

Université Lumière Lyon2
Année 2003

Thèse
pour obtenir le grade de
Docteur
en
Informatique

présentée et soutenue publiquement par

Pierre-Emmanuel JOUVE
le 10 décembre 2003

Apprentissage Non Supervisé et Extraction de Connaissances à partir de Données

préparée au sein du laboratoire ERIC
Equipe de Recherche en Ingénierie des Connaissances

sous la direction de
Nicolas Nicoloyannis

devant le jury, composé de:

Jean-Paul Rasson, Rapporteur	Professeur, Facultés Universitaires N.D. de la Paix, Namur
Gilles Venturini, Rapporteur	Professeur, Université de Tours
Mohand-Saïd Hacid, Examinateur	Professeur, Université Claude Bernard-Lyon 1
Michel Lamure, Examinateur	Professeur, Université Claude Bernard-Lyon 1
Gilbert Ritschard, Examinateur	Professeur, Université de Genève
Nicolas Nicoloyannis, Directeur de thèse	Professeur, Université Lumière-Lyon 2

Table des matières

1	Introduction, Préambule	1
2	Concepts, Notions et Notations Utiles	7
2.1	Données Catégorielles	7
2.1.1	Domaines et Attributs Catégoriels	8
2.1.2	Objets Catégoriels	9
2.1.2.1	Similarités, Dissimilarités entre Objets Catégoriels	10
2.1.3	Ensemble d'Objets Catégoriels	11
2.1.3.1	Mode d'un Ensemble d'Objets Catégoriels	11
2.1.3.2	Similarités et Dissimilarités entre Ensembles d'Objets Catégoriels	12
2.1.3.3	Similarités et Dissimilarités au sein d'un Ensemble d'Objets Catégoriels	12
2.1.3.4	Voisinage d'une Partition d'un Ensemble d'Objets Catégoriels	13
2.2	Le Nouveau Critère de Condorcet	13
3	Classification Non Supervisée	15
3.1	Introduction	15
3.1.1	Méthodologie Générale de la Classification Non Supervisée	16
3.1.2	Applications de la Classification Non Supervisée	16
3.1.3	Taxonomies des Méthodes de Classification Non Supervisée	17
3.1.4	Méthodes de Classification Non Supervisée pour Données Catégorielles	19
3.1.5	Challenges Actuels en Classification Non Supervisée	22
3.2	Une Nouvelle Méthode de Classification Non Supervisée "Orientée Utilisateur"	24
3.2.1	Critère d'Évaluation de l'Aspect Naturel d'une Partition d'Objets	24
3.2.2	La Méthode de Classification Non Supervisée "Orientée Utilisateur"	26
3.2.2.1	Travaux Liés et Spécificités du Travail	26
3.2.2.2	L'Algorithme de Classification Non Supervisée	27

3.2.2.3	Complexité de l’Algorithme	29
3.2.2.4	Qualités de la Méthode pour l’Utilisateur	30
3.2.2.5	Illustration du Fonctionnement de l’Algorithme	30
3.2.3	Evaluation de l’Algorithme de Classification non Supervisée	31
3.2.3.1	Evaluation de la Validité des Classifications	31
3.2.3.2	Evaluation de la Stabilité	37
3.2.3.3	Evaluation de l’Efficacité Algorithmique	40
3.2.4	Eléments Additionnels	42
3.2.4.1	Valeurs Spécifiques pour le Domaine des Variables Catégorielles	42
3.2.4.2	Gestion des Valeurs Manquantes:	44
3.2.4.3	Introduction de Contraintes:	44
3.2.4.4	De l’Apprentissage Non Supervisé à l’Apprentissage Supervisé : l’Apprentissage Non Supervisé sous Contraintes	50
3.3	Conclusion	54
4	Validité en Apprentissage Non Supervisé	57
4.1	Validité d’une Classification Non Supervisée:	
	Définition et Evaluation	58
4.1.1	Mode d’Evaluation par Critères Externes	59
4.1.1.1	Méthode de Monte Carlo	59
4.1.1.2	Mesures Statistiques	60
4.1.2	Mode d’Evaluation par Critères Internes	61
4.1.3	Modes d’Evaluation Relatifs	63
4.1.3.1	Cas 1 : Le nombre final de classes, nc , n’est pas contenu dans P_{alg}	63
4.1.3.2	Cas 2 : Le nombre final de classes, nc , est contenu dans P_{alg}	64
4.1.3.3	Indices	64
4.1.4	Autres Modes d’Evaluation	67
4.2	Nouveaux Indices et Nouvelle Méthodologie pour l’Evaluation et la Comparaison de la Validité de Classifications Non Supervisées	68
4.2.1	Concepts et Formalismes Introductifs	69
4.2.1.1	Evaluation de l’homogénéité interne des classes d’une cns	71
4.2.1.2	Evaluation de la séparation entre classes d’une cns (ou hétérogénéité entre classes)	
	72	
4.2.1.3	Notions Additionnelles	73
4.2.1.4	Remarques importantes concernant l’aspect calculatoire	73

