Université Lumière Lyon2
Année 2003

Thèse
pour obtenir le grade de
Docteur
en
Informatique

présentée et soutenue publiquement par

Pierre-Emmanuel JOUVE
le 10 décembre 2003

Apprentissage Non Supervisé et Extraction de Connaissances à partir de Données

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

sous la direction de
Nicolas Nicoloyannis

devant le jury, composé de:

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

Table des matières

4 Validité en Apprentissage Non Supervisé

"The popularized definition of KDD postulates it as "The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". We are unsure if valid is listed among the first characteristics in proportion to its importance, but certainly, patterns in data will be far from useful if they were invalid. A more cynical view would say that trivial processes can certainly deliver invalid, understandable and novel patterns."

- Vladimir Estevill-Castro -
"Why so many clustering algorithms - A Position Paper", SIGKDD Explorations, vol.4, issue 1, pages 65-75 (2002)

L'ECD est définie dans [FPSS96] comme le processus non trivial d'identification de connaissances valides, nouvelles, potentiellement utiles et intelligibles au sein d'un ensemble de données[1]. Estevill-Castro remarque dans [EC02] que, si l'on ne peut pas forcément expliquer la présence du terme valide en tête de définition par une connotation à l'importance majeure de la validité, il est par contre certain que la découverte de connaissances non valides est d'un intérêt nul. L'évaluation de la validité de la connaissance extraite (ou plutôt de l'éventuelle connaissance extraite puisqu'une connaissance non valide ne peut être considérée comme de la connaissance) constitue donc une étape clé du processus ECD que nous abordons ici dans le cadre restreint de la cns.

L'aspect fondamental de la classification non supervisée dans le processus d'ECD a mené ces dernières années à de multiples efforts de recherche dans ce domaine et plus particulièrement au développement de nouveaux algorithmes au coût calculatoire faible ainsi qu'à la mise au point de critères adaptés au traitement de type de données spécifiques. Il résulte de ces études le besoin de posséder des outils pour l'évaluation et la comparaison de la validité de résultats du processus de classification non supervisée. En effet, outre leur utilité pour la confirmation de la validité de résultats, ces outils peuvent également

1. Il s'agit ici de la définition anglo-saxone, la définition francophone insistant également sur la nature itérative et le caractère interactif de ce processus.

assister les utilisateurs dans le choix d'une classification parmi un ensemble de classifications, et ce, indépendamment de la méthode utilisée, des paramètres de la méthode et du nombre de classes. De tels outils pourraient encore permettre la comparaison objective de méthodes et ainsi conduire finalement à la définition d'un cadre pour la comparaison des méthodes de cns. [2]

Bien que nous regrettions que cette problématique ait été, selon nous, et, eu égard à son importance, l'objet de trop peu de travaux, nous pouvons cependant citer plusieurs travaux de recherche plus ou moins récente [BEF84], [Dav96], [Dom01], [Hal00a], [HBV01], [HV01], [HBV02b], [HBV02a], [RLR98b], [Sha96], [TK99], [XB91]... Notons cependant que cette problématique connaît un regain d'intérêt fort comme le démontre notamment les publications récentes de Maria Halkidi et Michalis Vazirgiannis ([Hal00a], [HBV01], [HV01], [HBV02b], [HBV02a]) ainsi que le tutoriel sur l'estimation de la qualité en fouille de données [3] qu'ils ont animé lors des conférences ECML/PKDD02 : *"An Introduction to Quality Assessment in Data Mining"* et leur très récent livre consacré intégralement à cette problématique [VHD03].

Dans un premier temps, nous nous référons à ces travaux pour introduire les critères existants pour l'évaluation de la validité de cns, puis proposons et expérimentons deux nouveaux critères associés à une méthodologie qui leur est propre pour l'évaluation de la validité de cns.

4.1 Validité d'une Classification Non Supervisée : Définition et Evaluation

L'objectif principal de la cns est la découverte de l'organisation (structuration) d'un ensemble d'objets selon des classes naturelles afin d'identifier des similarités et différences entre classes ainsi qu'inférer des spécificités intéressantes pour chaque classe. Or, la nature non supervisée de ce processus n'autorise pas une définition claire et directe de ce que sont des structures/organisations valides. Ainsi, les multiples algorithmes de cns se caractérisent par l'ensemble d'hypothèses qu'ils emploient afin de définir les propriétés devant être satisfaites par une structure valide. L'ensemble d'hypothèses déterminant la validité d'une structure n'étant pas universel et différant selon les méthodes, les résultats varient donc selon la méthode utilisée. Conséquemment, il est essentiel de définir une méthode d'évaluation des structures résultant d'un processus de cns. Ce type d'évaluation est nommé évaluation de la validité d'une cns.

On peut considérer qu'il existe, de manière générale, trois modes d'évaluation de la validité de cns. Le premier est basé sur des critères dits externes, qui

2. Sans l'utilisation de ce type d'outils l'évaluation de résultats provenant de cns ainsi que la comparaison de tels résultats peut être difficile et hasardeux.

3. La validité constitue une des composantes les plus importantes pour l'estimation de la qualité en fouille de données.

impliquent l'évaluation de cns au moyen d'une structure définie a priori, cette structure traduisant les connaissances et intuitions de l'utilisateur sur la structure des données. Le second mode est lui fondé sur des critères dits internes, ces derniers impliquent quant à eux une évaluation sur l'unique base des données à traiter et ne font aucunement intervenir des informations exogènes. Le dernier mode est dit relatif ; l'idée clé est ici d'évaluer une cns en se référant à d'autres cns obtenues par l'intermédiaire de la même méthode mais avec des paramétrages différents. Notons enfin que, majoritairement, les critères évalués ont un rapport avec l'homogénéité des classes ou avec la séparation entre classes, plus rarement, ces deux notions sont intégrées simultanément au sein d'un critère.

4.1.1 Mode d'Evaluation par Critères Externes

Le mode d'évaluation impliquant un critère externe est utilisé implicitement dans la plupart des publications traitant de l'évaluation expérimentale de méthodes de cns. Ces critères de validité évaluent dans quelle mesure une cns correspond à des connaissances et intuitions établies a priori sur les données. On admet généralement que ces informations ne peuvent être directement calculées à partir des données initiales. La forme la plus classique d'informations externes est un ensemble de classes et d'étiquettes associées à chacun des objets[4].

L'idée clé est donc de tester si l'ensemble d'objets est structuré de manière aléatoire ou non, et ce, en se référant à une structure pré-définie. Cette analyse se base alors sur l'hypothèse nulle H_0 d'une structure aléatoire. Pour tester cette hypothèse, les tests statistiques peuvent être utilisés, cependant, ils peuvent mener à des procédures calculatoirement coûteuses. Aussi, la méthode de Monte Carlo est parfois utilisée pour résoudre ce type de problèmes.

