# Classification Non Supervisée pour Données Catégorielles

Pierre-Emmanuel JOUVE *, Nicolas NICOLOYANNIS *

*LABORATOIRE ERIC, Université Lumière - Lyon2
Bâtiment L, 5 av. Pierre Mendès-France
69 676 BRON cedex FRANCE
{pierre.jouve, nicolas.nicoloyannis}@eric.univ-lyon2.fr,
http://eric.univ-lyon2.fr

**Résumé.** La classification non supervisée (CNS) constitue l'une des problématiques centrales de l'Extraction de Connaissances à partir de Données (E.C.D.). Le cadre spécifique de la CNS pour données catégorielles a été l'objet de multiples travaux ces dernières années, les principaux challenges associés à ses recherches sont d'une part la définition de critères bien adaptés à ce cadre particulier et d'autre part la mise au point d'algorithme au coût calculatoire relativement faible. Le propos de cet article n'est pas de poursuivre dans ces directions de recherche mais plutôt de s'appuyer sur ces travaux afin de proposer une méthode efficace exhibant de nombreux avantages pour l'utilisateur et utilisable par un non spécialiste. Nous proposons et évaluons donc une méthode de CNS pour données catégorielles permettant la mise à jour relativement rapide d'une classification pertinente d'un ensemble d'objets, tout en facilitant la tâche de l'utilisateur : aucun paramètre obscur ni nombre final de classes à fixer, description compréhensible de la classification, possibilité d'intervention de l'utilisateur dans le processus de CNS...

## 1 Introduction

Nous présentons dans cet article une nouvelle méthode de classification non supervisée (CNS) pour données catégorielles dont le principal attrait réside dans la résolution de plusieurs problèmes auquel un utilisateur se trouve confronté en pratique (détermination du nombre de classes, compréhensibilité de la taxonomie construite...) tout en n'impliquant pas une perte de qualité dans les résultats fournis. L'ensemble des terminologies et formalismes que nous utiliserons afin d'introduire le processus de CNS pour 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 [8] (ce choix est motivé par le souci de conserver une certaine uniformité avec certains travaux notoires du domaine).Afin de faciliter la présentation nous nous appuierons sur des exemples (notés Exemple :) basés sur un jeu de données décrivant un ensemble de 3 votes de motions différentes par 9 nations lors de sessions à l'O.N.U.(voir Tableau 1 page suivante).

| Pays | $M_1$ | $M_2$ | $M_3$ | Pays | $M_1$ | $M_2$ | $M_3$ | Pays | $M_1$ | $M_2$ | $M_3$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| USSR | A | A | C | PORT | D | C | B | FRAN | C | B | C |
| POLA | A | A | C | DENM | C | B | C | SWED | C | B | C |
| CUBA | A | D | C | FINL | B | B | C | NORW | C | B | C |

TAB. 1 – *Votes à l'O.N.U. pour 9 pays différents pour 3 Motions différentes*

# 2 Données Catégorielles

On définit les données catégorielles comme les données décrivant des objets ne possédant que des caractéristiques catégorielles. Les objets, par conséquent nommés objets catégoriels, ne peuvent posséder des caractéristiques numériques (quantitatives) [1].

## 2.1 Domaines Et Attributs Catégoriels

**Notation 1** *$V_1$,$V_2$,...,$V_p$ p variables décrivant un espace V*
EXEMPLE : $V = \{M_1,M_2,M_3\}$.

**Notation 2** *$Dom(V_1)$,...,$Dom(V_p)$ sont les domaines respectifs des variables de V.*

**Définition 1** *Un domaine $Dom(V_j) = \{v_{j1},...,v_{jk}\},k \in N^*$ 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 alors appelée variable catégorielle.*
EXEMPLE : $Dom(M_1) = \{A,B,C,D\}$,$Dom(M_2) = \{A,B,C,D\}$,$Dom(M_3) = \{B,C\}$.

**Définition 2** *V est un espace catégoriel si $\forall V_j,j \in 1..p,V_j$ est catégorielle.*

Notons, que les domaines catégoriels sont définis par des singletons ainsi des valeurs provenant de combinaisons ne sont pas autorisées contrairement à [4]. Afin de simplifier la présentation de notre méthode nous ne considérerons ni les relations d'inclusions conceptuelles entre variables contrairement à [9], ni les valeurs manquantes (dans ce cas nous définirons une valeur supplémentaire pour les domaines des attributs présentant des valeurs manquantes, valeur qui sera considérée comme une valeur classique). Il sera cependant possible d'adapter avantageusement notre méthode (et en particulier la mesure de l'aspect naturel d'une partition utilisée) au traitement de ces cas particuliers.

