$\wedge [\text{vitesse vers 14h}(u_2) > \text{vitesse vers 15h}(u_4)].$

Et, $F(u_1\ u_2)$ affirme que u_1 est plus petit et plus léger que u_2.

En donnant une signification à u_1, u_2, u_3, u_4, nous pouvons avoir, par exemple:

$h_p(chat, souris, chien, puce) = [\text{couleur}(chat) = \{gris, noir\}]$
$\wedge [F(chat, souris)] \wedge [\text{taille}(chat) < \text{taille}(chien)]$
$\wedge [\text{vitesse vers 14h}(souris) > \text{vitesse vers 15h}(puce)].$

1.5.4 Objets de synthèse munis de méthodes et de propriétés

1.5.4.1 Introduction

Soient k ensembles $\Omega_1 \ldots \Omega_k$ associés respectivement à k ensembles de variables $J_1 \ldots J_k$, un objet de synthèse est une conjonction de k objets hordes h_p^i définis sur (Ω_i, J_i) munis de méthodes et de propriétés. Il est encore possible d'ajouter des méthodes et propriétés liant ces objets hordes entre eux.

1.5.4.2 Forme générale

Finalement, la forme générale d'un objet de synthèse muni de méthodes et de propriétés est donc:

$$h_{sp} = \underbrace{h_p^1 \wedge \ldots \wedge h_p^q}_{(1)} \wedge \underbrace{[meth(h_p^i) \ldots (h_p^i)]}_{(2)} \wedge \ldots \wedge \underbrace{[h_p^l\ R_k\ h_p^m]}_{(3)}.$$

- (1) exprime la conjonction de k objets hordes munis de méthodes et de propriétés;

- (2) et (3) désignent les méthodes et les propriétés reliant les objets hordes munis de méthodes et de propriétés.

1.5.4.3 Exemple

Si nous reprenons l'exemple du paragraphe 1.4.3 concernant la description de maisons, une méthode serait une manière de situer les fenêtres par rapport aux portes et une propriété pourrait être: [la taille d'une fenêtre est plus petite ou égale à celle d'une porte.]

1.6 Quelques propriétés des objets symboliques

1.6.1 Proposition 1

Les objets élémentaires assertions, hordes et de synthèse peuvent se mettre sous une forme algébrique équivalente.

- *Si E, A, H, S sont respectivement des ensembles d'objets (événements) élémentaires, assertions, hordes et de synthèse*

$$\Longrightarrow \quad E \subset A \subset H \subset S\,;$$

- *Il est toujours possible de trouver un ensemble d'objets élémentaires $\Omega^\star$,un ensemble d'observations $O^\star$ et une variable $y : \Omega^\star \to O^\star$ tels que tout objet symbolique appartenant à E, A, H, S soit considéré comme un événement élémentaire.*

Conséquences : comme nous pouvons tout ramener sous forme d'évènement élémentaire, il est possible de compliquer la représentation symbolique indéfiniment. Ceci permettra de représenter la réalité multidimensionnelle. Il résulte aussi que toute propriété algébrique pour un type d'objet et non spécifique à celui-ci sera satisfaite par les autres types d'objets.

1.6.2 L'ensemble des objets symboliques et l'extension symbolique

1.6.2.1 Ensemble des objets symboliques

Supposons (pour simplifier) que nous ayons un ensemble S d'objets symboliques définis sur un ensemble Ω caractérisé par des variables $y_i : \Omega \to O_i$.

D'après la proposition 1, si nous choisissons correctement y_i , Ω_i et O_i, S pourra aussi bien être un ensemble d'objets élémentaires, assertions, hordes ou de synthèse.
Nous prendrons comme convention de dire que S est un ensemble d'objets assertions (ces derniers étant plus explicites que les objets élémentaires tout en ne comportant pas la complexité des objets hordes ou de synthèse).

1.6.2.2 Extension symbolique

Avant de donner la définition de l'extension symbolique, nous devons introduire une autre notion : la bijection φ entre Ω et S.

Soit l'application $\varphi : \Omega \rightarrow S$ qui à *tout* élément (donc partout définie) $\omega \in \Omega$ associe l'assertion

$$\varphi(\omega) = [y_1 = y_1(\omega)] \wedge \ldots \wedge [y_p = y_p(\omega)].$$

Si nous posons ensuite la condition, deux éléments différents de Ω ne peuvent pas prendre des valeurs identiques pour toutes les variables (intervenant dans la formulation de φ), alors nous saurons que

$$\varphi : \Omega \rightarrow \varphi(\Omega)$$

est une bijection.
En effet, la condition ci-dessus nous assure l'injection et le fait que nous ayons réduit l'ensemble d'arrivée de S de φ à son ensemble image $\varphi(\Omega)$ entraine le caractère surjectif.
Nous noterons $\omega^s \in S$ l'objet symbolique associé à $\omega \in \Omega$ au moyen de l'application φ $(\omega^s = \varphi(\omega))$.

Nous sommes en mesure maintenant de donner la définition de l'extension symbolique :

L'extension symbolique *d'un objet symbolique* $s \in S$ *sera notée* s' *avec*

$$s' = \{\varphi(\omega) \in S \ tq \ \omega(\in \Omega) \ \in | \ s \ |_\Omega\}.$$

Exemple Soit le tableau de données

	y_1	y_2	y_3
ω_1	[1, 15]	*espagnol*	1.65
ω_2	]15, 30]	*belge*	1.80
ω_3	]30, 45]	*belge*	1.29
ω_4	]45, 60]	*suisse*	1.30
ω_5	]60, 75]	*espagnol*	1.58

et prenons

$$s = [y_2 = \{espagnol, suisse\}] \wedge [y_3 = [1.30, 1.60]],$$

nous trouverons alors

$$| s |_\Omega = \{\omega_4, \omega_5\} \Longrightarrow s' = \{\omega_4^s, \omega_5^s\}.$$

s' est constitué des objets symboliques

$$\omega_4^s = [y_2 = suisse] \wedge [y_3 = 1.30] \text{ et}$$
$$\omega_4^s = [y_2 = espagnol] \wedge [y_3 = 1.58].$$

Dans le paragraphe qui suit, nous nous baserons sur cette définition pour construire un préordre partiel sur S.

1.6.3 Ordre, Héritage, Généralisation des objets symboliques

1.6.3.1 Rappel

Nous supposons qu'en règle générale un objet symbolique est un objet assertion (de toute façon en vertu du théorème ci-dessus, on peut facilement passer d'un type à un autre).

1.6.3.2 Ordre symbolique

$$\forall s_1, s_2 \in S \quad s_1 \leq s_2 \iff s_1' \subseteq s_2'$$

avec $s' = \{\varphi(\omega) \in S$ tels que $\omega(\in \Omega) \in | s |_\Omega\}$ est l'extension symbolique de $s \in S$

où φ est l'application $\varphi : \Omega \to S$

$$\omega \to \varphi(\omega) = [y_1 = y_1(\omega)] \wedge \ldots \wedge [y_p = y_p(\omega)]$$

et S est l'ensemble des objets symboliques définis sur Ω et caractérisés par les variables $y_i : \Omega \to O_i$.

Nous dirons alors que s_2 est plus *général* que s_1
ou de manière équivalente que,

s_1 *hérite* des propriétés de s_2.

Il s'agit d'une *relation de préordre partiel* entre objets symboliques :

- *Préordre* : car la propriété d'antisymétrie (deux objets ayant une extension identique ne sont pas toujours égaux) n'est pas vérifiée;
- *Partiel* : car l'inclusion est une relation partielle. Nous définissons notre préordre au moyen d'une inclusion ensembliste.

1.6.3.3 Définition de l'union et de l'intersection symbolique

Union symbolique *L'union symbolique $s_1 \cup s_2$ est la conjonction de tous les objets symboliques de S dont l'extension symbolique contient l'ensemble des objets de s'_1 et s'_2.*

Exemple Soit les données suivantes :

	y_1	y_2	y_3
ω_1	[1, 15]	*espagnol*	1.65
ω_2	]15, 30]	*belge*	1.80
ω_3	]30, 45]	*belge*	1.29
ω_4	]45, 60]	*suisse*	1.30
ω_5	]60, 75]	*espagnol*	1.58

Prenons

$s_1 = [y_1 =]1, 15] \cup]45, 60]] \wedge [y_2 = \{espagnol, suisse\}] \wedge [y_3 = [1.30, 1.65]]$

et

$s_2 = [y_1 =]15, 45]] \wedge [y_2 = \{belge, espagnol\}] \wedge [y_3 = \{[1.29, 1.65], 1.80\{]$.

Nous trouvons alors que

$$| s_1 |_\Omega = \{\omega_1, \omega_4\} \Longrightarrow s'_1 = \{\omega_1^s, \omega_4^s\}.$$

et

$$| s_2 |_\Omega = \{\omega_1, \omega_2, \omega_3\} \Longrightarrow s'_2 = \{\omega_1^s, \omega_2^s, \omega_3^s\}.$$

Ainsi

$$s'_1 \cup s'_2 = \{\omega_1^s, \omega_2^s, \omega_3^s, \omega_4^s\}.$$

L'union symbolique de s_1 et s_2 sera

$$s_1 \cup s_2 = \{t_1, t_2, t_3, \ldots\}$$

avec

$t_1 = [y1 = [1, 60]]$
$t_2 = [y_3 = \{[1.25, 1.30], [1.65, 1.80]\}]$;
$t_3 = [y1 = [1, 60]] \wedge [y_2 = \{belge, espagnol, suisse\}]$.

Intersection symbolique *L'intersection symbolique* $s_1 \cap s_2$ *est la conjonction de tous les objets symboliques de* S *dont l'extension symbolique contient les objets communs à* s'_1 *et* s'_2.

Exemple Nous reprenons les données du dernier exemple, nous obtenons alors :

$$s'_1 \cap s'_2 = \{\omega_1^s\}.$$

L'intersection symbolique de s_1 et s_2 sera

$$s_1 \cap s_2 = \{r_1, r_2, r_3, \ldots\}$$

avec

$r_1 = [y1 =]1, 15]]$;
$r_2 = [y_3 = [1.65]]$;
$r_3 =]y_2 = \{espagnol\}] \wedge [y_3 = [1.60, 1.65]]$.

1.6.3.4 Conventions

Nous observerons deux conventions :

1. $s \wedge [y_i = O_i] = s$
   et
   $[y_1 = O_1] \wedge \ldots \wedge [y_p = O_p] = \Omega^s$ où Ω_s est l'objet symbolique plein ;

2. $[y_i = v_i] \wedge [y_i = V_i] = \begin{cases} [y_i = v_i] & si \quad v_i \subset V_i \\ y_i = \emptyset & si \quad v_i \cap V_i = \emptyset \end{cases}$
   et
   $[y_i = \emptyset] \wedge \ldots \wedge [y_p = \emptyset] = \emptyset^s.$

Nous constatons donc que

$s'_1 \cap s'_2 = \emptyset \iff | s_1 |_\Omega \cap | s_2 |_\Omega = \emptyset$
$\iff s_1 \cap s_2$ est la conjonction de tous les évènements d'extension vide dans Ω.

De même, nous avons que

$s'_1 \cup s'_2 = \Omega \iff | s_1 |_\Omega \cup | s_2 |_\Omega = \Omega$
$\iff s_1 \cup s_2$ est la conjonction de tous les évènements élémentaires dont la conjonction est Ω
$\iff s_1 \cup s_2 = [y_1 = O_1] \wedge \ldots \wedge [y_p = O_p] = \Omega^s.$

Si nous prenons (comme nous le ferons à partir de maintenant) des variables $y_i : \Omega \to O_i$ surjectives (au besoin en donnant des valeurs aux données manquantes et à celles qui n'existent pas), alors $s_1 \cup s_2$ sera l'objet symbolique plein. Nous résumons ce qui vient d'être dit par la proposition qui suit.

1.6.3.5 Proposition 2

1. *L'union (resp. l'intersection) de 2 objets symboliques* s_1 *et* s_2 *est la conjonction de tous les évènements élémentaires de plus petite extension dont l'extension contient* $s'_1 \cup s'_2$ *(resp.* $s'_1 \cap s'_2$*)*

2. *L'union (resp. l'intersection) de 2 objets symboliques* s_1 *et* s_2 *est l'objet symbolique plein (resp. vide)*

$$\Longleftrightarrow$$

$$s_1' \cup s_2' = \varphi(\Omega) \quad (resp.\ s_1' \cap s_2' = \emptyset).$$

1.6.3.6 Exemple

Soit le tableau de données suivant :

	y_1	y_2	y_3	y_4
ω_1	1	0	0	0
ω_2	1	0	1	0
ω_3	1	1	0	1
ω_4	1	1	1	1
ω_5	0	1	0	0

Les variables y_i sont des applications

$$y_i : \Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\} \rightarrow O_i = \{0, 1\}.$$

Nous avons que $\varphi(\omega_1) = \omega_1^s = [y_1 = 1] \wedge [y_2 = 0] \wedge [y_3 = 0] \wedge [y_4 = 0]$
$\varphi(\omega_2) = \omega_2^s = [y_1 = 1] \wedge [y_2 = 0] \wedge [y_3 = 1] \wedge [y_4 = 1]$

De plus, par la définition de s' et celle de $\mid s \mid_\Omega$, nous trouvons que

$$(\omega_1^s)' \cup (\omega_2)' = \{\omega_1^s, \omega_2^s\}$$

ce qui entraine, en tenant compte du point 1 de la proposition 2, l'égalité suivante :

$$\omega_1^s \cup \omega_2^s = [y_1 = 1] \wedge [y_2 = 0] \wedge [y_4 = 0].$$

Et, par application du point 2 de cette même proposition, nous avons les implications ci-dessous :

$$[y_1 = 1]' \cup [y_2 = 1]' = \varphi(\Omega) \Rightarrow [y_1 = 1] \cup [y_2 = 2] = \Omega^s ;$$
$$[y_1 = 0]' \cap [y_4 = 1]' = \emptyset \Rightarrow [y_2 = 0] \cap [y_4 = 1] = \emptyset^s.$$

1.6.3.7 Proposition 3

1. *L'union et l'intersection d'objets symboliques sont associatives, commutatives et existent toujours;*

2. *Si* $s = s_1 \cup s_2$ *(resp.* $s_1 \cap s_2$ *)*
   on a : $\underbrace{s' \supseteq s'_1 \cup s'_2}_{(1)}$ *(resp.* $\underbrace{s' \equiv s'_1 \cap s'_2}_{(2)}$*)*

   - (1) l'union booléenne est généralisante;
   - (2) signifie $(s_1 \cap s_2)' = s'_1 \cap s'_2$.

Conséquence : en définissant l'intersection symbolique de s_1 et s_2 (notée $(s_1 \cap s_2)'$) comme la disjonction de tous les objets symboliques dont l'extension symbolique est contenue dans

$$s'_1 \cap s'_2 \, (i.e \, (s_1 \cap s_2)' = \{\omega^s \in S \ tq \ (\omega^s)' \subset (s'_1 \cap s'_2)\} \, ;$$

nous remarquons assez facilement qu'elle possède la même extension que $s_1 \cap s_2$ car parmi les éléments de cette disjonction celui qui a la plus grande extension est $s_1 \cap s_2$. Ce dernier résultat vient de la proposition 3 qui affirme que

$$(s_1 \cup s_2)' = s'_1 \cap s'_2.$$

Ceci peut encore s'exprimer comme suit : puisque

$$s'_1 = \{\omega^s \in S \, \text{tq} \, \omega \in | \, s_1 \, |_\Omega\} \ et \ s'_2 = \{\omega^s \in S \ \text{tq} \ \omega \in | \, s_2 \, |_\Omega\} \, ,$$

nous aurons que

$$s'_1 \cap s'_2 = \{\omega^s \in S \, \text{tq} \, \omega \in (| \, s_1 \, |_\Omega \cap | \, s_2 \, |_\Omega)\} \, .$$

Et, par la proposition 3, cette dernière quantité est égale à :

$$(s_1 \cap s_2)' = \{\omega^s \in S \, \text{tq} \, \omega \in | \, s_1 \cap s_2 \, |_\Omega\}$$

1.6.3.8 Rappel

Un *treillis* est un ensemble muni d'une relation pour laquelle tout couple d'évènements admet une borne supérieure et une borne inférieure.

1.6.3.9 Proposition 4

Muni de l'ordre symbolique, l'ensemble S des objets symboliques est un treillis et la borne supérieure de tout couple d'éléments est leur union symbolique; la borne inférieure de tout couple d'éléments est leur intersection symbolique.

1.7 Qualités des objets

1.7.1 Complétude

1.7.1.1 Principe

Si je vois, dans ma rue, un groupe de chats noirs et que je dis: 'je vois des chats. Cette assertion ne fournit pas une description complète de mon observation puisque j'ai omis de dire qu'ils sont de couleur noire.
C'est cette idée que veut traduire la notion de complétude d'un objet symbolique. Formellement, cela revient à mesurer l'écart entre l'ensemble des évènements élémentaires dont la conjonction définit un objet symbolique s et l'ensemble de tous les évènements élémentaires dont l'extension contient s'.

- $d(s)$ ***Définition*** est l'ensemble des événements élémentaires dont la conjonction définit s;

- $c(s)$ ***Complet*** est l'ensemble des événements de plus petite extension contenant s';

- s^c est la conjonction de tous les éléments de $c(s)$.

1.7.1.2 Exemple

Nous reprenons le tableau de données du paragraphe précédent concernant les données booléennes.

	y_1	y_2	y_3	y_4
ω_1	1	0	0	0
ω_2	1	0	1	0
ω_3	1	1	0	1
ω_4	1	1	1	1
ω_5	0	1	0	0

Nous considérons l'objet symbolique

$$s = [y_1 = 1] \wedge [y_4 = 1],$$

nous aurons :

$$d(s) = \{[y_1 = 1], [y_4 = 1]\}$$

$$c(s) = \{\overbrace{[y_1 = 1]}^{e_1}, \overbrace{[y_2 = 1]}^{e_2}, \overbrace{[y_4 = 1]}^{e_3}\}$$

et par conjonction des éléments de $c(s)$, on obtient :

$$s^c = [y_1 = 1] \wedge [y_2 = 1] \wedge [y_4 = 1].$$

En effet, nous savons que :

$$| s |_\Omega = \{\omega_3, \omega_4\}$$

$$| e_1 |_\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\} \; ; \; | e_2 |_\Omega = \{\omega_3, \omega_4, \omega_5\} \; ; \; | e_3 |_\Omega = \{\omega_3, \omega_4\}$$

1.7.1.3 Critère

La complétude d'un objet symbolique se calcule à l'aide de critères.

En voici deux, mais il en existe d'autres :

- $c_1 = \#(c(s) \setminus d(s))$;
- $c_2 = \#(\{(e_i)' \; tq \; e_i \in [c(s) \setminus d(s))]\})$.

1.7.1.4 Exemple

Nous reprenons l'objet s défini dans l'exemple qui précède.
Nous calculons que

$$\begin{aligned} c(s) \setminus d(s) &= \{[y_1 = 1], [y_2 = 2], [y_4 = 1]\} \setminus \{[y_1 = 1], [y_4 = 1]\} \\ &= \{[y_2 = 1]\} \end{aligned}$$

ainsi
$c_1 = 1$: car il ne reste qu'un seul événement dans la différence : $e_2 = [y_2 = 1]$
$c_2 = 2$: car l'extension de e_2 contient deux éléments.

1.7.1.5 Définition

Nous dirons que :

$$s \text{ est complet} \iff c_1(s) = 0 \iff c(s) = d(s)$$

Cette définition se base sur le critère 1.

1.7.1.6 Propriétés des objets symboliques.

Proposition 1

1. $d(s_c) = c(s)$ *et* $(s^c)')$
2. s^c *est complet* $(i.e\ d(s^c) = c(s^c))$
3. $s^c = \{ \cup \varphi(\omega_i) \ \ tq \ \ \omega_i \in \mid s \mid_\Omega \}$
4. $s^c = \{ \cap e_i \ \ tq \ \ e_i \in c_1(s)\}$

où s^c est l'objet défini par la conjonction de tous les éléments de $c(s)$ [2].

2. Ici s est quelconque.

Proposition 2

1. *Si un objet symbolique s est l'union ou l'intersection d'objets symboliques alors s est complet.*

   Ce qui a pour conséquence que :

2. *l'ensemble des objets symboliques complets muni de l'ordre symbolique est un treillis.*

1.7.2 Affinement d'un objet symbolique

1.7.2.1 Principe

Un objet symbolique est d'autant plus affiné que les événements élémentaires qui le définissent ont une extension proche de l'extension de s.

1.7.2.2 Critère

Critère (mesurant l'affinement) $A(s) = \dfrac{\#(\mid \cup e'_i(s) \setminus \cap e'_i(s) \mid_\Omega)}{\#(\cup e'_i(s))}$

où les e_i sont les évènements élémentaires qui définissent s.

Un objet sera dit affiné si $A(s)=0$ $(i.e.\ si\ \cup e'_i(s) = \cap e'_i(s))$

1.7.2.3 Exemple

Nous reprenons à nouveau le tableau :

	y_1	y_2	y_3	y_4
ω_1	1	0	0	0
ω_2	1	0	1	0
ω_3	1	1	0	1
ω_4	1	1	1	1
ω_5	0	1	0	0

Nous trouvons que l'affinement de l'OAB

$$s = [y_1 = 1] \wedge [y_2 = 1]$$

vaut :

$$A(s) = \frac{1}{5}\#(\Omega \setminus \{\omega_3 , \omega_4\}) = \frac{3}{5} \star$$

En effet,

$$\Omega = \{\omega_1 , \omega_2 , \omega_3 , \omega_4 , \omega_5\}$$

et

$$\begin{aligned} | \cup e'_i(s) |_\Omega &= | e'_1(s) \cup e'_2(s) |_\Omega \\ &= | [y_1 = 1]' \wedge [y_2 = 1]' |_\Omega \\ &= \{\omega_1 , \omega_2 , \omega_3 , \omega_4\} \cup \{\omega_3 , \omega_4\} \\ &= \{\omega_1 , \omega_2 , \omega_3 , \omega_4 , \omega_5\} \\ &= \Omega \end{aligned}$$

$$\text{et } \#(| \cup e'_i(s) |_\Omega) = 5$$

$$\begin{aligned} | \cap e'_i(s) |_\Omega &= \{\omega_1 , \omega_2 , \omega_3 , \omega_4\} \cap \{\omega_3 , \omega_4 , \omega_5\} \\ &= \{\omega_3 , \omega_4\} \end{aligned}$$

$$\text{et } \#(| \cap e'_i(s) |_\Omega) = 2$$

Si nous reprenons la définition du critère d'affinement et que nous y remplaçonsn les termes $\#(| \cap e'_i(s) |_\Omega)$ et $\#(| \cup e'_i(s) |_\Omega)$ par les valeurs trouvées ci-dessus, nous obtenons bien la formule $\star$.

1.7.3 Simplicité d'un objet symbolique

1.7.3.1 Principe

Un objet symbolique est d'autant plus simple que l'ensemble des événements élémentaires qui le décrivent est proche d'un ensemble d'événements élémentaires de cardinal minimun et dont la conjonction s_p a la même extension .
Exprimé autrement : plus le nombre d'événements élémentaires décrivant s est proche du nombre d'élément de s_p , plus l'objet symbolique s sera simple ; s_p est défini par le nombre minimal d'événements élémentaires tel que l'extension de s_p

soit identique à celle de s.
Il n'y a pas unicité de s_p (plusieurs ensembles dévénements élémentaires conviennent).
Nous mesurons la simplicité de s au moyen de critères.

En voici un noté $S(s)$:

$$S(s) = \#[d(s)] - \#[d(s_p)]$$

1.7.3.2 Remarque

Il est possible de regrouper simplicité et complétude d'un objet symbolique, nous parlerons alors de redondance dont voici un critère :

$$|\underbrace{\#(s_p)}_{(1)} - \underbrace{\#(s^c)}_{(2)}|$$

(1) traduit la simplicité (ce qui semble assez intuitif);
(2) représente la complétude.

Un objet à la fois complet et simple a une redondance nulle.
(C'est le cas d'un objet élémentaire $\varphi(\omega) = \omega^s$.)

1.8 Qualités des classes d'objets symboliques

1.8.1 Principe

Comme dans le cas des objets symboliques, si nous désirons étudier les qualités d'une classe, il faut définir ce que sont l'ordre, l'union et l'intersection entre les classes.
Une manière simple de procéder serait d'associer à chaque classe un objet symbolique la représentant (en prenant par exemple l'union ou l'intersection des objets de la classe) puis à appliquer sur ces objets les différentes notions, propriétés et qualités établies pour les objets symboliques.
Nous dirons alors qu'une classe est complète, affinée ou simple si son représentant l'est.

1.8.2 Extension

L'*extension d'une classe* peut se définir comme l'union (ou l'intersection) des objets qui la composent.

1.8.2.1 Ordre

Partant de cette première définition, nous dirons que l' ***ordre symbolique interclasse*** *est déterminé au moyen de l'une des deux expressions suivantes :*

1. *La classe 1 est plus petite ou égale à la classe 2 pour l'odre symbolique interclasse si l'extension de la classe 1 est incluse dans l'extension de la classe 2 ;*

2. *La classe 1 est plus petite ou égale à la classe 2 pour l'ordre symbolique interclasse si le plus grand élément de la classe 1 est inférieur ou égal au plus grand élément de la classe 2.*

1.8.2.2 Propriétés spécifiques aux classes

Il existe encore des propriétés spécifiques à la notion de classe. Citons, par exemple, l'effritement et la stabilité.

1.8.2.3 Stabilité d'une classe

Nous la définissons comme la capacité d'une classe à être représentée par l'objet symbolique d'extension minimale et qui contient l'union des extensions des objets de la classe.
Nous exprimerons le degré de stabilité au moyen d'un critère :

$$s_t(C) \;=\; \#(\,|\cup c_i\,|_\Omega \;-\cup|\,c_i\,|_\Omega\,)$$

où les c_i sont les éléments de la classe C; ce sont donc des objets symboliques.

