# Formulation Condorcéenne du critère de la modularité

Lazhar Labiod, Nistor Grozavu, Younès Bennani

LIPN UMR 7030, Université Paris 13
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
Prénom.Nom@lipn.univ-paris13.fr,

**Résumé.** La mesure de modularité a été utilisée récemment pour la classification de graphes (Newman et Girvan, 2004), (Agarwal et Kempe, 2008). Dans ce papier, nous montrons que la mesure de modularité peut être formellement étendue pour la classification non supervisée des données catégorielles. Nous établissons également des connexions entre le critère de modularité et celui de l'analyse relationnelle qui est basé sur le critère de Condorcet. Nous développons ensuite un algorithme efficace inspiré de l'heuristique de l'analyse relationnelle pour trouver la partition optimale maximisant le critère de modularité. Les résultats expérimentaux montrent l'efficacité de notre approche.

## 1 Introduction

La classification automatique est une méthode d'apprentissage non supervisé permettant le partitionnement d'un ensemble d'observations en classes. les méthodes de classification automatique conduisent à une partition de la population initiale en groupes disjoints, tels que, selon un critère choisi a priori, deux individus d'un même groupe aient entre eux un maximum d'affinité et deux individus de deux groupes différents aient entre eux un minimum d'affinité. La classification automatique a été largement étudiée en apprentissage automatique, en bases de données et en statistique de divers points de vue.

De nombreuses applications de la classification automatique ont été discutées et de nombreuses techniques ont été développées. Une étape importante dans la conception d'une technique de classification consiste à définir un critère pour mesurer la qualité de partitionnement en termes des deux objectifs cités ci-dessus. Pour la classification des données numériques continues, il est naturel de penser à utiliser une mesure basée sur une distance géométrique. Étant donnée une telle mesure, une partition appropriée peut être calculée par l'optimisation de certaines quantités (par exemple, la somme des distances des observations à leurs centroïdes). Toutefois, si les vecteurs de données contiennent des variables catégorielles, le problème de la classification devient plus difficile et d'autres stratégies doivent être développées. C'est souvent le cas dans de nombreuses applications où les données sont décrites par un ensemble d'attributs descriptifs ou binaires, dont beaucoup ne sont pas numériques. Des exemples de tels attributs sont le pays d'origine et la couleur des yeux dans les données démographiques. De nombreux algorithmes ont été développés pour la classification des données catégorielles, par exemple, (Barbara et al, 2002), (Gibson et al, 1998), (Huang, 1998) et ( Ganti et al, 1999).

La mesure de modularité a été utilisée récemment pour la classification de graphes (Agarwal et Kempe, 2008), (Newman et Girvan, 2004) et (White et Smyth, 2005). Dans ce papier,

nous montrons que le critère de modularité peut être formellement etendu pour la classification des données catégorielles. Nous avons également établi des liens entre le critère de modularité et celui de l'analyse relationnelle (AR) (Marcotorchino et Michaud, 1978) (Marcotorchino, 2006), qui est basée sur le critère de Condorcet. Nous développons ensuite une procédure efficace inspirée de l'heuristique de l'AR pour trouver la partition optimale maximisant le critère de modularité. Les résultats expérimentaux montrent l'efficacité de notre approche. La première contribution de ce papier est l'introduction d'une mesure de modularité étendue pour la classification des données catégorielles. La deuxième contribution est la présentation de la mesure de modularité étendue comme un critère de Condorcet modifié. En particulier, nous montrons que notre nouveau critère de modularité introduit une pondération en fonction du profil de chaque observation.

Le reste du papier est organisé comme suit: la section 2 introduit quelques notations et définitions et nous présentons l'approche de l'analyse relationnelle dans la section 3. La section 4 présente deux variantes de la mesure de modularité étendue et sa connexion avec le critère de l'AR. Des discussions sur la procédure d'optimisation proposée sont décrites à la section 5. La section 6 montre nos résultats expérimentaux et enfin, la section 7 présente des conclusions et certains travaux futurs.

# 2 Définitions et notations

Soit $I$ un ensemble de données avec $N$ objets $\{O_1,O_2,...,O_N\}$ décrit par l'ensemble $V$ de $M$ attributs (ou variables catégorielles) $\{V^1,V^2,..,V^m,...,V^M\}$ chacun ayant $p_1,..,p_m,...,p_M$ catégories, respectivement, et soit $P = \sum_{m=1}^{M} p_m$, désigne le nombre total de catégories de toutes les variables. Chaque variable catégorielle peut être décomposée en une collection de variables indicatrices. Pour chaque variable $V^m$, considérons les $p_m$ valeurs qui correspondent naturellement aux nombres de 1 à $p_m$ et $V_1^m,V_2^m,...,V_{p_m}^m$ sont des variables binaires telles que, pour chaque $j$, $1 \leq j \leq p_m$, $V_j^m = 1$ si et seulement si $V^m$ prend la $jieme$ valeur. Ainsi la matrice de données peut être exprimée comme une collection de $M$ matrices $K^m$, $(m = 1,..,M)$ de terme général $k_{ij}^m$ tel que :

$$k_{ij}^m = \begin{cases} 1 & \text{si l'objet } i \text{ possède la catégorie } j \text{ de } V^m \\ 0 & \text{sinon} \end{cases} \quad (1)$$

ce qui donne la matrcie disjonctive $K$ de dimensions $N \times P$; $K = (K^1|K^2|...|K^m|...|K^M)$.