## 2.2 Objets Catégoriels et Ensemble d'Objets Catégoriels

**Notation 3** *O est l'ensemble des objets d'un jeu de données.*

Comme dans [4] un objet catégoriel $o_i \in O$ est représenté par une conjonction logique de paires attributs-valeurs $[V_1 = o_{i_1}] \bigcap [V_2 = o_{i_2}] \bigcap ... [V_p = o_{i_p}]$. (Une paire attribut-valeur est dénommée sélecteur dans [13].) Nous représenterons chaque objet $o_i \in O$ par un vecteur $[o_{i_1},o_{i_2},...,o_{i_p}]$ (chaque objet possède exactement $p$ valeurs d'attributs). (EXEMPLE : $FRAN = [C,B,C]$ est un objet catégoriel.)
On a alors $o_i = o_j$ si $\forall k, o_{i_k} = o_{j_k}$. Cette dernière relation n'implique toutefois pas que

1. si tel est le cas on devra s'astreindre à une phase de discrétisation de ces caractéristiques afin d'uniformiser la description de ces objets

$o_i$ et $o_j$ représentent le même objet, mais elle signifie qu'ils possèdent les même valeurs catégorielles pour les attributs $V_1$,$V_2$,...,$V_p$.
EXEMPLE : "FRAN"$\neq$ "DENM" mais $[C,B,C] = [C,B,C]$.

Nous introduisons maintenant la notion de mode d'un ensemble d'objets (notion notamment définie dans [8]) qui représente l'objet virtuel le moins dissimilaire (ou le plus similaire) "en moyenne" de la totalité des objets de cet l'ensemble. Le mode d'un ensemble d'objets constitue donc en quelque sorte le profil type de ses objets.

**Notation 4** *$E = \{o_1,o_2,...,o_h\}$ un ensemble de $h$ objets catégoriels et $E \subseteq O$*
*$n_{E_{k,j}}$ le nombre d'objets de $E$ ayant la valeur $k$ pour la variable $V_j \in V$*
*$f_r(V_j = k|E) = n_{E_{k,j}}/card(E)$ la fréquence relative de la valeur $k$ pour $V_j$ pour $E$.*

**Définition 3** *Le mode d'un ensemble d'objet $E$ est l'objet virtuel $mode^E = \{mode_j^E, j = 1..p\}$ tel que pour toute variable $V_j \in V$ la valeur d'attribut de $mode^E$ est, celle, la plus représentée pour cette variable au sein de la classe $E$:*
$\forall j = 1..p, \forall o_i \in E, f_r(V_j = mode_j^E|E) \geq f_r(V_j = o_{i_j}|E)$.
*(Le mode d'un ensemble d'objet $E$ n'est ni forcément un objet de $E$ ni forcément unique.)*
EXEMPLE : $E = \{FRAN,SWED,DENM,NORW,FINL\}$, $mode^E = [C,B,C]$.

### 2.3 Voisinage d'une Partition d'un Ensemble d'Objets Catégoriels

**Notation 5** *$P_h = \{E_1,...,E_h\}$ une partition de $O$ en $h$ groupes*

**Définition 4** *Nous dirons qu'une partition $P_z$ appartient à $Vois(P_h)$ l'ensemble des partitions voisines d'une partition $P_h$ si*

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

EXEMPLE : Soient $O = \{CUBA,POLA,USSR,FRAN,SWED\}$, et la partition
$P_1 = \{\{CUBA,POLA,USSR\},\{FRAN,SWED\}\}$ on a alors $V(P_1) = \{P_1,P_2,P_3\}$ avec
$P_2 = \{\{CUBA,POLA,USSR,FRAN,SWED\}\}$ obtenue par fusion des 2 classes de $P_1$,
$P_3 = \{\{CUBA\},\{POLA,USSR\},\{FRAN,SWED\}\}$ obtenue par segmentation de la classe $\{CUBA,POLA,USSR\}$ selon la variable $M_2$. (aucune autre segmentation n'est possible)

## 3 Aspect Naturel d'une Partition d'Objets : le Nouveau Critère de Condorcet

Le problème de la CNS est, étant donné un ensemble d'objets $O$, de déterminer une partition $P_{nat}$ de $O$ que l'on dénommera naturelle. Cette partition doit être telle que ses classes soient constituées d'objets présentant une relative forte similarité et que les objets de classes différentes présentent une relative forte dissimilarité. Pour ce faire on doit disposer d'un critère permettant de capturer l'aspect naturel d'une partition,

nous présentons et utilisons ici un critère relativement peu exploité dans la littérature : le nouveau critère de Condorcet ($NCC$) défini dans [14], article auquel on se référera pour avoir une comparaison de ce critère au critère classique de l'inertie intra-classe. Le principal avantage de ce critère est une détermination automatique du nombre final de classes pour la CNS. Voici sa définition formelle, en adoptant les notations de la section précédente : $NCC(P_h)$ la mesure de l'aspect naturel d'une partition $P_h$

$$NCC(P_h) = \sum_{i=1..h, j=1..h, i\neq j} Sim(E_i,E_j) + \alpha \times \sum_{i=1}^{h} Dissim(E_i) \quad (1)$$