Ce critère a toujours un sens (i.e $s_t \geq 0$). En effet, d'une part, nous savons que l'union de deux objets symboliques contient l'union de leur extension (voir proposition 3 de 1.6) et d'autre part,

en utilisant l'associativité de l'union symbolique, nous pouvons généraliser celle-ci à plusieurs objets en disant que :

$$\cup \mid c_i \mid_\Omega \subseteq \mid \cup c_i \mid_\Omega .$$

Définition *Nous dirons qu'une classe C est* ***stable*** *si et seulement si* $s_t(C) = 0$.

En raison du caractère bijectif de φ , il est encore équivalent d'exiger que :

$$\varphi(\mid \cup c_i \mid_\Omega) = \varphi(\cup \mid c_i \mid_\Omega)$$

et donc que

$$\underbrace{(\cup c_i)'}_{(1)} = \cup(\varphi(\mid c_i \mid_\Omega)) = \underbrace{\cup c_i'}_{(2)}$$

Il est bon de constater que

(1) est une union symbolique et

(2) est une union ensembliste.

On notera C_t l'ensemble des classes stables constituées d'éléments de $\varphi(\Omega)$ et telles que $\forall \omega^s \in C \in C_t$, on ait $(\omega^s)' \subseteq C$ (voir exemple 2).

Exemple 1 Nous nous servons du tableau

	y_1	y_2	y_3	y_4
ω_1	1	0	0	0
ω_2	1	0	1	0
ω_3	1	1	0	1
ω_4	1	1	1	1
ω_5	0	1	0	0

ainsi que de la classe d'objets C telle que

$$C = \{\underbrace{\omega_1^s}_{c_1}, \underbrace{[y_1 = 1] \wedge [y_3 = 1]}_{c_3}, \underbrace{[y_4 = 1]}_{c_4}\}$$

pour laquelle, nous avons que

$$\begin{aligned} \cup c_i = [y_1 = 1] \text{ puisque } \cup \mid c_i \mid_\Omega &= \{\omega_1\} \cup \{\omega_2 , \omega_4\} \cup \{\omega_3 , \omega_4\} \\ &= \{\omega_1 , \omega_2 , \omega_3 , \omega_4\} \end{aligned}$$

$$\begin{aligned} \mid \cup c_i \mid_\Omega &= \mid \omega_2^s \cup [y_1 = 1] \wedge [y_3 = 1] \cup [y_4 = 1] \mid_\Omega \\ &= \{\omega_1 , \omega_2 , \omega_3 , \omega_4\} \end{aligned}$$

ce qui entraine que

$$\begin{aligned} s_t(C) &= \#(\{\omega_1 , \omega_2 , \omega_3 , \omega_4\} \setminus \{\omega_1 , \omega_2 , \omega_3 , \omega_4\}) \\ &= 1. \end{aligned}$$

d'où nous conlurons que C n'est pas stable.

Exemple 2

Sur base des mêmes données, nous saurons que les classes

$$C_1 = \{\omega_3^s , \Omega_4^s\} \text{ et } C_2 = \{\omega_4^s , \omega_2^s\}$$

sont stables car constituées d'objets symboliques élémentaires.

Proposition 8

Il existe une bijection entre l'ensemble des classes stables C_t et l'ensemble des objets complets.

1.8.2.3 Effritement d'une classe

Nous définissons cette qualité comme le plus petit nombre d'objets symboliques $a \in A \subseteq S$ dont la réunion des extensions est contenue dans l'extension

de l'union des éléments d'une classe (tout en s'en écartant le moins possible).

Etant données : une classe $C \subseteq S$ d'éléments génériques c_i ;
une classe A d'éléments génériques a_i ;
nous utiliserons le critère suivant :

$$E_1(C) = min\{[1+\#(\cup \mid c_i \mid_\Omega \setminus \cup \mid a_i \mid_\Omega)]\#(A) \quad tq \quad A \subset S : \mid a_i \mid_\Omega \subseteq \mid \cup c_i \mid_\Omega\}$$

Si nous imposons à C de vérifier que $\cup \mid c_i \mid_\Omega = \cup \mid a_i \mid_\Omega$, nous obtenons le critère E_2 donné par :

$$E_2(C) = \{\#(A) \ tq \cup \mid a_i \mid_\Omega = \cup \mid c_i \mid_\Omega\}$$

- E_1 *et* E_2 sont minimales s'il existe un objet symbolique A tq $A = \cup c_i$
  A est alors complet et $E_1 = E_2 = 1$;

- E_2 est maximal lorsque $A \equiv C$ et vaut alors #(C).

Exemple Nous utilisons comme données le tableau

	y_1	y_2	y_3	y_4
ω_1	1	0	0	0
ω_2	1	0	1	0
ω_3	1	1	0	1
ω_4	1	1	1	1
ω_5	0	1	0	0

Soit la classe d'objets symboliques

$$C = \{\underbrace{\omega_2^s}_{c_1}, \underbrace{\omega_3^s}_{c_2}, \underbrace{[y_2 = 1] \wedge [y_3 = 1]}_{c_3}\}.$$

L'effritement minimun est donné par

$$A = \{[y_4 = 1]; \omega_2^s\},$$

nous avons

$$\#(A) = 2.$$

Puisque

$$\cup \mid a_i \mid = \{\omega_2 , \omega_3 , \omega_4\}$$

et

$$\cup \mid c_i \mid = \{\omega_2 , \omega_3 , \omega_4\}$$

nous avons donc que $\#(\cup \mid a_i \mid_\Omega) = \#(\cup \mid c_i \mid_\Omega)$, d'où nous pouvons conclure que

$$E_1(C) = E_2(C) = 2 \ \star\star$$

Si $A = \{\omega_2^s\}$, nous avons :

$$(1 + \#(\cup \mid c_i \mid_\Omega - \cup \mid a_1 \mid_\Omega))\#(A) = (1 + 2)1 = 3.$$

La meilleure solution $\star\star$ n'est pas unique, nous aurions pu encore prendre

$$A = \{[y_3 = 1], \omega_3^s\} \Rightarrow E_1(C) = E_2(C) = 2$$

car nous savons ici aussi que

$$\cup \mid a_i \mid_\Omega = \{\omega_2 , \omega_4\} \cup \{\omega_3\} = \{\omega_2 , \omega_3 , \omega_4\} = \cup \mid c_i \mid_\Omega .$$

1.9 Qualités des classifications

1.9.1 Définition

La classification d'une classe $C \subseteq S$ est définie comme étant un ensemble de classes qui recouvrent cet ensemble C d'objets symboliques.

Un recouvrement sera, par exemple, une pyramide, une hiérarchie ou encore une partition.

1.9.2 Propriétés

Une classification pourra être complète, affinée, simple mais aussi avoir une bonne stabilité et un bon effritement suivant que ses classes (ou leur représentant) ont ces qualités.

1.9.3 Qualités propres aux classifications

En outre, il existe des qualités propres aux classifications :

- le *degré de recouvrement* des extensions des classes de la classification ;
- la *qualité* de l'héritage des classes entre elles.

Chapitre 2

Les objets symboliques modaux

Introduction

2.0.4 Présentation générale

L'information véhiculée par les objets symboliques booléens est assez restreinte : ils ne prennent que deux valeurs ($Vrai$ ou $Faux$). La représentation booléenne est insuffisante pour une description acceptable de la réalité multidimensionnelle. En effet, dans la pratique, on est souvent amené à faire intervenir des *jugements* ou *modes* portant sur les évènements élémentaires et servant à traduire la connaissance dont on dispose sur le modèle (on dit aussi sémantique).

En voici quelques exemples :

- la sémantique alétique : *il se peut ; il est certain*;
- la sémantique déontique : *il est permis ; il est interdit.*

Dans ces deux exemples, nous n'avons considéré que deux modes opposés, il existe cependant plusieurs variantes. Pour celles-ci, les modes ne sont pas toujours contradictoires. Ainsi, dans le cas de la sémantique de la certitude, on peut avoir les modes suivants :

m_1 = *il est absolument vrai*;
m_2 = *il est vrai sous conditions*;
m_3 = *il est faux sous conditions*;
m_4 = *il est absolument faux*,

qui permettent d'émettre des jugements plus nuancés.

Il serait intéressant de diversifier les valeurs prises par nos objets symboliques. On introduit la notion d'*objet modal.* Ce type d'objet permettra de moduler les valeurs prises par les descripteurs (exemple: un chat est *rarement* blanc et *il se peut* qu'un chien soit roux) et ainsi de tenir compte des connaissances du domaine étudié. Il sera possible, en les utilisant, d'exprimer des notions telles que la variation, le doute, le vague, l'incertain...

2.0.5 Description des deux sortes d'objets modaux.

Il existe deux sortes d'objets symboliques modaux qui ne sont pas logiquement équivalentes: on ne passe pas de l'une à l'autre sans en modifier la signification. Elles se différencient comme suit:
les modes portent soit sur l'évènement élémentaire e_i tout entier, soit simplement sur l'ensemble de valeurs V_i de la variable y_i qui décrit e_i.
Dans le premier cas, nous parlerons *d'objet modal de l'extérieur* pour désigner un OAB qui est la conjonction d'évènements sur lesquels le mode porte entièrement et nous le noterons:

$$a_x = \wedge_i M_i[y_i = V_i]$$

2.1.2.1 Exemple

Il est possible qu'une tortue aille vite et *il arrive parfois* qu'un lièvre soit lent.

Dans le second cas, nous parlerons *d'objet modal de l'intérieur* afin de désigner un OAB qui est la conjonction d'événements élémentaires pour lesquels le mode porte sur l'ensemble de valeurs du descripteur. Nous noterons cet objet:

$$a_x = \wedge_i [y_i = M_i V_i].$$

2.1.2.2 Exemple

Une souris est *rarement* de couleur verte et *souvent* de petite taille.

2.1.2.3 Remarque

1. L'indice x désigne la sémantique avec laquelle on travaille. Cette sémantique est choisie en fonction de la connaissance du domaine étudié : on prend celle qui semble le mieux convenir aux types de données dont on dispose.

2. Un objet modal prend ses valeurs dans un ensemble ordonné, le plus souvent [0,1]. En d'autres termes, un objet modal est une application $a_x(\omega) \in [0\,,\,1]$ qui exprimera un degré de certitude compris entre 0 et 1. Ce degré qualifiera le niveau d'appartenance de ω à la classe décrite par a. Nous examinerons uniquement les objets modaux de l'intérieur. Pour les mêmes raisons que dans le chapitre précédent, nous choisirons de ne travailler qu'avec les assertions. On emploiera donc indifféremment les termes assertion et objet.

2.1 Objet modal de l'intérieur

2.1.1 Description au moyen d'un exemple

Supposons que l'on veuille utiliser un objet symbolique afin de représenter un ensemble d'individus (objets élémentaires) satisfaisant à la proposition suivante : *leur poids est probablement compris entre 5 et 10 kg et leur couleur est souvent brune et rarement beige.* Cette affirmation contient deux évènements : e_1 = [poids = [5, 10]] et e_2 = [couleur = {*brune*, *beige*}] auxquels manquent les modes *probablement, souvent et rarement.* Nous introduisons alors un type d'évènement f_i tel que :

$$f_1 = [\text{poids} = \textit{probablement}\,[5,\,10]];$$
$$f_2 = [\text{couleur} = \{\textit{souvent}\ \text{brune}, \textit{rarement}\ \text{beige}\}].$$

On voit que f_1 et f_2 contiennent des modes *internes* affectant les valeurs prises par e_1 et e_2 respectivement.
On peut donc à présent décrire notre proposition de manière informelle au moyen de l'assertion modale notée $a = f_1 \wedge_x f_2$ où $\wedge_x$ désigne une sorte de conjonction définie en fonction de la connaissance du domaine.
Nous définirons aussi une série d'opérateurs $\wedge_x$, $\cup_x$... adaptés aux diverses sémantiques et qui permettront de généraliser, spécialiser, mesurer ... les assertions.

En fait, nous calculerons la valeur $a_x(\omega) = \wedge_{i,x}[y_i = q_i] : P(\Omega)^1 \rightarrow [0, 1]$ (qui est le degré d'appartenance de ω à la classe d'individus décrite par a)
en

1. récrivant ω sous la forme d'un objet symbolique $\omega^s = \wedge_i[y_i = r_i]$;

   Exemple :
   Si ω est un chien *peut-être* vieux et *parfois* sale, on écrira
   $\omega^s =$ [âge $= r_1$] $\wedge$ [propreté $= r_2$] avec
   r_1(vieux) $=$ *peut-être* et
   r_2(sale) $=$ *parfois*.

2. comparant les q_i et r_i à l'aide d'une fonction de comparaison g_x;

   Exemple :
   Soit $a =$ [âge $= q_1$] $\wedge$ [propreté $= q_2$] avec
   q_1(vieux) $=$ *pas du tout* et
   q_2(sale) $=$ *de temps en temps*.
   Une fonction de comparaison g_x donnera, par exemple, comme résultats :
   g_x(*peut-être* vieux, *pas du tout* vieux) $=$ *correspond pas*;
   g_x(*de temps en temps* sale, *parfois* sale) $=$ *correspond*.

3. faisant le bilan des résultats de g_x sur tous les i en se servant d'une fonction d'agrégation f_x;

   Exemple :

   $$f_x(correspond,\ correspond\ pas) = correspond \wedge_x correspond\, pas$$
   $$= correspond\ pas.$$

1. $P(\omega)$ désigne l'ensemble des parties de Ω. Ceci signifie que a décrit bien une classe d'éléments de Ω.

2.1.2 définition d'un objet modal de l'intérieur (objet mi).

Soient :

- M^x un ensemble de noms ou de nombres traduisant les modes associés à une sémantique (connaissance du domaine) notée x;

  Exemples :
  $M_x = [0, 1]$,
  $M_x = \{jamais, rarement, parfois, souvent, toujours\}$.

- $Q_i = \{q_i^j\}_j$ [2] l'ensemble des applications $q_i^j : O_i \rightarrow M^x$;

  Exemple :
  q_i^j est une mesure de probabilité.

- OP_x qui désigne les trois opérations 'ensemblistes' définies dans Q_i : $\cup_x$, $\cap_x$, c_x qui sont les opérations d'union, d'intersection et de complémentation exprimant la sémantique (on note parfois $c_x(q_i)$, $\bar{q}_i$);

  Exemples :
  $q_i^1 \cup_p q_i^2 = max(q_i^1, q_i^2)$;
  $q_i^1 \cap_p q_i^2 = min(q_i^1, q_i^2)$;
  $q_i^1 \cup_{pr} q_i^2 = q_i^1 + q_i^2 - q_i^1 q_i^2$ où $q_i^1 q_i^2(v) = q_i^1(v) q_i^2(v)$;
  $q_i^1 \cap_{pr} q_i^2 = min(q_i^1, q_i^2)$.

---

2. **Remarque:** L'indice j de q_i^j est utilisé lorsque plusieurs modes portent sur une même valeur prise par la variable.
Par exemple, a = [poids = *parfois* lourd, *souvent* lourd], on aura que q_1^1(lourd) = *parfois* et q_1^2(lourd) = *souvent*.
Nous omettrons d'écrire le j lorsqu'il n'existera qu'un seul mode portant sur chaque valeur prise par le descripteur.

- g_x une application dite de comparaison :

$$g_x : Q_i \times Q_i \rightarrow L^x$$

où L^x est l'espace d'interprétation, il est ordonné et parfois identique à M^x.

Exemples :
$g_x(q_i^1, q_i^2) = sup\{min(q_i^1(v), q_i^2(v)) \text{ tq } v \in O_i\} \in M^x$;
$g_x(q_i^1, q_i^2) =< q_i^1, q_i^2 > \in [0, 1]$ si $<,>$ est le produit scalaire classique et $M^x = [0, 1]$.

- f_x est une application symétrique dite d'agrégation :

$$f_x : \underbrace{P(L^x)}_{\text{parties de } L^x} \rightarrow L^x$$

Exemple :
$\forall L \subset L^x \quad f(L) = min\{L_i \text{ tq } L_i \in L\}$.

- On note par $Y = \{y_i\}$ l' ensemble des descripteurs et
$V = \{V_i\} = \{(q_i^j)\}_j \subseteq Q_i$ un ensemble de parties V_i de Q_i.

2.2.2.1 Définition

Etant donnés OP_x, g_x, f_x, un objet mi est une application $a_{YV} : \Omega \rightarrow L^x$ notée $a_{YV} = \wedge_{i,x}[y_i = \{q_i^j\}_j]$ tel que si $\omega \in \Omega$ est décrit pour chaque i par $y_i(\omega) = \{r_i^j\}$ alors,

$$a_{YV}(\omega) = f_x(\{g_x(\cup_{j,x} q_i^j, \cup_{j,x} r_i^j)\}_i).$$

2.2.2.2 Exemple

Dans une anonce d'offre d'emploi, une entreprise indique qu'elle recherche une personne *parfaitement* bilingue et de caractère *très* aimable.

$$a = [\text{langue} = q_1] \wedge [\text{caractère} = q_2].$$

avec

$q_1(\text{bilingue}) = parfaitement;$
$q_2(\text{aimable}) = très.$

Une jeune femme se présente, elle est *pratiquement* bilingue et de caractère *fort peu* aimable.

$$\omega^s = [\text{langue} = r_1] \wedge [\text{caractère} = r_2].$$

avec

$r_1(\text{bilingue}) = pratiquement;$
$r_2(\text{aimable}) = fort\ peu.$

La question est de savoir si elle convient ou non pour cet emploi.
Il nous faut donc calculer $a(\omega)$. Si on suppose g_x et f_x connus (définis à l'aide de tables), on pourrait trouver par exemple :

$g_x(parfaitement \text{ bilingue}, pratiquement \text{ bilingue}) = convient$
$g_x(très \text{ aimable}, fort\ peu \text{ aimable}) = ne\ convient\ pas.$

$f_x(convient, convient\ pas) = convient\ pas.$

Ce qui signifierait que cette demoiselle n'est pas la personne adéquate pour ce poste suivant les critères de sélection employés.

2.2.2.3 Conventions

Nous noterons par A_x l'ensemble des objets *mi* associés à la sémantique x et par[3]

$$\varphi(\omega) = \omega^s = \wedge_{i,x}[y_i = y_i(\omega)].$$

3. **Rappel :** la définition de φ se trouve dans le chap1 1.6.2.2.

Nous pouvons induire sur A_x les opérations OP_x définies sur les Q_i en posant :

$$a_1 \star_A a_2 = \wedge_{i,x}[y_i = q_i \star_x q_j] \quad \forall \star_x (\star_A) \in OP_x(OP_A).$$

De même, on peut associer à Ω des opérations OP_Ω grâce à φ^{-1} en posant :

$$\varphi^{-1}(\omega_1^s \star_A \omega_2^s) = \varphi^{-1}(\omega_1^s) \star_\Omega \varphi^{-1}(\omega_2^s) \quad \forall \star_A (\star_\Omega) \in OP_A(OP_\Omega).$$

Le choix des opérateurs ensemblistes, d'une part, et des fonctions de comparaison et d'agrégation, d'autre part, nécessite une certaine cohérence. Certaines sémantiques particulières ont été étudiées en détail; c'est sur elles que l'on peut se baser pour faire ces choix. Nous en examinerons quelques-unes dans la section suivante.

2.1.3 Extension d'un objet *mi*

Il y a au moins deux manières de définir l'extension d'un objet *mi* a.
La première consiste à considérer que chaque élément $\omega \in \Omega$ est plus ou moins dans l'extension de a d'après la valeur de son appartenance donnée par $a(\omega)$. Dans ce cas, l'extension de a est définie par

$$| a_x |_x = \{(\omega, a(\omega)) \; tq \; \omega \in \Omega\}$$

et notée $| a |_\Omega$.

La seconde nécessite la spécification d'un seuil α et est définie par

$$| a |_{\Omega,\alpha} = \{\omega \in \Omega \text{ tq } a(\omega) \geq \alpha\}$$

On l'appelle la α–extension.

2.1.4 Autres notions exprimées par les objets symboliques modaux

Sur base des définitions ci-dessus, nous pouvons expliquer en détail les notions précipitées en début de section : doute, incertain, variation, vague, imprécision.

1. *L'incertain :*
   Les valeurs prises par les assertions sur les $\omega \in \Omega$, ne sont plus uniquement

Vrai ou *Faux*, mais peuvent être une valeur quelconque dans un ensemble ordonné

$$L = \{L_1, L_2, ..., L_k\}.$$

Exemples :

(a) $L = \{1, 2, 3, 4, 5\}$

(b) L_1 = *convient tout à fait;*
L_2 = *peut convenir;*
L_3 = *ne convient pas;*
L_4 = *ne convient pas du tout.*

mais aussi si k est infini, nous pouvons travailler avec :

$$L = [0, 1].$$

2. *La variation :*
L'événement $[y_i = V_i]$ signifie que chaque élément de l'extension de $a = \wedge_i [y_i = V_i]$ est décrit par des valeurs des variables y_i telles que ces valeurs appartiennent aux V_i correspondants. Ceci signifie qu'au sein de la même classe, on peut retrouver des objets ne présentant pas exactement les mêmes caractéristiques.
Ainsi, par exemple, si a est une équipe de foot décrite par a = [taille = [1.60, 2]] $\wedge$ [nombre enfants = {0, 1, 2}] $\wedge$ [nationalité = {*italien*, *belge*}]., il pourra y avoir dans cette équipe un *italien* ayant *2* enfants et mesurant *1 mètre 60* mais aussi un *belge* n'ayant *pas* d'enfants et mesurant *1 mètre 86*. Nous constatons donc qu'il y de la *variation* dans les caractéristiques (valeurs prises par les descripteurs) des membres de la *même* équipe.
Dans le cas des objets modaux de l'intérieur, ces valeurs sont quantifiées par un mode. (Ce mode traduisant, par exemple, de la possibilité ou de la fréquence.)

3. *Le doute :*
On le rencontre principalement au niveau des objets élémentaires (c'est-à-dire sur les éléments de $\varphi(\Omega)$). Pour des objets de ce genre, chaque descrip-

teur n'a qu'une seule valeur possible. S'il y en a plusieurs (c'est-à-dire que V_i n'est pas réduit à un singleton), cela signifie que que l'on hésite entre ces valeurs. Il y a donc un doute (pas une variation) parce qu'un individu ne peut être décrit que par une seule valeur pour chaque variable.

Exemple :
Soit $\omega^s = [y_1 = 4,5] \wedge [y_2 = jaune, rouge]$ qui décrit une fleur. On remarque qu'il existe un doute quant au nombre de pétales (4 ou 5) et à la couleur (jaune ou rouge).

4. *L'imprécision :*
Elle survient au niveau des valeurs prises par les descripteurs (c'est-à-dire les éléments de V_i).

Exemple :
L'évènement [taille $= 12 \pm 0.5\,, 21 \pm 1$] exprime une variation ou un doute entre deux valeurs imprécises de contours connus. Ainsi, la première de ces quantités varie dans [11.5, 12.5] et la seconde [19, 21].

5. *Le vague :*
Il se produit au niveau des descripteurs lorsque les contours de l'imprécision sont mal connus (contrairement au cas précédent où ils étaient totalement spécifiés).

Exemple :
Soit l'évènement [y_i = grand] où grand est une application de O_i dans [0, 1] prenant une valeur un quand on atteint ou dépasse 1.95m et diminue si la taille diminue à partir de cette limite.

2.1.5 Exemple

La sémantique du domaine (x est ici noté i pour intensité) concerne les formes données par

$$O_1 = \{ \text{étalée, arrondie} \} \text{ et } O_2 = \{ \text{lourde} \}$$

et d'intensités diverses

$$M^i = \{ \textit{très, assez, peu, très peu, nil} \}$$

On considère que, si $nil(v_i) = \emptyset$, cela signifie que le caractère v_i n'intervient pas dans la description de a.
Soit un objet modal de l'intérieur

$$a = [y_1 = \textit{peu} \text{ étalée}, \textit{assez} \text{ arrondie}] \wedge_i [y_2 = \textit{peu} \text{ lourde}]$$

qui décrit un ensemble de pièces d'usinage dont la forme est soit peu étalée, soit assez arrondie mais qui sont toujours peu lourdes.
Nous disposons d'une pièce ω décrite par

$$\omega^s = [y_1 = \textit{assez} \text{ arrondie}] \wedge_i [y_2 = \textit{très} \text{ lourde}, \textit{assez} \text{ lourde }].$$

Ce qui signifie que cette pièce est assez arrondie et qu'il existe un doute sur le fait qu'elle soit *trés* ou *assez* lourde. On hésite entre ces deux modes pour le même objet ω. Il y a un doute.
Définissons à présent pour a les applications $q_j : O_j \rightarrow M^i$:

$$q_1(\text{étalée}) = \textit{peu};$$
$$q_1(\text{arrondie}) = \textit{assez};$$
$$q_2(\text{lourde}) = \textit{peu}.$$

Exprimons de manière semblable (pour ω^s) les applications $r_j : O_j \rightarrow M^i$:

$$r_1^1(\text{étalée}) = \textit{nil};$$
$$r_1^1(\text{arrondie}) = \textit{assez};$$
$$r_2^1(\text{lourde}) = \textit{très};$$
$$r_2^2(\text{lourde}) = \textit{assez}.$$

Nous allons calculer la valeur de $a(\omega)$ afin de voir si la pièce appartient ou non à l'ensemble d'objets décrits par a.
Pour simplifier, nous regroupons les éléments de M^i *très* et *assez* sous la dénomination *plutôt.*
Dès lors, nous avons

$$r_2^1 \cup_i r_2^2(\text{lourd}) = (\textit{très, assez}) = \textit{plutôt.}$$

Nous définirons l'ensemble ordonné $L^i = \{L_1, L_2, L_3\}$ où

$L_1 = \textit{pas acceptable};$
$L_2 = \textit{acceptable};$
$L_3 = \textit{complètement acceptable.}$

Ensuite, nous nous servirons de l'application de comparaison g_i et de celle d'agrégation f_i.
La première de ces fonctions est :

$$g_i : Q_i \times Q_i \to L^i.$$

Elle est spécifiée par une table T_{gi} de telle sorte que :

$g_i(q_1, r_1^1) = T_{gi}((\textit{peu}$ étalée, $\textit{assez}$ arrondie), ($\textit{nil}$ étalée, $\textit{assez}$ arrondie))
$= \textit{acceptable}$

$g_i(q_2, r_2^1 \cup_i r_2^2) = T_{gi}(\textit{peu}$ lourd, $\textit{plutôt}$ lourd$) = \textit{pas acceptable}$

Si on définit la fonction d'agrégation f_i par :

$$f_i(L_k) = Min_k\, L_k\,,$$

en sachant que $L_1 < L_2 < L_3$, nous trouvons donc que

$a(\omega) = f_i(g_i(q_1^1, r_1^1), g_i(q_2^1, r_2^1 \cup_i r_2^2))$
$= f_i(\textit{acceptable, pas acceptable.})$
$= \textit{pas acceptable.}$

2.2 Objets modaux de l'intérieur associés à diverses sémantiques

2.2.1 Les objets possibilistes

2.3.2.1 Introduction

Les possibilités expriment une sémantique différente de celle des probabilités. Par exemple, pour un dé, la probabilité qu'un lancé prenne la valeur 3 est de $\frac{1}{6}$ tandis ce que sa possibilité est de 1. Si le dé est pipé, cette probabilité peut être nulle et la possibilité correspondante faiblement positive.