## 2.1 Représentation relationnelle des données

Si les données se composent de $N$ objets $\{O_1,O_2,...,O_N\}$ décrits par $M$ attributs (ou variables) $\{V^1,V^2..,V^k,...,V^M\}$ sur les quelles ont été mesurés, le principe de comparaison par paires consiste à transformer les données, qui sont habituellement représentées par une matrice rectangulaire de dimension $N \times M$ en deux matrices carrées, $S$ et $\bar{S}$. La matrice $S$ est appelée la matrice de Condorcet de terme général $s_{ii'}$, représentant la mesure de similarité globale entre les deux objets $O_i$ et $O_{i'}$, mesurée sur tous les $M$ attributs. La matrice $\bar{S}$ de terme général $\bar{s}_{ii'}$ représente la mesure de dissimilarité globale entre ces deux objets. Pour obtenir la matrice $S$, chaque attribut $V^m$ est transformé en une matrice carré $S^m$ de taille

$N \times N$ et de terme général $s_{ii'}^m$, représentant la mesure de similarité entre deux objets $O_i$ et $O_{i'}$ pour l'attribut $V^m$. Pour obtenir la matrice $\bar{S}$, une mesure de dissimilarité $\bar{s}_{ii'}^m$ entre les objets $O_i$ et $O_{i'}$ pour l'attribut $V^m$ est alors calculée comme le complément à la mesure de similarité maximale possible entre ces deux objets. Comme la similarité entre deux objets différents est inférieure ou égale à leur auto-similarités: c.a.d $s_{ii'}^m \leq min(s_{ii}^m, s_{i'i'}^m)$, alors il vient, $\bar{s}_{ii'}^m = \frac{1}{2}(s_{ii}^m + s_{i'i'}^m) - s_{ii'}^m$. Cela nous amène à une matrice de dissimilarité $\bar{S}^m$. Les matrices $S$ et $\bar{S}$ sont alors obtenues en additionnant, respectivement, toutes les matrices $S^m$ et $\bar{S}^m$, soit $S = \sum_{m=1}^{M} S^m = KK^t$ et $\bar{S} = \sum_{m=1}^{M} \bar{S}^m$. La similarité globale entre chaque deux objets $O_i$ et $O_{i'}$ est donc $s_{ii'} = \sum_{m=1}^{M} s_{ii'}^m$ et leur dissimilarité globale est $\bar{s}_{ii'} = \sum_{m=1}^{M} \bar{s}_{ii'}^m$.

### 2.1.1 Graphe non orienté et matrice de similarité

La mesure de modularité a été utilisée pour la classification des graphes, un lien intéressant entre une matrice de données et la théorie des graphes peut être établi ici. Une matrice de similarité peut être représentée par un graphe non orienté pondéré $G = (V,E)$, où $V = I$ représente l'ensemble des sommets et $E$ l'ensemble des arêtes. La matrice de données $S$ peut être considérée comme une matrice de poids associée au graphe $G$ où chaque noeud $i$ dans $V$ correspond à une ligne. Le lien entre deux sommets $i$ et $i'$ a le poids $s_{ii'}$, désignant l'élément de la matrice située à l'intersection entre la ligne $i$ et la colonne $i'$.

### 2.1.2 Relation d'équivalence

Considérons un ensemble de données divisé en $L$ classes $C = \{C_1, C_2, ...C_L\}$. On peut modéliser la partition $C$ dans un esapce relationnel par une relation d'équivalence $X$, qui doit respecter les propriétés suivantes

$$\begin{cases} x_{ii'} \in \{0,1\}, \forall(i,i') & \text{binarité} \\ x_{ii} = 1, \forall ii & \text{réflexivité} \\ x_{ii'} - x_{i'i} = 0, \forall(i,i') & \text{symétrie} \\ x_{ii'} + x_{i'i''} - x_{ii''} \leq 1, \forall(i,i',i'') & \text{transitivité} \end{cases} \tag{2}$$

# 3 Analyse relationnelle

L'Analyse Relationnelle a été développée en 1977 par Marcotorchino et Michaud, s'inspire des travaux du Marquis de Condorcet, qui s'est intéressé au 18ème siècle au résultat collectif d'un vote à partir de votes individuels. Cette méthodologie est basée sur la représentation relationnelle (comparaison par paires) des différentes variables et l'optimisation sous contraintes du critère de Condorcet.

D'une manière générale, la fonction objective correspond au critère d'adéquation de la solution aux données. Le choix de ce critère est un point fondamental puisque c'est lui qui induit la nature de l'intensité des ressemblances que l'on veut mettre en évidence. L'approche relationnelle permet de choisir parmi une vaste gamme de critères celui qui répond le mieux au problème posé par les données en présence. Certains critères opèrent sur des données binaires, d'autres sont plus appropriés à des données de fréquences; la plupart sont basés sur des règles de majorité qui déterminent le niveau de relation seuil au delà duquel on considère que deux objets sont regroupables. Rappelons que l'un des atouts majeurs de l'approche relationnelle

réside dans le fait que l'on ne doit pas fixer a priori le nombre de classes de la partition. Ce paramètre caractéristique de la solution est directement issu du traitement, réflétant ainsi le potentiel classificatoire inhérent aux données.