$\alpha$ scalaire appelé facteur de granularité, fixé par défaut à 1 mais modifiable ($\alpha \geq 0$)

$$Sim(E_i,E_j) = \sum_{o_a\in E_i, o_b\in E_j} sim(o_a,o_b), \quad Dissim(E_i) = \textstyle\sum_{o_a\in E_i, o_b\in E_i} dissim(o_a,o_b) \quad (2)$$

$$sim(o_a,o_b) = \sum_{i=1}^{p} \delta_{sim}(o_{a_i},o_{b_i}), \quad dissim(o_a,o_b) = \textstyle\sum_{i=1}^{p} 1 - \delta_{dissim}(o_{a_i},o_{b_i}) \quad (3)$$

$$\delta_{sim}(o_{a_i},o_{b_i}) = \delta_{dissim}(o_{a_i},o_{b_i}) = \begin{cases} 1 \text{ si } o_{a_i} = o_{b_i} \\ 0 \text{ si } o_{a_i} \neq o_{b_i} \end{cases} \quad (4)$$

Ainsi, $NCC(P_h)$ mesure simultanément les dissimilarités entre objets de même classe de la partition $P_h$, et les similarités entre objets de classes différentes (on peut donc dire que $NCC(P_h)$ est définie comme une fonction de l'homogénéité interne des classes et de l'hétérogénéité entre classes). Donc, les partitions présentant une forte homogénéité intra-classe et une forte disparité inter-classes posséderont une faible valeur pour $NCC$ et constitueront les partitions apparaissant comme les plus naturelles.

**Définition 5** *Une partition $P_1$ est dite plus naturelle qu'une partition $P_2$ (ou encore représentant mieux la structure interne des données) si $NCC(P_1) < NCC(P_2)$.*

**Définition 6** *Une partition d'un ensemble d'objets $O$ est nommée Partition Naturelle de $O$ et est notée $P_{nat}$ si elle minimise $NCC$ : $\wp$ ensemble des partitions de $O$, $\forall P_i \in \wp, NCC(P_{nat}) \leq NCC(P_i)$.*
Notons le rôle du facteur de granularité $\alpha$ (non présent dans la définition initiale du $NCC$) : celui ci permet soit de privilégier l'influence de l'homogénéité intra-classe ou de la disparité inter-classe pour la détermination de l'aspect naturel d'une partition. En effet, plus $\alpha$ est élevé (resp. faible) plus une partition doit présenter une forte homogénéité intra-classe (resp. une forte disparité inter-classes) pour apparaître naturelle.

# 4 Une Méthode de Classification Non Supervisée Orientée Utilisateur

## 4.1 Travaux Liés et Spécificités du Travail

La CNS pour données catégorielles a été l'objet de multiples travaux ces dernières années, les principaux challenges associés à ses recherches sont d'une part la définition

de critères bien adaptés à ce cadre particulier (dans ROCK [6] les auteurs utilisent une mesure basée sur le nombre de voisins communs que possède deux objets, [1] utilisent quant à eux une mesure basée sur l'entropie généralisée, enfin [14] utilise quant à lui le $NCC$) et d'autre part la mise au point d'algorithmes au coût calculatoire relativement faible ([8] propose les K-Modes une adaption de la méthode des K-Means dans le cas de données catégorielles, [3] utilisent les systèmes dynamiques pour STIRR).

Notre propos n'est pas ici de poursuivre dans ces directions de recherche mais plutôt de s'appuyer sur ces travaux afin de proposer une méthode efficace exhibant de nombreux avantages pour l'utilisateur et utilisable par un non spécialiste. On peut ainsi lister un ensemble de qualités pratiques désirables par un utilisateur et qui distinguerait largement une méthodes les possédant des approches existantes :

- générer une description des classes définies aisément compréhensible et ne nécessitant aucun post-traitement contrairement à [6], [3] [14]
- coût calculatoire relativement faible contrairement à [6]
- ne pas nécessiter de fixer a priori le nombre final de classe contrairement à [8] mais toutefois permettre à l'utilisateur de jouer sur la finesse de la partition s'il considère que le nombre de classes obtenues est trop élevé ou trop faible.
- permettre la découverte de structures ne présentant pas une régularité dans le nombre d'objets par classe.contrairement à [8]
- permettre la gestion de données manquantes, l'introduction de contraintes et l'intervention de l'utilisateur au cours du processus de CNS et ce de manière aisée.

L'objectif du travail mené est donc la mise au point d'une méthode de CNS permettant la mise à jour de partitions pertinentes et possédant ces qualités pratiques.

## 4.2 L'Algorithme de Classification Non Supervisée

L'algorithme que nous proposons consiste en une mise en oeuvre astucieuse et nouvelle de principes et techniques existants afin d'atteindre les objectifs détaillés précédemment. Pour cela nous utilisons le critère $NCC$ très peu exploité dans la littérature, et une technique de types graphes d'induction pour découvrir une partition $P_{\sim nat}$ proche ou égale à $P_{nat}$ ce qui conférera un aspect non hiérarchique à la méthode ([17] avaient proposé une technique de type arbre d'induction ce qui conférait un aspect hiérarchique à la méthode et utilisaient un critère de type khi2).