Les objets possibilistes permettent de modéliser plusieurs sémantiques (connaissances) du domaine.
Parmi celles-ci, on trouvera :

1. *Possibilité physique :* aptitude (capacité) matérielle à réaliser une tâche. Par exemple : un chien *peut* courir, une souris ne *peut* pas transporter 250 kg.

2. *Possibilité en concordance avec une connaissance actuelle.* Par exemple, nous ne *pourrons* pas aller skier demain car la météo a annoncé le dégel cette nuit.

3. *Possibilité exprimant le non-étonnement :* elle peut se construire à l'aide d'une densité de probabilité de telle sorte que son mode (i.e l'évènement de non étonnement maximal) ait une possibilité égale à un et que les inégalités $\Pi(A) \geq P(A) \geq N(A)$ soient satisfaites. Elle peut encore se décrire grâce aux typicicités c'est-à-dire par les termes *typique*, *atypique* ...Par exemple : la couleur typique d'une plante donnée est jaune et brune.

Remarquons que 2 et 3 expriment un jugement qui engage plus ou moins son auteur tandis ce que 1 est une possibilité indépendante de celui qui l'énonce.
Nous illustrerons plus loin la première de ces sémantiques.
A la notion de possibilité une autre notion est étroitement associée : la nécessité. Nous les examinerons en même temps.

Avant de donner la définition exacte d'une assertion possibiliste et nécessitiste, rappelons les axiomes de la théorie des possibilités et de celle des nécessités.

2.3.2.2 Rappel

- Une *mesure de possibilité* est une application

$$\Pi : P(\Omega) \rightarrow [0, 1]$$

telle que les axiomes suivants soient vérifiés

1. $\Pi(\Omega) = 1$ et $\Pi(\emptyset) = 0$;
2. $\forall A, B \subseteq \Omega \ \Pi(A \cup B) = max(\Pi(A), \Pi(B))$.

- Une *mesure de nécessité* associée à Π est une application

$$N : \Omega \rightarrow [0, 1]$$

telle que

$$N(A) = 1 - \Pi(\bar{A})$$

où $\bar{A} = c(A)$, le complémentaire de A dans Ω.

On montre facilement que

1. $N(\emptyset) = 0$;
   $N(A \cap B) = min(N(A), N(B))$.

2.3.2.3 Définitions

Soit Q_i un ensemble de mesures de possibilité (nécessité sur O_i :

1. *Une assertion possibiliste est une assertion mi qui prend ses valeurs dans* $L^p = [0, 1]$ *et qui se définit par :*

   (a) OP_p *qui comprend les trois opérations suivantes :*

   i. $\forall q_i^1, q_i^2 \in Q_i \ q_i^1 \cup_p q_i^2 = max(q_i^1, q_i^2)$;

ii. $\forall\, q_i^1, q_i^2 \in Q_i \;\; q_i^1 \cap_p q_i^2 = min(q_i^1, q_i^2)$;

iii. $\forall\, q \in Q_i \;\; c_p(q) = 1 - q$.

(b) g_p : $\forall\, q_i^1, q_i^2 \in O_i \;\; g_p(q_i^1, q_i^2) = sup\{min(q_i^1(v), q_i^2(v))\ tel\ que\ v \in O_i\}$;

(c) f_p : $\forall\, L_i \subset L^p = [0, 1] \;\; f_p(L_i) = max\{l\ tq\ l \in L_i\}$.

2. *Une assertion possibiliste est une assertion mi qui prend ses valeurs dans* $L^n = [0, 1]$ *et qui se définit par :*

(a) OP_n *qui comprend les trois opérations suivantes :*

i. $\forall\, q_i^1, q_i^2 \in Q_i \;\; q_i^1 \cup_n q_i^2 = min(q_i^1, q_i^2)$;

ii. $\forall\, q_i^1, q_i^2 \in Q_i \;\; q_i^1 \cap_n q_i^2 = max(q_i^1, q_i^2)$;

iii. $\forall\, q \in Q_i \;\; c_n(q) = 1 - q$.

(b) g_n : $\forall\, q_i^1, q_i^2 \in O_i \;\; g_n(q_i^1, q_i^2) = inf\{max(q_i^1(v), q_i^2(v))\ tel\ que\ v \in O_i\}$;

(c) f_n : $\forall\, L_i \subset L^n = [0, 1] \;\; f_n(L_i) = min\{l\ tq\ l \in L_i\}$.

2.3.2.4 Remarques

1. On dit que les assertions *mi* qui viennent d'être définies sont possibilistes car

   (a) elles se mettent sous la forme $a = \wedge_{i,p}[y_i = \{q_i^j\}_j]$ et que les q_i^j sont des mesures de possibilités sur O_i;

   (b) une assertion possibiliste est une mesure de possibilité sur l'ensemble Ω dont les éléments, considérés comme des mesures de possibilités, sont munis des opérations OP_p idempotentes[4] sur les Q_i.

2. La remarque précédente reste valable si nous remplaçons partout le terme possibiliste par nécessitiste.

4. voir annexe

3. Pour passer de la définition d'un objet possibiliste à celle d'un objet nécessitiste, nous pouvons procéder ainsi :

$$\forall \omega \in \Omega \ a_n(\omega) = \varphi^{-1}(c(\omega^s)) = c(\varphi^{-1}(\omega^s)) = c(\omega).$$

avec la dernière égalité qui découle directement de la définition.
Il en résulte que :

$$a_n(\omega) = 1 - \wedge_{i,p} g_p(q_i, r_i),$$

ce qui entraîne que :

$$\begin{aligned} a_n(\omega) &= 1 - f_p[g_p(q_i, \bar{r_i})] \\ &= 1 - max_i[g_p(q_i, \bar{r_i})] \\ &= min(q_i, r_i). \end{aligned}$$

2.3.2.5 Exemple

Il s'agit d'un exemple concernant une possibilité physique.
Soit Ω, un ensemble de repas. Nous allons chercher à savoir si ces repas satisfont à un régime. Celui-ci est décrit par la consommation en oeufs et sucres.
La variable

$$y_1 : \Omega \rightarrow Q_1 : \omega \rightarrow y_1(\omega) = q_1$$

donne pour chaque repas la quantité d'oeufs consommés : v_j^1 et la possibilité $q_1(v_j^1) = m_1^j$ associée à v_j^1.

Et la variable

$$y_2 : \Omega \rightarrow Q_1 : \omega \rightarrow y_2(\omega) = q_2$$

donne pour chaque repas le nombre de sucres consommés : v_j^2 et la possibilité $q_2(v_j^2)$ associée à v_i^2.

Si le nombre d'oeufs ou de sucres absorbés au cours du repas est connu, nous lui associons une possibilité égale à 1 ainsi qu'aux nombres inférieurs et 0 aux autres. (P.ex : si le nombre d'oeufs est 3, nous donnerons une possibilité 1 au fait

d'en avoir mangé 3, 2 ou 1. Par contre, nous donnerons à 4, 5, 6 ... la possibilité 0.)

Un menu de régime peut s'écrire sous la forme d'une asssertion $a_p = \wedge_{i,p}[y_i = q_i]$ avec q_i qui est la possibilité se rapportant au nombre d'oeufs (i = 1) et au nombre de sucres (i = 2) consommables si on désire rester fidèle au régime.

S'il y a un doute (P.ex : si le nombre de sucres = $\{4, 5, 6\}$), nous nous ramènons au cas précédent par la procédure ci-dessous :
Si on hésite entre plusieurs valeurs $D = \{v_i\}$ pour ce nombre, on remplacera, par prudence, D par son v_i de plus faible possibilité déterminée par le q_i défini par le régime.

Afin que a_p soit bien une assertion possibiliste, il faut que q_i soit une mesure de possibilité sur O_i. Ce choix traduit-il bien la connaissance du domaine? OUI.

En effet,

- $q_i(O_i) = 1$ car la possibilité de manger un certain nombre d'oeufs ou de sucres (y compris 0) est égale à 1.
- $q_i(v_1 \cup_p v_2)$ où $v_1, v_2 \in O_i$ est la plus grande des deux possibilités. (P.ex si la possibilité de manger un oeuf est de 1 et celle d'en manger dix est de 0, la possibilité d'en manger un ou dix est 1.)

Le calcul de $a_p(\omega)$ se déroule ainsi : si

$$a_p(\omega) = [y_1 = q_1] \wedge [y_2 = q_2]$$

et

$$\omega^s = [y_1 = q_1] \wedge [y_2 = r_2]$$

avec les q_i définis ainsi :

$$q_1(0) = q_1(1) = 1; \quad q_2(0) = q_2(1) = q_2(2) = 1$$
$$q_1(2) = 0.5; \qquad q_2(3) = 0.8;$$

$$q_1(3) = 0.5; \qquad q_2(4) = 0.4;$$
$$q_1(j) = 0 \text{ si } j > 3; \quad q_2(j) = 0 \text{ si } j \geq 3,$$

et les r_i :

$$r_1(0) = r_1(1) = r_1(2) = 1; \quad r_2(0) = r_2(1) = r_2(2) = 1;$$
$$r_1(j) = 0 \text{ si } j \geq 3; \qquad r_2(j) = 0 \text{ si } j \geq 3.$$

On a alors que comme :

$$g_p(q_i, r_i) = max\{min(q_i(v), r_i(v)) \quad \forall v \in O_i\}^5$$

$$\begin{aligned} a_p(\omega) &= \underbrace{g_p(\omega)(q_1, r_1)}_{1} \wedge_p \underbrace{g_p(q_2, r_2)}_{2} \\ &= 1 \wedge_p 1 \\ &= f_p(1, 1) \\ &= max(1, 1) \\ &= 1. \end{aligned}$$

On a, en effet, étant donné que $O_i = \{1, 2, 3 \ldots\}$:

- (1) vaut le maximun des valeurs suivantes
  - $v = 0$ $min(q_1(0), r_1(0)) = min(1, 1) = 1$
  - $v = 1$ $min(q_1(1), r_1(1)) = min(1, 1) = 1$
  - $v = 2$ $min(q_1(2), r_1(2)) = min(0.5, 1) = 0.5$
  - $\vdots$

  On voit qu'il suffit qu'un des minima soit égal à 1 pour que $g_p(q_1, r_1)$ vaille 1.

- Même raisonnement dans le cas (2).

5. Pour chaque $v \in O_i$, on prend la valeur minimale entre $q_i(v)$ et $r_i(v)$. Réitérant cette comparaison pour chacun des v, on obtient ainsi un ensemble de valeurs dont on prend le maximun.

Nous pouvons ensuite calculer $a_n(\omega)$

$$a_n(\omega) = g_n(q_1, r_1) \wedge_n g_n(q_2, r_2)$$

avec les $\bar{q}_i$ définis par :

$$\begin{array}{lll} \bar{q}_1(0) = 1 - q_1(0) = \bar{q}_1(1) = 0; & \bar{q}_2(0) = \bar{q}_2(1) = \bar{q}_2(2) = 0; \\ \bar{q}_1(2) = \bar{q}_1(3) = 0.5; & \bar{q}_2(3) = 0.2; \\ \bar{q}_1(j) = 1 \text{ si } j \geq 3; & \bar{q}_2(3) = 0.6; \\ & \bar{q}_2(j) = 1 \text{ si } j \geq 3. \end{array}$$

On a alors que

$$\begin{aligned} a_n(\omega) &= g_n(q_1, r_1) \wedge_n g_n(q_2, r_2) \\ &= \wedge_{i,n} min\{max(\bar{q}_i(v), r_i(v)) \ v \in O_i\} \\ &= 0.5 \wedge_n 0.2 \\ &= min(0.5, 0.2) \\ &= 0.2. \end{aligned}$$

La nécessité exprime l'écart entre le bord de r (i.e ses faibles valeurs) et q. Ici, il s'agira des faibles possibilités du repas et celles du régime.
Autrement dit, si le repas a suivi les possibilités du régime, la nécessité est d'autant plus élevée que l'on a mangé ce qui était permis[6]

2.2.2 Les objets probabilistes

2.3.2.1 Introduction

Dans le cas probabiliste, mesurer dans l'espace des objets symboliques revient à étendre les mesures de probabilité à des ensembles d'objets symboliques munis d'opérateurs d'union et d'intersection adéquats de manière à retrouver les axiomes de Kolmogorov (rappelés ci-dessous) sur de tels objets. En d'autres termes, on

6. En fait, la nécessité vaut le minimun du maximun de $(1 - q_i, r_i)$. On commence donc par prendre le minimun des deux valeurs $1 - q_i(v)$ et $r_i(v)$ pour chaque v. On rassemble ensuite les valeurs ainsi obtenues et on en prend le maximun. Dans le cas de la nécessité, on s'intéressera aux *faibles* valeurs puisque l'on termine avec la prise d'un minimun. Et, là où q_i et r_i, on choisira r_i. Tandis que si q_i et r_i sont petits, on sélectionnera la valeur de $\bar{q}_i$

doit définir l'union et l'intersection probabilistes de telle sorte que les axiomes soient vérifiés. On doit donc étendre l'axiomatique de Kolmogorov à des ensembles constitués de mesures de probabilité classiques, notées q_i^j munies des opérations archimédiennes (bien différentes des opérations ensemblistes habituelles). Nous travaillerons donc avec des objets modaux dont les q_i^j sont des lois de probabilité. La théorie des probabilités modélise plusieurs sortes de connaissances et au moins les trois suivantes :

1. *La chance*:
   Les origines des probabilités se situent dans les calculs de chances au jeu. Si on suppose que tous les éléments de Ω ont une probabilité identique de survenir[7], nous aurons la loi simple :

$$P(\text{évènement}) = \frac{nombre\ de\ cas\ favorables}{nombre\ de\ cas\ possibles}.$$

   Le calcul de la probabilité d'un évènement quelconque[8] se fera par calcul combinatoire grâce à la loi simple ci-dessus.

2. *La fréquence*:
   La probabilité d'un évènement est la limite de la fréquence de ce résultat lorsque que l'on réitère un grand nombre de fois l'expérience.

---

7. **Exemple :** si on jette une pièce, la probabilité d'obtenir pile ou face est égale et vaut $\frac{1}{2}$.

8. **Exemple :** Un comité de 5 personnes doit être choisi parmi les 6 hommes et les 9 femmes d'un groupe. Si le choix est le résultat du hasard, quelle est la probabilité que le comité soit composé de 3 hommes et 5 femmes?
Solution : Admettons que 'choix dû hasard' signifie que chacune des $\binom{15}{5}$ combinaisons possibles a les mêmes chances d'apparaître. La probabilité cherchée sera donc égale à :

$$\frac{\binom{6}{3}\binom{9}{2}}{\binom{15}{5}} = \frac{240}{1001}.$$

Cet exemple est tiré de [13].

Exemple :

La probabilité d'obtenir pile est donnée par la limite du rapport

$$\frac{nombre\ de\ piles}{nombre\ total\ de\ jets}$$

lorsque le nombre de jets devient très important.

3. *L'incertitude*:
   Nous désirons mesurer un degré certitude de l'occurence d'un évènement ne survenant qu'une seule fois.

   Exemple :

   (a) Le nom de cette personne est *probablement* Paul;

   (b) La couleur de cet objet est *probablement* verte.

   Dans le premier cas, l'incertitude provient d'un mauvais souvenir et, dans le second, d'un éclairage insuffisant.

2.3.2.2 Rappel : Axiomes de Kolmogorov

Soit $C(\Omega)$ une σ-algèbre d'événements sur Ω

Une probabilité p sur $(\Omega.\ C(\Omega))$ est une application telle que :

1. $p(\Omega) = 1$ *où* Ω *est l'évènement certain;*
2. *Pour tout ensemble dénombrable d'évènements élémentaires e_i d'extensions disjointes $A_i = \mid e_i \mid_\Omega$, on a $p(\cup_i A_i) = \sum_i p(A_i)$.*

Les conditions 1 et 2 désignent les axiomes de Kolmogorov.
De cette définition, on tire les propriétés suivantes :

- $p(\emptyset) = 0$;
- $p(\bar{A}) = 1 - p(A)$;
- $A \subseteq B \Rightarrow P(A) \leq p(B)$;

- $p(A \cup B) = p(A \cup B) = p(A) + p(B) - p(A) \cap p(B)$;
- $p(\sum_i A_i) \leq \sum_i p(A_i)$.

Voici la définition exacte des objets probabilistes.

2.3.2.3 Définition

Si chaque Q_i est un ensemble de mesures de probabilités sur O_i :

Une assertion probabiliste est une assertion mi qui prend ces valeurs dans $L^{pr} = [0, 1]$, *elle se définit par :*

1. OP_{pr} *qui comprend les trois opérations suivantes :*
   (a) $\forall q_i^1, q_i^2 \in Q_i \; q_i^1 \cup_{pr} q_i^2 = q_i^1 + q_i^2 - q_i^1 q_i^2$
   tel que $q_i^1 q_i^2$ *associe à chaque* $v \in O_i \; q_i^1(v) q_i^2(v)$;
   (b) $\forall q_i^1, q_i^2 \in Q_i \; q_i^1 \cap_{pr} q_i^2 = q_i^1 q_i^2$
   tel que $q_i^1 q_i^2$ *associe à chaque* $v \in O_i \; q_i^1(v) q_i^2(v)$;
   (c) $\forall q_i^j \in Q_i \; c(q_i^j) = 1 - q_i^j$.
2. $g_{pr} : \forall q_i^1, q_i^2 \in Q_i \; g_{pr}(q_i^1, q_i^2) = < q_i^1, q_i^2 > = \sum\{q_i^1(v) q_i^2(v) \; v \in O_i\}$;
3. $f_{pr} : f_{pr}(L_i) =$ *moyenne des* L_i
   avec $L_i \in L^{pr}$.

2.3.2.4 Remarques

1. $\bar{q}_i(v)$ vaut $1 - q_i$ car $\bar{q}_i(v)$ est la probabilité de ne pas obtenir v (P.ex : lors d'un tirage).
2. Il est possible de montrer que OP_{pr} définit des opérations dites archimédiennes[9] Nous pouvons encore démontrer que les assertions probabilistes sont des mesures satisfaisant aux axiomes de Kolmogorov sur Ω. La seule condition est que Ω soit muni des opérations archimédiennes induites des opérations de OP_{pr} sur les Q_i. Pour cette raison, nous dénommons parfois

9. voir annexe.

les assertions probabilistes *probabilités archimédiennes*. Ainsi, dans le cadre symbolique, les probabilités ne sont plus calculées avec les opérations ensemblistes habituelles mais avec des opérations représentant la sémantique du domaine. Ceci est assez intuitif : nous ne pouvons unifier des mesures de probabilité comme nous unifions les parties d'un ensemble classique. Notons ainsi que, par exemple, l'idempotence[10] n'est pas satisfaite par l'union archimédienne.

3. Selon la sémantique désirée, d'autres choix pour g_{pr} et f_{pr} sont admissibles. Par exemple :
   Définir f_{pr} comme le produit des L_i est un choix plus adapté que de prendre la moyenne dans des problèmes d'identification de pièces d'usinage ou d'espèces de plantes.

4. En ce qui concerne le choix de $\cup_{pr}$, nous pouvons montrer par un contre-exemple que $q_i^1 \cup_{pr} q_i^2$ n'est pas une loi de probabilité.
   Contre-exemple :
   Soit deux pièces de monnaie lancées de manière indépendante et de lois de probabilité respectives q_i^1, q_i^2, nous avons alors que $q_i^1 \cup_{pr} q_i^2(v)$ est la probabilité que le résultat v (pile ou face) se produise avec l'une ou (non exclusif) l'autre de ces pièces. D'après ce que nous avons dit ci-dessus, $q_i^1 \cup_{pr} q_i^2 = q_i^1 + q_i^2 - q_i^1\, q_i^2$. Cette formule n'est pas une mesure de probabilité car les résultat v =pile et v =face sont compatibles. En effet, ils peuvent se produire en même temps, il suffit qu'une des pièces tombe sur le côté face et l'autre sur le côté pile. Dès lors, la (probabilité de face) est différente de 1 - (probabilté de pile). Ce qui ne respecte pas une des conséquences des lois de Kolmogorov.

2.3.2.5 Exemple

Soient

- un objet ω décrit par sa couleur $y_1(\omega) = \{\text{rouge, bleu}\}$ et sa forme $y_2(\omega) = \{\text{rond, plat}\}$;

10. voir annexe.

- $a = [y_1 = q_1^1, q_1^2] \wedge_{pr} [y_2 = q_2]$

où $q_1^1(\text{rouge}) = 0.9$; $q_1^1(\text{bleu}) = 0.1$;
$q_1^2(\text{rouge}) = 0.5$; $q_1^2(\text{bleu}) = 0.5$;
$q_2(\text{rond}) = 0.2$ et $q_2(plat) = 0.8$;

Il découle de ces choix que a décrit deux sortes d'objets:

1. souvent rouge et rarement bleu, parfois rond mais généralement plat;
2. rouge ou bleu avec une probabilité égale, parfois rond mais généralement plat.

- $\omega^s = [y_1 = r_1] \wedge_{pr} [y_2 = r_2]$

où $r_1(\text{rouge}) = 1$; $r_1(\text{bleu}) = 0$;
$r_2(\text{rond}) = 1$ et $r_2(\text{plat}) = 0$.

Nous pouvons, à présent, calculer le degré d'appartenance de ω^s à l'ensemble des objets décrits par a.
En utilisant la formule:

$$q_1^3 = q_1^1 \cup_{pr} q_1^2 = q_1^1 + q_1^2 - q_1^1 q_1^2,$$

nous obtenons que:

$$q_1^3(\text{rouge}) = 0.9 + 0.5 - 0.9 \times 0.5 = 0.95$$
$$q_1^3(\text{bleu}) = 0.1 + 0.5 - 0.1 \times 0.5 = 0.55$$

$$\begin{aligned} a(\omega) &= f_{pr}[g_{pr}(q_1^3, r_1), g_{pr}(q_2, r_2)] \\ &= f_{pr}[(0.95 \times 1 + 0.55 \times 0), (0.2 \times 1 + 0.8 \times 0)] \\ &= f_{pr}[0.95, 0.20] \\ &= \frac{1}{2}[0.95 + 0.20] \\ &= 0.575. \end{aligned}$$

0.575 est le degré d'appartenance de ω à l'objet *mi* défini par a. On dit encore que 0.575 est la probabilité archimédienne que ω appartienne à l'extension de a.

2.2.3 Les objets booléens

Bien que les objets booléens ne soient pas à proprement parler modaux, on peut leur associer des objet mi.
Etant donnée une assertion booléenne

$$a = \wedge_i[y_i = V_i],$$

on lui fait correspondre l'assertion modale

$$a_b = \wedge_i[y_i = q_i]$$

où q_i est une fonction caractéristique de $V_i \subseteq O_i$.

Les opérations ensemblistes $OP_p = \{\cup_b, \cap_b, c_b\}$ seront :

1. $q_1 \cup_b q_2 = max(q_1, q_2)$;
2. $q_1 \cap_b q_2 = min(q_1, q_2)$;
3. $c_b(q) = 1 - q$.

Les applications de comparaison et d'agrégation seront :

1. $g_b(q_i, r_i) = < q_i, r_i >$
2. $f_b[g_b(q_i, r_i)] = min\{l \in L^b = \{0, 1\}\}$

Il est aisé de démontrer que

$$a_b(\omega) = 1 \iff a(\omega) = Vrai;$$
et $$a_b(\omega) = 0 \iff a(\omega) = Faux.$$

Remarque

Dans le cas booléen, la seconde définition de l'extension[11] s'applique avec $\alpha = 1$.

11. voir paragraphe 1.2.3

2.2.4 Tableau de synthèse.

Voici un tableau reprenant les choix de OP_x, g_x et f_x pour les diverses objets mi.

Objets $a_x = \wedge_{i,x}[y_i = q_i]$	Booléens	Possibilistes	Nécessitistes	Probabilistes
q_i	f.caractéristique	mesure de poss.	mesure de nécess.	mesure de proba
$q_1 \cup_x q_2$	$max(q_1, q_2)$	$max(q_1, q_2)$	$min(q_1, q_2)$	$q_1 + q_2 - q_1 q_2$
$q_1 \cap_x q_2$	$min(q_1, q_2)$	$min(q_1, q_2)$	$max(q_1, q_2)$	$q_1 q_2$
c_q	$1 - q$	$1 - q$	$1 - q$	$1 - q$
$g_x(q_i, r_i)$	$< q_i, r_i >$	$sup\{min(q_i(v), r_i(v))\}$	$inf\{max(\bar{q}_i(v), r_i(v))\}$	$< q_i, r_i >$
f_x	min	max	min	$moyenne$

2.3 Propriétés des objets modaux

2.3.1 Propriétés générales

Toutes les propriétés des objets symboliques booléens (et celles de leurs classes et classsifcations) peuvent être généralisées aux objets modaux (complétude, affinement ...) et ainsi qu'à leurs classes et classifications.
Il est encore possible de généraliser les objets hordes et de synthèse booléens à des objets hordes et de synthèse modaux.

