Consensus de classifications basé sur les regroupements fréquents

Bruno Leclerc

École des Hautes Études en Sciences Sociales
Centre d'Analyse et de Mathématique Sociales (CNRS UMR 8557)
54 boulevard Raspail, 75270 Paris cedex 06, France
leclerc@ehess.fr

Résumé. Les classifications considérées ici sont des ensembles de classes deux à deux incomparables pour l'inclusion, par exemple des partitions, ou simplement une classe unique. Soit $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k)$ un profil de telles classifications sur un ensemble fixé fini E, que l'on veut agréger en une classification unique $\mathcal{D}$. Pour un entier p compris entre 1 et k (inclus), on définit un *consensus par regroupements fréquents* en considérant les classes maximales incluses dans des éléments d'au moins p des $\mathcal{D}_i$. On étudie les propriétés de cette règle de consensus et on en donne trois caractérisations.

1 Introduction

Nous nous intéressons dans cet article à des regroupements d'objets apparaissant fréquemment dans une collection de classifications (produits, par exemple, par des itérations successives d'un même algorithme). On se propose d'étudier les systèmes de classes obtenus de cette façon, que nous appelons ici *regroupements fréquents*.

Des considérations provenant de divers domaines vont être évoquées. En particulier, un résultat obtenu précédemment aura un rôle important. Soient E un ensemble fini fixé et $R \subseteq (\mathcal{P}(E))^2$ une relation binaire sur l'ensemble $\mathcal{P}(E)$ des parties de E. Nous avons montré dans plusieurs travaux antérieurs (Domenach et Leclerc 2004b, Leclerc 2004, Leclerc 2005) l'unicité d'une classification $\mathcal{D}$, sous forme d'une famille de Moore (la définition d'une telle famille est donnée ci-dessous) vérifiant deux conditions relatives à R et généralisant celles posées par Adams (1986) pour le consensus d'arbres de classification ; ces conditions portent sur l'ajustement à R de la relation d'emboîtement (définie au paragraphe 5 ci-dessous) de $\mathcal{D}$. On a alors un problème d'existence, car une telle classification $\mathcal{D}$ n'existe pas pour toute relation R. En fait, Adams établit cette existence dans le cas particulier qu'il considère, celui du consensus de hiérarchies selon une forme de règle d'unanimité. Nous allons montrer que les regroupements fréquents correspondent à des fonctions de consensus du type de celle d'Adams, mais appliquées à des objets différents, et moins contraignantes que des fonctions d'unanimité. De plus, la détermination par ces fonctions d'une classification consensus $\mathcal{D}$ est proche de celle des "motifs fréquents", qui constitue un thème majeur en fouille des données pour la recherche de règles d'association (cf., e.g., Hipp et al. 2000, Han et Kamber 2001).

Les principales définitions, dont celle des fonctions de consensus par regroupements fréquents, sont données au paragraphe 2, avec quelques propriétés de ces fonctions. Au paragraphe 3, nous signalons le lien des regroupements fréquents avec les motifs fréquents de la littérature, et les conséquences potentiellement intéressantes du point de vue algorithmique de ce lien. Au paragraphe 4, nous présentons une première caractérisation des fonctions de consensus par regroupements fréquents dans l'esprit du théorème d'Arrow. Au paragraphe 5, nous rappelons d'abord les définitions des relations d'implication et d'emboîtement associées à tout ensemble de classes. Nous donnons alors une caractérisation des fonctions de consensus par regroupements fréquents en termes d'emboîtements (donc, dans la suite d'Adams), puis sa variante en termes d'implications (autrement dit, de règles d'association exactes).

2 Définitions et propriétés de base

Les classifications que nous considérons dans cet article sont des ensembles de classes d'objets pris dans un ensemble fini E à n éléments ($n \geq 2$). Une classification est donc ici une famille $\mathcal{D}$ de parties de E. La principale restriction est que l'on suppose que $\mathcal{D}$ est une *famille de Sperner*, i.e. dont les classes sont deux à deux incomparables pour l'inclusion, qui de plus est *propre*, soit $\mathcal{D} \neq \emptyset$ et $\mathcal{D} \neq \{E\}$. Alors, la classification $\mathcal{D}$ est un *recouvrement de* E si l'union de ses classes est E et une *partition de* E si, de plus, ses classes sont deux à deux disjointes. On s'intéressera aussi au cas particulier d'une famille $\mathcal{D}$ réduite à une classe unique C (avec $C \subset E$ puisque $\mathcal{D}$ est une famille de Sperner propre).