4.2.2	La nouvelle méthodologie pour l'évaluation et la comparaison de validité de cns	75
4.2.2.1	Caractérisation statistique des valeurs de : <i>LM</i> et <i>NLD</i>	76
4.2.2.2	Méthodologie	77
4.2.2.3	Expérimentations	82
4.2.2.4	Expérimentations sur le jeu de données Small Soybean Disease	82
4.2.3	Expériences sur le jeu de données <i>Mushrooms</i>	92
4.2.3.1	Description	92
4.2.3.2	Analyse des Résultats	95
4.2.4	Résumé et Informations Supplémentaires	96
5	Sélection de Variables, Contributions pour l'apprentissage supervisé et non supervisé	105
5.1	Sélection de Variables pour l'Apprentissage Supervisé	107
5.1.1	Caractéristiques de la Sélection de Variables	107
5.1.2	Les Types de Méthodes	107
5.1.3	Directions de Recherche	108
5.1.3.1	Forward Selection (FS) (Ajout de variables) . .	108
5.1.3.2	Backward Elimination (BE) (Suppression de variables)	109
5.1.3.3	Méthodes Bidirectionnelles	109
5.1.4	Stratégie de Recherche	109
5.1.5	Fonction d'Evaluation	110
5.1.6	Critère d'Arrêt	111
5.1.7	Approches Filtres	111
5.1.8	Approches Enveloppes	114
5.1.9	Autres Approches	115
5.2	Contribution à la Sélection de Variables pour l'Apprentissage Supervisé: Une Nouvelle Méthode Efficiente et Rapide	118
5.2.1	Hypothèses et Idées Fondamentales	118
5.2.2	Evaluation de la Validité d'une Partition dans un Sous-Espace de l'ERD	119
5.2.3	La Nouvelle Méthode de Sélections de Variables	120
5.2.3.1	La Méthode de Base : une Méthode Exhaustive .	121
5.2.3.2	Réduction de la Complexité par Introduction d'un AG	124
5.2.4	Evaluation Expérimentale	126
5.2.4.1	Présentation de l'Evaluation Expérimentale .	126
5.2.4.2	Analyse de l'Evaluation Expérimentale	127
5.2.5	Conclusion	131
5.3	Contribution à la Sélection de Variables pour l'Apprentissage Non Supervisé: Une Nouvelle Méthode Efficiente et Rapide . .	143