4.1.1.1 Méthode de Monte Carlo

Cette méthode est utilisée afin de calculer la fonction de densité de probabilité d'un indice statistique par l'intermédiaire de la simulation [TK99] : on procède alors tout d'abord par génération aléatoire d'un grand nombre de partitions des objets du jeu de données considéré (ces partitions correspondent pour ce jeu de données à des cns potentielles) ; puis pour chacune de ces partitions, on calcule la valeur de l'indice dont on recherche la fonction de densité de probabilité ; puis, en utilisant les différentes valeurs obtenues pour l'indice on peut déterminer une approximation de la fonction de densité de probabilité de l'indice. Enfin, les tests statistiques classiques peuvent être employés.

4. Ce type d'informations peut éventuellement être obtenu par une classification manuelle

4.1.1.2 Mesures Statistiques

Considérons $O = \{o_i, i = 1..n\}$ un ensemble d'objets, $P_h = \{C_1,...,C_z\}$ une partition de O en z classes et $P_{spec} = \{C_{spec_1},...,C_{spec_x}\}$ une cns pré-spécifiée (une partition de O en x classes). Par la suite, afin de faire référence à la cardinalité de différents ensembles composés de paires d'objets (o_i,o_j) , nous utilisons les notations suivantes :

- **SS** la cardinalité de l'ensemble composé des paires d'objets de O telles que les deux objets appartiennent à la même classe dans chacune des partitions P_h et P_{spec}
- **SD** la cardinalité de l'ensemble composé des paires d'objets de O telles que les deux objets appartiennent à la même classe dans la partition P_h et à des classes différentes dans la partition P_{spec}.
- **DS** la cardinalité de l'ensemble composé des paires d'objets de O telles que les deux objets appartiennent à la même classe dans la partition P_{spec} et à des classes différentes dans la partition P_h
- **DD** la cardinalité de l'ensemble composé des paires d'objets de O telles que les deux objets appartiennent à des classes différentes dans chacune des partitions P_h et P_{spec}.

REMARQUES :

- $M = SS + SD + DS + DD = \frac{n \times (n-1)}{2}$
- *Si l'on considère l'ensemble d'objets O, que l'on effectue une correspondance entre la partition P_{spec} (resp. P_h) et une variable catégorielle et que l'on considère que cette variable catégorielle est l'unique descripteur de chacun de ses objets alors il existe une relation entre les notations **SS**, **SD**, **DS**, **DD** et la valeur du nouveau critère de Condorcet pour la partition P_h (resp. P_{spec}) : $NCC(P_h) =$ **SS** + **DD** (resp. $NCC(P_{spec}) =$ **SS** + **DD**).*

Des mesures statistiques classiques sont alors définies, en employant ces notations, de la manière suivante :

- *Statistique de Rand :* $R = \frac{SS+DD}{M}$
- *Coefficient de Jaccard :* $J = \frac{SS}{SS+SD+DD}$
- *Indice de Folkes et Mallows :* $FM = \sqrt{\frac{SD}{SS+SD} \frac{SS}{DS+DD}}$
- *Statistique* Γ *de Hubert* : $\Gamma = \frac{SS}{M}$
- *Statistique* Γ *Normalisée :* $\overline{\Gamma}$

REMARQUES :

- *Les deux premières statistiques (R et J) prennent des valeurs entre 0 et 1 et sont maximales pour $z = x$ (i.e. lorsque les partitions P_h et P_{spec} ont le même nombre de classes).*

- *Il a été prouvé que, pour les 3 premières mesures (R, J et FM), de fortes valeurs indiques une grande similarité entre P_h et P_{spec} (i.e. plus fortes sont les valeurs, plus similaires sont les deux partitions)*
- *Pour les indices Γ et $\overline{\Gamma}$, on considère les matrices d'adjacence $n \times n$ MC_{P_h} et $MC_{P_{spec}}$ décrivant les relations d'équivalence associées respectivement aux partitions P_h et P_{spec}. Ces matrices sont définies comme suit :*

  $(MC_{P_{spec}})_{i,j} = \begin{cases} 1 \text{ si } \exists k \in \{1,..,x\} \text{ tel que } o_i \in C_{spec_k} \text{ et } o_j \in C_{spec_k} \\ 0 \text{ sinon} \end{cases}$

  $(MC_{P_h})_{i,j} = \begin{cases} 1 \text{ si } \exists k \in \{1,..,z\} \text{ tel que } o_i \in C_k \text{ et } o_j \in C_k \\ 0 \text{ sinon} \end{cases}$

  avec ces notations on a :

$$\Gamma = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} MC_{P_{h_{i,j}}} MC_{P_{spec_{i,j}}}$$

$$\overline{\Gamma} = \frac{\left(\frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (MC_{P_{h_{i,j}}} - \mu_{MC_{P_h}})(MC_{P_{spec_{i,j}}} - \mu_{MC_{P_{spec}}}) \right)}{(\sigma_{MC_{P_h}} \sigma_{MC_{P_{spec}}})}$$

- *Quel que soit l'indice que nous venons de présenter, on doit connaître sa fonction de densité de probabilité sous l'hypothèse H_0 de structuration aléatoire afin de procéder à des tests statistiques visant à évaluer la validité de la cns. Or comme nous l'avons annoncé plus tôt, déterminer cette fonction peut impliquer un coût calculatoire trop important et nécessiter l'adoption de techniques de Monte Carlo.*

Nous pouvons enfin noter l'existence de critères basés sur la théorie de l'information qui ne nécessitent pas l'utilisation de la méthode de Monte-Carlo tels celui proposé par Dom [Dom01], l'information mutuelle normalisée, ou encore la pureté moyenne ou des mesures basées sur l'entropie [BGG+99].

4.1.2 Mode d'Evaluation par Critères Internes

L'idée est ici d'évaluer une cns résultant de l'application d'un algorithme particulier par utilisation d'une mesure ne considérant que l'information comprise dans les données et aucune information additionnelle. L'utilisation de ce type de mesure conduit à la question suivante : "La mesure que j'utilise permet-elle réellement de capturer l'adéquation entre la classification obtenue et ce qui est particulier dans les données et que je souhaite découvrir ? ".

Les critères internes, tels que la somme du carré des erreurs (SSE : sum of squared errors), ont été utilisés de manière extensive car, la cns peut être vue comme un problème d'optimisation de la valeur d'une mesure interne de validité donnée, (par exemple, les k-means optimisent de manière gloutonne le

critère SSE). Nous listons maintenant quelques unes des mesures internes les plus courantes :