On note $\mathbf{S}$ l'ensemble de toutes les familles de Sperner propres sur E, et $\mathbf{S}^* = \mathbf{S} \cup \{\emptyset\}$. Une partie A de E est dite un *regroupement de* $\mathcal{D}$ s'il existe au moins une classe C de $\mathcal{D}$ contenant A.

Nous nous intéresserons aussi à un autre type de familles de parties. Une famille $\mathcal{M}$ de parties de E est une *famille de Moore* (ou un *système de fermeture*) s'il vérifie les deux propriétés suivantes :

(i) $E \in \mathcal{M}$,

(ii) pour tous $A, B \in \mathcal{M}$, $A \cap B \in \mathcal{M}$.

Une telle famille de Moore $\mu(\mathcal{D}) = \{\bigcap \mathcal{D}' : \mathcal{D}' \subseteq \mathcal{D}\}$ est associée à toute famille (de Sperner ou non) $\mathcal{D}$ de parties de E. On a notamment $\mathcal{M} = \{E\}$ pour $\mathcal{D}' = \emptyset$.

Soit $\boldsymbol{\mathcal{D}} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k) \in \mathbf{S}^k$ un *profil* de classifications. Nous souhaitons agréger $\boldsymbol{\mathcal{D}}$ en une famille de Sperner unique $\mathcal{D}$. Posons $K = \{1, 2, \ldots, k\}$, et associons au profil $\boldsymbol{\mathcal{D}}$ un *indice de regroupement* $g_{\boldsymbol{\mathcal{D}}}$ défini sur l'ensemble $\mathcal{P}(E)$ des parties de E. On pose, pour tout $A \subseteq E$,

$g_{\boldsymbol{\mathcal{D}}}(A) = |\{i \in K : A \subseteq C \text{ pour au moins une classe } C \text{ de } \mathcal{D}_i\}|$.

La valeur de $g_{\boldsymbol{\mathcal{D}}}(A)$ est donc entière et correspond au nombre des classifications du profil $\boldsymbol{\mathcal{D}}$ dont A est un regroupement. Pour $p \in K$, une partie A de E est dite un *regroupement p-fréquent* si on a $g_{\boldsymbol{\mathcal{D}}}(A) \geq p$.

On définit alors, pour chaque entier $p \in K$, une fonction de consensus $F_p : \mathbf{S}^k \rightarrow \mathbf{S}^*$ à partir de l'indice $g_{\boldsymbol{\mathcal{D}}}$. Le *consensus par regroupements p-fréquents* de $\boldsymbol{\mathcal{D}}$, noté $F_p(\boldsymbol{\mathcal{D}})$ est la famille de Sperner des regroupements p–fréquents maximaux ; il est à noter que $F_p(\boldsymbol{\mathcal{D}})$ peut être vide, d'où la nécessité d'élargir l'ensemble d'arrivée de F_p à $\mathbf{S}^*$. Cependant, en prenant

l'entier p suffisamment petit, $F_p(\mathcal{D})$ appartient à $\mathbf{S}$. Ainsi, $F_1(\mathcal{D})$ est l'ensemble des classes de $\bigcup_{1\le i\le k} \mathcal{D}_i$ qui sont maximales pour l'inclusion. A l'opposé, la fonction F_k correspond à une sorte de règle d'unanimité, $F_k(\mathcal{D})$ étant l'ensemble des parties C de E de la forme $C = \bigcap_{1\le i\le k} C_i$, avec $C_i \in \mathcal{D}_i$ pour tout $i \in K$, et maximales avec cette propriété. Ainsi, si $\mathcal{D}$ est un profil de partitions de E, on retrouve le croisement des partitions de $\mathcal{D}$.