2.3.2 Propriétés spécifiques

2.4.2.1 Caractéristique de l'union probabiliste et booléenne.

Définition *On dit qu'une union $(A \cup B)$ est généralisante si*

$$A \cup B \geq max(A, B).$$

Application L'union probabiliste et booléenne est généralisante. En effet, puisque :

1. $q_1 \cup_{pr} q_2 = q_1 + q_2 - q_1 q_2$

et $q_1(v), q_2(v) \in [0, 1]$

2. et que par définition :
   $q_1 \cup_b q_2 = max(q_1, q_2)$.

1.4.2.2 Indice de distance

Nous avons déjà rencontré un indice de proximité entre OAB. Cet indice est construit à partir de

1. la fonction de comparaison g_b;
2. la fonction d'agrégation f_b.

En fait, nous affirmons que :

1. g_b est un indice de proximité basé sur une mesure positive appelée Potentiel de Description d'un événement élémentaire booléen (à ne pas confondre avec celui du OAB)[12].
2. f_b est un indice de proximité basé sur la distance de Minkowski[13] qui agrège les résultat fournis par g_b.

2.3.3 Espace dual

2.4.3.1 Introduction

Outre la généralisation de deux objets que l'on avait définie en ces termes :

Définition (rappel) *On définit une relation de préordre partiel entre objets symboliques en disant que :*

$a_1 \leq a_2 \iff$ *l'extension* ***symbolique*** *de* a_1 *est contenue dans celle de* a_2.

Dans ce cas, a_2 *est dit plus général que* a_1.

12. Nous reviendrons sur ces notions dans la partie 3.
13. Nous reviendrons sur cette notion dans la partie 3.

Objectif Nous allons maintenant suggérer une autre manière pour généraliser. Par le biais de deux théorèmes, nous montrerons que ces objets peuvent être étendus afin d'exprimer des métaconnaissances (connaissances sur les connaissances). Plus précisément, si nous disons que les ensembles classiques représentent des connaissances de niveau 0; les probabilités, possibilités, .. des connaissances de niveau 1 (en fournissant des mesures portant sur des combinaisons algébriques d'ensembles de niveau 0); ces théorèmes affirmeront qu'il existe un niveau 2. A ce niveau, nous trouverons des mesures sur des combinaisons algébriques d'ensembles de niveau 1. Celles-ci satisferont à des propriétés analogues à celles des probabilités, possibilités ... du niveau 1.
Le niveau 2 contient des sortes de probabilités sur les probabilités, possibilités sur les possibilités ...

2.4.3.2 Description

Notre but est d'étendre une assertion $mi\ a = \wedge_{i,x}[y_i = q_i]$ (où q_i dépend du choix de la sémantique x et peut être une mesure de probabilité, possibilité ...), à une assertion mi duale. On la notera $a^\star$. Elle sera définie sur les sous-ensembles de a_x qui est l'ensemble des assertions mi associé à x. Nous donnerons ensuite, deux théorèmes affirmant que $a^\star$ est elle même une sorte de probabilité, possibilité dépendant de x.

Plus précisément.

soient donnés

- $A \subseteq a_x$;
- $a_j^\star$ une mesure duale de $a_j = \wedge_{i,x}[y_i = q_i^j]$ et
- Q_i^A, l'ensemble des q_i^l tels que $a_l = \wedge_{i,x}[y_i = q_i^l]$,

alors on posera que:

$$a_j^\star(A) = f_x(\{g_x(q_i^j, \{\cup_{l,x}\, q_i^l \;\; q_i^l \in Q_i^A\})\}_i)$$

et, si $\star_x \in \{\cup_x, \cap_x\}$ alors

$$a_j^\star(A \star_x B) = f_x(\{g_x(q_i^j, \{\star_x q_i^l \ q_i^l \in Q_i^{A \cup B}\})\}_i).$$

Cas possibiliste a_p est l'ensemble des assertions possibilistes, on étend leur espace de défintion (qui était Ω) sur a_p en transformant

$$a_1 : \Omega \rightarrow [0, 1]$$

en

$$a_1^\star : a_p \rightarrow [0, 1]$$

telle que

$$a_1^\star(a_2) = f_p(\{g_p\{q_i^1, q_i^2\}\}_i)$$

où $a_j = \wedge_{i,p}[y_i = q_i^j]$ et les q_i^j sont des mesures de possibilité sur O_i. Le théorème

suivant généralise les axiomes des possibilités à l'espace dual :

Théorème 1

$$i)\ a^\star(a_p) = 1\ ,\ \ a^\star(\emptyset) = 0\,;$$

$$ii)\ a^\star(A_1 \cup_p A_2) = max(a^\star(A_1, A_2)),$$

où $A_i = \cup_p(a \in A_i \subseteq a_p)$.

Cas probabiliste a_{pr} est l'ensemble des assertions probabilistes, on étend leur espace de définition (qui était Ω) sur a_{pr} en transformant

$$a_1 : \Omega \rightarrow [0, 1]$$

en

$$a_1^\star : a_{pr} \rightarrow [0, 1]$$

telle que

$$a_1^\star(a_2) = f_{pr}(\{g_{pr}\{q_i^1, q_i^2\}\}_i)$$

où

$a_j = \wedge_{i,pr}[y_i = q_i^j]$ et les q_i^j sont des mesures de probabilité sur 0_i.

Le théorème suivant généralise les axiomes de Kolmogorov à l'espace dual :

Théorème 2

$$i)\ a^\star(a_{pr}) = 1\ ,\quad a^\star(\emptyset) = 0\,;$$

$$ii)\ a^\star(A_1 \cup_{pr} A_2) = a^\star(A_1) + a^\star(A_2) - a^\star(A_1 \cap_{pr} A_2),$$

où $A_i = \cup_{pr}(a \in A_i \subseteq a_{pr})$.

2.4.3.3 Autre généralisation grâce à l'espace dual

Description Une autre manière, plus précise, de généraliser plusieurs assertions $a_1, \ldots, a_n$ par $b_1, \ldots, b_k$ $k \leq n$ assertions, est obtenue en résolvant le problème suivant :
Soit

$$i)\ b^\star = \cup_{j,x} b_j^\star,$$

maximiser W tel que

$$W(b) = \Pi_{i=1}^n b^\star(a_i).$$

On peut aussi choisir

$$ii)\ W(b) = min_i\, b^\star(a_i).$$

Dans ces deux cas, cela revient à chercher k assertions b_j dont l'extension contienne les $a_1, \ldots, a_n$ au seuil le plus élevé possible.
$b^\star(a)$ atteint sa valeur maximale lorsque les q_i^a et les q_i^b sont les plus proches possible ou si les q_i^b sont plus généraux que les q_i^a.
Afin de maximiser les formule $i)$ et $ii)$, on devra maximiser tous les termes $b^\star(a_i)$.
Il y a cependant une contrainte à satisfaire : $b_i^\star(b_j) = 0$. En d'autres mots, on exige qu'il n'y ait pas de redondance entre les b_j.

Adéquation d'un objet symbolique avec un ensemble, par décomposition de mélanges de lois en lois Soit un ensemble d'assertions symboliques $A = \{a_1, \ldots, a_n\}$, il s'agit de trouver

$$b^\star = P_1\, b_1^\star + \ldots + P_k\, b_k^\star$$

où P_i est proportionnel à l'extension de $b_i^\star$ dans A[14] et $\Sigma_i\, P_i = 1$ tel que le critère $W(b) = \Pi_i\, b^\star(a_i)$ soit maximun avec des contraintes sur b et (ou) sur les b_i. Ces contraintes permettront d'éviter les solutions triviales.

Dans le cas particulier des assertions probabilistes caractérisées par un seul événement[15] suivant une loi donnée[16], on obtient un problème dual par rapport au problème de probabilité classique en statistique de la décomposition de mélanges de lois de probabilité.

Exemple Soit Ω, un ensemble de bébés nés dans une commune en 1992. Cette commune est représentée par l'événement probabiliste $a_i = [y_i = q_i]$ où q_i est la densité de probabilité (de loi normale) définie sur O_i. O_i est l'ensemble des poids possibles des bébés à la naissance.

Si l'on associe à un bébé ω l'événement $\omega^s = [y = r]$ où

$$r : Q_i \to \{0, 1\} : r(v) = \begin{cases} 1 \iff v \text{ est le poids du bébé } \omega \\ 0 \text{ sinon.} \end{cases}$$

Nous avons alors

$$a_i = < q_i, r > = \Sigma\{q_i(v)\, r(v) \;\; v \in O_i = q_i(v)\}$$

où v est le poids du bébé ω.
Un ensemble de communes est une province.
Si

$$a_i^\star = \Sigma\{q^\star(v)\, q_i(v) \quad v \in O_i\}$$

est 'faible', cela signifie que la commune est excentrée dans la province représentée par $a_i^\star$. Par contre, si $a_i(\omega)$ est faible, cela veut dire que le bébé est excentré

14. c'est-à-dire l'ensemble $E = \{a \in A \;\; b_i^\star \geq \alpha\}$

15. i.e $\omega^s = [y_i = r]$ avec r qui est une mesure de probabilité.

16. P.ex, r suit une loi normale.

dans la commune[17] décrite par a_i.

En probabilité classique de décomposition, un mélange, étant données k communes, consisterait à décomposer la loi de probabilité dans chacune de ces communes. Par opposition, le problème que nous posons ici se situe à un niveau dual. En effet, il s'agit ici de décomposer $b^\star$ qui est une loi sur des lois.

Ainsi, la problématique augmente d'un niveau en passant des données aux connaissances sur les lois portant sur les données. Nous passons d'une distribution de poids à une distribution sur les distributions des poids.

Nous pourrions passer à un niveau supérieur : les pays. Cette fois, il faudra organiser en loi une distribution de distributions de probabilité sur des poids. Si nous désirons compliquer encore un peu les calculs, nous nous intéresserons aux continents. A ce stade, nous devrons organiser en loi des distributions de distributions de distributions de probabilité sur les poids et ainsi de suite pour les niveaux supérieurs...

17. Un individu est dit excentré quand il décalé par rapport au centre (moyenne) de l'ensemble des individus de la même catégorie. On dira encore que ce centre est l'individu moyen parfait.

Deuxième partie

Comparaison avec les objets et l'analyse classiques

Introduction

Dans ce qui va suivre, nous allons comparer l'analyse de données usuelles à celle des objets symboliques. Nous commencerons par donner quelques propriétés distinguant les objets de l'analyse de données symboliques de ceux de l'analyse classique. Ensuite, nous examinerons les quatre types d'analyse de données.

Chapitre 1

Comparaison : données usuelles, données symboliques

1.1 Introduction

Maintenant que nous avons précisé ce qu'étaient des objet symboliques, il serait intéressant de les comparer (situer par rapport) aux données (numériques) de l'analyse de données classique.

Nous allons présenter deux manières de les distinguer.
La première se basera principalement sur des 'définitions' et la seconde sur la description de propriétés.

1.2 Première méthode

Un objet (individu) est dit *classique (numérique)* s'il peut se représenter par un point de $\Re^p$ [1].
Sinon, il sera dit symbolique.
Un ensemble de données classiques se présentera donc sous forme d'un tableau

1. $\Re^p$ est considéré comme un espace vectoriel muni des opérations usuelles.

(n, p) où p est le nombre de variables[2] quantitatives (et parfois qualitatives) servant à décrire les n individus[3] sur l'ensemble desquels porte l'analyse.
Les objets symboliques 'évoquent' des données plus complexes. Comme nous l'avons souligné dans la partie 1, ces outils sont bien adaptés pour la représentation de connaissances et s'écrivent sous forme de conjonctions de propriétés décrites à l'aide de variables de l'analyse classique.
Il découle de ceci que l'analyse de données classiques traite (déjà depuis longtemps) des objets symboliques particuliers.

1.3 Seconde méthode

Nous nous proposons de différencier les objets symboliques des données classiques au moyen de propriétés des objets symboliques. Nous les classerons selon deux points de vue : la description et la manipulation.

1.3.1 Niveau de la description

Valuation des variables

Dans le cas des objets symboliques, chaque variable peut prendre des valeurs multiples pour un même objet symbolique afin de représenter soit des classes définies en intension (exprimant ainsi de la variation parmi les membres de ces classes), soit un individu (dénotant du doute à propros de celui-ci).
Ainsi, par exemple, [couleur $=$ $\{roux, brun\}$] exprimera que la couleur d'un *chien* pourra être *rousse* ou *brune* et [température $=$ [38, 39]] signifiera qu'un *chien* possède une température comprise entre *38* et *39* degrés.
Par contre, un objet classique ne prend qu'une *seule* valeur pour chacune des variables servant à le décrire.
De façon équivalente, nous dirons que les objets symboliques sont définis en intension plutôt qu'en extension. Ce sont des objets qui *unifient*[4] contrairement

2. Les valeurs prises par chacune des variables sur l'ensemble des individus est un vecteur colonne $(n, 1)$.

3. Un individu est représenté par un vecteur ligne $(1, p)$.

4. Ainsi, un même objet désignera parfois plusieurs individus.

aux objets de l'analyse de données classique qui caractérisent un seul individu à la fois.
Voici deux exemples pour aider à la compréhension.

- Nous parlerons des habitudes *des* clients de mon épicerie à la place de décrire les marottes *d'un* seul client bien précis.

- Nous décrirons toute une collection de portraits au lieu d'étudier uniquement la Joconde.

Liens logiques et liens entre individus

- Il sera possible de traduire des liens entre les valeurs prises par les descripteurs. En d'autres termes, si des implications logiques existent entre V_i et V_j, elles pourront apparaître dans l'expression de l'objet symbolique utilisé. Par exemple, il se produit que certains descripteurs n'aient aucun sens si d'autres en ont.

- Au moyen des objets hordes, des liens connus entre plusieurs individus pourront s'exprimer dans un seul objet.

1.3.2 Manipulation

En utilisant des opérations d'union, d'intersection et de complémentation particulières, les objets modaux permettent de prendre en compte des informations sur le domaine d'application.
Ceci permettra, par exemple, d'éviter de généraliser *un chien qui mange des hamburgers et du pain* et *un chien qui mange du jambon et du pain* par *un chien qui mange du pain.* Nous généraliserons plutôt par *un chien qui mange de la viande et du pain.*

Chapitre 2

Les différents types d'analyse de données

Nous pouvons citer quatre types. Les frontières entre ceux-ci ne sont cependant pas clairement définies.

2.1 Analyse de données classiques

Nous traitons des données quantitatives et qualitatives par des méthodes numériques utilisant l'algèbre linéaire et les outils statistiques usuels.

2.2 Analyse numérique de données symboliques

2.2.1 Principe

Une mesure est introduite sur les valeurs prises par les variables. La théorie des probabilités et celle de la mesure deviennent alors utilisables.

2.2.2 Exemple

Utilisation des distances entre objets pour faire une classification ou encore une analyse en composantes principales.

2.2.3 Inconvénients

Comment tenir compte des connaissances supplémentaires (méthodes, propriétés, sémantiques ...)?

2.3 Analyse symbolique de données classiques

2.3.1 Principe

Les tableaux de l'analyse de données classiques sont traités par l'approche symbolique. Ainsi, on utilise sur ces données des notions telles l'extension, la généralisation ...
Il se peut encore que nous ne l'employons pas directement l'approche symbolique sur les données initiales mais que nous commencions d'abord par leur appliquer une méthode classique de l'analyse des données (ce qui est le cas dans l'exemple qui suit).

2.3.2 Exemple

Extraire les variables les plus explicatives d'un axe factoriel et les deux classes d'individus les plus contributifs à chaque extrémité de l'axe.
Considérer l'ensemble des objets symboliques associés à ces variables.
Déterminer dans cet ensemble des objets symboliques complets et d'effritement minimun qui soient caractéristiques à chacune des classes.
Trouver les objets (toujours dans cet ensemble) de meilleure stabilité qui minimisent le recouvrement de la partition associée à ces classes.

2.4 Analyse symbolique des données symboliques

2.4.1 Principe

L'approche symbolique est usitée pour étudier des données qui le sont aussi.

2.4.2 Exemple

Soient S un ensemble d'objets symboliques et $S_1 \subset S$,
on demande de déterminer une partition de S_1 qui maximise la stabilité et minimise le recouvrement. On désire aussi une hiérarchie de classes stables et de meilleure héritance ainsi que des assertions d'effritement minimun permettant de recouvrir chaque classe.

2.4.3 Remarques

En réalité, il serait idéal de pouvoir utiliser les objets symboliques en entrée comme en sortie des opérations ce qui permettrait de perdre le moins d'informations possible.
Cette perte est généralement causée par des modélisations ou codages arbitraires. Par exemple, pour représenter l'ensemble des valeurs contenues dans un intervalle, on choisit habituellement d'en prendre la moyenne et la variance.
De plus, les résultats exprimés sous forme symbolique ont un très grand pouvoir explicatif par eux-mêmes (leur interprétation est assez facile). Les objets symboliques utilisés tant en entrée qu'en sortie constituent les éléments de l'analyse des données symboliques.

Tableau récapitulatif des quatres approches.

Données	classiques	symboliques
Analyse numérique	1	2
Analyse symbolique	3	4

'Perspectives'

L'objectif de ces descriptions était de souligner que les techniques classiques (l'ACP et les méthodes de classification) utilisées sur des objets symboliques dans la partie 4 constitueront un exemple d'analyse de type 2.
Avant d'en arriver là, il nous semble utile de rappeler les définitions de ces outils

de l'analyse classique et de les illustrer au moyen de quelques exemples.
Ce sera l'objectif de la partie suivante.

Troisième partie

Outils de l'analyse de données classique

Introduction

Nous allons rappeler quelques techniques employées en analyse de données classique.
Nous commencerons par la description globale de deux modèles de classification hiérarchique agglomérative. Ensuite, nous exposerons les grands principes de l'Analyse en Composantes Principales (ACP). Ces techniques nécessitent la définition d'une distance.
L'étape suivante (partie 3) sera la construction d'une distance sur les objets symboliques (booléens) afin de pouvoir leur appliquer ces méthodes

Chapitre 1

Méthodes de classification

1.1 Introduction

1.1.1 Problème posé

Le problème posé est de répartir une population donnée d'individus ou d'objets, décrite par un ensemble de caractéristiques (variables), en un nombre optimal de groupes homogènes. Ces groupes seront désignés par le terme *classes*. Les membres de chaque classe ont en commun certaines caractéristiques qui les distinguent des membres des autres classes.

Exemples

Voici quelques exemples de classes que l'on peut retrouver dans différentes disciplines

- *Psychologie :*
  - classe des psychopathes;
  - classe des psychotiques;
  - classe des névrosés;
  - etc ...

- *Zoologie*
  - classe des vertébrés;
  - classe des invertébrés.

1.1.2 Utilité

Cette répartition en classes rend réalisable une synthèse de l'information contenue dans les données. De plus, l'appartenance d'un individu à une de ces classes permet d'en préciser les caractéristiques, d'en prévoir le comportement ...

Exemples

- *Géologie :*
  classification d'un ensemble d'échantillons de roches dans le but de les dater;
- *Enseignement :*
  classification des élèves afin de prédire leurs chances de réussite;
- *Banques :*
  classification des différents types de clients pour savoir lesquels ont la plus grande prédisposition à honorer leur prêt.

1.1.3 Caractéristiques générales

1. **But.** Simplifier une réalité complexe par la constitution de groupes d'objets ou d'individus *semblables.*
2. **Inexistence d'à priori.** Il n'existe pas de classification à priori. Les classes sont inconnues ainsi que leur nombre et effectif.
3. **Attributs.** Chaque objet ou individu est caractérisé par un ensemble de variables ou attributs mesurables.
4. **Structure.** L'analyse des données essaie de rendre explicite la structure interne des données avec le maximun d'objectivité et le minimun d'hypothèses. Il ne pourra pas y avoir d'hypothèse privilégiant une variable aux dépens des autres.

5. **Simplification.** La structure mise à jour devrait permettre de formuler des hypothèses propres à simplifier une tâche répétitive (ex: grands groupes, grands échantillons, ...), à améliorer l'efficacité d'une action entreprise (exemples: statistique sur le nombre de réussites universitaires, statistique sur la datation de roches, ...).

1.1.4 Etapes de la classification

1. Collecte des données sous forme d'une matrice reprenant les individus décrits par des variables;

$$\begin{pmatrix} x_{11} & x_{12} & \dots & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & \dots & x_{2p} \\ \vdots & & & & \vdots \\ x_{np} & x_{n2} & \dots & \dots & x_{np} \end{pmatrix}$$

   avec les lignes $x_{i.}$ $i = 1, 2, \dots, n$ désignant les individus et
   les colonnes $x_{.j}$ $j = 1, 2, \dots, p$ désignant les variables (attributs).

2. Calcul de la proximité entre ces individus et construction d'une matrice de proximité;

3. Constitution des groupes au moyen de la méthode sélectionnée;

4. Interprétation des résultats et description des groupes;

5. Validation des résultats: nous examinons les propriétés et qualités de la classification obtenue (ex: tester la stabilité).

1.1.5 'Perspectives'

Les méthodes de classification étant assez nombreuses, nous nous pencherons uniquement sur le cas des techniques dites hiérarchiques agglomératives et en particulier la méthode du voisin le plus proche et celle du plus éloigné. Ce choix se justifie par le fait que nous les appliquerons sur les objets symboliques booléens dans la partie 3.

1.2 Classification hiérarchique agglomérative

1.2.1 Description générale

On part du principe suivant :
Soit un ensemble B comprenant N éléments (ou classes).
A chaque étape de la classification, nous regrouperons les deux classes les plus proches.
Ainsi, nous aurons :

N classes à l'étape 0 (chaque classe constituée d'un seul individu);
N-1 classes à l'étape 1;
... ...
... ...
1 classe (B tout entier) à l'étape N-1.

Etant donné qu'il existe plusieurs notions de proximité, nous utiliserons différentes procédures. Parmi ces techniques, nous trouvons celle du voisin le plus proche, du voisin le plus éloigné, du centroïde, de Ward, ... Nous étudierons plus en détail les deux premières. Nous rappelerons les distances utilisées pour quelques-unes des autres.

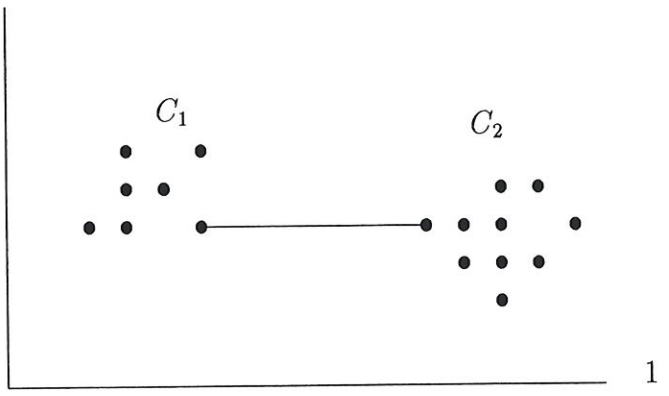
1.2.2 Voisin le plus proche

1.2.2.1 Distance

Soient les classes C_i, C_j, la distance interclasse du *Voisin le plus proche* sera

$$d_{C_i,C_j} = min\, d(x, y) \quad x \in C_i; y \in C_j$$

où $d(x, y)$ est une distance (exemple : distance euclidienne, distance définie par la norme infinie, ...).



Distance interclasse du voisin le plus proche.

La distance (entre classes) utilisée est celle construite à partir de la norme infinie (entre individus). Ainsi,

$$\begin{aligned} d(C_1, C_2) &= min\, d(x, y) \\ &= min(max_{i=1,\dots N} \mid x_i - y_i \mid). \end{aligned}$$

Ici $N = 2$.

1.2.2.2 Application

Soit $E = \{a, b, c, d, e\}$ l'ensemble des individus, la matrice de distance est donnée par

$D_0 =$

	$\{a\}$	$\{b\}$	$\{c\}$	$\{d\}$	$\{e\}$
$\{a\}$	0	2	6	10	9
$\{b\}$	2	0	5	9	8
$\{c\}$	6	5	0	4	6
$\{d\}$	10	9	4	0	10
$\{e\}$	9	8	6	10	0

Etape 0 :

$$P_0 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}\}.$$

Etape 1 :

$$d_{\{a\},\{b\}} = min\, d(\{i\}, \{j\}) \quad i, j \in \{a, b, c, d, e\}, \ i \neq j$$

Nous regroupons donc les objets a et b.

$$P_1 = \{\{a, b\}, \{c\}, \{d\}, \{e\}\}.$$

Etape 2 :

Nous calculons les distances :

$$\begin{aligned} d_{\{a,b\},\{c\}} &= min\{d_{\{a\},\{c\}}, d_{\{b\},\{c\}}\} = d_{\{a\},\{c\}} = 5 \\ d_{\{a,b\},\{d\}} &= min\{d_{\{a\},\{d\}}, d_{\{b\},\{d\}}\} = d_{\{b\},\{d\}} = 9 \\ d_{\{a,b\},\{e\}} &= min\{d_{\{a\},\{e\}}, d_{\{b\},\{e\}}\} = d_{\{b\},\{e\}} = 8 \end{aligned}$$

Nous obtenons alors la matrice

$D_2 =$

	$\{a, b\}$	$\{c\}$	$\{d\}$	$\{e\}$
$\{a, b\}$	0	5	9	8
$\{c\}$	5	0	4	6
$\{d\}$	9	4	0	10
$\{e\}$	8	6	10	0

La plus petite distance est $d_{\{c\},\{d\}}$, nous fusionnons donc $\{c\}$ et $\{d\}$. Ainsi :

$$P_2 = \{\{a, b\}, \{c, d\}, \{e\}\}$$

Etape 3 :

Nous calculons les distances

$$\begin{aligned} d_{\{a,b\},\{e\}} &= 8 \\ d_{\{a,b\},\{c,d\}} &= min\{d_{\{a\},\{c\}}, d_{\{a\},\{d\}}, d_{\{b\},\{c\}}, d_{\{b\},\{d\}}\} = d_{\{b\},\{c\}} = 5 \\ d_{\{c,d\},\{e\}} &= min\{d_{\{c\},\{e\}}, d_{\{d\},\{e\}}\} = d_{\{c\},\{e\}} = 6 \end{aligned}$$

Nous obtenons alors la matrice

$D_3 =$

	$\{a, b\}$	$\{c, d\}$	$\{e\}$
$\{a, b\}$	0	5	8
$\{c, d\}$	5	0	6
$\{e\}$	8	6	0

La plus petite distance est $d_{\{a,b\},\{c,d\}}$, nous fusionnons donc $\{a, b\}$ et $\{c, d\}$. Ainsi :

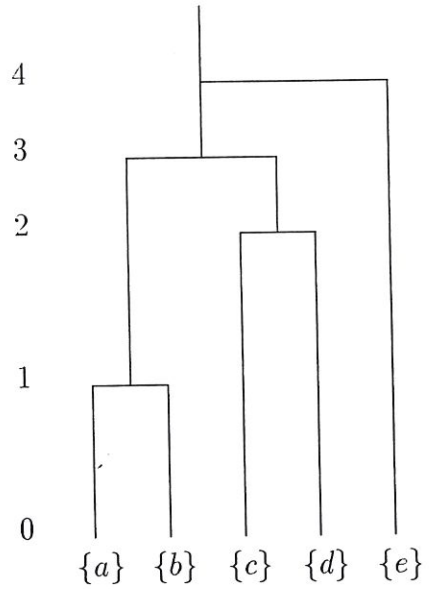
$$P_3 = \{\{a, b, c, d, \{e\}\}.$$

Etape 4 :

Nous rassemblons les deux dernières classes obtenues.