Remarques :
-La découverte de $P_{nat}$ (problème combinatoire) peut être résolue par des méthodes au coût calculatoire élevée : par une approche de type Programmation en nombre entier, par une approche basée sur les méthodes de Plans de Coupe et Branch and Bound [5].
-La découverte d'une partition $P_{\sim nat}$ proche ou égale à $P_{nat}$ peut être effectuée par l'intermédiaire d'heuristiques de coût calculatoire en $O(n^2)$ : [15] a proposé une approche itérative basée sur la prétopologie, [16] utilisent une méthode itérative utilisant le recuit-simulé permettant la découverte de $P_{\sim nat}$, [11] utilisent une approche de type programmation linéaire.
-Utiliser ces méthodes ne permet malheureusement pas d'atteindre les objectifs visés.

Nous avons donc adopté une heuristique gloutonne de type graphe d'induction (on procède par segmentation/fusion successives de classes de partitions). En définitive, à partir de la partition grossière de $O$, l'algorithme va évoluer itérativement de partition en partition en suivant le principe d'évolution suivant : on passe d'une partition $P_t$ vers la partition $P_{t+1}$ appartenant à son voisinage (cf. Définition 4) telle qu'elle réduise au plus le $NCC$. L'algorithme s'achève lorsque $P_t = P_{t+1}$. Le pseudo-code de l'algorithme est donc le suivant:

1. soit $P_0$ la partition grossière de $O$
2. i:=0
3. Déterminer $V(P_i)$
4. Déterminer $P_{i+1} \in V(P_i)$ la meilleure partition de $V(P_i)$ selon le $NCC$
5. Si $P_{i+1} = P_i$ aller en 6), sinon i:=i+1 et aller en 3)
6. $P_{\sim nat} = P_i$

Remarques :
Pour éviter des minima locaux, on peut autoriser dans le cas $P_{i+1} = P_i$, un nombre fixé a priori d'évolution vers la partition du voisinage de $P_i$ obtenue par segmentation selon une variable et impliquant la plus faible augmentation du $NCC$.

## 4.3 Qualités de la Méthode pour l'Utilisateur

- Chacune des classes de la partition résultat ($P_{\sim nat}$) est caractérisée par une règle logique (règle formée de disjonctions de conjonctions de sélecteurs) correspondant à la suite de mécanismes (fusion / segmentation) lui ayant donné le jour. Cette règle correspond alors pour un objet de $O$ à une condition nécessaire et suffisante pour appartenir à la classe qu'elle décrit.
- On associera également à chaque classe de $P_{\sim nat}$ son mode, celui-ci correspondant en définitive à une sorte de profil ou individu type de la classe.
- Ces deux premiers points simplifient largement la compréhension et l'interprétation des résultats.
- La description de chaque classe par une règle logique peut également permettre l'assignation d'un nouvel objet à l'une des classes sans pour autant connaître l'ensemble de ses caractéristiques.
- le nombre de classe est déterminé automatiquement par la méthode, toutefois l'utilisateur peut influer sur la finesse de la partition finalement produite par l'intermédiaire du facteur de granularité. (si le nombre de classes apparaît trop faible (resp. trop fort) à l'utilisateur, ce dernier peut procéder à une nouvelle CNS en augmentant (resp. en diminuant) la valeur du facteur de granularité).

## 4.4 Illustration du Fonctionnement de l'Algorithme

Considérons le jeu de données classique : les données mushrooms[12]. Ce jeu de données est composé de 8124 objets (en l'occurrence des champignons), chacun décrit par 23 variables. Chaque objet est, de plus, identifié par sa comestibilité. Le jeu de données est ainsi composé de champignons comestibles et de champignons vénéneux.

La figure 1 présente le processus de CNS sur le jeu de données "Mushrooms" pour un facteur de granularité $\alpha$ valant 1. Elle détaille ainsi l'ensemble des processus de segmentation/fusion, permettant l'obtention de la CNS . Voici pour exemple les règles logiques caractérisant les classes C8 et C10 (rappelons que chacune de ces règles détermine l'appartenance ou la non appartenance de tout objet à la classe à laquelle elle est associée) :

- C8 [Ring_Type = evanescent] ET [Gill_Size = broad] ET [Bruises? = bruises]
- C10 [Ring_Type = pendant] ET [ [Bruises? = bruises] OU [ [Bruises? = no] ET [stalk-color-above-ring = white] ET [Gill_Size = narrow] ] ]