Bien qu'il paraisse assez naturel de s'y intéresser, les regroupements fréquents ne semblent pas avoir été souvent pris en considération dans la littérature. Les "formes fortes" de Diday (1971, dans le cadre des "nuées dynamiques") constitutent une exception remarquable.

La fonction de consensus F_p de $\mathbf{S}^k$ dans $\mathbf{S}^*$ associe la famille de Sperner $F_p(\mathcal{D})$ à tout profil $\mathcal{D}$ de familles de Sperner sur E. On vérifie facilement qu'elle a les propriétés suivantes :

- si tous les $\mathcal{D}_i$ sont des recouvrements de E, alors $F_p(\mathcal{D})$ est un recouvrement de E (car dans ce cas, tout élément de E est dans au moins une classe de $F_p(\mathcal{D})$, qui n'est pas vide),
- si tous les $\mathcal{D}_i$ sont des familles d'intervalles d'un ordre total fixé L sur E, alors $F_p(\mathcal{D})$ est une famille d'intervalles de L,
- si tous les $\mathcal{D}_i$ sont des partitions de E, alors $F_k(\mathcal{D})$ est une partition de E.

Pour $p < k$, le consensus $F_p(\mathcal{D})$ d'un profil $\mathcal{D}$ de partitions de E n'est pas nécessairement une partition, aussi proche de k que soit p. Par exemple, prenons $E = \{a, b, c, d\}$ et $k \ge 3$, avec un profil $\mathcal{D}$ de partitions dont k-2 sont égales à $\{\{a, b, c\}, \{d\}\}$, et les deux dernières à $\{\{a, b\}, \{c\}, \{d\}\}$ et à $\{\{a\}, \{b, c\}, \{d\}\}$. On obtient $F_{k-1}(\mathcal{D}) = \{\{a, b\}, \{b, c\}, \{d\}\}$, qui n'est pas une partition.

3 Regroupements fréquents et motifs fréquents

Considérons le cas particulier où, dans le profil $\mathcal{D}$, chaque famille $\mathcal{D}_i$ est réduite à une classe unique C_i, pour tout $i = 1,..., k$. Une situation équivalente est celle où l'on considère une base de données $\mathcal{D}$ dont les transactions sont les classes C_i, $i = 1,..., k$. Dans un tel cas, les regroupements fréquents se ramènent aux *motifs fréquents* de $\mathcal{D}$.

La détermination des motifs fréquents est un thème important en fouille des données (précisément en recherche de règles d'association). Beaucoup d'algorithmes ont été proposés pour les obtenir, même dans des ensembles de données de grande taille. Selon la remarque précédente, les regroupements fréquents, tels qu'ils ont été définis plus haut, constituent une généralisation des motifs fréquents. Du point de vue algorithmique, ceci a des conséquences intéressantes. Donnons simplement ici l'exemple de l'adaptation de l'algorithme "prototypal" *Apriori* (Agrawal et Srikant 1994).

Cet algorithme comprend une exploration arborescente de l'ensemble $\mathcal{P}(E)$. L'élagage d'un grand nombre de branches permet en général de surmonter la nature exponentielle du problème (et permet donc de traiter des données de grande taille). Il correspond à la sélection de motifs potentiellement fréquents. Pour tout motif B de ce type, la base de données $\mathcal{D}$ est parcourue pour déterminer si B est contenu dans au moins p des C_i. L'adaptation de cette procédure aux regroupements fréquents est immédiate : on parcourt successivement les familles $\mathcal{D}_i$, avec l'instruction supplémentaire de sauter à la famille $\mathcal{D}_{i+1}$ dès que l'on a trouvé une classe de $\mathcal{D}_i$ contenant B.

D'autres algorithmes de la littérature (dont de nombreuses variantes de *Apriori*) peuvent être examinés au cas par cas.