5.3.1	Evaluation de l'Adéquation entre deux Ensembles de Variables	144
5.3.2	Remarques Importantes Concernant l'Aspect Calculatoire	145
5.3.3	Evaluation de l'adéquation entre EV un Ensemble de Variables et EV_* un Sous Ensemble de EV ($EV_{*} \subseteq EV$)	146
5.3.4	Evaluation/Comparaison de l'Adéquation entre un Ensemble de Variables (EV) et des Sous Ensembles de EV	148
5.3.5	La Nouvelle Méthode de Sélection de Variables	148
5.3.6	Evaluations Expérimentales	149
5.3.6.1	Expérience #1 : Evaluation expérimentale sur jeux de données synthétiques	149
5.3.6.2	Expérience #2 : Evaluation Expérimentale sur Jeux de Données de l'UCI	154
5.3.7	Conclusion	162
6	Agrégation de Classifications Non Supervisées : La Problématique "Cluster Ensembles"	165
6.1	Introduction	165
6.1.1	Illustration de la Problématique "Cluster Ensembles"	167
6.1.2	Motivations, Objectifs de la Problématique "Cluster Ensembles"	168
6.1.2.1	Réutilisation de Connaissances	169
6.1.2.2	Calcul Distribué pour la cns	169
6.1.3	Travaux Liés	172
6.1.4	Principaux Challenges pour la Problématique "Cluster Ensembles"	175
6.2	Mesures d'Adéquation	175
6.2.1	Adéquation entre Classifications Non Supervisées	176
6.2.2	Adéquation pour un Couple de Classification Non Supervisée	176
6.2.3	Adéquation entre une Classification Non Supervisée et un Ensemble de Classifications Non Supervisées	177
6.3	Contribution à la Problématique "Cluster Ensembles" : Trois Méthodes pour l'Agrégation de Classifications Non Supervisées	177
6.3.1	Première Méthode pour l'Agrégation de cns : Une Méthode Intuitive	178
6.3.2	Seconde Méthode pour l'Agrégation de Classifications Non Supervisées : Utilisation de la Méthode KEROUAC	179
6.3.2.1	Utilisation de KEROUAC pour la cns en considérant des Méta-Variables	179
6.3.2.2	Relation entre P_{*} and P_{β}	180
6.3.2.3	Conclusion	181
6.3.2.4	Illustration	181

6.3.2.5	Propriétés de la Méthode	182
6.3.3	Troisième Méthode pour l’Agrégation de Classifications Non Supervisées : Utilisation de la Méthode K-Modes . .	183
6.3.3.1	Illustration	184
6.3.3.2	Propriétés de la Méthode	184
6.3.4	Evaluations Expérimentales	184
6.3.4.1	Evaluations, Comparaisons et Discussions Pré- liminaires	184
6.3.4.2	Evaluations, Comparaisons et Discussions Com- plémentaires	191
6.3.4.3	Comportement de la méthode KEROUAC face à des cns à agréger possédant des nombre de classes très différents	204
6.4	Conclusion	207
7	Conclusion	211
8	Données Utilisées pour les Expérimentations	217
8.1	Jeu de Données ADULT	217
8.2	Jeu de Données MUSHROOMS	218
8.3	Jeu de Données BREAST CANCER	220
8.4	Jeu de Données CAR	222
8.5	Jeu de Données : ADULT	224
8.6	Jeu de Données Contraceptive Method Choice	225
8.7	Jeu de Données FLAGS	226
8.8	Jeu de Données GERMAN	227
8.9	Jeu de Données HOUSE VOTES 84	229
8.10	Jeu de Données IONOSPHERE	230
8.11	Jeu de Données MONKS	231
8.12	Jeu de Données NURSERY	232
8.13	Jeu de Données PIMA	234
8.14	Jeu de Données SICK	235
8.15	Jeu de Données SMALL SOYBEAN DISEASES	236
8.16	Jeu de Données VEHICLE	237
8.17	Jeu de Données WINE	240
8.18	Jeu de Données SPAM	241
Bibliographie		243
Table des figures		254
Liste des tableaux		257

2 Concepts, Notions et Notations Utiles

"Un concept est une invention à laquelle rien ne correspond exactement mais à laquelle nombre de choses ressemblent."

- Friedrich Nietzsche -
Posthumes

Ce chapitre est l'occasion de présenter un ensemble de concepts, notions et notations qui seront utilisés tout au long de ce document. Les raisons sous-jacentes à la rédaction d'un tel chapitre sont doubles : il constituera une base de théorie nécessaire ultérieurement et son introduction préalable permettra une approche plus directe et intuitive des développements proposés plus tard. De plus, il nous semble utile et intéressant de regrouper l'ensemble de ces informations en une unique entité à laquelle on se référera facilement. Les notions introduites ici sont relatives tout d'abord au concept de données catégorielles, et enfin au Nouveau Critère de Condorcet.

2.1 Données Catégorielles

L'ensemble des terminologies et formalismes que nous utiliserons pour introduire les données catégorielles proviennent de multiples références de la littérature que nous ne manquerons pas d'évoquer. La forme de cette présentation s'inspire quant à elle de [Hua97].

Afin d'assurer une plus grande clarté nous nous appuierons sur des exemples basés sur un jeu de données décrivant un ensemble de 3 votes de motions différentes par 54 nations lors de sessions à l'O.N.U.(voir Tableau 2.1).