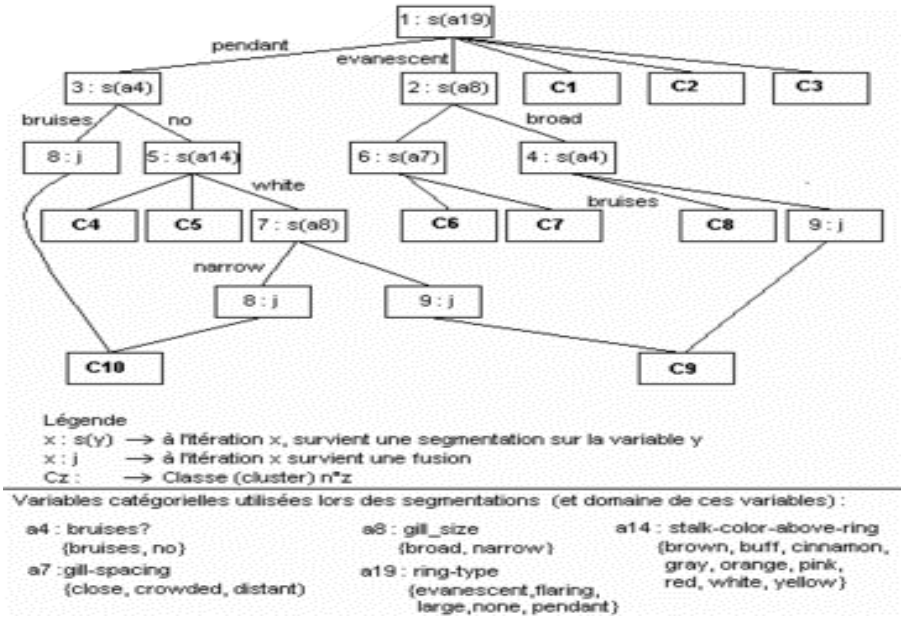


FIG. 1 – *Illustration du Fonctionnement de l'Algorithme*

# 5 Evaluation de l'Algorithme de Classification non Supervisée

Classiquement, une méthode de CNS s'évalue selon : la validité et la stabilité des classifications qu'elle propose, et selon son efficacité algorithmique.

## 5.1 Evaluation de la Qualité des Classifications

L'évaluation de la validité d'une CNS est généralement réalisée par utilisation de mesures de validité de CNS[7] [2]. Ces mesures sont de deux types : externes et internes (les modes d'évaluation habituels correspondant en définitive à l'utilisation implicite d'une mesure externe). Les critères externes de validité évaluent dans quelle mesure le résultat du processus de CNS correspond à des connaissances avérées sur les données.

De manière assez générale, on admet que ces informations ne sont pas calculables à partir des données. La forme la plus commune de données de ce type est un ensemble d'étiquettes que l'on associe à chacun des objets (ce dernier type d'information peut éventuellement être obtenu par une classification manuelle). Les critères internes de validité consistent quant à eux en une mesure basée uniquement sur le traitement des données servant au processus de CNS et leur utilisation n'est que rarement envisageable.

Nous avons utilisé ici une évaluation de type mesure externe largement utilisée dans la littérature. Nous avons considéré le jeu de données mushrooms (composé de champignons comestibles et de champignons vénéneux) et avons réalisé plusieurs CNSs, pour des valeurs de $\alpha$ différentes, enfin nous avons utilisé le concept "comestibilité" et le taux de correction T.C. des CNSs par rapport à ce concept afin de caractériser la qualité de la CNS résultant (la variable définissant la "comestibilité" n'étant évidemment pas introduite dans le processus de CNS). Les résultats obtenus pour 3 valeurs différentes de $\alpha$ ($\alpha = 1,2,3$) présentés dans le tableau 2 montrent que les différentes classifications réalisées permettent bien d'obtenir des partitions reflétant correctement la structure impliquée par le concept comestibilité ainsi que la capacité de l'algorithme à déterminer une structure présentant des irrégularités dans le nombre d'objets par classe. (Notons de plus que pour des valeurs de $\alpha$ supérieures à 3, les classifications obtenues présentaient un nombre de classes strictement supérieur à 24, chacune étant homogène du point de vue de la comestibilité des champignons la constituant.)

Nous présentons également une comparaison des résultats obtenus par notre méthode et ceux obtenus par les k-modes pour différentes CNSs. Nous avons, pour cela, lancé plusieurs processus de CNS avec notre méthode en utilisant des facteurs de granularité différents. Cela nous a permis d'obtenir des CNSs en 6, 10, 21, 24, 58, 141 et 276 classes. Les taux de correction de ces CNSs par rapport au concept "comestibilité" sont ainsi reportés sur la figure 2. Ensuite nous avons lancé des séries de 10 CNSs en utilisant les k-modes paramétrés de manière telle qu'on obtienne des CNSs en 6, 10, 21, 24, 58, 141 et 276 classes. La valeur moyenne du taux de correction par rapport au concept "comestibilité" de chacune de ces 7 séries de 10 CNSs sont ainsi reportées sur la figure 2. L'ensemble de ces tests montrent une qualité légèrement supérieure pour les CNSs obtenues par l'intermédiaire de notre méthode.