L'analyse relationnelle est utilisée pour résoudre de nombreux problèmes rencontrés dans des domaines comme: le classement des préférences, les systèmes de vote, la classification, etc. L'approche de l'analyse relationnelle est un modèle de classification non supervisée qui fournit automatiquement le nombre approprié de classes et qui prend en entrée une matrice de similarité. Dans notre contexte, nous voulons regrouper les objets de l'ensemble $I$ en classes disjointes, la matrice de similarité $S$ est donnée, le but est donc de maximiser la fonction objective suivante :

$$\mathcal{R}_{AR}(S,X) = \sum_{i}\sum_{i'}(s_{ii'} - m_{ii'})x_{ii'} \tag{3}$$

Où $\mathcal{M} = [m_{ii'} = \frac{s_{ii}+s_{ii'}}{4} = \frac{M}{2}]_{i,i'=1,\ldots,N}$ est la matrice des valeurs seuils. Notons que le critère de Condorcet repose sur la notion de majorité, c'est à dire deux individus $i$,$i'$ seront a priori affectés dans une même classe si et seulement si leur similarité $s_{ii'}$ est supérieure ou égale à la valeur de majorité $\frac{M}{2}$. $X$ est la solution recherchée, elle modélise une partition dans un espace relationnel (une relation d'équivalence), et doit vérifier les propriétés données en (2).

# 4 Extension de la mesure de modularité à la classification des données catégorielles

Cette section explique comment adapter la mesure de modularité à la classification des données catégorielles.

## 4.1 Graphe et modularité

La modularité est une mesure récemment utilisée pour mesurer la qualité d'une classification de graphes, elle a immédiatement reçu une attention considérable comme en témoignent les articles (Newman et Girvan, 2004), (Agarwal et Kempe, 2008). Comme dans le cas de la classification relationnelle, la maximisation de la mesure de modularité peut être exprimée sous la forme d'un problème de programmation linéaire en nombres entiers. Étant donné le graphe $G = (V,E)$, soit $A$ une matrice binaire et symétrique où chaque entrée $a_{ii'} = 1$ s'il existe une arête entre les noeuds $i$ et $i'$, s'il n'y a pas de lien entre les noeuds $i$ et $i'$, $a_{ii'}$ est égal à zéro. $A$ est une matrice contenant toutes les informations sur le graphe $G$, est souvent appelée matrice d'adjacence. Trouver une partition de l'ensemble des noeuds $V$ en sous-ensembles homogènes conduit à la résolution du programme linéaire en variables bivalentes suivant :

$$\max_{X} Q(A,X) \tag{4}$$

où

$$Q(A,X) = \frac{1}{2|E|}\sum_{i,i'=1}^{N}(a_{ii'} - \frac{a_{i.}a_{i'.}}{2|E|})x_{ii'} \tag{5}$$

avec, $2|E| = \sum_{i,i'} a_{ii'} = a_{..}$ est le nombre totald'arêtes (liens) et $a_{i.} = \sum_{i'} a_{ii'}$ le degré de l'objet $i$.

La modularité évalue la densité des arêtes dans les classes de façon relative à la densité attendue en cas d'indépendance entre les extrémités des arrêtes. Elle prend ses valeurs entre $-1$ et $1$ et des valeurs positives, quand les classes ont plus d'arêtes observées que dans le cas d'indépendance des extrémités des arêtes. Ce critère vaut $0$ dans les deux cas d'une partition triviale; le cas d'une seule classe et le cas où chaque noeud est isolé dans une classe. La modularité comme le critère de Condorcet possède une propriété intéressante : elle ne nécessite aucun paramètre comme par exemple le nombre de classes.

## 4.2 Première extension : Intégration a priori

L'intégration a priori consiste en une combinaison directe de graphes obtenus à partir de toutes les variables dans un seul ensemble de données (graphe) avant d'appliquer l'algorithme d'apprentissage. Prenons la matrice de Condorcet $S$, (où chaque entrée $s_{ii'} = \sum_{m=1}^{M} s_{ii'}^m$), qui peut être considérée comme une matrice de poids associée au graphe $G = (V,E)$, où chaque arête $e_{ii'}$ a le poids $s_{ii'}$. Par analogie avec la mesure de modularité classique, nous définissons l'extension $Q_1(S,X)$ comme suit (voir figure 1) :

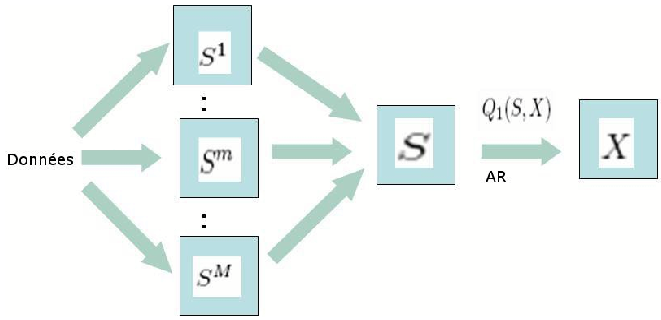


FIG. 1 – – *Intégration a priori*

$$Q_1(S,X) = \frac{1}{2|E|} \sum_{i,i'=1}^{N} (s_{ii'} - \frac{s_{i.}s_{i'.}}{2|E|}) x_{ii'} \tag{6}$$

où $2|E| = \sum_{i,i'} s_{ii'} = s_{..}$ est le poids total et $s_{i.} = \sum_{i'} s_{ii'}$ le degré de l'objet $i$.