- *SSE (inertie intra-classe, homogénéité intra-classe)* il s'agit certainement de la mesure la plus populaire. Elle est définie de la manière suivante : soit une partition (une cns) $P_h = \{C_1,...,C_z\}$, nous notons n_{C_i} le nombre d'objets de la classe C_i et définissons $c_i = \{c_{i_1},...,c_{i_p}\}$ le centroïde de C_i par $c_{i_j} = \frac{1}{n_{C_i}} \sum_{o_a \in C_i} o_{a_j}$. Ainsi, $SSE(P_h) = \sum_{k=1..z} \sum_{o_a \in C_k} ||o_a - c_k||_2^2$.
  On peut étendre cette définition à d'autres mesure de dissimilarité s entre les objets et leurs centroïdes respectifs : $SSE(P_h) = \sum_{k=1..z} \sum_{o_a \in C_k} s(o_a,c_k)$.
- *Nombre d'arêtes coupées* : lorsque la cns est posée comme un problème de partitionnement de graphe, l'objectif est alors de minimiser le nombre d'arêtes coupées.
- *CU (Category Utility)* [GC85][Fis87], ce critère est une fonction de la prédictabilité des valeurs des attributs impliquées dans une cns. La mesure *CU* est définie comme la différence entre le nombre de valeurs d'attributs pouvant être correctement prédits grâce à l'établissement d'une partition des objets d'un jeu de données et le nombre de valeurs d'attributs pouvant être correctement prédits sans une telle connaissance. (Récemment, il a été montré que ce critère est lié aux critères de type SSE pour un type de codage spécifique [Mir01].) La mesure *CU* est donc définie pour maximiser la prédictabilité des attributs pour une cns, ce qui limite son champ d'utilisation à des problèmes de cns possédant une dimensionalité faible (et touchant de préférence des attributs catégoriels). En effet, pour des problèmes à forte dimension, tels que ceux posés en cns sur textes, l'objectif n'est évidemment pas d'être capable de prédire la présence d'un mot dans un document associé à une classe particulière
- *Coefficient de Corrélation CoPhénétique (CPCC)*: dans le cadre de cns hiérarchiques, une matrice, appelée matrice cophénétique, représente le diagramme hiérarchique (dendrogramme) produit par l'algorithme : chaque élément $M_{coph_{i,j}}$ de la matrice cophénétique représente le niveau pour lequel les objets o_i et o_j se retrouvent pour la première fois dans la même classe. Un indice de proximité entre une matrice cophénétique ($M_{coph_{i,j}}$) et la matrice d'adjacence MC_{P_h} d'une partition P en z classes a été établi et appelé coefficient de corrélation cophénétique (CPCC):

$$CPCC = \frac{\frac{1}{M}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} M_{coph_{i,j}} MC_{P_{h_{i,j}}} - \mu_{M_{coph}}\mu_{MC_{P_h}}\right)}{\sqrt{\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} MC^2_{P_{h_{i,j}}} - \mu^2_{M_{coph}}\right)\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} M^2_{coph_{i,j}} - \mu^2_{MC_{P_h}}\right)}}$$

$$-1 \leq CPCC \leq 1$$

Avec, n le nombre d'objets, $M = \frac{n(n-1)}{2}$ le nombre de paires d'objets différents,

$$\mu_{M_{coph}} = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} M_{coph_{i,j}} \qquad \mu_{MC_{P_h}} = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} MC_{P_{h_{i,j}}}$$

Une procédure de type Monte Carlo peut être employée afin d'évaluer la fonction densité de probabilité de l'indice CPCC et procéder ultérieurement aux tests statistiques.

Dans la plupart des cas, l'utilisation de critères internes permet des comparaisons entre cns résultant de processus de classification relativement similaires et notamment en ce qui concerne la mesure de distance/similarité sous jacente à ce processus. Ainsi, dans de nombreuses situations (comme, par exemple, la comparaison de la validité de cns résultant de deux algorithmes employant des mesures de similarité très largement différents) un consensus sur la mesure de validité interne à utiliser ne peut pas être trouvé. Cela explique alors d'une part que, lors d'évaluations expérimentales d'un algorithme de cns, les mesures externes soient privilégiées si l'on dispose des informations nécessaires à leur mise en œuvre, et d'autre part que lorsqu'on est confronté à un problème d'évaluation de validité de cns sans posséder ces informations on privilégie souvent les modes d'évaluation relatifs qui impliquent un effort calculatoire moindre.

4.1.3 Modes d'Evaluation Relatifs

Les deux approches précédentes sont souvent basées sur des tests statistiques pouvant nécessiter un effort calculatoire important. L'approche évaluation de la validité par critères relatifs est différente et s'appuie sur le principe "choisir la meilleure cns parmi un ensemble de cns selon un critère prédéfini".

Plus précisément, le problème peut être posé de la manière suivante : "Soit P_{alg} l'ensemble des paramètres associés à un algorithme particulier (le nombre final de classes par exemple); parmi un ensemble de cns $\{P_i, i = 1..k\}$ obtenues par l'intermédiaire d'un même algorithme mais avec différents paramètres de P_{alg}, choisir celui traduisant le mieux la structure des données."

Ce problème possède diverses déclinaisons selon la constitution de l'ensemble P_{alg}, chacune de ces déclinaisons possédant une solution pratique propre :

- **Cas 1** : Le nombre final de classes, nc, n'est pas contenu dans P_{alg}
- **Cas 2** : Le nombre final de classes, nc, est contenu dans P_{alg}.

4.1.3.1 Cas 1 : Le nombre final de classes, nc, n'est pas contenu dans P_{alg}.

Dans ce cas, la détermination des valeurs optimales pour les paramètres se déroule comme suit :

- l'utilisateur lance l'algorithme pour une large gamme de valeurs de paramètres
- la plage de valeurs la plus large pour laquelle le nombre final de classes reste constant est ensuite sélectionnée.
- les valeurs correspondant au centre de cette plage sont alors choisies comme paramètres valides.

Cette procédure permet donc de déterminer les paramètres de l'algorithme de cns et le nombre de classes de la cns correspondant au mieux à la structure interne des données.

4.1.3.2 Cas 2 : Le nombre final de classes, nc, est contenu dans P_{alg}.

Dans ce cas, la procédure d'identification de la meilleure cns implique l'utilisation d'un indice de validité ; la détermination des valeurs optimales pour les paramètres se déroule quant à elle comme suit :

- l'utilisateur spécifie dans un premier temps un intervalle de valeurs $[nc_{min}; nc_{max}]$ qui doit comprendre le bon nombre final de classes nc
- pour chaque valeur de $nc \in [nc_{min}; nc_{max}]$l'algorithme est alors lancé x fois avec des valeurs distinctes pour les autres paramètres de P_{alg} (par exemple, des conditions initiales différentes dans le cas de méthodes de type K-means). A chaque jeu de paramètres correspond alors une cns et une valeur pour l'indice de validité employé.
- les meilleures valeurs de l'indice de validité obtenues pour chaque nombre final de classes sont alors représentées graphiquement.

L'étude de ce graphique permet de choisir la meilleure cns. Dans la mesure où certains indices ne sont pas indépendants du nombre de classes de la cns (certains indices, tels ceux basés uniquement sur l'inertie interne des classes par exemple, ont tendance à croître ou décroître pour des nombres de classes croissants), on ne se contente pas de rechercher la cns possédant la plus faible (ou plus forte) valeur pour l'indice :

- Si l'indice ne montre pas de tendance croissante (ou décroissante) on se contente de rechercher la valeur la plus forte (la plus faible) pour l'indice qui correspond dans ce cas à la cns la plus en adéquation avec la structure sous jacente aux données.
- Si, par contre, l'indice exhibe une tendance croissante (ou décroissante) on recherche alors la valeur correspondant au changement local le plus marquant : il s'agit ici de rechercher un "coude" dans le graphique. Cette valeur correspond alors à la cns la plus en adéquation avec la structure sous jacente aux données. L'absence d'un tel changement peut ici être considérée comme le signe d'une absence de structure interne dans les données.