On définit (de manière tautologique) les données catégorielles comme les données décrivant des objets par l'intermédiaire de caractéristiques catégorielles. Les objets décrits par un ensemble de données catégorielles, sont nommés en conséquence objets catégoriels, ils correspondent à une version très simplifiée des objets symboliques définis dans [GD91]. Ces objets ne peuvent posséder de caractéristiques numériques (quantitatives), si tel est le cas on doit

Pays	M_1	M_2	M_3	Pays	M_1	M_2	M_3	Pays	M_1	M_2	M_3
DOMI	A	A	A	PANA	C	A	A	FRAN	C	B	C
POLA	A	A	C	VANE	C	A	A	SWED	C	B	C
HUNG	A	A	C	PERU	C	A	A	NORW	C	B	C
CZEC	A	A	C	CHIL	C	A	A	DENM	C	B	C
YUGO	A	A	C	ARGE	C	A	A	USA	C	C	A
BULG	A	A	C	GREE	C	A	A	UK	C	C	A
ROMA	A	A	C	CYPR	C	A	A	NETH	C	C	A
USSR	A	A	C	CANA	C	B	A	BELG	C	C	A
UKRA	A	A	C	HOND	C	B	A	LUXE	C	C	A
BYEL	A	A	C	ELS	C	B	A	URUG	C	D	A
CUBA	A	D	C	NICA	C	B	A	EQUA	C	D	B
ALBA	A	D	C	BRAZ	C	B	A	HAIT	D	A	A
FINL	B	B	C	PARA	C	B	A	GUYA	D	A	A
JAMA	C	A	A	IREL	C	B	A	BOLI	D	A	A
TRIN	C	A	A	SPAIN	C	B	A	BARB	D	A	B
MEXI	C	A	A	ITAL	C	B	A	COLU	D	B	A
GUAT	C	A	A	ICEL	C	B	A	PORT	D	C	B
COST	C	A	A	AUST	C	B	B	MALT	D	D	A

TAB. 2.1 –: Votes à l’O.N.U. de 54 pays différents pour 3 motions différentes

s’astreindre à une phase de discréétisation de ces caractéristiques afin d’uniformiser la description de ces objets¹.

2.1.1 Domaines et Attributs Catégoriels

Dans l’ensemble de ce document, nous considérons qu’un jeu de données est caractérisé par un ensemble de p variables notées V_1, V_2, \dots, V_p décrivant un espace EV ($EV = \{V_1, V_2, \dots, V_p\}$). Nous notons $Dom(V_1), Dom(V_2), \dots, Dom(V_p)$ les domaines respectifs des variables de EV .

EXEMPLE : $EV = \{V_1, V_2, V_3\}$ (V_1, V_2, V_3 correspondent respectivement aux variables nommées M_1, M_2 et M_3).

Définition 1 Un domaine $Dom(V_j) = \{v_{j1}, \dots, v_{jk}\}$, ($k \in N^*$) est défini comme catégoriel s’il est fini, et non ordonné.

Ainsi $\forall a, b \in Dom(V_j)$ les seules relations pouvant exister entre a et b sont : $a = b$ ou $a \neq b$.

V_j est en conséquence appelée variable catégorielle.

(concernant l’aspect non ordonné, il n’est pas nécessaire que l’aspect non ordonné soit réel mais on ne tiendra pas compte ultérieurement de cet ordre s’il existe)

1. Notons qu’une perte d’information plus ou moins forte est associée à ce processus

EXEMPLE : $Dom(V_1) = \{A, B, C, D\}$, $Dom(V_2) = \{A, B, C, D\}$, $Dom(V_3) = \{A, B, C\}$.

Définition 2 $EV = \{V_1, \dots, V_p\}$ est un espace catégoriel si $\forall V_j, j \in 1..p$, V_j est une variable catégorielle.

Notons, que les domaines catégoriels sont définis par des ensembles de singltons ainsi des valeurs provenant de combinaisons ne sont pas autorisées a contrario des travaux présentés dans [GD91].

Nous définissons un ensemble de valeurs additionnelles spécifiques pour les domaines des variables catégorielles. Cet ensemble de valeurs (noté $E_\varepsilon = \{\varepsilon_i\}$) permet de représenter des cas particuliers comme ceux de la présence de valeurs manquantes (nous reviendrons ultérieurement sur les spécificités, l'intérêt et la signification de ces valeurs).