## 4.3 Deuxième extension : Intégration intermédiaire

Pour cette extension, l'idée principale est de calculer une mesure de modularité combinée à partir des mesures de modularités claculées séparément sur chaque graphe, et d'appliquer ensuite l'algorithme d'apprentissage (voir figure 2).

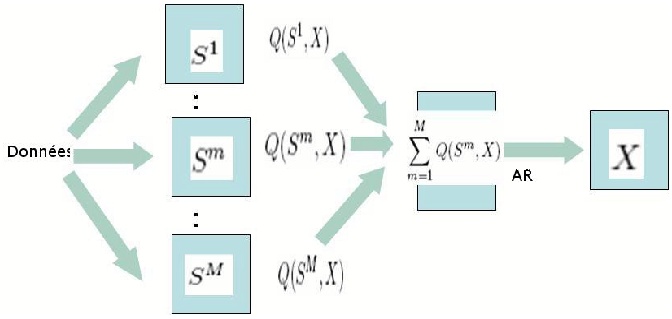


FIG. 2 – – *Intégration intermédiaire*

L'intégration intermédiaires peut être considérée comme une variante de la technique dite "Ensemble Clustering". Considérons un ensemble $V$ de $N$ points. Un ensemble de classifications est une collection de $M$ solutions de classification: $C = \{S^1,S^2,...,S^M\}$. Chaque solution de classification $S^m$ pour $m = 1,...,M$, est une partition (une relation d'équivalence) de l'ensemble V, à savoir $C^m = \{C_1^m,C_2^m,...,C_L^m\}$, où $\bigcup_l C_l^m = V$. Étant donné un ensemble de solutions de classification $C$, l'objectif est de combiner les différentes solutions de classification et de calculer une nouvelle partition de $V$ en groupes disjoints. Le défi dans le problème dit "Ensemble Clustering" est la construction d'une fonction de consensus appropriée qui combine les différentes solutions de classification dans une seule solution finale la plus représentative de la collection des différentes classifications. La fonction objective à maximiser est la suivante :

$$\begin{aligned} Q_2(S^m_{\{m=1,..,M\}},X) &= \sum_{m=1}^{M} Q(S^m,X) \\ &= \sum_{m=1}^{M} \frac{1}{2|E^m|} \sum_{i,i'=1}^{n} (s^m_{ii'} - \frac{s^m_{i.} s^m_{i'.}}{2|E^m|}) x_{ii'} \\ &= \frac{M}{\mathcal{H}} \sum_{i,i'=1}^{n} (s_{ii'} - \frac{\sum_{m=1}^{M} s^m_{i.} s^m_{i'.}}{\mathcal{H}_{ii'}}) x_{ii'} \end{aligned} \quad (7)$$

où $2|E^m| = \sum_{i,i'} s^m_{ii'} = s^m_{..}$ est le poids total dans le graphe $G^m$ et $s^m_{i.} = \sum_{i'} s^m_{ii'}$ le degré de l'objet $i$

$\mathcal{H}$ and $\mathcal{H}_{ii'}$ Sont les moyennes harmoniques des suites $(s^1_{..},s^2_{..},..,s^m_{..},..,s^M_{..})$ et $(\frac{s^1_{i.}s^1_{i'.}}{s^1_{..}},...,\frac{s^m_{i.}s^m_{i'.}}{s^m_{..}},..,\frac{s^M_{i.}s^M_{i'.}}{s^M_{..}})$ respectivement.

$$\frac{M}{\mathcal{H}} = \sum_{m=1}^{M} \frac{1}{s^m_{..}} \tag{8}$$

et

$$\frac{\sum_{m=1}^{M} s^m_{i.}s^m_{i'.}}{\mathcal{H}_{ii'}} = \sum_{m=1}^{M} \frac{s^m_{i.}s^m_{i'.}}{s^m_{..}} \tag{9}$$

## 4.4 Relation entre la mesure de Modularité et le critère de Condorcet

Nous pouvons établir une relation entre les deux extensions de la mesure de modularité et le critère de l'analyse relationnelle, en effet les fonctions $Q_1(S,X)$ et $Q_2(S,X)$ peuvent être exprimées comme étant un critère modifié de celui de l'AR de la façon suivante :

### 4.4.1 $Q_1(S,X)$ comme un critère de Condorcet modifié

$$Q_1(S,X) \quad = \quad \frac{1}{2|E|}(\mathcal{R}_{AR}(S,X) + \psi_1(S,X)) \tag{10}$$

et

$$\psi_1(S,X) = \sum_{i=1}^{n}\sum_{i'=1}^{n}(m_{ii'} - \frac{s_{i.}s_{i'.}}{2|E|})x_{ii'} \tag{11}$$

est le terme de pondération qui dépend du profil de chaque paire d'objets $(i,i')$.