4.1.3.3 Indices

Les indices que l'on peut utiliser dans le cadre de cns non floue[5] sont :

- *Statistique* Γ *de Hubert Modifiée et* Γ *de Hubert normalisée* $\overline{\Gamma}$

5. nous n'introduisons pas les indices employer dans le cadre de la cns floue, et invitons le lecteur intéresser à ce référer au références données au début de ce papier pour un approfondissement ou encore à [XB91], [BEF84], ...

– *Les Mesures Dunn et Dunn apparentées* [Dun74] L'indice de Dunn [Dun74] tente de permettre l'identification de classes homogènes et bien séparées. Cet indice se définit pour un nombre fixé de classe comme suit

$$D_{nc} = min_{i=1..nc}\left(min_{j=i+1..nc}\left(\frac{d(c_i,c_j)}{max_{k=1..nc}diam(c_k)}\right)\right)$$

avec $d(c_i,c_j)$ la distance entre les classes c_i et c_j : $d(c_i,c_j) = min_{x\in c_i, y\in c_j} d(x,y)$ et $diam(c_k)$ le diamètre de la classe c_k : $diam(c_k) = max_{x,y\in c_k} d(x,y)$. $d(c_i,c_j)$ peut être considérée comme une mesure de la séparation de deux classes, et $diam(c_k)$ une mesure de l'hétérogénéité d'une classe. Ainsi, s'il existe des classes homogènes et bien séparées l'indice doit présenter une forte valeur. Notons tout d'abord que cet indice présente l'avantage de ne pas présenter de tendance (croissante ou décroissante) en rapport avec nc, il peut donc être utilisé pour déterminer le nombre de classes de la cns la plus en adéquation avec les données. Par contre, le coût calculatoire qui lui est associé est relativement important et il présente une forte sensibilité au bruit (qui peut impliquer un accroissement des valeurs de $diam(c_k)$). Trois indices proposés dans [PB97] constituent des adaptations plus robustes de cet indice. Ces trois indices utilisent des graphes de voisinage : l'arbre recouvrant minimal (arm), le graphe des voisins relatifs (gvr), le graphe de Gabriel. L'introduction des graphes de voisinage permet de redéfinir $diam(c_k)$. Si l'on considère par exemple l'adaptation de l'indice de Dunn impliquant l'arm, on associe à chaque classe c_k l'arm lui correspondant et $diam(c_k)$ est alors défini comme le poids de l'arête la plus fortement valuée et que l'on note $diam^{arm}(c_k)$.

$$D_{nc} = min_{i=1..nc}\left(min_{j=i+1..nc}\left(\frac{d(c_i,c_j)}{max_{k=1..nc}diam^{arm}(c_k)}\right)\right)$$

Ainsi, la robustesse de l'indice est accrue, toutefois le coût calculatoire associé constitue là encore un point faible de ce type d'indice. (Les adaptations de l'indice de Dunn utilisant les graphes des voisins relatifs ou de Gabriel sont basées sur une adaptation similaire de celle présentée pour l'arm).

– *Indice de Davies Bouldin*[DB79] [PB97] Afin d'introduire cet indice, on définit une mesure de similarité R_{ij} entre deux classes c_i et c_j basée sur une mesure s_i d'hétérogénéité interne d'une classe c_i et une mesure de dissimilarité d_{ij} entre deux classes c_i et c_j. Cette mesure est définie de manière telle que les propriétés suivantes soient respectées :
  - $R_{ij} \geq 0$
  - $R_{ij} = R_{ji}$
  - si $s_i = 0$ et $s_j = 0$ alors $R_{ij} = 0$
  - si $s_j > s_k$ et $dij = d_{ik}$ alors $R_{ij} > R_{ik}$
  - si $s_j = s_k$ et $dij < d_{ik}$ alors $R_{ij} > R_{ik}$

Davies et Bouldin ont proposé la mesure R_{ij} suivante : $R_{ij} = \frac{s_i+s_j}{d_{ij}}$. L'indice DB est alors défini comme suit :

$$DB_{nc} = \frac{1}{nc} \sum_{i=1..nc} R_i, \qquad R_i = max_{j=1..nc, i \neq j} R_{ij}$$

Selon cette définition l'indice DB_{nc} correspond donc à la similarité moyenne entre chaque classe et sa classe la plus similaire. Dans la mesure où l'on cherche des classes telles qu'elles soient le moins similaires possible des autres classes, on cherche donc à minimiser DB_{nc}, de plus cet indice ne présente pas de tendance en relation avec le nombre de classes. Des définitions pour la dissimilarité entre classes ainsi que pour l'hétérogénéité interne d'une classe sont proposées dans [DB79]. Enfin, [PB97] ont introduit 3 variantes de cet indice utilisant les graphes de voisinage de manière analogue à l'indice de Dunn.

- *Indice de validité SD*, dans [Hal00a] Halkidi et al. proposent un indice basé sur les concepts de dispersion moyenne des classes et de séparation totale entre classes.
- *SDbw*, récemment proposé par Halkidi et al. [HV01], cet indice exploite les caractéristiques inhérentes aux classes d'une cns pour en estimer la validité et permettre la détermination du partitionnement optimal des données. Tout comme l'indice SD, cet indice est basé sur les concepts de dispersion moyenne des classes et de séparation totale entre classes et introduit la notion supplémentaire de densité. Plus récemment encore, [KL03] proposent $SDbw^{\star}$ une évolution de cet indice.
- *RMSSTD, SPR, RS, CD*[Sha96] Ces 4 indices nécessitent une utilisation simultanée, et doivent être appliqués à chaque étape d'un processus de cns hiérarchique. Ils sont définis de la manière suivante :
  - *Root Mean Square STandard Deviation (RMSSTD)* (racine carrée de la moyenne des carrés des écarts types) de la cns : cet indice mesure l'homogénéité de la cns associée à une étape de la cns hiérarchique par l'intermédiaire de la moyenne de la variance de chaque variable au sein de chaque classe de la cns. Cette mesure se doit donc d'être la plus faible possible afin de montrer une forte homogénéité des classes de la cns. Si l'on observe, lors d'un passage d'une cns vers la cns suivante possédant plus de classes, un accroissement de la valeur de la mesure cela signifie alors un problème d'homogénéité dans cette dernière cns.
  - *Semi-Partial R squared (SPR)* : cet indice mesure la différence d'homogénéité locale sur deux cns successives de la cns hiérarchique : il s'agit de mesurer la différence d'homogénéité entre une classe d'une cns (ca) et celle des deux classes de la cns précédente qui ont été fusionnée afin de "créer" ca. Ainsi, une valeur faible indiquera qu'il y a eu fusion de deux classes relativement homogènes

alors qu'une valeur élevée signifie que les deux classes fusionnées ne sont pas homogènes.

- *R Squared (RS)* mesure l'hétérogénéité entre classes, ses valeurs sont comprises entre 0 et 1 : une valeur proche de 0 indiquant une faible hétérogénéité entre classes alors qu'une valeur proche de 1 signifie une hétérogénéité significative entre classes.
- *Distance between two Clusters (CD)* (Distance entre deux classes), cet indice mesure la distance entre les classes qui sont fusionnées à un niveau donnée de la cns hiérarchique.