4 Une caractérisation arrowienne

Soit un profil $\mathcal{D} \in \mathbf{S}^k$. Pour conclure si une partie A de E est ou non un regroupement de $F_p(\mathcal{D})$, on n'a besoin de connaître que la valeur de l'indice $g_{\mathcal{D}}(A)$, tandis que les éléments ou parties de $E\backslash A$ (extérieurs à A) n'ont aucun rôle dans cette conclusion. Pour formaliser cette remarque, considérons les trois conditions suivantes pour une fonction de consensus F de $\mathbf{S}^k$ dans $\mathbf{S}^*$ (cf., e.g., Monjardet 1990, Day et McMorris 2003). Selon la propriété (S) de *symétrie*, le résultat de la règle de consensus F ne dépend pas de l'ordre des éléments d'un profil ; étant donnés une permutation σ de K et un profil $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k)$, on pose $\mathcal{D}^\sigma = (\mathcal{D}_{\sigma(1)}, \mathcal{D}_{\sigma(2)}, \ldots, \mathcal{D}_{\sigma(k)})$. Les propriétés (UG) d'*unanimité pour les regroupements* et (NMG) de *neutre-monotonie pour les regroupements* sont de type "arrowien", avec les regroupements comme constituants élémentaires d'une famille de Sperner :

(S) Pour toute permutation σ de K, $F(\mathcal{D}) = F(\mathcal{D}^\sigma)$.

(UG) $[A \subseteq E$ et $g_{\mathcal{D}}(A) = k] \Rightarrow [A$ est un regroupement de $F(\mathcal{D})]$,

(NMG) $[\mathcal{D}, \mathcal{D}' \in \mathbf{S}^k, A, A' \subseteq E$ et $\{i \in K : A \in \mathcal{D}_i\} \subseteq \{i \in K : A' \in \mathcal{D}'_i\}] \Rightarrow$
$[A$ est un regroupement de $F(\mathcal{D}) \Rightarrow A'$ est un regroupement de $F(\mathcal{D}')]$.

Théorème 1. *Une fonction de consensus $F : \mathbf{S}^k \rightarrow \mathbf{S}^*$ est une fonction de consensus par regroupements p-fréquents pour un entier $p \in K$ si et seulement si elle vérifie les trois propriétés (S), (UG), et (NMG).*

Preuve. Il est immédiat que toute fonction de consensus par regroupements p-fréquents F_p vérifie les propriétés (S), (UG) et (NMG).

Pour la réciproque, considérons une fonction de consensus F vérifiant les conditions (S), (UG) et (NMG). Une partie J de K est dite *décisive* pour un profil $\mathcal{D} \in \mathbf{S}^k$ et pour une partie A de E si $J = \{i \in K : A$ est un regroupement de $\mathcal{D}_i\}$ et si A est un regroupement de $F(\mathcal{D})$. Selon (NMG), on a alors, pour tout profil $\mathcal{D}' \in \mathbf{S}^k$ et pour tout $A' \subseteq E$, $J \subseteq \{i \in K : A'$ est un regroupement de $\mathcal{D}'_i\}$ entraîne $[A'$ est un regroupement de $F(\mathcal{D}')]$. Donc J est décisive pour toute partie et pour tout profil, et il en est de même de toute partie J' de K contenant J. Nous pouvons simplement dire que J est une partie décisive de K.

Il reste à déterminer ces parties décisives. Il y en a puisque, selon (UG), K en est une. Soit J une partie décisive de cardinal minimum $p \leq k$. Si $p = k$, on trouve immédiatement $F = F_k$. Sinon, soit J' une autre partie de K de cardinal p. Considérons une permutation σ de K qui envoie exactement les éléments de J sur J'. Soit $A \subset X$ et $\mathcal{D}$ un profil pour lequel on a $J = \{I \in K : A$ est un regroupement de $\mathcal{D}_i\}$. Puisque J est une partie décisive, A est un regroupement de $F(\mathcal{D})$. Selon la condition (S), A est aussi un regroupement de $F(\mathcal{D}^\sigma)$, et J' est encore une partie décisive de K pour F. Alors, toute partie de K de cardinal p (et donc, de cardinal au moins p) est décisive tandis que, selon l'hypothèse de minimalité de p, toute partie de K à moins de p éléments ne l'est pas. Autrement dit, $F = F_p$. ❑

Remarque. Le résultat précédent s'obtient aussi comme conséquence d'un théorème sur le consensus dans les treillis distributifs (Monjardet 1990 ; ici, il s'agit du treillis des parties commençantes – ou idéaux – du treillis $(\mathcal{P}(E), \subseteq)$). Bien qu'il soit intéressant d'obtenir un résultat comme cas particulier d'un théorème plus général, nous nous en sommes tenus à la

preuve directe ci-dessus, l'équivalence avec un problème d'agrégation de parties commençantes étant sans difficulté mais un peu hors sujet.