Afin de simplifier la présentation nous ne considérerons pas les relations d'inclusions conceptuelles pouvant exister au sein de bases de données provenant de la pratique contrairement aux travaux de Kodratoff et Tecuci [KT88].

2.1.2 Objets Catégoriels

Dans l'ensemble de ce document, nous considérons qu'un jeu de données est également caractérisé par un ensemble O de n objets, ces objets sont notés o_i ($O = \{o_1, \dots, o_n\}$). Ainsi un jeu de données est caractérisé par les ensembles $EV = \{V_1, V_2, \dots, V_p\}$ et $O = \{o_1, \dots, o_n\}$.

Cette section est consacrée à l'introduction des objets catégoriels, qui, comme dans [GD91], sont représentés par une conjonction logique de paires attributs-valeurs (Une paire attribut-valeur est dénommée sélecteur dans [MS83].). L'objet catégoriel o_i est ainsi décrit par la règle $[V_1 = o_{i_1}] \cap [V_2 = o_{i_2}] \cap \dots \cap [V_p = o_{i_p}]$.

EXEMPLE : $FRAN = [V_1 = C] \cap [V_2 = B] \cap [V_3 = C]$

En conséquence, nous représenterons chaque objet $o_i \in O$ par l'ensemble de p valeurs $\{o_{i_1}, o_{i_2}, \dots, o_{i_p}\}$ (chaque objet possède exactement p valeurs d'attributs).

EXEMPLE : $FRAN = [C, B, C]$.

REMARQUES :

- Si la valeur d'un attribut V_j est non disponible pour un objet o_i , on fixe alors $o_{i_j} = \varepsilon_k$, $\varepsilon_k \in E_\varepsilon$. Afin de simplifier ce chapitre introductif, nous considérerons en cas de valeur manquante pour un objet o_i qu'il lui est assigné la valeur ε_1 dont le comportement est identique aux autres modalités de cet attribut.
- $o_i = o_j$ si $\forall k, o_{i_k} = o_{j_k}$. Cette dernière relation n'implique toutefois pas que o_i et o_j représentent le même objet du jeu de données, mais elle signifie qu'ils possèdent les même valeurs catégorielles pour les attributs

V_1, V_2, \dots, V_p .

EXEMPLE : "FRAN" \neq "DENM" mais on a $[C, B, C] = [C, B, C]$.

- Soient $C = \{o_1, o_2, \dots, o_n\}$ un ensemble de n objets catégoriels au sein duquel p objets sont distincts, N la cardinalité du produit cartésien $Dom(V_1) \otimes Dom(V_2) \otimes \dots \otimes Dom(V_p)$. On a alors $p \leq N$, n peut quant à lui être inférieur, égal ou supérieur à N , ce dernier cas impliquant obligatoirement la présence de "doublons" dans C .

2.1.2.1 Similarités, Dissimilarités entre Objets Catégoriels

Nous définissons maintenant les notions de similarité et dissimilarité entre objets catégoriels. Nous venons d'indiquer qu'une notion de similarité (ou de dissimilarité) globale entre objets catégoriels existe : deux objets peuvent être similaires sans pour autant être les mêmes (cf. exemple précédent concernant les objets FRAN et DENM). Cette similarité globale implique un ensemble de similarités locales (au niveau de chaque variable), nous pouvons définir naturellement deux fonctions $\delta_{sim}(o_{a_i}, o_{b_i})$ et $\delta_{dissim}(o_{a_i}, o_{b_i})$ qui mesurent la similarité de deux objets o_{a_i} et o_{b_i} au niveau de la variable V_i :

$$\delta_{sim}(o_{a_i}, o_{b_i}) = \begin{cases} 1 & \text{si } o_{a_i} = o_{b_i} \\ 0 & \text{si } o_{a_i} \neq o_{b_i} \end{cases} \quad (2.1)$$

$$\delta_{dissim}(o_{a_i}, o_{b_i}) = \begin{cases} 0 & \text{si } o_{a_i} = o_{b_i} \\ 1 & \text{si } o_{a_i} \neq o_{b_i} \end{cases} \quad (2.2)$$

Ainsi, $\delta_{sim}(o_{a_i}, o_{b_i})$ (resp. $\delta_{dissim}(o_{a_i}, o_{b_i})$) vaut 1 si les objets o_{a_i} et o_{b_i} sont similaires (resp. dissimilaires) au niveau de la variable V_i .