L'utilisation simultanée de ces 4 indices permet de déterminer le nombre de classes le plus approprié pour une cns d'un jeu de données particulier. Pour cela on s'appuie sur une étude graphique des valeurs de ces différents indices et l'on recherche le "coude" le plus important pour les courbes associées à ces indices. Ces indices, et plus particulièrement les indices *RMSSTD* et *RS*, peuvent être utilisés pour des cns non hiérarchiques. L'idée étant ici de lancer l'algorithme pour des nombres de classes différents puis de procéder à une étude graphique similaire.

L'évaluation de la validité de cns par mode d'évaluation relatif correspond ainsi à l'utilisation d'indices de type mesure interne associée à une analyse graphique des valeurs de l'indice utilisé pour différentes cns.

4.1.4 Autres Modes d'Evaluation

D'autres modes d'évaluation sont proposés dans la littérature tels celui proposé par Smyth [Smy96] et les approches d'évaluation de la stabilité des algorithmes de cns :

- Smyth [Smy96] introduit en effet un algorithme de cns basé sur la cross-validation ainsi que la méthode de Monte-Carlo. Plus précisément, cet algorithme consiste en γ cross validations sur γ échantillons d'apprentissage/test du jeu de données. Pour chaque échantillon du jeu de données ech, l'algorithme EM est utilisé pour déterminer une cns en nc classes de la partie d'apprentissage de l'échantillon, cette étape est réitérée nc_{max} fois pour des valeurs de nc allant de 1 à nc_{max}. La log-vraisemblance $L^u_{nc}(D)$ est ensuite calculée pour chaque cns à nc classes, elle est formellement définie en utilisant la fonction de densité de probabilité des données : $Lk(D) = \sum_{i=1..n} log(f_k(x_i|\varphi_k))$; avec f_k la fonction de densité de probabilité des données et φ_k l'ensemble des paramètres estimés à partir des données. Cela est ainsi répété γ fois, puis on calcule la moyenne des γ estimations réalisées par cross-validation pour chaque valeur de nc. Sur la base de ces estimations, il est alors possible de déterminer les probabilités associées à chaque valeur nc : $P(nc|D)$, l'étude de ces probabilités permettant de mettre en évidence le nombre de classes correspondant au mieux au jeu de données (s'il existe). Cette approche est basée sur des concepts probabilistes afin d'estimer le nombre de classes correspondant

au mieux aux données mais n'utilise pas des concepts plus directement liés aux données tels que la séparation entre classes ou l'homogénéité interne des classes.

– Les approches utilisées pour l'évaluation de la stabilité des algorithmes de cns peuvent également être utilisées pour évaluer dans quelle mesure une modification de l'ensemble des objets sur lequel est réalisée la cns implique une modification de la partition résultat. Des méthodes d'échantillonnage et de comparaison des partitions obtenues sont disponibles [LD01] pour ce type d'évaluation. L'idée est donc ici d'évaluer la stabilité de l'algorithme de cns pour des nombres de classes différents et de choisir le nombre de classes impliquant la plus forte stabilité.

Notons que ces approches permettent essentiellement la détermination du nombre de classes le plus valide pour une cns d'un jeu de données spécifique et qu'elles ne permettent donc pas vraiment la comparaison de validité de deux partitions. De plus, il faut également noter la dépendance de ces approches avec les méthodes de cns employées.

4.2 Nouveaux Indices et Nouvelle Méthodologie pour l'Evaluation et la Comparaison de la Validité de Classifications Non Supervisées

Nous proposons maintenant deux nouveaux indices de type critère interne (qui ne nécessitent donc aucune connaissance ou intuition a priori sur les données) utilisés au sein d'une méthodologie particulière de type évaluation relative pour l'évaluation et la comparaison de la validité de cns. Cette méthodologie bien qu'utilisant des caractérisations statistiques n'implique pas l'utilisation de la méthode de Monte Carlo et son coût calculatoire est relativement faible. Elle ne nécessite qu'une seule passe sur le jeu de données, et, son coût calculatoire est linéaire selon le nombre de variables du jeu de données et soit linéaire selon le nombre d'objets dans le cas de données catégorielles, soit quadratique selon le nombre d'objets pour des données quantitatives. Notons enfin que, l'encombrement mémoire associé à cette méthodologie est très faible quel que soit le type de données traité (nécessite le stockage d'un ensemble de tables de contingences). Ces indices peuvent être utilisés pour tous types de données mais sont toutefois spécialement adaptés au traitement de données catégorielles. Enfin, l'utilisation de cette méthodologie permet une représentation graphique de la validité des cns qui est particulièrement intéressante dans la mesure où par un simple artifice de visualisation, il est possible de dériver des connaissances additionnelles concernant la structure des données.

4.2.1 Concepts et Formalismes Introductifs

Notation 1
Nous rappelons ici les notations utilisées :
$O = \{o_i, i = 1..n\}$ *un ensemble d'objets*
$EV = \{V_1,..,V_p\}$ *l'espace, constitué de* p *variables décrivant les objets de* O.
$o_i = \{o_{i_1},..,o_{i_p}\}$ *un objet de* O, o_{i_j} *correspond à la valeur de* o_i *pour la variable* V_j *(cette valeur peut être numérique, catégorielle...)*
C_k *un ensemble d'objets de* O ($C_k \subseteq O$),
$P_h = \{C_1,...,C_z\}$ *une partition de* O *en* z *classes*

De manière classique, l'évaluation de la validité d'une cns s'appuie sur l'étude d'un critère traduisant l'homogénéité interne de ses classes ou la séparation de ses classes. Certains indices tirent toutefois partie d'une évaluation combinée de ces deux notions au sein d'un unique indice (les indices SD, $SDbw$, $SDbw^\star$ par exemple), d'autres indices s'utilisent quant à eux au travers d'évaluations graphiques simultanées mais séparées (les indices *RMSSTD, SPR, RS, CD* par exemple). La méthodologie que nous proposons prend, elle aussi, en compte ces deux notions, mais plutôt que d'introduire un critère combinant ces notions, nous proposons d'utiliser simultanément deux critères rendant compte pour l'un de l'hétérogénéité interne des classes et de la séparation entre classes pour l'autre et ce au travers d'une unique représentation graphique et non plusieurs.

Nous utilisons en fait, deux mesures traduisant respectivement l'homogénéité interne des classes et la séparation entre classes. Ces mesures sont en relation directe avec le Nouveau Critère de Condorcet (*NCC*)[6].