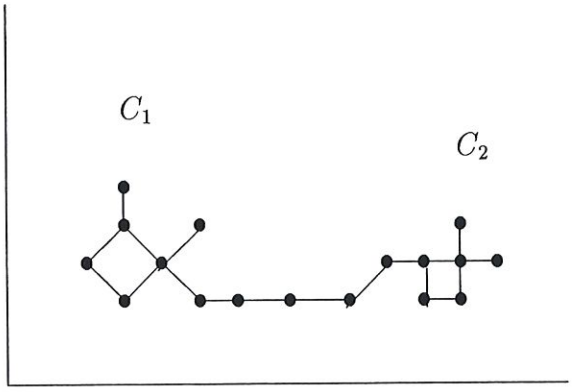
$$P_4 = \{\{a, b, c, d, e\}\}.$$

Arbre hiérarchique :



1.2.2.3 Inconvénient

Il existe un effet de chaînage. Comme l'illustre la figure ci-dessous, lorsqu'il y a un pont entre deux classes, cette méthode n'en trouve qu'une.



Il existe un pont entre ces 'deux' classes.

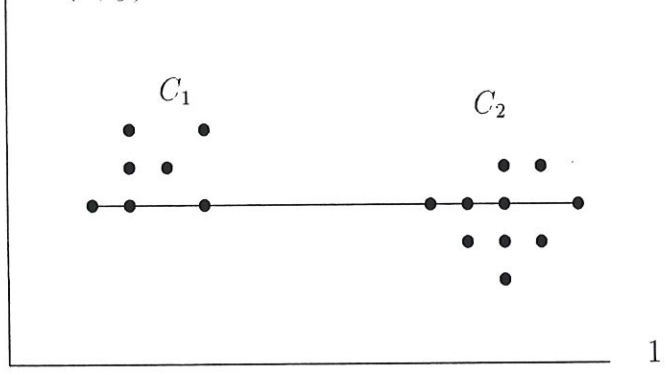
1.2.3 Voisin le plus éloigné

1.2.3.1 Distance

Soient les classes C_i, C_j, la distance interclasse du *Voisin le plus éloigné* sera

$$d_{C_i,C_j} = max\, d(x,y) \quad x \in C_i; y \in C_j$$

où $d(x, y)$ est une distance.



Distance interclasse du voisin le plus éloigné.

La distance (entre classes) utilisée est celle construite à partir de la norme infinie (entre individus). Ainsi,

$$\begin{aligned} d(C_1, C_2) &= max\, d(x,y) \\ &= max(max_{i=1,\ldots N} \mid x_i - y_i \mid \end{aligned}$$

Ici $N = 2$.

1.2.3.2 Application

Soit $E = \{a, b, c, d, e\}$, l'ensemble des individus. La matrice des distances est donnée par :

$D_0 =$

	$\{a\}$	$\{b\}$	$\{c\}$	$\{d\}$	$\{e\}$
$\{a\}$	0	2	6	10	9
$\{b\}$	2	0	5	9	8
$\{c\}$	6	5	0	4	6
$\{d\}$	10	9	4	0	10
$\{e\}$	9	8	6	10	0

Etape 0 :

$$P_0 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}\}.$$

Etape 1 :

$$d_{\{a\},\{b\}} = min\, d(\{i\}, \{j\}) \quad i, j \in \{a, b, c, d, e\},\ i \neq j$$

Nous regroupons donc les objets a et b.

$$P_1 = \{\{a, b\}, \{c\}, \{d\}, \{e\}\}.$$

Etape 2 :

Nous calculons les distances :

$$d_{\{a,b\},\{c\}} = max\{d_{\{a\},\{c\}}, d_{\{b\},\{c\}}\} = d_{\{a\},\{c\}} = 6$$
$$d_{\{a,b\},\{d\}} = max\{d_{\{a\},\{d\}}, d_{\{b\},\{d\}}\} = d_{\{a\},\{d\}} = 10$$
$$d_{\{a,b\},\{e\}} = max\{d_{\{a\},\{e\}}, d_{\{b\},\{e\}}\} = d_{\{a\},\{e\}} = 9$$

Nous obtenons alors la matrice

$D_2 =$

	$\{a, b\}$	$\{c\}$	$\{d\}$	$\{e\}$
$\{a, b\}$	0	6	10	9
$\{c\}$	6	0	4	5
$\{d\}$	10	4	0	3
$\{e\}$	9	5	3	0

La plus petite distance est $d_{\{c\},\{d\}}$, nous fusionnons donc $\{c\}$ et $\{d\}$. Ainsi :

$$P_2 = \{\{a, b\}, \{c, d\}, \{e\}\}$$

Etape 3 :

Nous calculons les distances

$$\begin{aligned} d_{\{a,b\},\{c,d\}} &= max\{d_{\{a\},\{c\}}, d_{\{a\},\{d\}}, d_{\{b\},\{c\}}, d_{\{b\},\{d\}}\} \\ &= max\{6, 10, 5, 9\} = 10 \\ d_{\{c,d\},\{e\}} &= max\{d_{\{c\},\{e\}}, d_{\{d\},\{e\}}\} = max\{5, 10\} = 10 \end{aligned}$$

Nous obtenons alors la matrice

$D_3 =$

	$\{a, b\}$	$\{c, d\}$	$\{e\}$
$\{a, b\}$	0	10	9
$\{c, d\}$	10	0	10
$\{e\}$	9	10	0

La plus petite distance est $d_{\{a,b\},\{e\}}$, nous fusionnons donc $\{a, b\}$ et $\{e\}$. Ainsi :

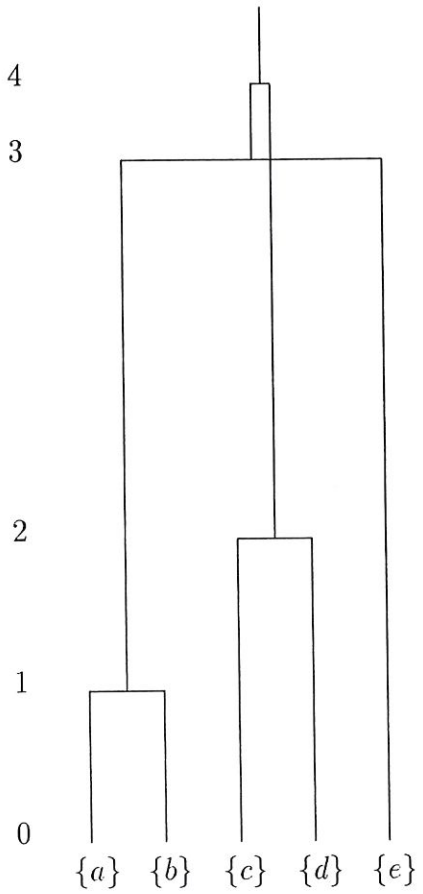
$$P_3 = \{\{a, b, e\}, \{c, d\}\}.$$

Etape 4 :

Nous rassemblons les deux dernières classes obtenues.

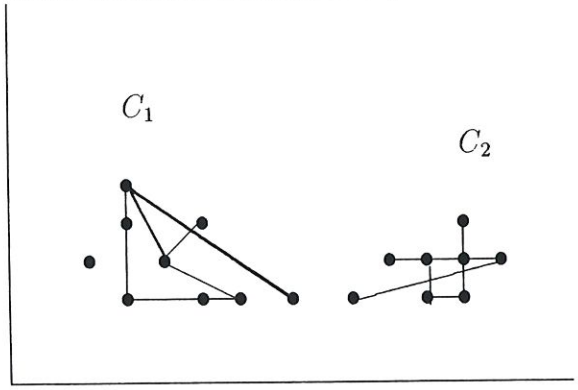
$$P_4 = \{\{a, b, c, d, e\}\}.$$

Arbre hiérarchique :



1.2.3.3 Inconvénient

Cette méthode construit toujours des classes hypersphériques.



Les classes sont 'sphériques'.

1.2.4 Remarque

Les classifications fournies par ces deux méthodes sont différentes.
Ceci est assez intuitif puisque nous avons utilisé des distances interclasses différentes.

1.2.5 Description sommaire d'autres méthodes disponibles

1.2.5.1 Centroïde

Dans ce cas, la distance interclasse utilisée est la distance entre leur centroïde (centre de gravité). En d'autres mots, à chaque étape, nous fusionnons les deux groupes dont les centroïdes sont les plus proches.

1.2.5.2 Médiane

Elle part du même principe que la précédente. Toutefois, nous ajoute une condition : tous les groupes doivent être de taille identique.

1.2.5.2 Ward

A chaque étape, nous assemble les classes pour lesquelles nous observons le plus petit accroissement de *la somme des carrés des déviations de chaque individu par rapport à la moyenne de la classe.*

Chapitre 2

Analyse en Composantes Principales (ACP)

2.1 Introduction

L'ACP est parmi les plus anciennes et les plus usitées des techniques multivariées. Elle nécessite comme données de départ une matrice $n \times p$ notée $\tilde{X}$ qui représente n individus qualifiés par p variables.

2.1.1 Exemple

$$\tilde{X} = \begin{pmatrix} 160 & 28 & 141 \\ 572 & 537 & 748 \\ 441 & 404 & 434 \\ 783 & 1114 & 1464 \end{pmatrix}$$

Si nous désignons par $x_{i.}$ les lignes de cette matrice et par $x_{.j}$ les colonnes, les $n = 4$ individus seront notés $x_{i.}$ $i = 1, 2, 3, 4$ et les $p = 3$ variables seront notées $x_{.j}$ $j = 1, 2, 3$. Dans cet exemple, les individus représentent des types de professions et les variables le nombre de personnes ayant choisi un style identique de logement de vacances.

$x_{1.}$:Agriculteurs;

$x_{2.}$:Cadres;

$x_{3.}$:Employés;

$x_{4.}$:Ouvriers;

$x_{.1}$:Hotel, pension de famille;

$x_{.2}$:Maison louée;

$x_{.3}$:Camping (tente, caravane).

2.1.2 Idée de base

Le principe initial de l'ACP est de décrire les variations contenues dans les données en termes d'un ensemble de nouvelles variables *non corrélées*. Chacune est une combinaison linéaire des variables originales. Ces nouvelles variables sont obtenues par ordre décroissant d'importance de telle sorte que la première *composante principale* prenne en compte la plus grande part possible de la variation des données originales. L'objectif de l'ACP est alors de voir si quelques-unes seulement des premières composantes prennent en compte la quasi-totalité de cette variation. Si tel est le cas, ces premières composantes seront employées afin de synthétiser les données avec une perte minime d'information.

2.2 Etapes de l'ACP

2.2.1 Etape 1 : centrer, réduire

Dans la plupart des cas, il est préférable de centrer et réduire la matrice des données. Cette démarche est primordiale lorsqu'on est confronté à des données hétérogènes : certaines variables peuvent être des effectifs alors que d'autres sont des pourcentages.

Pour ce faire, il suffit de donner même moyenne (nulle) et même variance (unité) à toutes les colonnes. Les colonnes auront ainsi toutes une importance identique; ceci permettra d'éviter que les celles correspondant aux variables de plus grande variance masquent les autres.

Nous procèdons comme suit :

2.2.1.1 Marche à suivre

1. **Centrer**

(a) Nous calculons la matrice *moyenne* $\tilde{X}_0$. Cette matrice exprimera la situation de n individus idèntiques dans p variables de telle manière que ces derniers prennent tous les valeurs moyennes de ces variables. Ainsi :

$$\tilde{X}_0 = \frac{1}{n} I_n \tilde{X}.$$

avec I_n qui est une matrice carrée de dimension n telle que

$$I_n = \begin{pmatrix} 1 & 1 & \dots & \dots & 1 \\ 1 & 1 & \dots & \dots & 1 \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ 1 & 1 & \dots & \dots & 1 \end{pmatrix}.$$

(b) L'analyse porte sur les écarts à la moyenne, colonne par colonne. Nous calculons ceux-ci par la formule $X = \tilde{X} - \tilde{X}_0$. Cette opération correspond à un centrage des variables. Nous remarquerons qu'une ligne *nulle* caractérise *l'individu moyen parfait* parmi l'ensemble des individus étudiés.

(c) La norme utilisée après centrage des variable est la norme euclidienne usuelle. Dès lors, la variation totale est représentée par *la somme des carrés des écarts à la moyenne*, chaque variable participant selon la somme des carrés de ses composantes (ces valeurs sont obtenues en calculant le produit matriciel $X'X$).
En divisant X par $\sqrt{n}$, le produit $\frac{X'X}{\sqrt{n}\sqrt{n}} = X'X$ sera une matrice de variance-covariance. Ce résultat nous simplifiera la tâche. La matrice X aura la forme suivante :

$$X = \frac{1}{\sqrt{n}}(\tilde{X} - \tilde{X}_0) = (x_{ij}),$$

où

$$x_{ij} = \frac{\tilde{x}_{ij} - \bar{x}_{.j}}{\sqrt{n}},$$

avec

- $\tilde{x}_{ij}$ la valeur à l'origine de la variable j sur l'individu i,
- $\bar{x}_{.j}$ la moyenne des valeurs prise par la variable j sur les n individus.

2. **Réduire**

Nous ramènons la variance de toutes les variables à 1.

Cela se fait en divisant les termes génériques de la matrice à réduire, ici $\frac{\tilde{X} - \tilde{X}_0}{\sqrt{n}}$, par la racine de la variance de la colonne s_j à laquelle ils appartiennent c'est-à-dire

$$x_{ij} = \frac{\tilde{x}_{ij} - \bar{x}_{.j}}{\sqrt{n}\, s_j}.$$

avec s_j^2 qui est la variance de la j-ème variable. Elle se calcule par la formule :

$$s_j^2 = \sum_{i=1}^{n} \frac{(\tilde{x}_{ij} - \bar{x}_{.j})^2}{n}.$$

2.2.1.2 Exemple

Nous reprenons la matrice de données de l'exemple introductif.

$$\tilde{X} = \begin{pmatrix} 160 & 28 & 141 \\ 572 & 537 & 748 \\ 441 & 404 & 434 \\ 783 & 1114 & 1464 \end{pmatrix}$$

Ici $n = 4$ et $p = 3$.

Nous commençons par calculer les $\bar{x}_{.j}$.

$$\bar{x}_{.1} = 489;$$
$$\bar{x}_{.2} = 520.75;$$
$$\bar{x}_{.3} = 696.75.$$

De ce résultat, nous pouvons tirer que

$$\tilde{X}_0 = \begin{pmatrix} 489 & 520.75 & 696.75 \\ 489 & 520.75 & 696.75 \\ 489 & 520.75 & 696.75 \\ 489 & 520.75 & 696.75 \end{pmatrix}$$

Et, puisque, $\frac{1}{\sqrt{n}} = \frac{1}{2}$, nous trouvons que :

$$X = \frac{1}{2}\begin{pmatrix} -329 & -492.75 & -555.75 \\ 83 & -16.25 & 51.25 \\ -48 & -116.75 & -265.75 \\ 296 & 593.25 & 767.25 \end{pmatrix}$$

Calculons à présent les variances s_j^2 :

$$\begin{aligned} s_1^2 &= 51262.50; \\ s_2^2 &= 152160.69; \\ s_3^2 &= 242695.06, \end{aligned}$$

et donc,

$$\begin{aligned} s_1 &= 226.41; \\ s_2 &= 390.08; \\ s_3 &= 492.64. \end{aligned}$$

Finalement, la matrice centrée réduite X vaudra :

$$X = \begin{pmatrix} -0.727 & -0.632 & -0.564 \\ 0.183 & -0.021 & 0.052 \\ -0.106 & -0.150 & -0.270 \\ 0.654 & 0.760 & 0.779 \end{pmatrix}$$

2.2.2 Etape 2 : Calcul des composantes principales

Les composantes principales sont de nouvelles variables composites construites à partir des vecteurs propres de la matrice de corrélation. Elles permettent de récupérer de manière hiérarchique le maximun de la variance totale des n individus dans l'espace à p dimensions. En d'autres termes, la première de ces composantes (vecteurs) est celle qui reprend la plus grande partie de la variance, la seconde la plus grande partie de ce qu'il en reste après retrait de celle prise en compte par la première ... et la dernière est celle qui contient le reste (non nul) de variance des données non encore prise en compte.

La première composante sera donc égale au vecteur Xu_1 à n lignes où u_1 est le vecteur propre unitaire correspondant à la plus grande valeur propre de la matrice

$X'X$ de corrélation. Nous terminerons par la projection des données dans l'espace dont les axes sont les composantes principales. Ceci permettra peut-être de trouver une structure et donc de classifier et synthétiser les informations fournies par ces données.

2.2.2.1 Marche à suivre

1. **Calcul de la matrice de corrélation**
   Il s'agit simplement du produit $X'X$ avec X *centrée réduite.* La matrice ainsi obtenue est carrée de dimension p.

2. **Valeurs propres**
   Nous cherchons les p valeurs propres λ_i de la matrice de corrélation et nous les rangeons par ordre croissant.
   Elles existent toujours et sont non-négatives. En effet, la définition de cette matrice implique qu'elle est définie positive.

3. **Vecteurs propres**
   A chaque valeur propre λ_i correspond un vecteur propre (ou k vecteurs propres si la multiplicité algèbrique de λ est k). Deux vecteurs associés à des valeurs différentes ont la propriété d'être orthogonaux.

4. **Calcul des composantes principales**
   Une fois les vecteurs propres u_i déterminés, il ne reste plus qu'à faire les produits matriciels $Xu_i \quad i = 1, \ldots, p$ pour obtenir les p composantes principales.
   Nous aurons ainsi :

   première composante principale : Xu_1;
   seconde composante principale : Xu_2;
   $\vdots \quad \vdots$
   dernière composante principale : Xu_p.

   Etant donné que les vecteurs propres sont deux à deux orthogonaux, les différentes composantes principales sont non corrélées entre elles.

5. **Projection**

   Nous projettons ensuite les données dans le plan (Xu_1, Xu_2). Les vecteurs directeurs de ce plan sont les directions dans lesquelles l'ensemble des données est le plus *étendu.* La figure obtenue par projection est dès lors telle que nous minimisons la superposition des points. Elle permettra donc de rendre compte de la stucture générale des données et d'entrevoir le mieux possible les groupes qui pourraient exister au sein de la population.

Remarques

1. Nous aurions pu nous contenter de projeter les données sur les variables initiales de plus grande variance. Mais il n'est pas certain qu'en se servant d'une combinaison linéaire de ces variables, nous ayons pu pu trouver une autre variable qui prenne en compte une plus grande part de la variance totale.

2. En général. cette méthode est appliquée sur des ensembles de données caractérisées par des variables *quantitatives.*

2.2.2 Exemple

Soit la matrice de données

$$\tilde{X} = \begin{pmatrix} 1 & 0 \\ 2 & 2 \\ 3 & 4 \\ 2 & 6 \end{pmatrix}.$$

$n = 4$, $p = 2$.
Nous calculons $\tilde{X}_0$

$$\tilde{X}_0 = \begin{pmatrix} 2 & 3 \\ 2 & 3 \\ 2 & 3 \\ 2 & 3 \end{pmatrix}.$$

Ainsi, la matrice centrée est

$$X = \frac{1}{2}\begin{pmatrix} -1 & -3 \\ 0 & -1 \\ 1 & 1 \\ 0 & 3 \end{pmatrix}.$$

Les variances sont :

$$\begin{aligned} s_1^2 &= \frac{1}{2}; \\ s_2^2 &= 5. \end{aligned}$$

Ainsi,

$$\begin{aligned} s_1 &= \frac{1}{\sqrt{2}}; \\ s_2 &= \sqrt{5}. \end{aligned}$$

La matrice centrée et réduite est donc :

$$X = \begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{3}{2\sqrt{5}} \\ 0 & -\frac{1}{2\sqrt{5}} \\ \frac{\sqrt{2}}{2} & \frac{1}{2\sqrt{5}} \\ 0 & \frac{3}{2\sqrt{5}} \end{pmatrix}$$

et celle de corrélation :

$$X'X = \begin{pmatrix} 1 & \frac{\sqrt{2}}{\sqrt{5}} \\ \frac{\sqrt{2}}{\sqrt{5}} & 1 \end{pmatrix}.$$

Les valeurs propres λ_1, λ_2 obtenues en annulant le déterminant suivant :

$$\begin{vmatrix} 1-\lambda & \frac{\sqrt{2}}{\sqrt{5}} \\ \frac{\sqrt{2}}{\sqrt{5}} & 1-\lambda \end{vmatrix}.$$

seront

$$\begin{aligned} \lambda_1 &= 1 + \sqrt{\frac{2}{5}}; \\ \lambda_2 &= 1 - \sqrt{\frac{2}{5}}. \end{aligned}$$

Les vecteurs propres associés sont

$$u_1 = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$$

et

$$u_2 = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix}.$$

Nous trouvons les composantes principales en faisant les produits matriciels Xu_1, Xu_2 :

$$Xu_1 = \begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{3}{2\sqrt{5}} \\ 0 & -\frac{1}{2\sqrt{5}} \\ \frac{\sqrt{2}}{2} & \frac{1}{2\sqrt{5}} \\ 0 & \frac{3}{2\sqrt{5}} \end{pmatrix} \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} - \frac{3\sqrt{2}}{4\sqrt{5}} \\ -\frac{\sqrt{2}}{4\sqrt{5}} \\ \frac{1}{2} + \frac{\sqrt{2}}{4\sqrt{5}} \\ \frac{3\sqrt{2}}{4\sqrt{5}} \end{pmatrix}$$

$$Xu_2 = \begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{3}{2\sqrt{5}} \\ 0 & -\frac{1}{2\sqrt{5}} \\ \frac{\sqrt{2}}{2} & \frac{1}{2\sqrt{5}} \\ 0 & \frac{3}{2\sqrt{5}} \end{pmatrix} \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} + \frac{3\sqrt{2}}{4\sqrt{5}} \\ -\frac{\sqrt{2}}{4\sqrt{5}} \\ \frac{1}{2} - \frac{\sqrt{2}}{4\sqrt{5}} \\ -\frac{3\sqrt{2}}{4\sqrt{5}} \end{pmatrix}$$

Chapitre 3

'Perspectives'

La question qui se pose à présent est : *Comment modifier ces techniques afin qu'elles deviennent applicables sur des données symboliques?*
Ainsi, par exemple sur des données de ce genre :

Objet	Poids spécifique	Température de solidification	Teneur en iode	Teneur en minerais	Acides gras principaux
1) Huile de lin	0.930-0.935	(-27)-(-18)	170-204	118-196	L, Ln, O, P, M
2) Huile dee Périlla	0.930-0.937	(-5)-(-4)	192-208	188-197	L, Ln, O, P, S
3) Huile de coton	0.916-0.918	(-6)-(-1)	99-113	189-198	L, O, P, M, S
4) Huile de sésame	0.920-0.926	(-6)-(-4)	104-116	187-193	L, O, P, S, A
5) Huile de camélia	0.916-0.917	(-21)-(-15)	80-82	189-193	L, O
6) Huile d'olive	0.914-0.919	0-6	79-90	187-196	L, O, P, S
7) Graisse de boeuf	0.860-0.870	30-38	40-48	190-199	O, P, M, S, C
8) Graisse de porc	0.858-0.864	22-32	53-77	190-202	L, O, P, M, S, Lu

Ceci fera l'objet de la partie suivant.

Quatrième partie

Notion de proximité, de distance entre objets booléens et représentation dans le plan

Introduction

En fait, l'élément qui nous manque pour parvenir à utiliser les méthodes vues dans la Partie 2, c'est une distance entre objets symboliques booléens. Nous allons donc essayer d'en construire une. Dans le Chapitre 1, nous ferons un premier essai qui n'aboutira pas sur une distance, mais simplement un indice de distance. Notre deuxième essai (Chapitre 2) sera plus concluant et nous obtiendrons la distance généralisée de Minkowski entre OAB. Afin de définir cette distance, il sera auparavant nécessaire d'introduire la notion de "Cartesian Space Model" (CSM). Nous poursuivrons en décrivant l'utilisation de la distance de Minkowski dans le cadre de la méthode du voisin le plus proche, du voisin le plus éloigné et de l'ACP. Ensuite, nous donnerons des façons graphiques de représenter les descriptions de classes relativement à d'autres classes. Il s'agira du MNG, des boîtes et de la silhouette. Enfin, nous terminerons par la présentation du théorème de prétendue simplicité.

Chapitre 1

Un indice de dissimilarité entre objets assertions booléens (OAB) basé sur l'extension

Notre indice de dissimilarité est basé sur une mesure positive: le **potentiel de description**. Il tiendra compte

- de la variabilité des V_i ; chacun de ceux-ci est soit un ensemble discret, soit une union d'intervalles selon le type de la variable y_i dont il est l'ensemble de valeurs; la conjonction de ces V_i définit l'OAB;
- des dépendances logiques entre variables.