REMARQUES :

- Les définitions de $\delta_{sim}(o_{a_i}, o_{b_i})$ et $\delta_{dissim}(o_{a_i}, o_{b_i})$ sont telles que $\delta_{sim}(o_{a_i}, o_{b_i}) = 1 - \delta_{dissim}(o_{a_i}, o_{b_i})$.
- Nous pourrions donc nous contenter de n'introduire qu'une seule de ces deux fonctions étant donnée la relation les unissant. Toutefois les développements futurs "cassant" cette relation, nous utiliserons tout au long du document ces deux fonctions afin de rendre plus intelligibles ces mêmes développements ultérieurs.

Ces mesures peuvent être étendues à l'ensemble des variables de EV de manière à rendre compte du degré de similarité globale entre les deux objets :

$$sim(o_a, o_b) = \sum_{i=1}^p \delta_{sim}(o_{a_i}, o_{b_i}), \quad 0 \leq sim(o_a, o_b) \leq p \quad (2.3)$$

$$dissim(o_a, o_b) = \sum_{i=1}^p \delta_{dissim}(o_{a_i}, o_{b_i}) \quad 0 \leq dissim(o_a, o_b) \leq p \quad (2.4)$$

Ainsi, plus $sim(o_a, o_b)$ (resp. $dissim(o_a, o_b)$) est proche de p plus o_a et o_b peuvent être considérés comme similaires (resp. dissimilaires).

REMARQUES :

- Les définitions de $sim(o_{a_i}, o_{b_i})$ et $dissim(o_{a_i}, o_{b_i})$ sont telles que $sim(o_{a_i}, o_{b_i}) = p - dissim(o_{a_i}, o_{b_i})$.
- Nous pourrions donc nous contenter de n'introduire qu'une seule de ces deux fonctions étant donnée la relation les unissant. Toutefois les développements futurs "cassant" cette relation, nous utiliserons tout au long du document ces deux fonctions afin de rendre plus intelligible ces mêmes développements ultérieurs.

2.1.3 Ensemble d'Objets Catégoriels

Nous introduisons maintenant un ensemble de notions relatives aux ensembles d'objets catégoriels.

Soit $C = \{o_a, o_b, \dots, o_h\}$ un ensemble de h objets catégoriels ($C \subseteq O$).

2.1.3.1 Mode d'un Ensemble d'Objets Catégoriels

Nous notons :

- $n_{C_{k,j}}$ le nombre d'objets de C ayant la valeur v_{jk} pour la variable $V_j \in EV$
- $f_r(V_j = v_{jk}|C) = n_{C_{k,j}}/card(C)$ la fréquence relative de la valeur v_{jk} pour V_j au sein de l'ensemble d'objets C .

Définition 3 [Hua97] *Le mode d'un ensemble d'objet C est l'objet virtuel $mode^C$ ($mode^C = [mode_j^C, j = 1..p]$) tel que pour toute variable $V_j \in EV$ la valeur d'attribut de $mode^C$ est, celle, la plus représentée pour cette variable au sein de la classe C : $\forall j = 1..p, \forall o_i \in C, f_r(V_j = mode_j^C|C) \geq f_r(V_j = o_i|C)$.*

En clair, le mode d'un ensemble d'objet C correspond au profil de cet ensemble, à l'objet type de cet ensemble

REMARQUES : Cette définition implique que :

- le mode d'un ensemble d'objets n'est pas forcément unique
EXEMPLE : le mode de l'ensemble $C = \{BARB, COLU, PORT, MALT\} = \{\{D, A, B\}, \{D, B, A\}, \{D, C, B\}, \{D, D, A\}\}$ peut être $mode^C = [D, A, A]$, ou $mode^C = [D, B, A]$, ou bien $mode^C = [D, C, A]$, ou $mode^C = [D, D, A]$, ou $mode^C = [D, A, B]$, ou bien $mode^C = [D, B, B]$, ou encore $mode^C = [D, C, B]$, ou finalement $mode^C = [D, D, B]$.
- le mode d'un ensemble d'objets n'est pas forcément un élément de cet ensemble, EXEMPLE : le mode de $C = \{COLU, FRAN, US\} = \{[D, B, A], [C, B, C], [C, C, A]\}$ est $[C, B, A]$.