Dans le cadre des données catégorielles[7] la notion de similarité entre objets est utilisée ; afin de proposer une méthodologie utilisable pour des données quantitatives, nous lui substituons une extension de cette notion. Cette extension, nommée lien (selon une variable), se définit comme suit :

Définition 9 ***Lien entre 2 objets***
A chaque variable V_i *est associée une fonction* $Lien_i$ *qui définit un lien (une sorte de similarité) ou un non-lien (une sorte de dissimilarité) selon* V_i *entre deux objets de* O *(*o_a *et* o_b*) :*

$$Lien_i(o_{a_i},o_{b_i}) = \begin{cases} 1 \textit{ si une condition particulière déterminant un lien} \\ \textit{(selon } V_i\textit{) entre les objets } o_a \textit{ et } o_b \textit{ est vérifiée} \\ 0 \textit{ sinon (non-lien)} \end{cases} \quad (4.1)$$

6. mesure de qualité d'une cns proposée par Michaud [Mic97] et utilisée pour la cns pour données catégorielle [Mic97], [JN03c], voir chapitre 2 pour des développements plus complets

7. qui constitue le cadre naturel pour la définition du NCC

EXEMPLES :

- *Pour une variable catégorielle V_i, on peut définir naturellement $Lien_i$ comme suit :*
  $Lien_i(o_{a_i},o_{b_i}) = \delta_{sim}(o_{a_i},o_{b_i}) = \begin{cases} 1 \text{ si } o_{a_i} = o_{b_i} \\ 0 \text{ sinon} \end{cases}$
- *Pour une variable quantitative V_i, on peut par exemple définir $Lien_i$ comme suit :*
  $Lien_i(o_{a_i},o_{b_i}) = \begin{cases} 1 \text{ si } |o_{a_i} - o_{b_i}| \leq \delta, \text{ avec } \delta \text{ un seuil fixé par l'utilisateur} \\ 0 \text{ sinon} \end{cases}$
- *Pour une variable quantitative V_i, on peut également utiliser la discrétisation puis utiliser une fonction $Lien_i$ définie pour les variables catégorielles.*

Nous illustrons le cas des variables quantitatives sur un jeu de données composé de 4 objets décrits par deux variables quantitatives V_1 et V_2 :
$o_1 = \{1; 2\}$, $o_2 = \{1.5; 1\}$, $o_3 = \{2; 1\}$, $o_4 = \{1; 3\}$.

On considère 2 cas :
*- **le cas 1** caractérisé par une discrétisation des 2 variables selon les intervalles :* $]-\infty; 1.25[; [1.25; +\infty[$ *pour V_1 et* $[-\infty; 1.5[; [1.5; 2.5[; [2.5; +\infty[$ *pour V_2*
*- **le cas 2** se caractérise quant à lui par l'utilisation d'une fonction de type seuil pour les 2 variables : un seuil de 0.5 pour V_1 et de 1 pour V_2.*

Les tableaux 4.1 et 4.2 ainsi que la figure 4.1 illustrent les différentes valeurs pour les fonctions $Lien_1$ et $Lien_2$ pour ces deux cas.

	o_1	o_2	o_3	o_4
o_1	x	0	0	1
o_2	0	x	1	0
o_3	0	1	x	0
o_4	1	0	0	x

	o_1	o_2	o_3	o_4
o_1	x	0	0	1
o_2	0	x	1	0
o_3	0	1	x	1
o_4	1	0	1	x

TAB. 4.1 –: *Fonctions $Lien_1$ (à gauche) et $Lien_2$ (à droite) pour le cas 1*

	o_1	o_2	o_3	o_4
o_1	x	0	0	0
o_2	0	x	1	0
o_3	0	1	x	0
o_4	0	0	0	x

	o_1	o_2	o_3	o_4
o_1	x	0	0	1
o_2	0	x	1	1
o_3	0	1	x	1
o_4	1	1	1	x

TAB. 4.2 –: *Fonctions $Lien_1$ (à gauche) et $Lien_2$ (à droite) pour le cas 2*

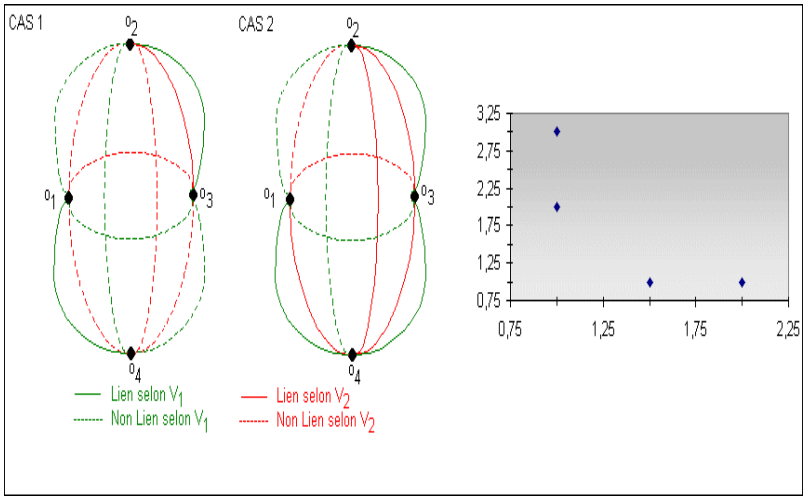

FIG. 4.1 –: *Représentation graphique des 4 objets dans l'espace 2D (à droite) et Représentation des liens/non-liens unissant les objets dans les 2 cas illustratifs (à gauche)*

4.2.1.1 Evaluation de l'homogénéité interne des classes d'une cns

Pour évaluer l'homogénéité interne d'une cns (une partition P_h de O), on peut utiliser la mesure LM (resp. NLM) qui dénombre le nombre de liens (resp. non-liens) entre objets de même classe de la cns. Ces mesures sont définies de la manière suivante :

$$LM(P_h) = \sum_{g=1..z} \left(\sum_{\substack{o_k \in C_g, o_l \in C_g, \\ k<l}} \left(\sum_{i=1..p} (lien_i\,(o_{k_i}, o_{l_i})) \right) \right) \tag{4.2}$$

$$0 \leq LM(P_h) \leq p \times \sum_{g=1..z} \frac{card(C_g)(card(C_g)-1)}{2} \tag{4.3}$$

$$NLM(P_h) = \sum_{g=1..z} \left(\sum_{\substack{o_k \in C_g, o_l \in C_g, \\ k<l}} \left(\sum_{i=1..p} (1 - lien_i\,(o_{k_i}, o_{l_i})) \right) \right) \tag{4.4}$$

$$0 \leq NLM(P_h) \leq p \times \sum_{g=1..z} \frac{card(C_g)(card(C_g)-1)}{2} \tag{4.5}$$

$$LM(P_h) + NLM(P_h) = p \times \sum_{g=1..z} \frac{card(C_g)(card(C_g)-1)}{2} \tag{4.6}$$

Ainsi, l'homogénéité interne d'une cns P_h est d'autant plus forte que $LM(P_h)$ (resp. $NLM(P_h)$) est élevée (resp. faible).