### 4.4.2 $Q_2(S,X)$ comme un critère de Condorcet modifié

De la même manière l'extension $Q_2(S,X)$ s'écrira

$$Q_2(S^m_{\{m=1,..,M\}},X) \quad = \quad \frac{M}{\mathcal{H}}(\mathcal{R}_{AR}(S,X) + \psi_2(S,X)) \tag{12}$$

où

$$\psi_2(S,X) = \sum_{i=1}^{n}\sum_{i'=1}^{n}(m_{ii'} - \frac{\sum_{m=1}^{M} s^m_{i.}s^m_{i'.}}{\mathcal{H}_{ii'}})x_{ii'} \tag{13}$$

Les deux extensions de la mesure modularité permettent d'introduire un système de pondération en fonction du profil de chaque paire d'objets.

# 5 Procédure d'optimisation

Comme la fonction objective est linéaire par rapport à $X$ et les contraintes que $X$ doit respecter sont des équations linéaires, théoriquement on peut résoudre le problème en utilisant un solveur de programmation linéaire en nombres entiers. Toutefois, ce problème est NP-difficile. En conséquence, dans la pratique, nous utilisons des heuristiques pour faire face aux grands ensembles de données.

## 5.1 Décomposition de la mesure de modularité

Les deux extensions de la mesure modularité peuvent être décomposées en termes de la contribution de chaque objet $i$ dans chaque classe $C_l$ de la partition recherchée de la manière suivante :

$$Q_1(S,X) = \sum_{l=1}^{L} \sum_{i=1}^{N} cont_{Q_1}(i,C_l) \tag{14}$$

où

$$cont_{Q_1}(i,C_l) = \frac{1}{2|E|} \sum_{i' \in C_l} (s_{ii'} - \frac{s_{i.} s_{i'.}}{2|E|}) \tag{15}$$

En utilisant les transformations : $s_{ii'} =< K_i,K_{i'} >$ et $s_{i.} = \sum_{i''} < K_i,K_{i''} >$ (où $K_i$ désigne la $ieme$ ligne du tableau disjonctif complet $K$), l'expression de la formule de contribution devient[1],

$$\begin{aligned} cont_{Q_1}(i,C_l) &= \frac{1}{2|E|} \sum_{i' \in C_l} (< K_i,K_{i'} > \\ &- \frac{\sum_{i''} < K_i,K_{i''} > \sum_{i''} < K_{i'},K_{i''} >}{2|E|}) \qquad (16) \\ &= \frac{1}{2|E|} < K_i,P_l > - \sum_{i' \in C_l} \delta_{ii'} \qquad (17) \end{aligned}$$

où

$$P_l = \sum_{i' \in C_l} K_{i'} \tag{18}$$

et

$$\delta_{ii'} = \frac{\sum_{i''} < K_i,K_{i''} > \sum_{i''} < K_{i'},K_{i''} >}{2|E|} \tag{19}$$

De la même façon la contribution $cont_{Q_2}(i,C_l)$ peut être réécrite ;

1. Rappelons que cette nouvelle écriture de la formule de contribution permet de réduire considérablement le coût de calcul lié à la matrice de similarité $S$ et de caractériser chaque classe $C_l$ avec son prototype $P_l$.

$$
\begin{aligned}
cont_{Q_2}(i,C_l) &= \frac{M}{\mathcal{H}} \sum_{i' \in C_l} (< K_i,K_{i'} > \\
&- \frac{\sum_{m=1}^{M} \sum_{i''} < K_i^m,K_{i''}^m > \sum_{i''} < K_{i'}^m,K_{i''}^m >}{\mathcal{H}_{ii'}}) \\
&= \frac{M}{\mathcal{H}}(< K_i,P_l > - \sum_{i' \in C_l} \tilde{\delta}_{ii'}) \qquad (20)
\end{aligned}
$$

où

$$
\tilde{\delta}_{ii'} = \frac{\sum_{m=1}^{M} \sum_{i''} < K_i^m,K_{i''}^m > \sum_{i''} < K_{i'}^m,K_{i''}^m >}{\mathcal{H}_{ii'}} \qquad (21)
$$

La nouvelle formule de contribution introduit une pondération automatique, la valeur de la nouvelle formule de contribution sera soit supérieure, inférieure ou bien égale à la contribution de l'AR en fonction des poids ($\delta_{ii'}$ ou $\tilde{\delta}_{ii'}$). Les formules de contributions $cont_{Q_1}$ et $cont_{Q_2}$ peuvent être écrites en terme de contribution $cont_{AR}$ par l'ajout d'un terme de pondération en fonction du profil de chaque paire d'objets $(i,i')$ :

$$
\begin{aligned}
cont_{Q_1}(i,C_l) &= \frac{1}{2|E|}[(< K_i,P_l > - \sum_{i' \in C_l} m_{ii'}) \\
&+ \sum_{i' \in C_l} (m_{ii'} - \delta_{ii'})] \qquad (22) \\
&= \frac{1}{2|E|}[cont_{AR}(i,C_l) + \sum_{i' \in C_l} (m_{ii'} - \delta_{ii'})] \qquad (23)
\end{aligned}
$$

**Remarque :** De la même manière, la contribution $cont_{Q_2}$ peut être réécrite en fonction de la contribution de l'AR, $cont_{AR} =< K_i,P_l > - \sum_{i' \in C_l} m_{ii'}$.