1.1 Base de connaissances

Une base de connaissances (BC) comprend :

- l'ensemble des classes d'individus $\mathcal{C} = \{C_1, \cdots, C_N\}$;
- l'ensemble des variables $\mathcal{Y} = \{y_1, \cdots, y_d\}$;
- l'ensemble des ensembles d'observation $\mathcal{O} = \{O_1, \cdots, O_d\}$;
- l'ensemble de règles $\mathcal{R} = \{R_1, \cdots, R_t\}$ décrivant les Dépendances Logiques (DL) entre variables de $\mathcal{Y}$;

- l'ensemble d'OAB $\mathcal{A} = \{a_1, \cdots, a_N\}$ décrivant l'ensemble $\mathcal{C}$ des classes d'individus.

Dans une BC où il y a des DL entre les variables, la description d'une classe d'individus par un OAB est cohérente si elle ne contredit pas ces DL.

1.2 Le potentiel de description (PD) d'un OAB

1.2.1 Définition

Soit un OAB a_j tel que

$$\begin{aligned} a_j &= [y_1 = V_1^j] \wedge \cdots \wedge [y_d = V_d^j] \\ &= \wedge_{i=1}^{d} e_i^j \end{aligned}$$

où $e_i^j = [y_i = V_i^j]$ est le i-ème Événement Élémentaire Booléen (noté EEB).

- Nous définissons le **potentiel de description** (PD) d'un OAB a_j , comme la part $V^1 \times \cdots \times V_d^j$ du volume du produit cartésien $(O_1^j \times \cdots \times O_d^j)$ formée par les descriptions d'individus données par a_j qui sont cohérentes.
- Le PD de a_j est noté $\pi(a_j)$.

1.2.2 Calcul du potentiel de description d'un OAB

2.2.1. Cas 1. Absence de DL dans la BC

S'il n'y a pas de DL dans la BC, toutes les descriptions d'individus (fournies par a_j) sont cohérentes et $\pi(a_j)$ peut être calculé par l'expression suivante :

$$\pi(a_j) = \prod_{i=1}^{d} \pi(e_i^j) \tag{1.1}$$

où $\pi(e_i^j)$ est le PD de l'EEB e_i^j :

$$\pi(e_i^j) = \mu(V_i^j) = \begin{cases} \#(V_i^j)\,, & \text{si } y_i \text{ est discrète,} \\ \text{écart } (V_i^j)\,, & \text{si } y_i \text{ est continue.} \end{cases} \tag{1.2}$$

Dans cette seconde partie, V_i^j étant un ensemble d'intervalles, écart (V_i^j) est la somme, sur chaque intervalle, de la valeur absolue de la différence entre sa limite supérieure et sa limite inférieure.

2.2.2. Cas 2. Présence de DL dans la BC

Il est possible de démontrer que l'équation (1.1) peut être aussi employée afin de calculer $\pi(a_j)$ dans le cas où il existe des DL entre variables de la BC. Il faut procéder aux modifications suivantes :

- Chaque ensemble de valeurs de la variable appartenant à un graphe connexe[1] qui est donc dans la prémisse ou dans la conclusion d'une règle exprimant une DL entre variables, est comparée avec l'ensemble des valeurs de l'EEB défini par cette variable. L'ensemble des valeurs de l'EEB doit alors être scindé de la façon suivante :

  "Les valeurs du sous-ensemble qui ne sont originaires d'aucune règle ou sont toutes originaires de la prémisse ou de la conclusion de la même règle."

- Après avoir séparé l'ensemble des valeurs de chaque EEB défini par les variables qui sont dans le graphe, nous décomposons l'OAB original en plusieurs OAB. Leur nombre est donné par le produit des nombres de sous-ensembles associés à chacune des variables appartenant au graphe connexe.

L'équation (1.1) est ensuite appliquée sur ces OAB et cette procédure est répétée pour chaque graphe connexe.

1.2.3 Proposition 1

Dans le cas où il n'y a pas de DL entre les variables dans la BC,

$$\forall\ (a_j, a_k)\ :\ \pi(a_j \cup_a a_k) \geq \pi(a_j) + \pi(a_k) - \pi(a_j \cap_a a_k)$$

1. Rappels :

- Un graphe $G(X, U)$ (où X est l'ensemble des noeuds et U celui des arcs) est dit **connexe** si pour tout x, y ($x \neq y$), il existe une chaîne dont les deux extrémités sont x et y .
- Une **chaîne** est une séquence d'arcs $(u_1, u_2, \cdots, u_i, \cdots, u_q)$ telle que chaque arc u_i ($\in U$) est attaché à u_{i-1} par une de ses extrémités et à u_{i+1} par l'autre de ses extrémités. Cette définition est tirée de [14].

où $a_j \cup_a a_k$ *et* $a_j \cap_a a_k$ *sont définis comme suit :*

si $a_j = \wedge_{i=1}^{d} e_i^j$ *et* $a_k = \wedge_{i=1}^{d} e_i^k$ *leur union est*

$$\begin{aligned} a_m = a_j \cup_a a_k &= \wedge_{i=1}^{d}(e_i^j \cup_e e_i^k) \\ &= \wedge_{i=1}^{d}[y_i = V_i^j \cup V_i^k] \end{aligned}$$

et leur intersection est

$$\begin{aligned} a_n = a_j \cap_a a_k &= \wedge_{i=1}^{d}(e_i^j \cap_e e_i^k) \\ &= \wedge_{i=1}^{d}[y_i = V_i^j \cap V_i^k] \,. \end{aligned}$$

1.3 Calcul de la proximité entre OAB

Soient a_j et a_k deux objets assertions booléens (OAB) tels que

$$\begin{aligned} a_j &= \wedge_{i=1}^{d}[y_i = V_i^j] \,, \\ a_k &= \wedge_{i=1}^{d}[y_i = V_i^k] \,. \end{aligned}$$

Dans cette approche, la proximité entre deux OAB est basée sur leur PD. Plus précisément. l'idée centrale de cette approche est l'hypothèse que la dissimilarité entre deux OAB est fonction du nombre de descriptions d'individus propre à chaque OAB (et à lui seul) et du nombre de descriptions d'individus qui ne sont dans les descriptions ni de l'un ni de l'autre.

Nous proposons comme indice de proximité entre deux OAB :

$$d_{\text{ext}}(a_j, a_k) = \underbrace{\pi(a_j \cup_a a_k)}_{(1)} - \underbrace{\pi(a_j \cap_a a_k)}_{(2)}$$

où (1) est le volume tout entier (l'union cartésienne des OAB);
(2) est la part du volume correspondant aux descriptions communes.

Voici une interprétation de cet indice : Dans la Figure 1.1, le "volume tout entier" est le rectangle tout entier;
A est la part de "volume" correspondant aux descriptions communes à a_j et a_k ;

B est la part de "volume" correspondant aux descriptions propres à a_j ;
C est la part de "volume" correspondant aux descriptions propres à a_k ;
D est la part de "volume" correspondant aux descriptions ne vérifiant ni a_j ni a_k .

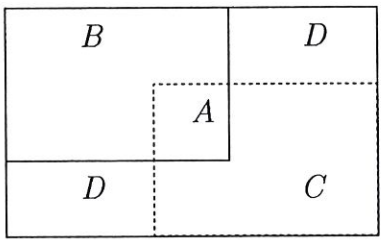


FIG. 1.1 – *Représentation cartésienne de deux OAB et de leur union symbolique.*

	"Accord" A	"Désaccord" A	Total A
"Accord" B	$A = \pi(a_j \cap_a a_k)$	$B = \pi(a_j) - \pi(a_j \cap_a a_k)$	$A + B = \pi(a_j)$
"Désaccord" B	$C = \pi(a_k) - \pi(a_j \cap_a a_k)$	$D = N - A - B - C$	$C + D = \pi(a_j \cup_a a_k) - \pi(a_j)$
Total B	$A + C = \pi(a_k)$	$B + D = \pi(a_j \cup_a a_k) - \pi(a_k)$	$N = A + B + C + D = \pi(a_j \cup_a a_k)$

TAB. 1.1 – *Comparaison entre a_j et a_k .*

Commentaires :

- Dans la case ("Accord" A,"Accord" B) se trouve le PD des descriptions pour lesquelles A **et** B sont vérifiés.
- Dans la case ("Accord" A,"Désaccord" B) se trouve le PD des descriptions telles que A est vérifié mais B ne l'est pas (même raisonnement pour ("Désaccord" A,"Accord" B)).
- Dans la case ("Désaccord" A,"Désaccord" B) se trouve le PD des descriptions telles que ni A ni B sont vérifiés.

- Dans la case (Total A,"Accord" B) se trouve le PD des descriptions telles que B est vérifié (même raisonnement pour (Total B,"Accord" A)).
- Dans la case (Total A,"Désaccord" B) se trouve le PD des descriptions telles que B n'est pas vérifié (même raisonnement pour (Total B,"Désacccord" A)).
- Dans la case (Total A,Total B) se trouve le PD de l'union cartésienne de ces deux OAB.

1.3.1 Proposition 2

Dans le cas où il n'y a pas de DL entre variables de la BC, d_{ext} *est un* ***indice de distance***.

Remarques :

1. Démontrer cette proposition implique de vérifier que, pour tout $(a_j\ a_k)$, d_{ext} répond aux propriétés suivantes :
   - $d_{\text{ext}}(a_j, a_k) = d_{\text{ext}}(a_k, a_j)$ (symétrie);
   - $d_{\text{ext}}(a_j, a_k) \geq 0$ (non négativité);
   - $d_{\text{ext}}(a_j, a_k) = 0 \Leftrightarrow a_j =_a a_k$,

   avec $a_j =_a a_k \Leftrightarrow \forall\, i \in \{1, \cdot, d\}$, $e_i^j =_e e_i^k \Leftrightarrow \forall\, i \in \{1, \cdot, d\}$, $\vee_i^j = \vee_i^k$.
2. En cas de DL, la propriété de non négativité n'est pas vérifiée.
3. Il ne s'agit pas d'une distance, mais bien d'un **indice** de distance : en effet, l'inégalité triangulaire n'est pas vérifiée.

1.3.2 Exemples

1. Cet exemple montre que d_{ext} ne vérifie pas l'inégalité triangulaire. Soient les OAB suivants qui caractérisent des villages :

$$a_j \quad = \quad [y_1 = [0, 50[] \wedge [y_2 = [0, 50[] \ ,$$

$$a_k = [y_1 = [100, 200[] \wedge [y_2 = [0, 50[] ,$$
$$a_l = [y_1 = [0, 50[] \wedge [y_2 = [100, 200[] ,$$

où y_1 désigne le nombre de chats dans le village, et
y_2 désigne le nombre de chiens dans ce village.

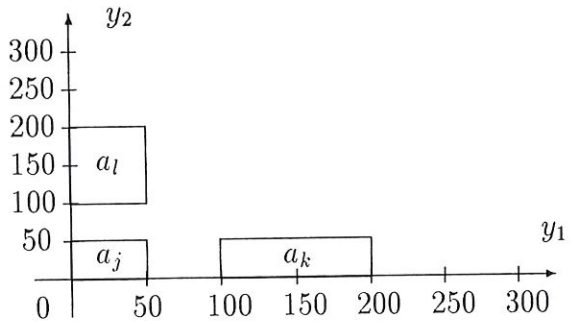


FIG. 1.2 – *Représentation dans le plan des OAB* a_j , a_k *et* a_l .

Alors.

$$a_j \cup_a a_k = [y_1 = [0, 50[\cup[100, 200]] \wedge [y_2 = [0, 50[] ,$$
$$a_j \cap_a a_k = [y_1 = \emptyset] \wedge [y_2 = [0, 50[] ;$$

dès lors.

$$\pi(a_j \cup_a a_k) = [(50 - 0) + (200 - 100)](50 - 0) = 7500 ,$$
$$\pi(a_j \cap_a a_k) = (0)(50 - 0) = 0 ;$$

de même,

$$a_j \cup_a a_l = [y_1 = [0, 50]] \wedge [y_2 = [0, 50[\cup[100, 200]] ,$$
$$a_j \cap_a a_l = [y_1 = [0, 50[] \wedge [y_2 = \emptyset] ;$$

dès lors,

$$\pi(a_j \cup_a a_l) = [(50 - 0)[(50 - 0) + (200 - 100)] = 7500 ,$$
$$\pi(a_j \cap_a a_l) = (50 - 0)(0) = 0 ;$$

enfin

$$\begin{aligned} a_k \cup_a a_l &= [y_1 = [0,50] \cup [100,200]] \wedge [y_2 = [0,50[\cup[100,200]] \ , \\ a_k \cap_a a_l &= [y_1 = \emptyset] \wedge [y_2 = \emptyset] \ ; \end{aligned}$$

dès lors,

$$\begin{aligned} \pi(a_k \cup_a a_l) &= [(50-0)+(200-100)][(50-0)+(200-100)] = 22500 \ , \\ \pi(a_k \cap_a a_l) &= (0)(0) = 0 \ . \end{aligned}$$

Par conséquent,

$$\begin{aligned} d_{\text{ext}}(a_j, a_k) &= 7500 - 0 = 7500 \ , \\ d_{\text{ext}}(a_j, a_l) &= 7500 - 0 = 7500 \ , \\ d_{\text{ext}}(a_k, a_l) &= 22500 - 0 = 22500 \ , \end{aligned}$$

ce qui entraîne

$$\exists \ (a_j, a_k, a_l) : d_{\text{ext}}(a_k, a_l) > d_{\text{ext}}(a_j, a_k) + d_{\text{ext}}(a_j, a_l)$$

c'est-à-dire d_{ext} ne satisfait pas l'inégalité triangulaire.

2. C'et indice exprime que deux OAB sont plus similaires sous l'hypothèse de dépendance logique que sous celle d'indépendance. En d'autres mots, la valeur de d_{ext} diminue si on ajoute des dépendances logiques (voir l'exemple 2 ci- dessous). Soient les OAB suivants :

$$\begin{aligned} a &= [y_1 = [10,50]] \wedge [y_2 = [100,400]] \ , \\ \hat{a} &= [y_1 = [30,50]] \wedge [y_2 = [200,500]] \ . \end{aligned}$$

Le côté gauche de la figure ci-dessous montre le produit cartésien $[10,50] \times [100,400]$, qui représente les descriptions d'individus cohérents donnés par a , et le produit cartésien $[30,50] \times [200,500]$, qui représente les descriptions d'individus cohérents données par $\hat{a}$ (lignes pointillés), sous l'hypothèse d'indépendance logique entre les variables y_1 et y_2 . Le côté droit montre les mêmes produits cartésiens mais sous l'hypothèse de dépendance logique entre les variables y_1 et y_2 exprimée par la règle

$$\forall \ \omega \in \Omega \ , \ y_1(\omega) \in [0,30[\quad \Rightarrow \quad y_2(\omega) \text{ est non applicable.}$$

D'après la figure ci-dessous, les deux OAB a et $\hat{a}$ sont plus similaires sous l'hypothèse de dépendance logique entre les variables y_1 et y_2 que sous l'hypothèse d'indépendance logique entre les mêmes variables.

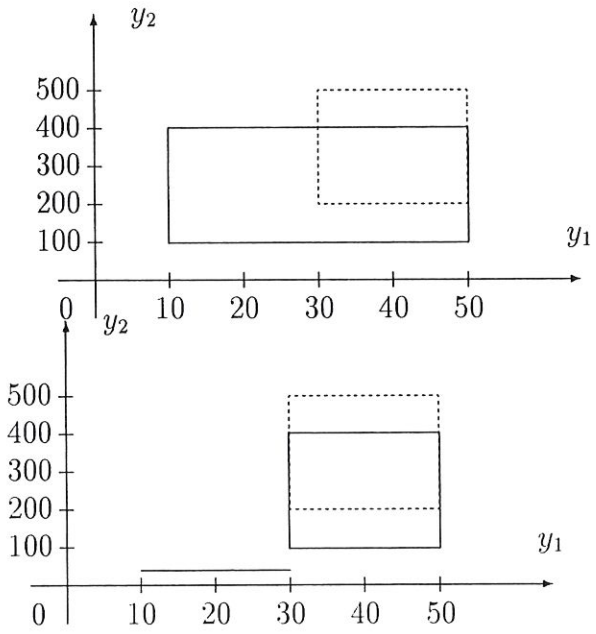


FIG. 1.3 – *Représentation dans le plan de a et $\hat{a}$.*

Vérifions par calcul que d_{ext} prend une valeur plus petite en cas d'existence de dépendance(s) logique(s) qu'en cas d'indépendance.

Calcul de d_{ext} sous l'hypothèse d'indépendance logique entre y_1 et y_2 : Nous avons

$$\begin{aligned} a \cup_a \hat{a} &= [y_1 = [10, 50]] \wedge [y_2 = [100, 500]] , \\ a \cap_a \hat{a} &= [y_1 = [30, 50]] \wedge [y_2 = [200, 500]] ; \end{aligned}$$

dès lors,

$$\begin{aligned} \pi(a \cup_a \hat{a}) &= (50 - 10)(500 - 100) = 16000 , \\ \pi(a \cap_a \hat{a}) &= (50 - 30)(400 - 200) = 4000 ; \end{aligned}$$

et on obtient

$$d_{\text{ext}} = 16000 - 4000 = 12000 .$$

Calcul de d_{ext} sous l'hypothèse de dépendance logique entre y_1 et y_2 : Nous avons toujours

$$\begin{aligned} a \cup_a \hat{a} &= [y_1 = [10, 50]] \wedge [y_2 = [100, 500]] , \\ a \cap_a \hat{a} &= [y_1 = [30, 50]] \wedge [y_2 = [200, 500]] . \end{aligned}$$

Comme nous l'avons remarqué auparavant, nous pouvons utiliser l'équation (1.1) pour calculer $\pi(a \cup_a \hat{a})$: nous décomposons $a \cup_a \hat{a} = [y_1 = [10, 50]] \wedge [y_2 = [100, 500]]$ dans

$$\begin{aligned} a_1 &= [y_1 = [10, 30[] \wedge [y_2 \text{ est non applicable}] , \\ a_2 &= [y_1 = [30, 50]] \wedge [y_2 = [100, 500]] . \end{aligned}$$

Alors.

$$\pi(a \cup_a \hat{a}) = \pi(a_1) + \pi(a_2) = (30 - 10) + (50 - 30)(500 - 100) = 8020 .$$

Comme nous n'avons pas besoin de décomposer $a \cap_a \hat{a}$, nous avons toujours

$$\pi(a \cap_a \hat{a}) = 4000$$

Alors

$$d_{\text{ext}} = 8020 - 4000 = 4020 .$$

Cet exemple montre que notre indice de similarité d_{ext} est capable d'exprimer ce que nous voyons dans la figure, c'est-à-dire que les deux OAB a et $\hat{a}$ sont plus similaires (et donc moins dissimilaires) sous l'hypothèse de dépendance logique entre les variables y_1 et y_2 que sous l'hypothèse d'indépendance logique entre les mêmes variables.

1.3.3 Remarques

1. Dans les applications concrètes où le nombre de variables est important, le calcul du PD d'un OAB peut rapidement dépasser la capacité de stockage informatique possible; nous utiliserons alors à la place l'indice

$$\hat{d}_{\text{ext}} = \begin{cases} 0 & \text{si } d_{\text{ext}} \leq 0\ , \\ \ln d_{\text{ext}} & \text{sinon.} \end{cases}$$

2. Tout comme dans la description d'individus par les objets habituels de l'analyse des données (où la variabilité est prise en compte par la moyenne, le mode ou encore l'écart-type), l'influence de la dépendance conditionnelle ne devient importante que si le nombre de variables en dépendance logique est suffisamment élevé par rapport au nombre total de variables dans la BC.

Chapitre 2

Distance généralisée de Minkowski

Nous commençons par décrire l'espace dans lequel nous définirons cette distance : le **Cartesian Space Model** (CSM).

2.1 Description du CSM

– Un événement ϵ_i est décrit au moyen de variables

$$y_i : \Omega \to V_i \subseteq O_i \,. \qquad \forall\, i = 1, \cdots, d \,.$$

où V_i est l'ensemble des valeurs prises par y_i . et
O_i est soit un intervalle (ou union d'intervalles),
soit un ensemble fini d'intervalles,
cela dépend du type de y_i .
Plus précisément,

y_i est quantitative (discrète ou continue) ou qualitative ordinale	O_i et V_i sont supposés être de la forme $[a_i, b_i]$,
y_i est qualitative nominale ou structurée en arbre[1]	O_i est un ensemble fini de valeurs qui sont pour un arbre l'ensemble de ses valeurs terminales.

1. On en verra une illustration plus loin.

- Un objet assertion

$$a = \wedge_{i=1}^{d} e_i$$

sera identifié au produit cartésien des ensembles de valeurs V_i des événements e_i qui le définissent. Au vu de cette convention, nous dirons donc qu'un OAB est donné par

$$a : V_1 \times \cdots \times V_d \, .$$

Nous désignerons encore

$$O^{(d)} = O_1 \times \cdots \times O_d$$

comme étant l'espace des descriptions cartésiennes[2].

2.1.1 Définition 1 : l'union cartésienne

Soient a_1 et a_2 deux OAB tels que

$$\begin{aligned} a_1 &= V_1^1 \times \cdots V_d^1 \, . \\ a_2 &= V_1^2 \times \cdots V_d^2 \, . \end{aligned}$$

Leur **union cartésienne** est définie par

$$a_1 \dot{+} a_2 = V_1^1 \dot{+} V_1^2 \times \cdots \times V_d^1 \dot{+} V_d^2$$

où $V_i^1 \dot{+} V_i^2$, l'union cartésienne des ensembles de valeurs V_i^1 et V_i^2 , est définie comme suit :

1. si la variable y_i est qualitative ordinale ou quantitative, $V_i^1 \dot{+} V_i^2$ est de la forme :

$$V_i^1 \oplus V_i^2 = [\min(V_{iL}^1, V_{iL}^2), \max(V_{iU}^1, V_{iU}^2)]$$

où V_{iL}^1 (resp. V_{iL}^2) est la borne inférieure de l'intervalle V_i^1 (resp. V_i^2), V_{iU}^1 (resp. V_{iL}^2) est la borne supérieure de l'intervalle V_i^1 (resp. V_i^2);

2. $O^{(d)}$ est l'espace cartésien et O^d représente l'espace euclidien.

2. si y_i est une variable qualitative nominale :

$$V_i^1 \oplus V_i^2 = V_i^1 \cup V_i^2$$

qui est l'union ensembliste habituelle;

3. si y_i est une variable structurée arbre et si le noeud parent le plus proche commun à toutes les valeurs appartenant à V_i^1 est noté $N(V_i^1)$:

(a) si $N(V_i^1) = N(V_i^2)$, $V_i^1 \oplus V_i^2 = V_i^1 \cup V_i^2$;

(b) si $N(V_i^1) \neq N(V_i^2)$, $V_i^1 \oplus V_i^2$ =ensemble de toutes les valeurs terminales issues du noeud $N(V_i^1 \cup V_i^2)$.

De plus, pour tout ensemble de valeurs V_i^k , on supposera que $V_i^k \oplus V_i^k = V_i^k$.

Notons, finalement, que l'union de deux OAB est un OAB.

2.1.2 Exemples

Cas 1 : y_i est quantitative ou qualitative ordinale :

Soient $V_i^1 = [1.3]$ et $V_i^2 = [4.7]$. Alors

$$V_i^1 \oplus V_i^2 = [1.7] .$$

Cas 2 : y_i est qualitative nominale :

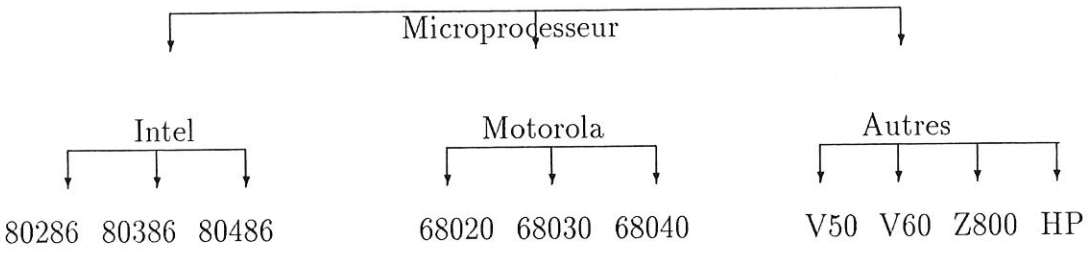
Soient $V_i^1 = \{\text{bleu, rouge, jaune}\}$ et $V_i^2 = \{\text{chien, chat}\}$. Alors

$$V_i^1 \oplus V_i^2 = \{\text{bleu, rouge, jaune, chien, chat}\} .$$

Cas 3 : la variable y_i est structurée arbre; nous prendrons pour y_i la variable microprocesseur.

Dans cette figure, les ensembles des valeurs V_i^k auront par exemple les formes

$$V_i^k = \{68020\} , \quad V_i^k = \{V50, 80286\} , \quad \cdots$$



Ce seront donc des ensembles de valeurs terminales de l'arbre.