5 Caractérisations à partir des emboîtements et des implications

Deux relations binaires sur $\mathcal{P}(E)$ sont associées à toute famille $\mathcal{D}$ de parties de E (on ne suppose pas ici que $\mathcal{D}$ est de Sperner).

- La relation *I d'implication* est définie par :

pour tous $A, B \subseteq E$, $[(A, B) \in I] \Longleftrightarrow$ [pour tout $C \in \mathcal{D}$, $A \subseteq C \Rightarrow B \subseteq C$].

Donc, $(A, B) \in I$ (ce que l'on note aussi $A \rightarrow B$) signifie que toute classe contenant A contient aussi B. On dit aussi que le couple $A \rightarrow B$ est une *règle d'association* (exacte), ou une *dépendance fonctionnelle*. On pourra consulter Caspard et Monjardet (2003) pour une revue et des résultats sur ces relations d'implication, par exemple leur caractérisation par les trois propriétés suivantes :

(I1) pour tous $A, B \subseteq E$, $B \subseteq A \Rightarrow A \rightarrow B$,

(I2) pour tous $A, B, C \subseteq E$, $A \rightarrow B$ et $B \rightarrow C \Rightarrow A \rightarrow C$,

(I3) pour tous $A, B, C, D \subseteq E$, $A \rightarrow B$ et $C \rightarrow D \Rightarrow A \cup C \rightarrow B \cup D$.

- L'ordre $\mathcal{E}$ *d'emboîtement* est défini par :

pour tous $A, B \subseteq E$, $[(A, B) \in \mathcal{E}] \Longleftrightarrow A \subset B$ et $(A, B) \notin I \Longleftrightarrow A \subset B$ et il existe $C \in \mathcal{D}$ avec $A \subseteq C$ et $B \not\subseteq C$.

Donc, $(A, B) \in \mathcal{E}$ (ce que l'on note aussi $A \mathcal{E} B$) signifie que la partie B est en un certain sens plus générale que A par rapport à la classification $\mathcal{D}$; voir Domenach et Leclerc (2004a) à propos de ces emboîtements (introduits d'abord par Adams (1986) dans le cas particulier des hiérarchies) et de leur caractérisation par les trois propriétés suivantes (les deux premières entraînant bien qu'il s'agit d'ordres stricts sur $\mathcal{P}(E)$).

(E1) pour tous $A, B \subseteq E$, $A \mathcal{E} B \Rightarrow A \subset B$;

(E2) pour tous $A, B, C \subseteq E$, $A \subset B \subset C \Rightarrow [A \mathcal{E} C \Longleftrightarrow A \mathcal{E} B$ ou $B \mathcal{E} C]$;

(E3) pour tous $A, B \subseteq E$, $A \mathcal{E} A \cup B \Rightarrow A \cap B \mathcal{E} B$.

Un constat important est que, par définition, une famille de Sperner $\mathcal{D}$ et la famille de Moore correspondante $\mu(\mathcal{D})$ ont les mêmes relation d'implication et ordre d'emboîtement.