Le changement dans la formule de contribution est intéressant car il introduit une pondération relative aux profils des objets de manière automatique sans nécessiter la présence d'un expert. On distingue trois scénarios :

1. Prenant $\delta_{ii'} = m_{ii'},\ \forall i,i'$, nous trouvons ainsi le cas de l'algorithme de l'AR
2. Si le poids $\delta_{ii'}$ est inférieure à $m_{ii'},\ \forall i,i'$, alors la valeur de contribution $cont_{Q_1}$ est supérieure à l'ancienne contribution $cont_{AR}$, et elle a donc plus de chance d'être positive que $cont_{AR}$; l'observation $i$ se trouvera alors affectée à une classe pré-existante. Ainsi, le nombre de classes sera petit.
3. Si le poid $\delta_{ii'}$ est supérieur à $m_{ii'},\ \forall i,i'$, alors la valeur de la contribution $cont_{Q_1}$ est inférieure à l'ancienne $cont_{AR}$, et elle a donc plus de chance d'être négative que $cont_{AR}$; l'observation $i$ est alors affectée à une nouvelle classe. Ainsi, le nombre de classes sera plus important.

### 5.2 Algorithme de l'Analyse Relationnelle

Le processus consiste à partir d'une classe initiale (une classe singleton) à construire de façon incrémentale une partition de l'ensemble $I$ en accentuant à chaque affectation la valeur du critère de la modularité. Nous donnons ci-dessous la description de l'algorithme d'analyse relationnelle qui a été utilisé par la méthodologie de l'analyse relationnelle (voir (Marcotorchino, 1978) pour plus de détails)

**Algorithme1: Algorithme de l'AR**

**Inputs**
*Initialisation :* $N_{iter}$= nombre d'itérations. $N$= nombre d'individus (observations). $L_{max} = N$ le nombre maximal de classes.
- Claculer les matrices des valeurs seuils ($M$,$\delta$ où $\tilde{\delta}$)
- Prendre le premier individu comme étant le premier élément de la classe $\mathcal{C}_1$
- $l = 1$, où $l$ est le nombre de classes

**for** t=1 to $N_{iter}$ **do**
  **for** $i = 1$ to $N$ **do**{*Affectation*}
    **for** $k = 1$ to $l$ **do**
      Calculer la contribution $cont(K_i,P_k)$
    **end for**
      $k^* = arg\max_k cont(K_i,P_k)$
      $cont(K_i,P_{k^*}) \leftarrow$ la contribution calculée
    **if** $cont(K_i,P_{k^*}) < 0$ **and** $l < L_{max}$ **then**
      Créer une nouvelle classe dont $i$ est le premier individu affecté à cette classe
      $l = l + 1$
    **else**
      Affecter $i$ à $\mathcal{C}_{k^*}$
    **endif**
  **endfor**
**endfor**
**Ouputs :** au plus une parition de $L_{max}$ classes.

Nous devons produire un certain nombre d'itérations afin d'obtenir une solution approchée dans un temps de traitement raisonnable. D'ailleurs, il est exigé un nombre maximal de classes mais puisque nous n'avons pas besoin de fixer ce paramètre, nous avons pris $L_{max} = N$ par défaut. Fondamentalement le coût de calcul de cet algorithme est en $O(N_{iter} \times L_{max} \times N)$ . En général, on peut supposer que $N_{iter} << N$, mais pas $L_{max} << N$. Ainsi, dans le pire des cas, l'algorithme a une complexité en $O(L_{max} \times N)$.

## 6 Expérimentation et validation

Afin de pouvoir évaluer la qualité de la classification obtenue, nous avons utilisé des bases de données qualitatives UCI (Asuncion et Newman, 2007) comportant un nombre variable d'observations (voir TAB. 1). Nous avons utilisé le taux de bonne classification (appelé aussi

pureté), les indices de Rand, de Jaccard et de Tanimoto en utilisant la classe connue de chaque observation. L'évaluation de la pureté ou de la qualité d'une partition obtenue consiste à évaluer si la partition résultat est cohérente par rapport à la connaissance disponible.

## 6.1 Mesures de performances

### 6.1.1 Indice de Pureté

Considérons $L$ clusters de l'ensemble de données $V$ et soit $|\mathcal{C}_l|$ la taille du cluster $\mathcal{C}_l$ . La pureté de ce cluster est donnée par l'expression Pureté($\mathcal{C}_l$) $= \frac{1}{|\mathcal{C}_l|} \max_k(|\mathcal{C}_l|_{cluster=k})$ où $|\mathcal{C}_l|_{cluster=k}$ désigne le nombre d'objets de la classe $k$ attribué au cluster $l$. La pureté globale d'une partition résultat peut être exprimée comme une somme pondérée des puretés individuelles des clusters.

$$\text{Pureté} = \sum_{l=1}^{L} \frac{|\mathcal{C}_l|}{|V|} \text{Pureté}(\mathcal{C}_l) \tag{24}$$

En général, plus la valeur de pureté est élevée, meilleure est la partition obtenue.