2.1.3.2 Similarités et Dissimilarités entre Ensembles d'Objets Catégoriels

Tout comme nous l'avons fait pour des couples d'objets catégoriels, nous introduisons maintenant des fonctions permettant de rendre de compte du degré de similarité ou de dissimilarité de deux ensembles d'objets catégoriels.

Afin de traduire le niveau de similarité (resp. dissimilarité) entre deux ensembles d'objets catégoriels, nous utiliserons la fonction $Sim(C_i, C_j)$ (resp. $Dissim(C_i, C_j)$) qui détermine le nombre de similarités (resp. dissimilarités) entre deux ensembles d'objets différents C_i et C_j ($C_i \neq C_j$).

$$Sim(C_i, C_j) = \sum_{o_a \in C_i, o_b \in C_j} sim(o_a, o_b) \quad (2.5)$$

$$0 \leq Sim(C_i, C_j) \leq \text{card}(C_i) \times \text{card}(C_j) \times p$$

$$Dissim(C_i, C_j) = \sum_{o_a \in C_i, o_b \in C_j} dissim(o_a, o_b) \quad (2.6)$$

$$0 \leq Dissim(C_i, C_j) \leq \text{card}(C_i) \times \text{card}(C_j) \times p$$

Ainsi, plus $Sim(C_i, C_j)$ (resp. $Dissim(C_i, C_j)$) est proche de $\text{card}(C_i) \times \text{card}(C_j) \times p$ plus C_i et C_j peuvent être considérés comme similaires (resp. dissimilaires).

2.1.3.3 Similarités et Dissimilarités au sein d'un Ensemble d'Objets Catégoriels

Les notions de similarités (resp. dissimilarités) au sein d'un ensemble d'objets catégoriels correspondent au nombre de similarités (resp. dissimilarités) entre objets d'un ensemble C_i et constituent une extension des définitions préalablement établies pour les similarités et dissimilarités entre ensembles d'objets catégoriels.

$$Sim(C_i) = \sum_{o_a \in C_i, o_b \in C_i, a < b} sim(o_a, o_b), \quad (2.7)$$

$$0 \leq Sim(C_i) \leq \frac{\text{card}(C_i) \times (\text{card}(C_i) - 1) \times p}{2}$$

$$Dissim(C_i) = \sum_{o_a \in C_i, o_b \in C_i, a < b} dissim(o_a, o_b) \quad (2.8)$$

$$0 \leq Dissim(C_i) \leq \frac{\text{card}(C_i) \times (\text{card}(C_i) - 1) \times p}{2}$$

Ainsi, plus $Sim(C_i)$ (resp. $Dissim(C_i)$) est proche de $\frac{\text{card}(C_i) \times (\text{card}(C_i) - 1) \times p}{2}$ plus C_i présente une forte homogénéité (resp. hétérogénéité) interne.

2.1.3.4 Voisinage d'une Partition d'un Ensemble d'Objets Catégoriels

Nous notons: $P_h = \{C_1, \dots, C_z\}$ une partition de O en z classes et P_g une partition de O en y classes.

Définition 4 Nous dirons qu'une partition P_g appartient à $\text{Vois}(P_h)$ le voisinage d'une partition P_h ou encore l'ensemble des partitions voisines d'une partition P_h si :

- P_g peut être obtenue à partir de P_h par segmentation d'une classe C_j de P_h selon une variable V_i (processus équivalent à la segmentation des arbres de décision)
- P_g peut être obtenue à partir de P_h par fusion de deux classes de P_h
- $P_g = P_h$

EXEMPLE : Soient $O = \{\text{BARB}, \text{COLU}, \text{EQUA}, \text{MALT}\}$;
 $P_1 = \{\{\text{COLU}, \text{MALT}\}, \{\text{BARB}, \text{EQUA}\}\}$.