Etant donné un profil $\boldsymbol{\mathcal{D}} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k)$ de familles de parties de E, on note $\mathcal{E}_i$ et $\rightarrow_i$, respectivement, les relations d'emboîtement et d'implication associées à la famille $\mathcal{D}_i$. Pour $p \in K$, $\mathcal{E}^{(p)} = \bigcup_{J \subseteq K,\, |J| \geq p} \bigcap_{i \in J} \mathcal{E}_i$ est l'ensemble des couples $(A, B) \in (\mathcal{P}(E))^2$ qui appartiennent à au moins p parmi les $\mathcal{E}_i$'. Les conditions (E1) à (E3) permettent de constater que l'ensemble des relations d'emboîtement est stable pour l'union, mais non forcément pour l'intersection, ce qui fait que, généralement, $\mathcal{E}^{(p)}$ n'est pas une relation d'emboîtement. Donc, si l'on souhaite baser l'agrégation d'un profil $\boldsymbol{\mathcal{D}}$ en une unique famille de Sperner propre $\mathcal{D}$ sur les relations d'emboîtement, on ne peut pas en général atteindre l'égalité $\mathcal{E} = \mathcal{E}^{(p)}$. A la place, on considère les deux conditions suivantes, inspirées de celles d'Adams :

(p-EP) $\mathcal{E}^{(p)} \subseteq \mathcal{E}$,

(p-EQ) pour tout $C \in \mathcal{D}$, $(C, E) \in \mathcal{E}^{(p)}$.

L'inclusion (p-EP) correspond à des *emboîtements préservés* (ceux apparaissant dans au moins p éléments du profil). La condition (p-EQ) des *emboîtements qualifiés* est une sorte de réciproque affaiblie de (p-EP), dans laquelle on demande seulement que les paires particulières (C, E), avec $C \in \mathcal{D}$, (ces paires sont évidemment dans $\mathcal{E}$) soient déjà des emboîtements pour au moins p éléments du profil $\boldsymbol{\mathcal{D}}$.

Théorème 2. *Soit $\boldsymbol{\mathcal{D}} \in \mathbf{S}^k$ un profil de familles de Sperner propres sur E. Alors, pour tout $p \in K$, la famille $\mathcal{D} = F_p(\boldsymbol{\mathcal{D}})$ est la seule famille de Sperner sur E vérifiant les conditions (p-EP) et (p–EQ).*

Preuve. Nous montrons essentiellement ici que $F_p(\boldsymbol{\mathcal{D}})$ vérifie les conditions (p-EP) et (p-EQ). Soient A, $B \subseteq E$, avec $(A, B) \in \mathcal{E}^{(p)}$. On a donc $A \subset B$, et il existe une partie J de K à au moins p éléments telle que, pour tout $i \in J$, il y a une classe C_i de $\mathcal{D}_i$ pour laquelle on a $A \subseteq C_i$ et $B \not\subseteq C_i$. En choisissant J de sorte que $C = \bigcap_{i \in J} C_i$ soit maximal pour l'inclusion avec ces propriétés, on a $C \in F_p(\boldsymbol{\mathcal{D}})$. On a alors $A \subseteq C$ et $B \not\subseteq C$, et donc $(A, B) \in \mathcal{E}$, ce qui correspond à la condition (p-EP).

Pour toute classe C de $\mathcal{D}$, il existe par définition une partie J de K, à au moins p éléments, telle que, pour tout $i \in J$, il existe une classe C_i de $\mathcal{D}_i$ contenant C. Ceci entraîne $(C, E) \in \mathcal{E}_i$ pour tout $i \in J$, d'où $(C, E) \in \mathcal{E}^{(p)}$. Par conséquent, la fonction F_p vérifie la condition (p–EQ).

L'unicité s'obtient comme conséquence de résultats précédents portant sur les familles de Moore (Domenach et Leclerc 2004b, Leclerc 2004), que nous ne détaillons pas ici. Ces résultats permettent de déduire l'unicité de la famille $\mathcal{D}$ de celle de la famille $\mu(\mathcal{D})$, qui a la même relation d'emboîtement que $\mathcal{D}$, en utilisant le fait que les éléments de $\mathcal{D}$ sont les "inf-irréductibles" du treillis $(\mu(\mathcal{D}), \subseteq)$. ❑

Comme les implications sont plus connues et utilisées que les emboîtements, il est intéressant de formuler une version du théorème 2 en termes d'implications. On commence par déduire deux conditions ((k-p)-IF) et ((k-p)-ID) des conditions (p-EP) et (p-EQ) précédentes. La condition ((k-p)-IF) indique que tout couple d'implication $A \rightarrow B$ de $\mathcal{D}$ est une *implication fréquente*, en ce sens que c'est un couple d'implication $A \rightarrow_i B$ pour un nombre minimum (precisément, pour au moins $k - p$) d'éléments $\mathcal{D}_i$ du profil $\boldsymbol{\mathcal{D}}$. La condition ((k-p)-ID) des *implications disqualifiées* signifie que, pour toute classe C de $\mathcal{D}$, le couple (C, E) (qui n'est pas un couple d'implication de $\mathcal{D}$) ne peut être un couple d'implication de nombreux éléments du profil.