Cas 3.1 : $N(V_i^1) = N(V_i^2)$

Soient $V_i^1 = \{80386\}$ et Alors
$V_i^2 = \{80486\}$.

$$N(V_i^1) = \text{Intel} = N(V_i^2)$$

et dès lors

$$V_i^1 \oplus V_i^2 = V_i^1 \cup V_i^2 = \{80386.80486\} .$$

Cas 3.2 : $N(V_i^1) \neq N(V_i^2)$

a) Soient $V_i^1 = \{80386\}$ et Alors
$V_i^2 = \{V50\}$.

$$N(V_i^1) = \text{Intel} \neq N(V_i^2) = \text{autres.}$$

On calculera donc $N(V_i^1 \cup V_i^2) =$ Microprocesseur, ce qui entraîne que

$$V_i^1 \cup V_i^2 = \{80286, 80386, 80486, 68020, 68030, V50, V60, Z800, HP\} .$$

b) Soient $V_i^1 = \{80486, 68040\}$, Alors
$V_i^2 = \{68030, 68040\}$ et
$V_i^3 = \{80386, 68030, 68040\}$.

$$\begin{aligned} V_i^1 \oplus V_i^2 &= \{80286, 80386, 80486, 68020, 68030, 68040, V50, V60, Z800, HP\} \\ &= V_i^2 \oplus V_i^3 \end{aligned}$$

bien que

$$V_i^1 \cup V_i^2 \neq V_i^2 \cup V_i^{3\ 3}\ .$$

Remarquons que, si nous appliquons la définition générale pour calculer :

$$\begin{aligned} V_i^1 \oplus V_i^1 &= \{80286, 80386, 80486, 68020, 68030, 68040, V50, V60, Z800, HP\} \\ &\neq V_i^1\ . \end{aligned}$$

2.1.3 Illustrations

1. Dans la figure ci-dessous, nous donnons trois exemples d'unions cartésiennes (dans le plan euclidien) de deux OAB décrits par les variables y_1 et y_2 . Étant

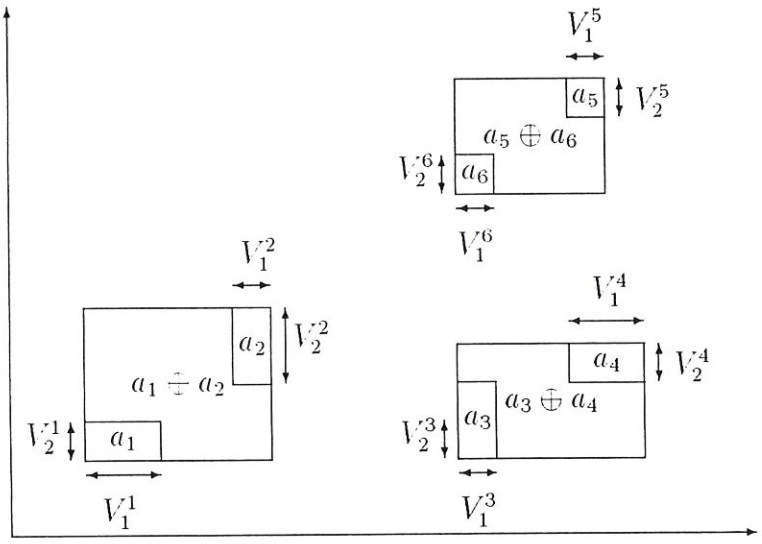


FIG. 2.1 – *Unions cartésiennes dans le plan euclidien.*

donné que les ensembles de valeurs de ces variables, les V_i^k ($i = 1, 2$, $k = 1, \cdots, 6$) sont sous forme d'intervalles, nous pouvons donc conclure que celles-ci sont soit quantitatives, soit qualitatives ordinales.

Ce dessin illustre assez bien que la notion d'union cartésienne (d'objets décrits par des variables quantitatives ou qualitatives ordinales) est inspirée de celle du produit cartésien.

2. Nous pouvons facilement constater à partir d'un dessin que l'union cartésienne de deux "sous-objets" d'un même OAB est encore contenue dans celui-ci (c'est-à-dire en est encore un "sous-objet").

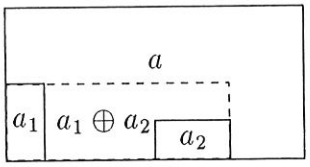


FIG. 2.2 – *Union de deux "sous-objets" a_1 et a_2 d'un OAB a .*

Cette propriété n'est plus vérifiée lorsque a n'est pas convexe, comme nous le constatons ci-dessous.

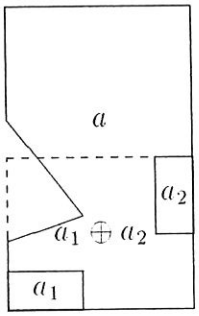


FIG. 2.3 – *Union de deux "sous-objets" d'un "OAB" non convexe.*

3. Voici un cas particulier et un contre-exemple d'OAB : d n'est pas le produit

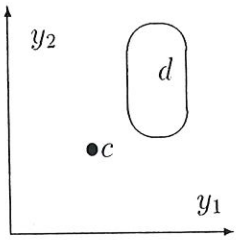


FIG. 2.4 – *Cas particulier et contre-exemple d'OAB.*

cartésien de deux OAB, ni d'ailleurs un OAB, même s'il est convexe.
c est un OAB particulier réduit à un point.

2.1.4 Lemme 1

Si y_i est une variable structurée arbre,

alors

$$N(V_i^k \oplus V_i^m) = N(V_i^k \cup V_i^m) .$$

2.1.5 Définition 2 : intersection cartésienne

- L'**intersection cartésienne** de deux OAB $a_1 = V_1^1 \times \cdots \times V_d^1$ et $a_2 = V_1^2 \times \cdots \times V_d^2$ est définie par un produit cartésien

$$a_1 \otimes a_2 = (V_1^1 \otimes V_2^1) \times \cdots \times (V_d^1 \otimes V_d^2)$$

où

$$V_i^1 \otimes V_i^2 = V_i^1 \cap V_i^2 . \tag{2.1}$$

- L'intersection cartésienne de deux OAB est un OAB.

- Si l'expression (2.1) est égale au vide pour au moins une valeur de i , alors a_1 et a_2 sont **disjoints** et

$$a_1 \otimes a_2 = \Phi$$

où Φ désigne l'OAB vide.

2.1.6 Illustration dans le plan euclidien

Voici deux exemples d'intersection cartésienne de deux OAB dans le plan euclidien.

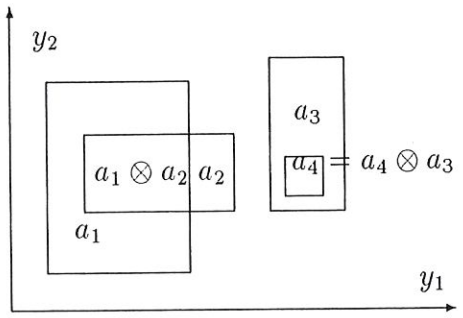
2.1.7 Lemme 2

Si y_i a une structure d'arbre, nous définissons un ordre sur les rangs de noeuds par

$$N(V_i^1) \leq N(V_i^2) \qquad \text{si } V_i^1 \subseteq V_i^2 .$$

Ainsi, on a

(i) $N(V_i^1), N(V_i^2) \leq N(v_i^3) \Rightarrow N(V_i^1 \oplus V_i^2) \leq N(V_i^3)$.



Remarquons que $a_2 \otimes a_3 = \Phi$.

(ii) Si $V_i^1 \otimes V_i^2$ est différent du vide, alors

$$N(V_i^1 \otimes V_i^2) = \min(N(V_i^1), N(V_i^2)) .$$

2.1.8 Modèle CSM

Nous appelons **Cartesian Space Model** le modèle mathématique $(O^{(d)}, \oplus, \otimes)$ et nous noterons $\Lambda(O^{(d)})$ la famille des OAB de $O^{(d)}$.

2.2 Distance généralisée de Minkowski

Sur base de la définition du CSM, nous allons construire une distance entre OAB.

Nous commencerons par la description d'une première distance que nous transformerons ensuite afin de tenir compte des unités respectives de chacune des variables y_i . Ensuite, nous utiliserons cette distance dans le cadre de la classification et de l'analyse en composantes principales.

2.2.1 Définition 1

Soient a_1 et a_2 deux OAB appartenant à $\Lambda(O^{(d)})$ tels que

$$\begin{aligned} a_1 &= V_1^1 \times V_2^1 \times \cdots \times V_d^1 \ , \\ a_2 &= V_1^2 \times V_2^2 \times \cdots \times V_d^2 \ . \end{aligned}$$

Nous définissons la mesure entre les événements V_i^1 et V_i^2 de la façon suivante :

$$\phi(V_i^1, V_i^2) = |V_i^1 \dot{+} V_i^2| - |V_i^1 \dot{\otimes} V_i^2| + \gamma \left(2\,|V_i^1 \dot{\otimes} V_i^2| - (|V_i^1| + |V_i^2|)\right) , \quad i = 1, 2, \cdots, d$$

où

- $\gamma \in [0.\, 0.5]$. et
- $|\cdot|$ désigne la longueur si $\cdot$ est un l'intervalle. et le nombre de ses éléments s'il s'agit d'un ensemble discret.

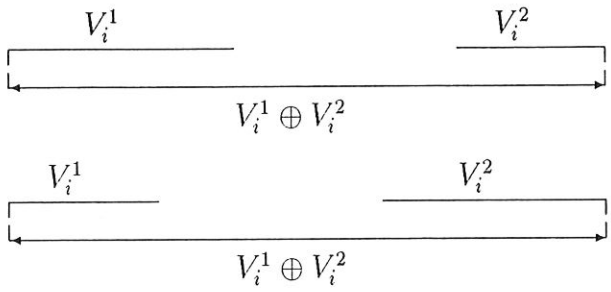
- Lorsque $\gamma = 0$, la fonction ϕ est simplifiée :

$$\phi(V_i^1, V_i^2) = |V_i^1 \oplus V_i^2| - |V_i^1 \otimes V_i^2| \ , \tag{2.2}$$

et si V_i^1 et V_i^2 sont deux intervalles disjoints, $\phi(V_i^1, V_i^2)$ ne représente que la distance extérieure entre V_i^1 et V_i^2 . En effet, nous remarquons que

$$|V_i^1 \otimes V_i^2| = |V_i^1 \cap V_i^2| = 0 \ .$$

Voyons sur un exemple que cela conduit à des imprécisions :



Sur cette figure, nous voyons que, malgré les différences de longueurs des intervalles respectifs, ϕ prend une valeur identique dans les deux situations.

- Lorsque $\gamma = 0.5$, alors

$$\phi(V_i^1, V_i^2) = |V_i^1 \oplus V_i^2| - \frac{(|V_i^1| + |V_i^2|)}{2} . \tag{2.3}$$

Ainsi, si nous reprenons les deux figures ci-dessus, nous aurons des valeurs de ϕ différentes.

- En prenant une valeur de $\gamma \in]0.0.5[$, nous obtiendrons des propriétés intermédiaires entre celles de (2.2) et de (2.3). Bien que le choix de γ dépende du but fixé, il semblerait que $\gamma = 0.5$ soit l'une des alternatives préférables (vu la simplicité et la précision qu'elle offre).

2.2.2 Lemme 3

Pour chaque événement[4] V_i^1, V_i^2, V_i^3 *des OAB* $a_1, a_2, a_3 \in \Lambda(O^{(d)})$, *la fonction* $\phi(\cdot,\cdot)$ *satisfait aux axiomes de distance, c'est-à-dire :*

1. $\phi(V_i^1, V_i^2) \geq 0$ *et est nul* $\Leftrightarrow$ $V_i^1 = V_i^2$ *(positivité);*
2. $\phi(V_i^1, V_i^2) = 0 \Leftrightarrow V_i^1 = V_i^2$ *(annulation);*
3. $\phi(V_i^1, V_i^2) = \phi(V_i^2, V_i^1)$ *(symétrie);*
4. $\phi(V_i^1, V_i^2) \leq \phi(V_i^1, V_i^3) + \phi(V_i^3, V_i^2)$ *(inégalité triangulaire).*

4. Comme auparavant, nous identifions un événement e_i à son ensemble V_i de valeurs et un OAB au produit des "événements" ainsi définis.

2.2.3 Distance de Minkowski

En utilisant la définition de $\phi(V_i^1, V_i^2)$, nous sommes à présent en mesure de définir la distance $d_P(a_1, a_2)$ entre les OAB $a_1 = V_1^1 \times \cdots \times V_d^1$ et $a_2 = V_1^2 \times \cdots \times V_d^2 \in \Lambda(O^{(d)})$ comme suit :

$$d_P(a_1, a_2) = \left[\sum_{i=1}^{d} (\phi(V_i^1, V_i^2))^P\right]^{\frac{1}{P}} .$$

Cette définition n'est cependant pas tout à fait satisfaisante car elle ne tient pas compte des unités respectives , ni de l'"étalement" des variables.

Exemples :

1. Si y_1 est exprimée en **cm** et y_2 en **kilos**, la distance $d_P(a_1, a_2)$ contient la somme de la P-ième puissance de $\phi(V_i^1, V_i^2)$ exprimée en **cm** et de $\phi(V_i^1, V_i^2)$ exprimée en **kilos**. Comment calculer une distance qui est la somme de termes d'unités différentes? Quelle peut être l'influence d'une des variables sachant qu'elles sont toutes deux exprimées en des unités incompatibles?

2. Si $V_1^1 = [0.190]$ et $V_1^2 = [-20.150]$. alors
   $V_2^1 = [0.10]$ et $V_2^2 = [3.9]$.

$$\Phi(V_1^1, V_1^2) = 210 - \tfrac{190+170}{2} = 30 .$$
$$\Phi(V_2^1, V_2^2) = 10 - \tfrac{10+6}{2} = 2 .$$

   Nous constatons que la variable 1 aura une influence beaucoup plus importante que la variable 2.

Afin de résoudre ces problèmes, nous prendrons comme nouvelle mesure **normalisée**

$$\Psi(V_i^1, V_i^2) = \frac{\phi(V_i^1, V_i^2)}{|O_i|} , \quad k = 1, \cdots, d , \tag{2.4}$$

où $|O_i|$ désigne

- la longueur du domaine O_i si y_i est quantitative continue, et
- le nombre de valeurs dans O_i si y_i est quantitative discrète, qualitative ou structurée en arbre.

Dès lors, pour chaque valeur de i , la fonction $\Psi(V_i^1, V_i^2)$ devient une quantité sans dimension et telle que

$$0 \leq \Psi(V_i^1, V_i^2) \leq 1 , \qquad i = 1, \cdots, d . \qquad \text{(normalisée)}$$

En utilisant ces quantités, nous obtenons la définition suivante : soient $a_1, a_2 \in \Lambda(O^{(d)})$ deux OAB telles que

$$\begin{aligned} a_1 &= V_1^1 \times \cdots \times V_d^1 , \\ a_2 &= V_1^2 \times \cdots \times V_d^2 . \end{aligned}$$

La distance d'ordre P entre ces deux OAB sera :

$$d_P(a_1, a_2) = \frac{1}{d} \left[\sum_{i=1}^{d} (\Psi(V_i^1, V_i^2))^P \right]^{\frac{1}{P}} . \qquad (2.5)$$

Cette distance ne rencontre plus les inconvénients cités auparavant; elle constitue une des formes de la mesure généralisée de Minkowski.

Cependant, il est encore possible de l'améliorer par l'introduction de coefficients (poids). Ceux-ci permettraient de contrôler l'importance relative de chaque variable. Nous noterons ces coefficients c_i , $i = 1, \cdots, d$. Dès lors, si nous connaissons au préalable le poids relatif des variables, nous utiliserons la définition suivante :

Définition 4 (Distance généralisée de Minkowski d'ordre P)

Soient une paire d'OAB $a_1, a_2 \in \Lambda(O^{(d)})$ tels que

$$\begin{aligned} a_1 &= V_1^1 \times \cdots \times V_d^1 , \\ a_2 &= V_1^2 \times \cdots \times V_d^2 , \end{aligned}$$

et une constante $P \geq 1$. Nous appellerons **distance généralisée de Minkowski d'ordre** P entre a_1 et a_2 , la formule suivante :

$$d_P(a_1, a_2) = \left[\sum_{i=1}^{d} (c_i \ \Psi(V_i^1, V_i^2))^P \right]^{\frac{1}{P}} \qquad (2.6)$$

où c_i , $i = 1, \cdots, d$[5], et $\sum_{i=1}^{d} c_1 = 1$[6].

Cette distance satisfait à

$$0 \leq d_P(a_1, a_2) \leq 1$$

et nous avons la proposition suivante:

Proposition 1 *La formule (2.6) satisfait aux axiomes des distances.*

Parmi les cas particuliers de la distance généralisée de Minkowski, nous trouvons

- $P = 1$: la distance en valeur absolue généralisée:

$$d_1(a_1, a_2) = \sum_{i=1}^{d} c_i \ \Psi(V_i^1, V_i^2) \ .$$

- $P = 2$: la distance euclidienne généralisée:

$$d_2(a_1, a_2) = \left[\sum_{i=1}^{d} c_i \ (\Psi(V_i^1, V_i^2))^2 \right]^{\frac{1}{2}} \ .$$

- $P = \infty$: la distance généralisée de Tchebyscheff:

$$d_\infty(a_1, a_2) = \max_{1 \leq i \leq d} [c_i \ \Psi(V_i^1, V_i^2)] \ .$$

Exemple de calcul de la distance généralisée de Minkowski : voici des données portant sur des huiles et graisses: soient

y_1 la densité (variable quantitative continue);

y_2 la température de solidification (variable quantitative discrète);

y_3 les acides gras composants (variable qualitative nominale).

Les symboles mentionnés dans cette dernière colonne représentent les acides suivants: L = acide linoléique, Ln = acide linolénique, O =acide oléique, P =

5. Dans la formule (2.4), tous les c_i sont égaux à $\frac{1}{d}$; nous donnons la même importance à toutes les variables.

6. Ceci est assez intuitif puisque l'influence totale des variables est de 100%.

acide palmique, M = acide myristique, S = acide séraïque, A = acide arachique, C = acide caprique, Lu = acide laurique.

	y_1	y_2	y_3
a_1) huile de lin	$[0.930, 0.935]$	$\{-27, \cdots, -18\}$	{L, LN, O, P, M}
a_2) huile de Périlla	$[0.930, 0.937]$	$\{-5, \cdots, -4\}$	{L, LN, O, P, S}
a_3) graisse de boeuf	$[0.860, 0.870]$	$\{30, \cdots, 38\}$	{O, P, M, S, C}
a_4) graisse de porc	$[0.858, 0.864]$	$\{20, \cdots, 32\}$	{L, O, P, M, S, Lu}

Ici,

$$\begin{aligned} O_1 = [0.850, 0.940] &\Rightarrow |O_1| = 0.090 , \\ O_2 = \{-30, \cdots, -40\} &\Rightarrow |O_2| = 71 , \\ O_3 = \{\text{L,Ln,O,P,M,S,A,C}\} &\Rightarrow |O_3| = 8 . \end{aligned}$$

Nous prendrons les c_i égaux à $\frac{1}{3}$, $\gamma = 0.5$ et $P = 1$. Evaluons la distance

$$\begin{aligned} d(a_1, a_2) &= \frac{|V_1^1 \oplus V_1^2| - \frac{|V_1^1|+|V_1^2|}{2}}{|O_1| \cdot d} + \frac{|V_2^1 \oplus V_2^2| - \frac{|V_2^1|+|V_2^2|}{2}}{|O_2| \cdot d} + \frac{|V_3^1 \oplus V_3^2| - \frac{|V_3^1|+|V_3^2|}{2}}{|O_3| \cdot d} \\ &= \frac{0.007 - \frac{0.007+0.005}{2}}{0.90 \bullet 5} + \frac{23 - \frac{10+2}{2}}{71 \bullet 5} + \frac{6 - \frac{5+5}{2}}{8 \bullet 5} \\ &= \tfrac{1}{5}\,(0.0111 + 0.2394 + 0.1250) \\ &= 0.07511 . \end{aligned}$$

Par le même procédé.

$$\begin{aligned} d(a_1, a_3) &= 0.21564 , \\ d(a_2, a_3) &= 0.17086 , \\ d(a_1, a_4) &= 0.187188 , \\ d(a_2, a_4) &= 0.139709 . \end{aligned}$$

2.3 Applications

2.3.1 Proposition 2

Nous avons donc trouvé une distance, sur l'espace des objets symboliques. Plus précisément,

Proposition 2

$\Lambda(o^{(d)}, d_P)$ est un espace muni d'une distance.

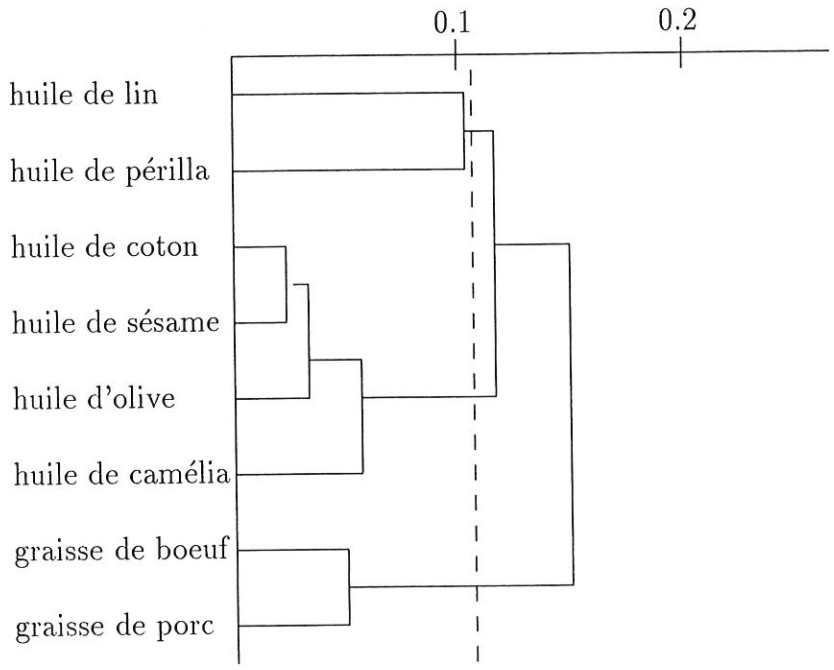
2.3.2 Classification

Nous reprenons le tableau de données de la partie précédente.

Objet	Poids spécifique	Température de solidification	Teneur en iode	Teneur en minerais	Acides gras principaux
1) Huile de lin	0.930-0.935	(-27)-(-18)	170-204	118-196	L, Ln, O, P, M
2) Huile de Périlla	0.930-0.937	(-5)-(-4)	192-208	188-197	L, Ln, O, P, S
3) Huile de coton	0.916-0.918	(-6)-(-1)	99-113	189-198	L, O, P, M, S
4) Huile de sésame	0.920-0.926	(-6)-(-4)	104-116	187-193	L, O, P, S, A
5) Huile de camélia	0.916-0.917	(-21)-(-15)	80-82	189-193	L, O
6) Huile d'olive	0.914-0.919	0-6	79-90	187-196	L, O, P, S
7) Graisse de boeuf	0.860-0.870	30-38	40-48	190-199	O, P, M, S, C
8) Graisse de porc	0.858-0.864	22-32	53-77	190-202	L, O, P, M, S, Lu

En principe, les paires d'objets (1,2), (3,4), (5,6) et (7,8) ont des propriétés similaires. Les objets 1 et 2 sont utilisés pour la peinture, 3 et 4 pour l'alimentation, 5 et 6 pour les produits cosmétiques, etc.

Lorsque nous appliquons sur ces données la méthode du voisin le plus proche (avec la distance généralisée de Minkowski), nous obtenons le graphe suivant :



Nous obtenons de même pour la méthode du voisin le plus éloigné :

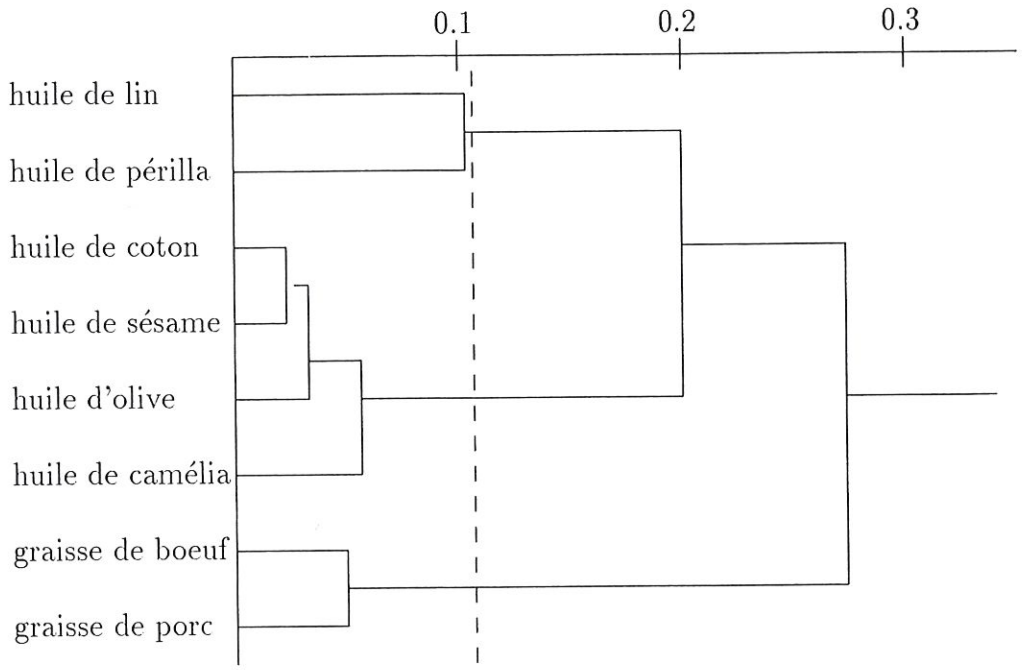
Commentaires : ces deux graphes ont même allure générale. Ils regroupent correctement les paires (1,2), (3,4) et (7,8); par contre, ils font la même erreur pour la paire (5,6). Dans l'ensemble, nous pouvons être satisfaits de la classification. En effet, si nous coupons en 0.115, nous trouvons trois classes, deux qui sont les bonnes classes et la troisième qui comprend les deux classes restantes. Cette coupe est indiquée en pointillés sur les deux graphes.