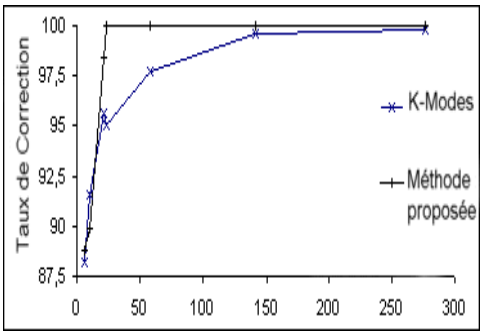


FIG. 2 - *Taux de Correction pour le concept "comestibilité"*

Des tests ont également été menés sur le jeu de données Soybean Disease [12]. Soybean Disease est un jeux de données standard en apprentissage symbolique (machine learning) composé de 47 objets, chacun étant décrit par 35 variables catégorielles. Chaque objet est caractérisé par une des 4 pathologies suivantes : Diaporthe Stem Canker (D1), Charcoal Rot (D2), Rhizoctonia Root Rot(D3), and Phytophthora Rot(D4).A l'exception de D4 qui est représentée par 17 objets, toutes les autres pathologies sont représentées par 10

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha = 1$ | N.C. | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 |
| | #C./#V. | 0/1296 | 48/0 | 0/36 | 192/0 | 16/0 | 192/0 | 1056/0 | 2656/816 |
| | N.C. | #9 | #10 | | | | | | |
| | #C./#V. | 48/1760 | 0/8 | | | | | | |
| $\alpha = 2$ | N.C. | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 |
| | #C./#V. | 0/1728 | 192/0 | 768/0 | 0/1296 | 1728/0 | 512/0 | 288/0 | 192/0 |
| | N.C. | #9 | #10 | #11 | #12 | #13 | #14 | #15 | #16 |
| | #C./#V. | 0/36 | 0/288 | 0/192 | 192/0 | 96/256 | 48/0 | 0/32 | 96/0 |
| | N.C. | #17 | #18 | #19 | #20 | #21 | | | |
| | #C./#V. | 32/72 | 16/0 | 0/8 | 48/0 | 0/8 | | | |
| $\alpha = 3$ | N.C. | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 |
| | #C./#V. | 92/0 | 0/1296 | 0/864 | 0/864 | 288/0 | 96/0 | 1728/0 | 768/0 |
| | N.C. | #9 | #10 | #11 | #12 | #13 | #14 | #15 | #16 |
| | #C./#V. | 192/0 | 0/72 | 48/0 | 512/0 | 0/192 | 0/32 | 0/36 | 0/288 |
| | N.C. | #17 | #18 | #19 | #20 | #21 | #22 | #23 | #24 |
| | #C./#V. | 192/0 | 96/0 | 0/256 | 0/8 | 48/0 | 0/8 | 32/0 | 16/0 |

TAB. 2 – *CNS sur le jeux de données "Mushrooms", $\alpha = 1$, $\alpha = 2$, $\alpha = 3$*
Légende : N.C. : numéro de la classe, #C./#V. : nb champignons comestibles/nb champignons vénéneux

objets chacune. Nous avons menés plusieurs CNSs pour différentes valeurs de $\alpha$ et utilisé le concept "pathologie" pour caractériser la qualité des CNSs obtenues (la variable "pathologie" n'étant évidemment pas introduite dans le procesus de CNS). Les résultats obtenus pour 4 valeurs différentes de $\alpha$ ($\alpha$ = 1,1.5,2,3) présentés dans le tableau 3. montrent que les CNSs obtenues reflètent correctement le concept "pathologie".( Pour $\alpha \geq 3$, les CNSs ont un nombre de classes strictement supérieur à 4, chaque classe étant homogène du point de vue du concept "pathologie". Nos résultats pour une CNS en 4 classes (taux de correction est égal à 100%) sont meilleurs que ceux des k-modes reportés dans [8], (taux de correction à peu près égal à 96% pour les k-modes.)

| | N.C. | #1 | #2 | #3 | #4 |
|---|---|---|---|---|---|
| $\alpha = 1$ | #D1/#D2/#D3/#D4 | 10/10/10/17 | | | |
| $\alpha = 1.5$ | #D1/#D2/#D3/#D4 | 10/0/0/0 | 0/10/10/17 | | |
| $\alpha = 2$ | #D1/#D2/#D3/#D4 | 10/0/0/0 | 0/10/0/0 | 0/0/10/17 | |
| $\alpha = 3$ | #D1/#D2/#D3/#D4 | 10/0/0/0 | 0/10/0/0 | 0/0/10/0 | 0/0/0/17 |

TAB. 3 – *Clusterings on "Soybean Diseases" data*

## 5.2 Evaluation de la Stabilité

Un autre point d'évaluation d'un algorithme CNS, est l'évaluation de sa stabilité, i.e. "aurai je obtenu une organisation des objets similaire ou très proche si l'ensemble d'objets que j'avais utilisé avait été légèrement différent (quelques objets supplémentaires

ou en moins) ?". Afin, de répondre à cette question des méthodes d'échantillonages et de comparaison des partitions obtenues ont été présentées, nous utiliserons ici celle présentée dans [10], son mode de fonctionnement est le suivant :