((k-p)-IF) pour tous $A, B \subseteq X$, $A \rightarrow B$ entraîne $|\{i \in K : A \rightarrow_i B\}| \geq k - p$,
((k-p)-ID) pour tout $C \in \mathcal{D}$, $|\{i \in K : C \rightarrow_i X\}| < k - p$.

Proposition 3. *On a les équivalences* (p-EP) $\Longleftrightarrow$ ((k-p)-IF) *et* (p-EQ) $\Longleftrightarrow$ ((k-p)-ID).

Preuve. (p-EP) entraîne ((k-p)-IF) : soient A, $B \subseteq E$ telles que l'on a $A \rightarrow B$. Avec l'inclusion $B \subseteq A$, on a $|\{i \in K : A \rightarrow_i B\}| = k$, ce qui s'accorde avec ((k-p)-IF). Autrement, on déduit des propriétés (I1) et (I3) des relations d'implication que $A \rightarrow B$ entraîne $A \rightarrow A \cup B$, avec $A \subset A \cup B$. Alors, $(A, A \cup B) \notin \mathcal{E}$ entraîne, à partir de (p-EP), l'inégalité $|\{i \in K :$

$A \mathrel{Œ}_i A \cup B\}| < p$ d'où $|\{i \in K : A \rightarrow_i A \cup B\}| \geq k - p$, ce qui entraîne, en utilisant (I1) et (I2), $|\{i \in K : A \rightarrow_i B\}| \geq k - p$. Donc, la condition (($k$-$p$)-IF) est vérifiée dans tous les cas.

((k-p)-IF) entraîne (p-EP) : soient $A, B \subseteq E$ tels que $(A, B) \in Œ^{(p)}$, c'est-à-dire $A \subset B$ et $|\{i \in K : A \mathrel{Œ}_i B\}| \geq p$. On a alors $|\{i \in K : A \rightarrow_i B\}| < k - p$, ce qui, avec (($k$-$p$)-IF), entraîne que l'on a pas l'implication $A \rightarrow B$, d'où $A \mathrel{Œ} B$.

(p-EQ) $\Longleftrightarrow$ ((k-p)-ID) : pour tout $C \in \mathcal{D}$, on a $C \subset E$, avec les équivalences $(C, E) \in Œ^{(p)} \Longleftrightarrow |\{i \in K : C \mathrel{Œ}_i E\}| \geq p \Longleftrightarrow |\{i \in K : C \rightarrow_i E\}| < k - p$. ❑

Corollaire 4. Soit $\mathcal{D} \in \mathbf{S}^k$ un profil de familles de Sperner propres sur E. Pour tout $p \in K$, la famille $\mathcal{D} = F_p(\mathcal{D})$ est l'unique famille de Sperner sur E vérifiant les conditions ((k-p)-IF) et ((k-p)-ID).

Remarque. Dans le cas particulier considéré au paragraphe 3, où chaque famille $\mathcal{D}_i$ est réduite à une classe unique C_i, les théorèmes 1 et 2 et le corollaire 4 fournissent des caractérisations de la famille de parties de E constituée par les motifs fréquents maximaux. On notera que, avec le corollaire, on montre que cette famille de parties a une relation d'implication qui constitue une sélection de règles d'association non plus exactes mais satisfaisant le critère "statistique" d'être attestées dans un nombre minimum de transactions.

6 Conclusion

On a défini une classe de règles de consensus par regroupements fréquents qui s'applique à tout profil $\mathcal{D} \in \mathbf{S}^k$ de familles de Sperner propres. On a donné trois caractérisations de ces règles. Il reste à généraliser ces résultats en les étendant à des types de classifications plus généraux. Comme les hiérarchies sur lesquelles portent les résultats de Adams ne sont pas des familles de Sperner, on peut s'attendre à l'existence de telles généralisations.