### 6.1.2 Indice de Rand (RI)

Indice de Rand (Rand, 1971) : mesure le nombre d'accords par paires entre une partition obtenue $U'$ et la vraie partition $U$ d'un même ensemble d'objets, normalisé de sorte que la valeur se situe entre 0 et 1 :

$$RI(U,U') = \frac{a+b}{a+b+c+d} \tag{25}$$

Où $a$ désigne le nombre de paires d'objets appartenant à la même classe de $U$ et affectés au même cluster de $U'$, $b$ désigne le nombre de paires dont les objets appartiennent à deux classes différentes de $U$ et à deux clusters différents de $U'$, $c$ désigne le nombre de paires d'objets appartenants à la même classe de $U$ et à deux clusters différents de $U'$, et $d$ désigne le nombre de paires dont les objets appartiennent à deux classes différentes de $U$ et affectées au même cluster de $U'$. Cet indice donne un résultat dans l'intervalle [0,1], où une valeur de 1 indique que $U$ et $U'$ sont identiques.

### 6.1.3 Indice de jaccard(JI)

Indice de Jaccard (Jaccard, 1912) a été couramment utilisé pour évaluer la similarité entre différentes partitions du même ensemble de données, le niveau d'accord entre la vraie partition $U$ et une partition résultat $U'$ est déterminé par le nombre de paires de points attribués à une même classe dans les deux partitions :

$$JI(U,U') = \frac{a}{a+d+c} \tag{26}$$

L'indice de Jaccard donne un résultat dans l'intervalle $[0,1]$, où une valeur de 1 indique que $U$ et $U'$ sont identiques.

#### 6.1.4 Indice de Tanimoto (TI)

La similarité entre différentes partitions d'un ensemble de données peut être mesurée par le ratio de leurs éléments communs au nombre de tous les différents éléments,

$$TI(U,U') = \frac{\frac{1}{2}(a+b)}{\frac{1}{2}(a+b)+d+c} \tag{27}$$

Cet indice donne un résultat dans l'intervalle [0,1].

## 6.2 Bases de données pour la validation

Dans cette section, nous évaluons la performance de l'heuristique proposée sur plusieures bases de données disponibles à l'UCI. La description des bases de données utilisées est donnée dans TAB. 1 :

TAB. 1 – – *Description des bases de données*

| Bases de données | # d'objets | # d'attributs | # de classes |
|---|---|---|---|
| Soybean small | 47 | 21 | 4 |
| Zoo | 101 | 16 | 7 |
| Soybean large | 307 | 35 | 19 |
| SPECTF | 267 | 22 | 2 |
| Post-Operative | 90 | 8 | 3 |
| Balance Scale | 625 | 4 | 3 |
| Audiology Normalized | 226 | 69 | 24 |

## 6.3 Résultats dans le cas d'une integration a priori

La méthode proposée est testée sur des bases de données obtenues à partir du référentiel de données UCI. Comme la méthode proposée est une modification de l'approche de l'AR, nous avons comparé les performances de l'algorithme proposé avec l'algorithme de l'AR. De la table TAB. 2 et FIG. 3 , il est clair que la performance de la méthode proposée qui repose sur la mesure modularité étendue est meilleure que l'approche AR pour toutes les bases de données. Cela signifie que le système de pondération introduit améliore les résultats de la pureté du clustering.

Afin de montrer la bonne performance de l'approche proposée, nous utilisons plusieures bases de données catégorielles de différentes tailles et nous indiquons dans TAB. 4 les valeurs des indices RI, JI et l'indice TI obtenus en utilisant le critère classique de l'AR et dans TAB. 3 les indices RI, JI et l'indice TI en utilisant la mesure de modularité étendue. Les résultats montrent que l'approche proposée augmente la valeur des indices par rapport à l'AR classique et permet d'introduire un système de pondération automatique relatif au profil de chaque objet dans la base de données.

TAB. 2 – – *Mesures de Pureté pour $\mathcal{R}_{AR}(S,X)$ et $Q_1(S,X)$*

| BD | Taille | $\mathcal{R}_{AR}(S,X)$ | $Q_1(S,X)$ |
|---|---|---|---|
| Soybean small | 47x21 | 78 % | 100 % |
| Zoo | 101x16 | 83.17% | 88.12 % |
| Soybean large | 307x35 | 70 % | 72.31 % |
| SPECTF | 267x22 | 61.25 % | 85 % |
| Post-Operative | 90x8 | 71.11 % | 73.33% % |
| Balance Scale | 625x4 | 63.52 % | 63.52 % |
| Audiology Normalized | 226x69 | 50.50 % | 58 % |

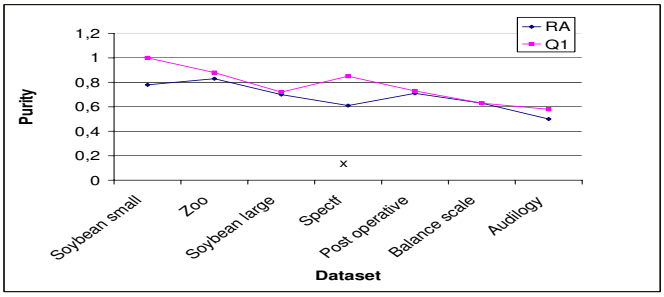


FIG. 3 – – *Mesures de Pureté sur différentes bases de données*

TAB. 3 – – *Résultats sur différentes bases de données en utilisant $Q_1(S,X)$*

| BD | Taille | RI | JI | TI |
|---|---|---|---|---|
| Soybean small | 47x21 | 100 % | 100 % | 100 % |
| Zoo | 101x16 | 94.2% | 79.9 % | 89.2 % |
| Soybean large | 307x35 | 91.2 % | 26.6 % | 83.9 % |
| SPECTF | 267x22 | 60.98 % | 38.28 % | 43.86 % |
| Post-Operative | 90x8 | 50.75% | 37.37% % | 34.01% |
| Balance Scale | 625x4 | 58 % | 20 % | 40 % |
| Audiology Normalized | 226x69 | 82 % | 20 % | 69% |