- on considère le jeu de données dans son intégralité et l'on réalise une première CNS qui constituera la classification de référence $C_{Ref}$ (le nombre d'objets du jeu de données est noté $n$).
- On réalise ensuite un ensemble $EC$ de $p$ CNSs ($C_i$,$i = 1..p$) sur des échantillons de taille $\mu \times n$ de ce jeu de données ($\mu \in ]0,1]$,$\mu$ est appelé facteur de dilution).
- Pour chaque CNS $C_i$,$i = 1..p$ on procède à une comparaison avec $C_{Ref}$ afin de calculer la proportion de paires d'objets (notée $prop_i$) traitées différemment par $C_{Ref}$. On dit qu'une paire d'objets est traitée différemment par $C_{Ref}$ et $C_i$, si les deux objets sont présents dans l'échantillon ayant permis de bâtir $C_i$ et si ces deux objets sont regroupés au sein d'une même classe dans $C_i$ alors qu'ils ne l'étaient pas pour $C_{Ref}$ ou si ces deux objets ne sont pas regroupés au sein d'une même classe dans $C_i$ alors qu'ils l'étaient pour $C_{Ref}$. $prop_i \in [0,1]$.
- On calcule la valeur d'un indicateur de stabilité de la CNS $Stab$ qui correspond à la moyenne des $prop_i$. $Stab \in [0,1]$.

Ainsi, une valeur élevé de $Stab$ (relativement proche de 1) correspondra à une forte différence entre les CNSs de $EC$ et $C_{Ref}$, et donc une valeur faible de $Stab$ (proche de 0) correspondra à une faible différence entre les CNSs de $EC$ et $C_{Ref}$. La valeur de $Stab$ permet alors de savoir si l'algorithme de CNS peut être considéré comme stable et son utilisation valable (la non stabilité impliquant la non utilisabilité de la méthode ou une recherche en profondeur des causes de la non stabilité).

Les tests de stabilité de l'algorithme réalisés sur le jeu de données "Mushrooms" sont présentés dans la figure. Les paramètres suivants ont été adoptés pour cet ensemble de tests : le nombre de CNSs réalisées pour chaque niveau du facteur de dilution est 100, la valeur du facteur de granularité $\alpha$ est 3. Ces tests montrent une excellente stabilité des CNSs obtenus par notre méthode, d'ailleurs, de tels résultats n'auraient pas été obtenus avec la plupart des méthodes existantes.

## 5.3 Evaluation de l'Efficacité Algorithmique

Nous ne présentons pas ici une étude poussée du coût calculatoire de la méthode, nous nous contentons de préciser que la complexité algorithmique de notre méthode est équivalente à celle des graphes d'induction largement utilisé en E.C.D. (cela s'expliquant notamment par un coût calculatoire relativement peu élevé) et présentons le temps de calcul assoocié à différentes CNSs pour le jeux de données "mushrooms". En fait, nous ne présentons pas explicitement les temps de calculs associés aux CNSs mais le rapport suivant : $R = \frac{temps\ de\ calcul\ associe\ a\ la\ CNS}{temps\ de\ calcul\ associe\ a\ la\ CNS\ en\ 6\ classes\ par\ les\ k-modes}$ .
Les rapports présentés pour les k-modes sont les valeurs moyennes de chacune des séries de 10 CNSs. Ces résultats montrent que les K-Modes (qui sont reconnus comme une méthode rapide et possédant une bonne scalabilité) sont plus rapides pour de faibles nombre de classes mais que les 2 méthodes semblent se comporter de manière similaire pour des nombres de classes plus élevés.