On a observé au paragraphe 2 que ces règles, et donc l'obtention de l'un ou l'autre des systèmes de conditions qui les caractérisent ont pour conséquence l'abandon de la stabilité sur les partitions. En fait, dans de nombreux domaines d'application, ce n'est pas du tout un inconvénient d'obtenir des classes empiétantes, souvent jugées plus réalistes. Ceci admis, les regroupements fréquents peuvent constituer un outil intéressant pour la classification de données décrites par variables qualitatives, moins contraignant que la recherche de partitions consensus initiée par Régnier (1965) et Mirkin (1975) (cf. aussi Barthélemy et Leclerc 1995).

Références

Adams III, E. N. (1986). *N*-trees as nestings: complexity, similarity and consensus. *Journal of Classification* 3, 299–317.

Agrawal, R., Srikant, R. (1994). Fast algorithms association rules. *Proceedings of the 20th VLDB Conference*, Santiago, Chile, 1-7.

Barthélemy, J. P., Leclerc, B. (1995). The median procedure for partitions. *In* I.J. Cox, P. Hansen and B. Julesz, eds., *Partitioning data sets*, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science 19*, Providence, RI, Amer. Math. Soc., 3-34.

Caspard, N., Monjardet, B. (2003). The lattices of Moore families and closure operators on a finite set: a survey. *Discrete Applied Mathematics* 127, 241–269.

Day, W. H. E., McMorris, F. R. (2003). *Axiomatic Consensus Theory in Group Choice and Biomathematics*. Philadelphia , SIAM.

Diday, E. (1971). Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques. *Revue de Statistique Appliquée* XIX, 19-33.

Domenach, F., Leclerc, B. (2004a). Closure Systems, Implicational Systems, Overhanging Relations and the case of Hierarchical Classification. *Mathematical Social Sciences* 47, 349-366.

Domenach, F., Leclerc, B. (2004b). Consensus of classification systems, with Adams' results revisited. *In* D. Banks, L. House, F.R. McMorris, P. Arabie and W. Gaul, editors, *Classification, Clustering and Data Mining Applications*, Berlin, Springer, 417-428.

Han, J., Kamber, M. (2001). *Data mining: concepts and techniques*. San Francisco, Morgan Kaufmann Publishers.

Hipp, J., Güntzer, U., Nakhaeizadeh, G. (2000). Algorithms for association rule mining – a general survey and comparison. *SIGKDD Explorations* 2, 58-64.

Leclerc, B. (2004), On the consensus of closure systems. *Annales du LAMSADE* 3, 237-247.

Leclerc, B. (2005). Implications, emboîtements et ajustements de classifications. *In* V. Makarenkov, G. Cucumel, F.-J. Lapointe (directeurs), *Comptes rendus des 12èmes rencontres de la Société Francophone de Classification*, Montréal, UQAM, 17-20.

Mirkin, B. G. (1975). On the problem of reconciling partitions. *In Quantitative Sociology, Intern.Persp.on Math.and Statist.Modelling*,New York,Academic Press,441-449.

Monjardet, B. (1990). Arrowian characterizations of latticial federation consensus functions. *Mathematical Social Sciences* 20 (1), 51-71.

Régnier, S. (1965). Sur quelques aspects mathématiques des problèmes de classification automatique. *ICC Bull.* 4, 175-191, repr. (1983) *Math. Sci. hum.* 82, 13-29.

Summary

Classifications considered here are Sperner families, for which classes are not pairwise included into each other. Partitions, or single classes are examples of such classifications. Let $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k)$ be a profile of classifications of a given set E. One aims to aggregate $\mathcal{D}$ into a unique consensus classification $\mathcal{D}$. To any integer p comprised between 1 and k (both included), one makes correspond a *frequent grouping consensus function* which returns the maximal subsets of E included in elements of at least p of the $\mathcal{D}_i$'s. Some properties and three characterizations of such consensus rules are given.