2.3.3 Analyse en composantes principales (ACP)

En général, l'ACP est appliquée sur des ensembles de données **quantitatives**. Ici, nous allons généraliser cette méthode à des données de type mixte (qualitatives, quantitatives). Voici les étapes principales de cette ACP généralisée :

(i) Déterminer l'OAB de référence :

Supposons que notre ensemble de données soit composé de N OAB, c'est-à-dire $B = \{a_1, \cdots, a_N\}$. Nous choisirons comme OAB de référence une



notion proche de celle du centre de gravité : nous prenons comme centre l'OAB dont la somme des distances aux autres OAB est minimale (distance de Minkowski), c'est-à-dire telle que ce centre a_j vérifie

$$\sum_{k=1}^{N} d_P(a_j, a_k) = \min_{1 \leq i \leq N} \sum_{k=1}^{N} d_P(a_i, a_k) \, .$$

(ii) Calculer les distances de Minkowski entre l'OAB de référence et chacun des autres OAB, en attribuant une valeur négative à la distance des OAB plus petits que celui de référence. Voici un exemple :

OAB	y_1	y_2	y'_1	y'_2
a_1	2	1	0.25	0.00
a_2	1	2	0.00	0.25
a_3	0	1	−0.25	0.00
a_4	1	0	0.00	−0.25
$a =$ référence	1	1	0.00	0.00

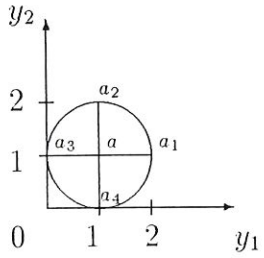
Calculons par exemple y_1' pour a_1 :

$$\Psi(V_1^1, V_1) = \frac{2 - \frac{3}{2}}{2} = 0.25 \ .$$

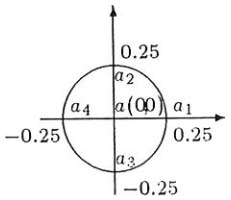
Le signe de y_1' est positif car nous remarquons que la valeur sur l'axe y_1 de a_1 (qui vaut 2) est plus élevée que celle de a pour cette même valeur (qui est 1). Calculons de même y_2' pour a_4 :

$$\Psi(V_2^4, V_2) = \frac{1 - \frac{1}{2}}{2} = 0.25 \ .$$

Ici, par contre, nous attribuons un signe négatif à 0.25 car, sur l'axe y_2 , a_4 a une valeur plus petite que a . Par cette méthode, le graphe



devient



(iii) Appliquer les techniques habituelles de l'ACP sur les objets définis en ii).

2.4 "Mutual Neighborhood Graph"

Le graphe présenté dans cette section fournit de l'information "interclasses" (entre classes d'objets symboliques booléens).

Soient C_1 et C_2 deux classes d'objets symboliques booléens tels que

$$\begin{aligned} C_1 &= \{a_1, \cdots, a_{N_1}\} \qquad \text{où } a_j = \Lambda_i[y_i = V_i^j] \,, \\ C_2 &= \{b_1, \cdots, b_{N_2}\} \qquad \text{où } b_k = \Lambda_k[y_k = V_k^j] \,. \end{aligned}$$

Comme pour les sections précédentes, on identifiera a_j (respectivement b_k) au produit cartésien $V_1^j \times \cdots \times V_d^j$ (respectivement $V_1^k \times \cdots \times V_d^k$).

Nous définirons le graphe de voisinage mutuel (MNG) de la manière suivante :

2.4.1 Définition

Soient deux OAB $a_\ell . a_m \in C_1$.

Ces deux objets seront dits **voisins mutuels opposés** à C_2 , si

$$\forall\, b_k \in C_2 \,, \qquad b_k \odot (a_j \oplus a_m) = \emptyset \,.$$

Le MNG de C_1 opposé à C_2 (noté MNG($C_1 \mid C_2$) est construit en joignant par une arête toutes les paires d'OAB. $a_\ell . a_m \in C_1$ qui sont voisins mutuels opposés à C_2 .

2.4.2 Remarque

On dira que deux objets $a = V_1^a \times \cdots \times V_d^a$ et $b = V_1^b \times \cdots \times V_d^b$ sont **complètement distingables** si

$$a \otimes b = \emptyset \,.$$

Il suffit pour cela qu'il existe une variable y_m telle que

$$V_m^a \cap V_m^b = \emptyset \,.$$

2.4.3 Exemple

La figure ci-dessous illustre un exemple de MNG dans le plan euclidien (avec $N_1 = 6$ et $N_2 = 5$).

Sur ce dessin,

$\{a_1, a_2\}$, $\{a_2, a_3\}$, $\{a_2, a_4\}$, $\{a_1, a_3\}$, $\{a_1, a_4\}$, $\{a_2, a_5\}$, $\{a_3, a_4\}$ et $\{a_4, a_5\}$ sont voisins mutuels dans C_1 opposés à C_2 et

$\{b_1, b_2\}$, $\{b_1, b_5\}$, $\{b_2, b_5\}$, $\{b_2, b_3\}$, $\{b_1, b_4\}$, $\{b_3, b_4\}$ et $\{b_4, a_5\}$ sont voisins mutuels dans C_2 opposés à C_1 .

Les pointillés forment des rectangles de côtés parallèles aux axes de coordonnées. On constate que les objets sont voisins mutuels si leur rectangle ne contient aucun objet de l'autre classe.

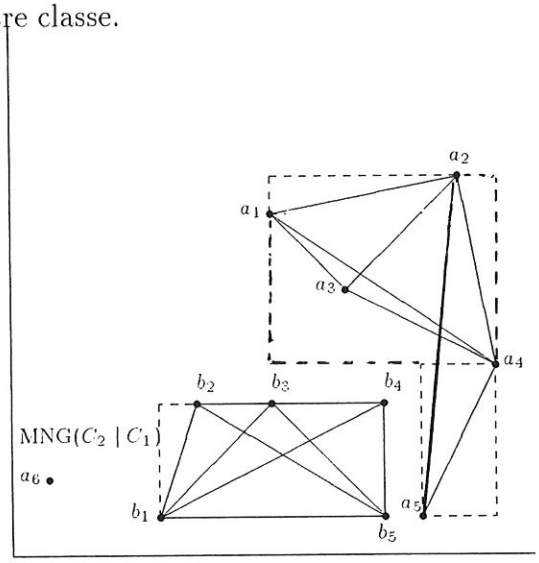


FIG. 2.5 – *Un exemple de MNG.*

En effet, donnons un contre-exemple où a_3 et a_5 ne sont pas voisins mutuels opposés à C_2 car leur rectangle contient b_4 et b_5 .

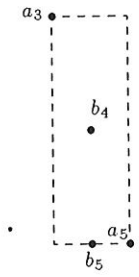


FIG. 2.6 – *Contre-exemple de voisins mutuels* $\{a_3, a_5\}$ *opposés à* C_2 .

2.4.4 Remarque

La proximité (étroitesse des connections) dans MNG($C_1 \mid C_2$) et MNG($C_2 \mid C_1$) est liée directement à la séparabilité de C_1 et C_2 .

Sur base de la proposition suivante, nous aurons une condition suffisante pour séparer linéairement C_1 et C_2 .

2.4.5 Proposition

Si MNG($C_1 \mid C_2$) et MNG($C_2 \mid C_1$) sont des graphes complets,
alors les classes C_1 et C_2 sont linéairement séparables.

Remarque : cette condition de séparabilité est très utile dans les espaces de dimension élevée. En effet, dans de tels espaces, il n'existe pas d'algorithme testant la séparabilité linéaire.

2.4.6 Application du MNG : "Generalized Box Classifier" (GBC)

Présentation

Voici une utilisation du MNG : la construction d'un classificateur boîte généralisé (GBC).

6. Un graphe est complet si, quelle que soit la paire de noeuds choisie (dans ce graphe), ces deux noeuds sont reliés au moyen d'une arête.

Ce classificateur range les OAB d'une même classe C_i dans un ensemble de boîtes $B_{i_1}, \cdots, B_{i_n}$. Ces boîtes sont construite à partir des unions cartésiennes des OAB de C_i opposés à C_j . Chacune de leurs faces est perpendiculaire à un axe de coordonnées de l'espace.

Une boîte B_{i_k} a la propriété de ne contenir que des OAB de la classe C_i .

On classifiera les OAB notés d_l en disant[7] que

(i) d_ℓ vient de la classe C_1 s'il existe B_{1_k} tel que $d_\ell \in B_{1_k}$ et $d_\ell \notin B_{2_j}$ pour tout j ;

(ii) d_ℓ vient de la classe C_2 s'il existe B_{2_j} tel que $d_\ell \in B_{2_j}$ et $d_\ell \notin B_{1_k}$ pour tout k ;

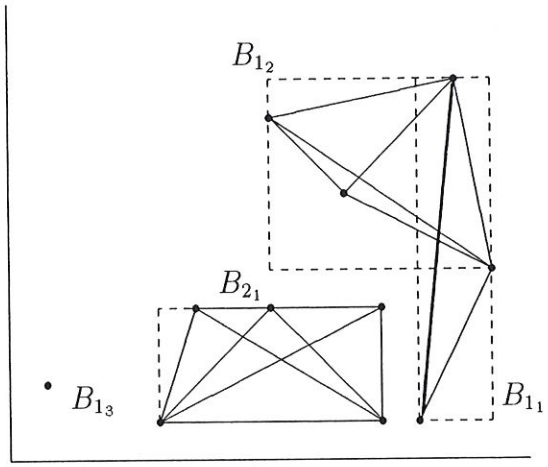
(iii) d_ℓ est rejeté de type I s'il existe B_{1_k} et B_{2_j} tels que $d_\ell \in B_{1_k}$ et $d_\ell \in B_{2_j}$;

(iv) d_ℓ est rejeté de type II s'il n'existe pas de boîte qui le contienne.

Exemple : la figure ci-dessous illustre les boîtes de la Figure 2.5.

- La boîte B_{2_1} a la propriété de contenir (couvrir) tous les objets qui forment le sous-graphe complet $\mathrm{MNG}(C_2 \mid C_1)$.
- B_{1_3} est une boîte ne contenant qu'un seul objet a_6 .

7. On peut, bien entendu, généraliser à plus de deux classes.



Les boîtes sont représentées en traits discontinus.

FIG. 2.7 – *Boîtes d'un MNG.*

Définition de la silhouette

On appelle **silhouette** de la classe C_i , notée $Sil(C_i)$, l'union des boîtes b_{i_k} de cette classe, c'est-à-dire

$$Sil(C_i) = \cup_k B_{i_k} .$$

ou, de façon équivalente,

$$Sil(C_1) = \cup_q \cup_q (a_q \oplus q_p)$$

où l'union est prise sur les couples de voisins mutuels (p, q) opposés à C_2 .

Exemple : les silhouettes C_1 et C_2 sont représentées en hachuré.

$Sil(C_1)$ est l'union des trois boîtes B_{1_1} , B_{1_2} et B_{1_3} .

$Sil(C_2)$ est la boîte B_{2_1} . Cette boîte a la propriété de contenir tous les objets formant le sous-graphe complet $MNG(C_2 \mid C_1)$.

Remarque : le MNG et la silhouette d'une classe sont des descriptions de classes **relativement** à d'autres classes (structure interclasse).

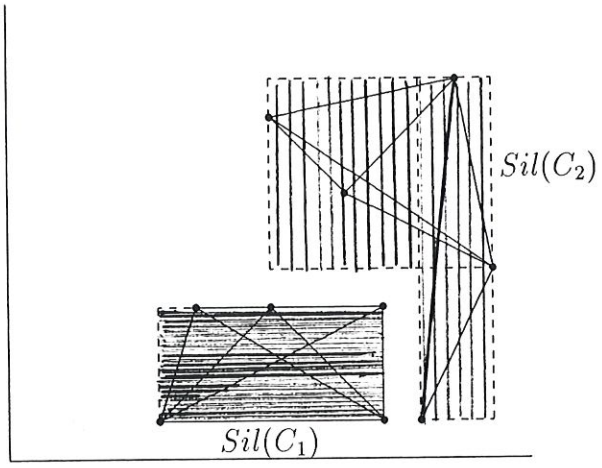


FIG. 2.8 – *Silhouettes de C_1 et C_2 .*

2.4.7 Théorème de prétendue simplicité

Présentation

On commence par énoncer une propriété concernant l'union cartésienne de deux OAB opposés à un troisième.

De celui-ci découle le théorème qui affirme l'inversabilité du voisinage mutuel par rapport à l'ajout de descripteurs. En tenant compte de ce résultat, on pourra alors déduire le théorème de Prétendue Simplicité qui avance que, sous certaines conditions, on peut rendre le $\mathrm{MNG}(C_1 \mid C_2)$ complet.

Propriété

Énoncé : *L'union cartésienne de deux objets a_i, a_j complètement distingables de b_k , est complètement distingable de b_k .*

Exemple :

$$a_1 = V_1^{a_1} \times V_2^{a_1} , \qquad a_2 = V_1^{a_2} \times V_2^{a_2} , \qquad b = V_1^{k} \times V_2^{b} .$$

Cette figure illustre (dans le plan euclidien) le cas où on a une distingabilité complète de $a_1 \oplus a_2$ de b . Cette distingabilité complète est due à la variable y_1

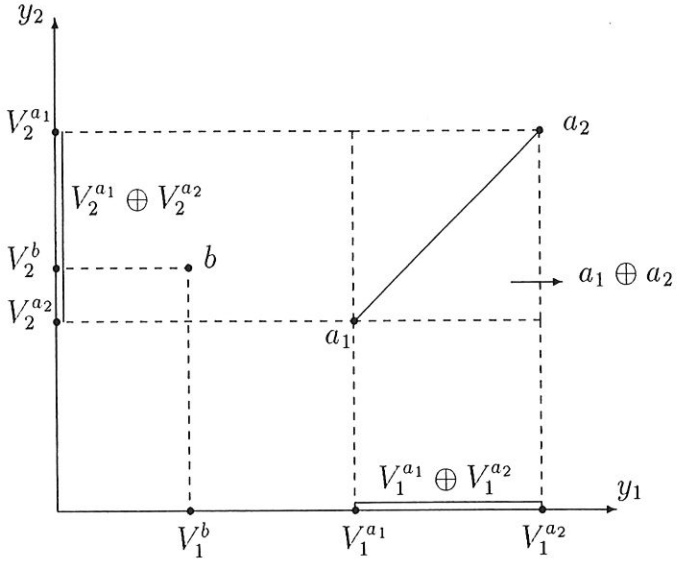


FIG. 2.9 – *Invariabilité des voisins mutuels par ajout de la variable* y_2 .

(i.e. $(V_1^{a_1} \dot{+} V_1^{a_2}) \odot V_1^b = \Phi$) et le fait que l'on ait ajouté la variable y_2 pour la description ne change rien à cette distingabilité.

Remarque : au vu de ces définitions[8] de la distinction complète de deux objets, on sait qu'il suffit que l'intersection des ensembles de valeurs d'une même variable soit vide.

Cette constatation nous amène au théorème suivant.

Théorème 1

Énoncé : *Une fois que les propriétés de voisinage mutuel entre deux OAB sont vérifiées, elles restent valables pour l'ajout d'autres descripteurs (à l'ensemble de ceux qui étaient déjà employés).*

Remarque : en tenant compte de ce résultat, on déduit le théorème suivant.

8. Voir paragraphe 2.4.2.

Théorème de prétendue simplicité

Énoncé : *si on suppose que, pour chaque paire d'OAB* a_j, a_i *de* C_1 *, il existe au moins un descripteur (* y_m *) tel que l'union cartésienne* $a_j \oplus a_i$ *est complètement distingable de tous les OAB* b_k *(* $k = 1, \cdots, N_2$ *) de* C_2 *.*

alors il sera toujours possible de rendre le MNG$(C_1 \mid C_2)$ *complet (ou presque)*[9]

9. Dans la Figure du §2.4.2, on ne pourra jamais rendre le MNG$(C_1 \mid C_2)$ tout à fait complet en raison, entre autres, de l'existence de a_6 .

Cinquième partie

Distance entre objets modaux

Introduction

Nous allons tenter de définir une distance entre objets symboliques modaux probabilistes.

Chapitre 1

Essai de définition

1.1 Présentation

Nous nous baserons sur la mesure d'information de discrimination de Kullback-Leibler. En voici la définition.

1.1.1 Définition de la mesure d'information de Kullback-Leibler

Soit Ω un espace de points ($\omega \in \Omega$).
Nous devons tester deux hypothèses :

1. *H_1 : la distribution de probabilité sur Ω est p:*
2. *H_2 : la distribution de probabilité sur Ω est π.*

L'information pour discriminer H_1 par rapport à H_2 est définie par:

$$I(p:\pi) = \sum_{\omega \in \Omega} p(\omega) ln \left(\frac{p(\omega)}{\pi(\omega)} \right) \tag{1.1}$$

dans le cas discret et

$$I(p:\pi) = \int_{\omega \in \Omega} p(\omega) ln \left(\frac{p(\omega)}{\pi(\omega)} \right) d\omega \tag{1.2}$$

dans le cas continu.

1.1.2 Remarques

1. Les formules (1.1) et (1.2) peuvent encore se récrire sous la forme commune

$$I(p:\pi) = E\left(\frac{ln(p(\omega))}{\pi(\omega)} \mid H_1\right). \tag{1.3}$$

2. L'information de discrimination mesure la divergence (dissemblance) entre les deux distributions: plus la valeur de cette mesure est élevée, plus la divergence entre ces deux distributions est importante.

1.1.3 Propriétés

Positivité

$$I(p:\pi) \geq 0.$$

Annulation

$$I(p:\pi) = 0 \iff p(\omega) = \pi(\omega) \ \forall\omega \in \Omega.$$

Ces deux propriétés découlent de l'inégalité de Jensens[1].

Caractère fini

$$I(p:\pi) < \infty \Rightarrow p(\omega) = 0 \text{ quand } \pi(\omega) = 0.$$

Nous définissons $0\,ln0 = 0$.

1. **Rappel (inégalité de Jensens):** Soit f une fonction convexe alors

$$E(f(x)) \geq f(E(x))$$

pour autant que ces espérances existent et soient finies.

1.2 Utilisation

1.2.1 Introduction

Cette mesure ne nous satisfait pas. Nous recherchons une distance et non pas simplement un indice de proximité. Il faudrait encore pour y parvenir vérifier la *symétrie* et l'*inégalité triangulaire.*

1.2.2 Symétrie

Nous allons modifier un peu la formulation précédente afin d'obtenir cette propriété.
Ainsi nous utiliserons dorénavant l'indice suivant :

$$v(p, \pi) = \frac{I(p : \pi) + I(\pi : p)}{2}. \tag{1.4}$$

Remarque

Dans le cas continu, nous pouvons encore reformuler (1.4) ainsi

$$I_v(p, \pi) = \frac{1}{2} \int_\Omega (p(\omega) - \pi(\omega)) ln \left(\frac{p(\omega)}{\pi(\omega)} \right). \tag{1.5}$$

1.2.3 Inégalité triangulaire

Malgré plusieurs tentatives de démonstration et la recherche de contre-exemples, nous ne sommes pas arriver à déterminer si l'inégalité triangulaire était ou non vérifiée.

Soient p, π, q trois distributions de probabilité sur Ω.

Conclusion et perspectives

Bien que nous ne soyons pas parvenus à démontrer s'il s'agissait ou non d'une distance, nous possédons quand même d'un indice de distance entre objets probabilistes.

Il serait toutefois très intéressant d'en construire une valable pour tous les types d'objets modaux. Ceci constitue un sujet de recherche pour l'avenir.

Conclusion

Nous avons introduit un nouveau type d'objets permettant de traiter des données plus complexes que celles étudiées par l'analyse de données classique. Ces objets sont bien adaptés pour extraire ou exprimer des informations à partir de données (non nécessairement numériques) dans un but explicatif ou décisionnel. Ces objets ont été classés en deux catégories (booléens et modaux).

Nous avons alors proposé une description sommaire de quatres types d'analyse de données et un rappel de certains outils de l'analyse de données classique. Suite à cela, nous nous sommes intéressés à l'application de ces techniques d'analyse classique sur des données symboliques. Pour y parvenir, la définition d'une distance (entre objets symboliques) était nécessaire. Dans le cas booléen, la distance généralisée de Minkowski convenait. Malheureusement, dans le cas modal, nous n'avons trouvé qu'**un indice** de distance.

La construction des techniques symboliques en est à ses débuts. Le champ de la recherche dans ce domaine est encore très vaste. Uniquement dans le cadre de ce travail, une distance entre objets symboliques modaux s'avèrerait déjà bien utile. Ajouté à cela, notons que les sujets 'analyse symbolique de données classiques' et 'analyse symbolique de données symboliques' n'ont pas été développés.

Le thème de ce mémoire m'a beaucoup intéressée. Par sa rédaction, je pense avoir un peu participé à la construction de cette nouvelle discipline. Je souhaite qu'elle continue à se développer. Il me semble qu'elle pourrait fournir de très bons résultats surtout grâce au nombre élevé d'informations qu'elle permet de prendre en compte.

Table des matières

Annexe

Chapitre 2

Idempotence

2.1 Définition.

L'ensemble des opérations

$$OP_x = \{\cup_x, \cap_x, c_x\}$$

est dit idempotent sur un ensemble Q_i si les propriétés suivantes sont satisfaites en plus de l'idempotence qui est définie par:

$$q_i \cup_x q_i = q_i$$
$$et$$
$$q_i \cap_x q_i = q_i.$$

1. Consistence :

$\emptyset \cup_x O_i^s = O_i^s \cup_x O_i^s = O_i^s \cup_x \emptyset = O_i^s$; $\emptyset \cup_x \emptyset = \emptyset$;
$\emptyset \cap_x O_i^s = \emptyset \cap_x \emptyset = O_i^s \cap_x \emptyset$; $O_i^s = O_i^s \cap_x O_i^s$.

2. Commutativité :

Commutativité de $\cup_x$ et $\cap_x$.

3. Associativité :

Associativité de $\cup_x$ *et* $\cap_x$.

4. Lois de Morgan :

il existe une loi de complémentation c_x *satisfaisant à*

(a) $c_x(\emptyset) = O_i^s$; $c_x(O_i^s) = \emptyset$.

(b) c_x *est strictement décroissante (une augmentation de* $E \subset \Omega$ *entraîne une diminution de* $c_x(E)$).

(c) c_x *est involutif :* $c_x(c_x(q_i)) = q_i$ *tel que*

$$c_x(q_i^1 \cup_x q_i^2) = c_x(q_i^1) \cap_x c_x(q_i^2)$$
$$c_x(q_i^1 \cap_x q_i^2) = c_x(q_i^1) \cup_x c_x(q_i^2).$$

5. éléments neutres :

$q_i \cup_x \emptyset = q_i$:
$q_i \cap_x O_i^s = q_i$.

6. Union et intersection :

L'union et l'intersection sont non décroissantes sur chaque argument.

2.2 Applications.

Seules les opérations

$$q_i^1 \cup_x q_i^2 = max(q_i^1, q_i^2)$$
et
$$q_i^x \cap_x q_i^2 = min(q_i^1, q_i^2)$$

sont idempotentes.
C'est le cas des sémantiques possibiliste, nécessitiste et booléenne.

Chapitre 3

Archimédien.

3.1 Définition.

Nous dirons que OP_x est stricte ou archimédien si la propriété 4 n'est pas nécessairement satisfaite, si la 6 est transformée en non décroissance stricte et si nous remplaçons l'idempotence par

$$q_i \cup_x q_i > q_i$$
$$q_i \cap_x q_i < q_i.$$

Nous exigerons encore que :

Pour $q_i \neq \{\Omega, \emptyset\}$ et $\forall\, q_i^1 > q_i^2$, on ait :

$$\begin{cases} q_i \cup_x q_i^1 > q_i^1 \cup_x q_i^2 \\ q_i \cap_x q_i^1 > q_i \cap_x q_i^2 \end{cases}$$

3.2 Application.

On a choisi OP_{pr} archimédien.
Dans ce cas, nous avions :

$$q_i^1 \cup_{pr} q_i^2 = q_i^1 + q_i^2 - q_i^1\, q_i^2$$
$$q_i^1 \cap_{pr} q_i^2 = q_i^1\, q_i^2$$
où $$q_i^1\, q_i^2(v) = q_i^1(v)\, q_i^2(v).$$

Références

3.3 Partie 1

Cette partie est construite à partir des documents [4], [3], [2], [1].

3.4 Partie 2

Nous avons utilisé les articles [4], [3].

3.5 Partie 3

Le chapitre 1 est en partie issu de [5] et le chapitre 2 se base principalement sur [6].

3.6 Partie 4

Le chapitre 1 se base sur le document [1]. Pour le chapitre 2. nous avons employé [9]. [8]. [7]. [11]. [12].

3.7 Partie 5

La mesure de Kullback-Leibler a été trouvée dans [10].

Bibliographie

[1] Francisco DE A.T. DE CARVALHO. Un indice de dissimilarité entre objets symboliques booléens.

[2] Edwin DIDAY. From data to knowledge : new objects for a statistical analysis. Rocquencourt(France).

[3] Edwin DIDAY. Quelques aspects de l'analyse des données symboliques. Rapport de recherche, 1993.

[4] Yves KODRATOFF et Edwin DIDAY. *Induction symbolique et numérique à partir de données.* Cepadue Editions. Toulouse, 1991.

[5] F. RASIR et C. REQUIER F. GÉRARD, N. JEANNÉE. L'analyse factorielle : théorie générale et cas particuliers. Examen d'Analyse Multivariée. juin 1996.

[6] André HARDY. Cours d'introduction à la classification automatique. Namur(Belgique), 1996. FUNDP.

[7] Manabu ICHINO. Feature selection for symbolic data classification. Paris. 1993. 4th Conference International Federation of Classification Society (IFCS).

[8] Manabu ICHINO and Jack SKLANSKY. The relative neighborhood graph for mixed feature variables. mai 1984.

[9] Manabu ICHINO and Hiroyuki YAGUCHI. *Generalized Minkowski Metrics for mixed feature-type data analysis.* Tokyo, avril 1994.

[10] S. KATZ and N.L. JONHSON. *Encyclopedia of statistical sciences*, volume 4, pages 421–425. 1981.

[11] Hiroyuki YAGUCHI Manabu ICHINO and Edwin DIDAY. A fuzzy symbolic pattern classifier. In *Pattern Recognition*, Tokyo,Paris, 1995. OSDA.

[12] Hiroyuki YAGUCHI Manabu ICHINO and Edwin DIDAY. A knowlegde acquisition system based on the cartesian space model. Tokyo,Paris, 1995.

[13] Sheldon M. ROSS. *Introduction to probability models.* Academic Press, New York, 1984. version française.

[14] Aimé SACHE. *La théorie des graphes*, volume 1554 of *Que sais-je?* Presse universitaires de France, Paris, 1974.