EXEMPLE : *Pour les 2 cas introduits précédemment et pour une partition* $P_h = \{\{o_1,o_4\},\{o_2,o_3\}\}$ *nous avons :* $LM(P_h) = 3$, $NLM(P_h) = 1$ *dans le cas 1 ; et* $LM(P_h) = 4$, $NLM(P_h) = 0$ *dans le cas 2.*

4.2.1.2 Evaluation de la séparation entre classes d'une cns (ou hétérogénéité entre classes)

Pour évaluer la séparation entre classes d'une cns (une partition P_h de O), on peut utiliser la mesure LD (resp. NLD) qui dénombre le nombre de liens (resp. non-liens) entre objets de classes différentes de la cns. Ces mesures sont définies de la manière suivante :

$$LD(P_h) = \sum_{\substack{f=1..z, g=1..z \\ f<g}} \left(\sum_{o_k \in C_f, o_l \in C_g,} \left(\sum_{i=1..p} (lien_i\,(o_{k_i}, o_{l_i})) \right) \right) \tag{4.7}$$

$$0 \leq LD(P_h) \leq p \times \sum_{\substack{f=1..z, g=1..z \\ f<g}} \frac{card(C_g)(card(C_f)}{2} \tag{4.8}$$

$$NLD(P_h) = \sum_{\substack{f=1..z, g=1..z \\ f<g}} \left(\sum_{o_k \in C_f, o_l \in C_g,} \left(\sum_{i=1..p} (1 - lien_i\,(o_{k_i}, o_{l_i})) \right) \right) \tag{4.9}$$

$$0 \leq NLD(P_h) \leq p \times \sum_{\substack{f=1..z, g=1..z \\ f<g}} \frac{card(C_g)(card(C_f)}{2} \tag{4.10}$$

$$LD(P_h) + NLD(P_h) = p \times p \times \sum_{\substack{f=1..z, g=1..z \\ f<g}} \frac{card(C_g)(card(C_f)}{2} \tag{4.11}$$

Ainsi, la séparation des classes d'une cns P_h est d'autant plus forte que $NLD(P_h)$ (resp. $LD(P_h)$) est élevée (resp. faible).

EXEMPLE : *Pour les 2 cas introduits précédemment et pour une partition* $P_h = \{\{o_1,o_4\},\{o_2,o_3\}\}$ *nous avons :* $LD(P_h) = 0$, $NLD(P_h) = 8$ *dans le cas 1 ; et* $LD(P_h) = 4$, $NLD(P_h) = 4$ *dans le cas 2.*

4.2.1.3 Notions Additionnelles

Nous définissons deux mesures additionnelles $M(P_h)$ et $D(P_h)$ qui correspondent respectivement :

- au nombre total de liens et non liens entre objets de même classe de P_h : $M(P_h) = NLM(P_h) + LM(P_h)$
- au nombre total de liens et non liens entre objets de classes différentes de P_h: $D(P_h) = NLD(P_h) + LD(P_h)$.

Finalement, nous notons $L(O)$ (resp. $NL(O)$) le nombre total de liens (resp. de non-liens) entre objets de O: $L(O) = LM(P_h) + LD(P_h)$ (resp. $NL(O) = NLM(P_h) + NLD(P_h)$).

REMARQUE : Pour des données catégorielles, le critère *NCC* est défini comme la somme de $NLM(P_h)$ et $LD(P_h)$.

EXEMPLE : *Pour les 2 cas introduits précédemment et pour une partition* $P_h = \{\{o_1,o_4\},\{o_2,o_3\}\}$ *nous avons :*
$M(P_h) = 4$, $D(P_h) = 8$, $L(O) = 3$, $NL(O) = 9$ *dans le cas 1 ; et* $M(P_h) = 4$, $D(P_h) = 8$, $L(O) = 8$, $NL(O) = 4$ *dans le cas 2.*

RÉSUMÉ :

$L(O) + NL(O) = \frac{n\times(n-1)}{2} \times p$

$M(P_h) + D(P_h) = \frac{n\times(n-1)}{2} \times p$

$M(P_h) = NLM(P_h) + LM(P_h)$

$D(P_h) = NLD(P_h) + LD(P_h)$

$L(O) = LM(P_h) + LD(P_h)$

$NL(O) = NLM(P_h) + NLD(P_h)$

Ces relations peuvent être synthétisées au sein d'une sorte de table de contingence de type comparaison par paires :

	liens	non liens	$Total$
même classes	LM	NLM	$M(C)$
classes diff.	LD	NLD	$D(C)$
$Total$	$L(O)$	$NL(O)$	$\frac{n(n-1)}{2}p$

4.2.1.4 Remarques importantes concernant l'aspect calculatoire

Bâtir cette table de contingence ne nécessite qu'une seule passe sur le jeu de données. Dans le cas de données catégorielles, cela ne requiert que $O(np)$ comparaisons afin de bâtir p tables de contingence (croisant les p variables avec la variable catégorielle virtuelle impliquée par la partition P_h)(intuitivement, les définitions formelles de LM, NLM, LD et NLD semblent impliquer $O(n^2p)$ comparaisons mais des astuces de calcul permettent de réduire ce nombre de comparaisons, voir l'exemple illustratif suivant), ce nombre de comparaisons peut atteindre $O(n^2p)$ en cas de présence de variables quantitatives et d'utilisation de fonctions $lien_i$ telles que définies dans le cas 2 des exemples illustratifs précédents.

Du point de vue de l'utilisation mémoire, quelle que soit la nature des données (catégorielles ou numériques), le stockage de p tables de contingence est nécessaire, ce qui correspond à un encombrement mémoire faible.

EXEMPLE :
Considérons un jeu de données synthétique composé de 4 objets ($o_1 = [y,y,n,n]$, $o_2 = [y,y,n,n]$, $o_3 = [n,y,y,y]$, $o_4 = [n,n,y,y]$) décrits par 4 variables catégorielles ($EV = \{V_1,V_2,V_3,V_4\}$) (voir table 4.3).
Considérons également la partition P_h suivante : $P_h = \{C_1,C_2\} = \{\{o_1,o_2\},\{o_3,o_4\}\}$.

Nous exposons maintenant comment calculer les valeurs des divers indices présentés dans la section précédente. (Nous utilisons ici la fonction $Lien$ telle qu'elle est définie dans l'exemple sur les données catégorielle de la définition 9).

	V_1	V_2	V_3	V_4
1	y	y	n	n
2	y	y	n	n
3	n	y	y	y
4	n	n	y	y

TAB. 4.3 –: *Jeu de données synthétique*

– *En une unique passe sur le jeu de données on peut bâtir les tables de contingence croisant la variable catégorielle virtuelle V_A impliquée par la partition P_h (V_A possède deux modalités a et b qui correspondent respectivement aux classes $\{o_1,o_2\}$ et $\{o_3,o_4\}$) avec les p variables de EV (V_1, V_2, V_3, V_4.)* :

$V_A \setminus V_1$	y	n	$V_A \setminus V_2$	y	n	$V_A \setminus V_3$	y	n	$V_A \setminus V_4$	y	n
a	2	0	a	2	0	a	0	2	a	2	0
b	0	2	b	1	1	b	2	0	b	0	2

Tables de Contingence croisant la variable virtuelle V_A et les variables de EV

– *le calcul de la valeur de chaque indice est alors réalisé à partir de ces tables. Si la table de contingence pour une variable V_i est notée:*

$V_A \backslash V_i$	V_{i_1}	...	$V_{i_{m_i}}$	
V_{A_1}	α_{1i1}	...	α_{1im_i}	$\alpha_{1i.}$
...	...	...	...	...
V_{A_z}	α_{zi1}	...	α_{zim_i}	$\alpha_{zi.}$
	$\alpha_{.i_1}$	...	$\alpha_{.im_i}$	n