Le voisinage de P_1 est alors constitué des partitions de O provenant de :

- la fusion de deux classes de P_1 , il s'agit uniquement ici de la partition $P_2 = \{\text{BARB}, \text{COLU}, \text{EQUA}, \text{MALT}\}$;
- la segmentation d'une classe de P_1 selon une des variables catégorielles, il s'agit ici des partitions :
 - $P_3 = \{\{\text{COLU}\}, \{\text{MALT}\}, \{\text{BARB}, \text{EQUA}\}\}$ (segmentation de la classe $\{\text{COLU}, \text{MALT}\}$ selon M_2),
 - $P_4 = \{\{\text{COLU}, \text{MALT}\}, \{\text{BARB}\}, \{\text{EQUA}\}\}$ (segmentation de la classe $\{\text{BARB}, \text{EQUA}\}$ selon M_1 ou M_2);
- P_1 elle-même.

Ainsi $\text{vois}(P_1) = \{P_1, P_2, P_3, P_4\}$.

2.2 Le Nouveau Critère de Condorcet

Nous présentons maintenant une mesure particulière pour l'évaluation de la qualité d'une partition : le Nouveau Critère de Condorcet² (*NCC*) introduit par Pierre Michaud. Le lecteur désirant une description de ce critère plus complète que celle que nous proposons peut se référer aux travaux de Michaud suivants [Mic82], [Mic83], [Mic85], [Mic87], [Mic97].

Dans le cadre de ces travaux, Michaud a proposé une nouvelle mesure de l'aspect naturel d'une partition : le *NCC*. Cette mesure est dérivée de la théorie de l'agrégation de votes, ce qui apparaît plus clairement si l'on considère la correspondance entre :

- une variable catégorielle définissant une partition sur un ensemble d'objets, et
- un juge donnant son opinion sur la ressemblance/similarité d'objets.

2. Le nom de ce critère est motivé par l'existence d'un critère similaire dans le cadre de l'agrégation de rangs, critère qui fut proposé par Condorcet en 1785

EXEMPLE : Considérons, par exemple, l'ensemble d'objets O ainsi que la variable M_1 ($O = \{DOMI, COLU, EQUA, MALT\}$). Cette variable définit la partition suivante de O : $\{\{DOMI\}, \{MALT\}, \{COLU, EQUA\}\}$. On peut ainsi l'associer naturellement à un juge pour qui l'ensemble d'objets s'organiseraient selon 3 groupes d'objets similaires.

Les variables dans le cadre de la classification non supervisé (cns³), ou les juges dans l'optique agrégation de votes, caractérisant les objets sont multiples. La recherche de la bonne classification s'apparente alors à un problème d'agrégation d'opinions des juges. Définir une méthode idéale pour l'agrégation des opinions peut être impossible à réaliser, cependant divers auteurs, tel K.J. Arrow [Arr63], ont axiomatisé les propriétés désirables pour une bonne agrégation. Le NCC vérifie certaines de ces propriétés que nous ne détaillons pas ici.

La mesure de l'aspect naturel d'une partition P_h (Nous notons $P_h = \{C_1, \dots, C_z\}$ une partition de O en z classes) $NCC(P_h)$ est définie comme suit :

$$NCC(P_h) = \sum_{i=1..z, j=1..z, i < j} Sim(C_i, C_j) + \sum_{i=1}^z Dissim(C_i) \quad (2.9)$$

$$0 \leq NCC(P_h) \leq \sum_{i=1..z, j=1..z, i < j} \text{card}(C_i) \times \text{card}(C_j) \times p + \sum_{i=1}^z \frac{\text{card}(C_i) \times (\text{card}(C_i) - 1) \times p}{2}$$

Le critère $NCC(P_h)$ comptabilise donc le nombre de dissimilarités internes à chacune des classes de la partition P_h ainsi que le nombre de similarités entre classes de P_h . Ainsi, plus il est faible (proche de 0) plus cela signifie que la partition présente un faible nombre de dissimilarités internes pour chacune des classes et un faible nombre de similarités entre classes ; cela signifie donc que plus la valeur du critère NCC est proche de 0 plus la partition apparaît comme naturelle et semblant traduire la structure interne des données.

3. Par la suite, l'acronyme cns remplacera l'expression classification non supervisée d'une part pour évoquer le processus de classification non supervisée et d'autre part pour évoquer le résultat de ce processus (i.e. une partition de l'ensemble des objets). Notons que cet acronyme est invariablement utilisé pour rendre compte d'un ou plusieurs processus ou de un ou plusieurs résultats de processus...