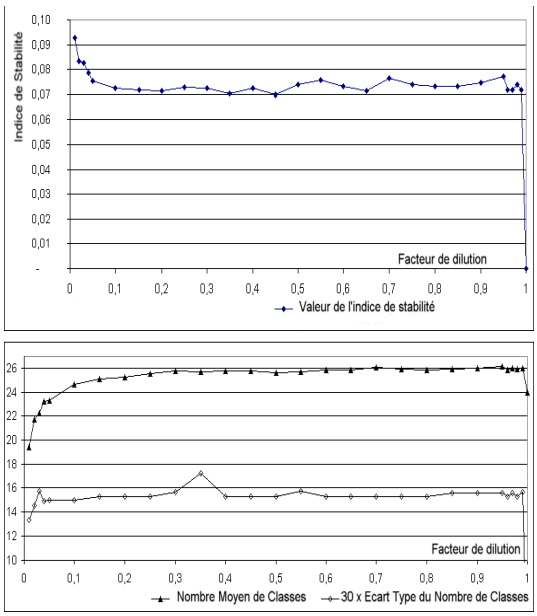


FIG. 3 - *Evaluation de la stabilité*

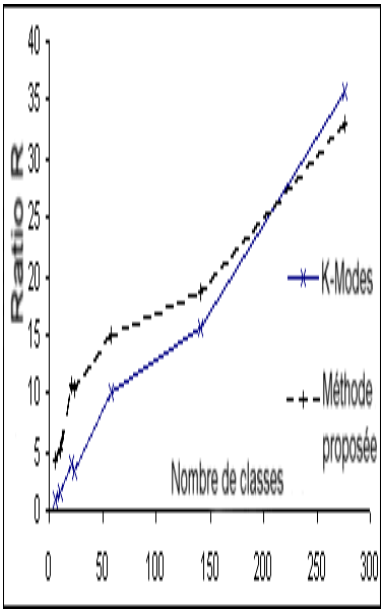


FIG. 4 - *Rapports R associés à différentes CNSs*

# 6 Conclusion

Nous venons de procéder à l'évaluation de notre méthode de CNS qui nous a permis de mettre en avant la qualité des clusters produits, la très bonne stabilité et le coût calculatoire relativement faible de la méthode. Ces différents points constituent plusieurs points forts, tout comme les avantages concernant l'utilisabilité cité à la section 4.3 nous listons maintenant un ensemble d'autres points non abordés qui participent également à rendre cette méthode très attrayante du point de vue de l'utilisateur : (1) la présence de données manquantes n'est pas gênante : leurs conséquences sur la classification est complètement paramétrable en codant l'implication particulière de la présence de données manquantes sur l'aspect naturel d'une partition, (2) l'introduction de contraintes est possible par l'intermédiaire de l'utilisation de variables supplémentaires sur lesquelles on n'autorisera pas de segmentation lors du processus de recherche de la partition naturelle approchée (Ainsi, l'interactivité entre utilisateur et processus de CNS est possible, par introduction de contraintes), (3) le nombre d'objets par classe peut varier très fortement d'une classe à l'autre, on peut ainsi découvrir des structures ne présentant pas une régularité dans le nombre d'objets par classe.

# Références

[1] Cristofor D., Simovici D., An information-theoretical approach to clustering categorical databases using genetic algorithms, 2nd SIAM ICDM, Workshop on clustering high dimensional data, 2002.

[2] Dom B., An Information-Theoretic External Cluster-Validity Measure , IBM, 2001.

[3] Gibson D., Kleinberg J. M., Raghavan P., Clustering Categorical Data : An Approach Based on Dynamical Systems, VLDB Journal : Very Large Data Bases, vol. 8, n°3–4, 2000, p. 222-236.

[4] Gowda K. C., Diday E., Symbolic Clustering using a New Dissimilarity Measure, Pattern Recognition, vol. 24, n°6, 1991, p. 567-578.

[5] Grötschel M., Wakabayashi., A Cutting Plane Algorithm for a Clustering Problem, Mathematical Programming, vol. 45, 1989, p. 59-96.

[6] Guha S., Rastogi R., Shim K., ROCK : A Robust Clustering Algorithm for Categorical Attributes, Information Systems, vol. 25, n°5, 2000, p. 345-366.

[7] Halkidi M., Batistakis Y., Vazirgiannis M., On Clustering Validation Techniques, Journal of Intelligent Information System, vol. 17, n°2-3, 2001, p. 107-145.

[8] Huang Z., A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining, Research Issues on Data Mining and Knowledge Discovery, 1997.

[9] Kodratoff Y., Tecuci G., Learning Based on Conceptual Distance, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 10, n°6, 1988, p. 897- 909.

[10] Levine E., Domany E., Resampling Method for Unsupervised Estimation of Cluster Validity, Neural Computation, vol. 13, n° 11, 2001, p. 2573-2593.

[11] Marcotorchino F., Michaud P., Heuristic Approach to the similarity Aggregation Problem, Methods of Operations Research, vol. 43, 1981, p. 395-404.

[12] Merz C., Murphy P., UCI repository of machine learning databases, http://www.ics.uci.edu/#mlearn/mlrepository.html, 1996.

[13] Michalski R. S., Stepp R. E., Automated Construction of Classifications : Conceptual Clustering Versus Numerical Taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 5, n° 4, 1983, p. 396-410.

[14] Michaud P., Clustering techniques, Future Generation Computer Systems, vol. 13, n° 2-3, 1997, p. 135 147.

[15] Nicoloyannis N., Structures Prétopologiques et Classification Automatique, PhD thesis, Université Lyon 1, 1988.

[16] Nicoloyannis N., Terrenoire M., Tounissoux D., An Optimisation Model for Aggregating Preferences : A Simulated Annealing Approach, Health and System Science, vol. 2, n° 1-2, 1998, p. 33-44.

[17] Williams W., Lambert J., Multivariate Methods in Plant Ecology, Journal of Ecology, vol. 47, 1959, p. 83-101.

## Summary

Clustering constitutes one of the central problem in Knowledge Discovery in Databases (K.D.D.). Categorical data clustering was the object of several work these last years, main challenges associated with these researches were the definition of well adapted criteria for this particular framework and the development of fast algorithms. The subject of this article is not to continue in these directions of research but rather to take into account those works in order to propose and evaluate an effective method that exhibits many advantages for the user and usable by a non-specialist.