TAB. 4 – – *Résultats sur différentes bases de données en utilisant* $\mathcal{R}_{AR}(S,X)$

| BD | Taille | RI | JI | TI |
|---|---|---|---|---|
| Soybean small | 47x21 | 86.66 % | 45.88 % | 76.47 % |
| Zoo | 101x16 | 72.9% | 46.09 % | 57.37 % |
| Soybean large | 307x35 | 85.03 % | 25.7 % | 73.97 % |
| SPECTF | 267x22 | 55.74 % | 38.69 % | 38.64 % |
| Post-Operative | 90x8 | 54.44% | 41.17% % | 37.4% |
| Balance Scale | 625x4 | 57 % | 19 % | 39 % |
| Audiology Normalized | 226x69 | 82 % | 20 % | 69 % |

# 7 Conclusions et perspectives

Dans ce papier, nous avons étudié deux extensions du critère de modularité pour la classification des données catégorielles et illustrer ses relations avec le critère de Condorcet. Une procédure itérative efficace d'optimisation est présentée. Les résultats expérimentaux montrent l'efficacité de la méthode de l'intégration a priori proposée par rapport à l'approche de l'analyse relationnelle classique. La validation de la démarche d'intégration intermédiaire sera établie dans des travaux futurs, une autre idée est d'adapter ces extensions pour la théorie des graphes en utilisant les matrices d'adjacence.

# Références

Agarwal, G. and Kempe, D. (2008). Modularity-maximizing graph communities via mathematical programming. The European Physical Journal B 66:33, 409-418.

Asuncion, A. Newman, D.J. (2007). "UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine", CA: University of California, School of Information and Computer Science.

Barbara, D., Couto, J., Li, Y. (2002). COOLCAT: an entropy-based algorithm for categorical clustering. Proceedings of the Eleventh ACM CIKM Conference (pp. 582-589).

Bock, H.-H. (1989). Probabilistic aspects in cluster analysis. In O. Opitz (Ed.), Conceptual and numerical analysis of data, 12-44. Berlin: Springer-verlag .

Celeux, G., Govaert, G. (1991). Clustering criteria for discrete data and latent class models. Journal of Classification, 8, 157-176.

Ganti, V., Gehrke, J., Ramakrishnan, R. (1999). CACTUS - clustering categorical data using summaries. Proceedings of the Fifth ACM SIGKDD Conference (pp. 73- 83).

Gibson, D., Kleinberg, J., Raghavan, P. (1998). Clustering categorical data: An approach based on dynamical systems. Proceedings of the 24rd VLDB Conference (pp. 311-322).

Guha, S., Rastogi, R., Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. Information Systems, 25, 345-366.

Gyllenberg, M., Koski, T., Verlaan, M. (1997). Classification of binary vectors by stochastic complexity. Journal of Multivariate Analysis, 47-72.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2, 283-304.

Li, T., Zhu, S., Ogihara, M. (2003). Efficient multi-way text categorization via generalized discriminant analysis. Proceedings of Twelfth ACM CIKM Conference (pp. 317-324).

Marcotorchino, J. F. (2006). Relational analysis theory as a general approach to data analysis and data fusion, in *Cognitive Systems with interactive sensors, 2006.*

Marcotorchino, J. F. and Michaud, P. *Optimisation en analyse ordinale des données.* (In Masson, 1978.)

Newman, M. and Girvan, M.(2004). Finding and evaluating community structure in networks. Physical Review E, 69, 026113.

Li, T. Ma, S., and Ogihara, M. (2004). Entropy-based criterion in categorical clustering. Proceedings of The 2004 IEEE International Conference on Machine Learning (ICML 2004). 536-543.

White, S. and Smyth, P. (2005). A spectral clustering approach to finding communities in graphs. In SDM, pages 76-84.

Zhao, Y., Karypis, G. (2001). Criterion functions for document clustering: Experiments and analysis (Technical Report). Department of Computer Science, University of Minnesota.

## Summary

This paper studies the extension of the Modularity measure for categorical data clustering. It first shows the relational data presentation and establishes the relationship between the extended Modularity and the Relational Analysis criterion. Two extensions are presented in this work: the early integration and the intermediate integration approaches. The proposed Modularity measure introduces an automatic weighting scheme which takes in consideration the profile of each data object. An iterative algorithm is then presented to search for the partitions maximizing this criterion. This algorithm deals linearly with large data sets and allows natural clusters identification, i.e. doesn't require fixing the number of clusters and the size of each cluster. For the early integration approach, several experiments are conducted in order to show the effectiveness of the proposed approach.