V_A la variable catégorielle virtuelle à z modalités (associée à P_h),
V_i une variable exogène à m_i modalités notées $V_{i_j}(j = 1..m_i)$.
α_{li_h} le nombre d'objets ayant la valeur V_{i_h} pour V_i et la valeur V_{A_l} pour V_A.
$\alpha_{.i_j} = \sum_{h=1..z} \alpha_{hi_j}$; $\alpha_{hi.} = \sum_{j=1..m_i} \alpha_{hi_j}$

– *on peut alors calculer :*

$$LM(P_h) = \sum_{\substack{i = 1..p \\ \text{tel que } V_i \in EV}} \sum_{j=1..m_i} \sum_{t=1..z} \frac{\alpha_{ti_j}(\alpha_{ti_j}-1)}{2}$$

$$M(P_h) = card(EV) \times \sum_{t=1..z} \frac{card(C_t)(card(C_t)-1)}{2}$$

$$L(O) = \sum_{\substack{i=1..p \text{ tel que} \\ V_i \in EV}} \sum_{j=1..m_i} \frac{\alpha_{.i_j}(\alpha_{.i_j}-1)}{2}$$

$$NLM(P_h) = M(P_h) - LM(P_h)\,;\, LD(P_h) = L(O) - LM(P_h)$$

$$D(P_h) = \frac{n(n-1)}{2} \times card(EV) - M(P_h)\,;\, NLD(P_h) = D(P_h) - LD(P_h)$$

– *Pour l'exemple cela donne :*
$LM(P_h) = \left(\frac{2\times1}{2} + \frac{0\times(-1)}{2} + \frac{0\times(-1)}{2} + \frac{2\times1}{2}\right) + \left(\frac{2\times1}{2} + \frac{1\times0}{2} + \frac{0\times(-1)}{2} + \frac{1\times0}{2}\right) +$
$\left(\frac{0\times(-1)}{2} + \frac{2\times1)}{2} + \frac{2\times1}{2} + \frac{0\times-1}{2}\right) + \left(\frac{2\times1}{2} + \frac{0\times(-1)}{2} + \frac{0\times(-1)}{2} + \frac{2\times1}{2}\right) = 7$

$$M(P_h) = 4 \times \left(\frac{2\times1}{2} + \frac{2\times1}{2}\right) = 8$$

$$L(O) = \left(\frac{2\times1}{2} + \frac{2\times1}{2}\right) + \left(\frac{3\times2}{2} + \frac{1\times0}{2}\right) + \left(\frac{2\times1}{2} + \frac{2\times1}{2}\right) + \left(\frac{2\times1}{2} + \frac{2\times1}{2}\right) = 9$$

$$NLM(P_h) = 8 - 7 = 1 \quad ; \quad LD(P_h) = 9 - 7 = 2$$

$$D(P_h) = \frac{4\times3}{2} \times 4 - 8 = 16 \quad ; \quad NLD(P_h) = 16 - 2 = 14$$

4.2.2 La nouvelle méthodologie pour l'évaluation et la comparaison de validité de cns

Il apparaît intuitivement, qu'une cns valide (P_h) doit être telle que les objets de même classe sont majoritairement reliés par des liens et que les objets de classes différentes sont majoritairement reliés par des non-liens, ainsi une cns valide doit présenter de fortes valeurs pour $LM(P_h)$ et $NLD(P_h)$ ce qui implique de faibles valeurs pour $NLM(P_h)$ et $LD(P_h)$ (cela signifie alors une forte homogénéité interne des classes ainsi qu'une forte séparation des classes de la cns).

Cependant, la signification de fortes et faibles valeurs n'est elle pas totalement intuitive. Nous devons de surcroît remarquer que ces valeurs ne sont pas indépendantes et que si elles sont fortement corrélées entre elles, elles sont aussi corrélées avec le nombre de classes de la cns. En outre, très peu de jeux de données possèdent une structure interne telle que $LM(P_h)$ et $NLD(P_h)$ sont simultanément maximisés pour une partition P_h. Nous décrivons donc maintenant une méthodologie pour l'évaluation et la comparaison de la validité de cns (d'un même jeu de données) par analyse de leurs valeurs respectives pour LM et NLD dans le but de déterminer l'unique ou le sous ensemble de cns les plus valides. Cette méthodologie constitue un outil utile pour les utilisateurs qui recherchent la meilleure, ou tout au moins, une bonne cns pour un

jeu de données. Cette méthodologie est indépendante des méthodes utilisées, des paramètres des méthodes, ainsi que du nombre de classes. Pour atteindre cet objectif, nous utilisons une approche statistique de manière à déterminer dans quelle mesure des valeurs LM et NLD exhibées par une cns peuvent être considérées comme significativement élevées.

4.2.2.1 Caractérisation statistique des valeurs de : *LM* et *NLD*

Faisons l'hypothèse H_0 d'une organisation aléatoire de l'ensemble d'objets O selon une partition P_h en z classes. Nous pouvons déterminer la loi statistique suivie par LM et NLD sous cette hypothèse :

- $LM(P_h)$ suit une loi binomiale de paramètres $M(P_h)$ et $\frac{L(O)}{L(O)+NL(O)}$, ce que nous notons :
  $LM(P_h) \hookrightarrow B(M(P_h), \frac{L(O)}{L(O)+NL(O)})$;
- $NLD(P_h)$ suit une loi binomiale de paramètres $D(P_h)$ et $\frac{NL(O)}{L(O)+NL(O)}$, ce que nous notons :
  $NLD(P_h) \hookrightarrow B(D(P_h), \frac{NL(O)}{L(O)+NL(O)})$.

Par approximation avec la loi normale, on obtient :

- $LM(P_h)$ suit une loi normale
  de moyenne : $E_1 = M(P_h) \times \frac{L(O)}{L(O)+NL(O)}$,
  et d'écart type : $SD_1 = \sqrt{M(P_h) \times \frac{L(O)}{L(O)+NL(O)} \times (1 - \frac{L(O)}{L(O)+NL(O)})}$,
  Nous notons : $LM(P_h) \hookrightarrow N(E_1, SD_1)$;

- $NLD(P_h)$ suit une loi normale
  de moyenne : $E_2 = D(P_h) \times \frac{NL(O)}{L(O)+NL(O)}$,
  et d'écart type : $SD_2 = \sqrt{D(P_h) \times \frac{NL(O)}{L(O)+NL(O)} \times (1 - \frac{NL(O)}{L(O)+NL(O)})}$,
  Nous notons : $NLD(P_h) \hookrightarrow N(E_2, SD_2)$.

Dès lors, par centrage-réduction on obtient deux indices $xv_1(P_h)$ et $xv_2(P_h)$ qui suivent une loi normale centrée réduite :

- $xv_1(P_h) = \frac{LM(P_h) - E_1}{SD1} = \frac{LM(P_h) - M(P_h) \times \frac{L(O)}{L(O)+NL(O)}}{\sqrt{M(P_h) \times \frac{L(O)}{L(O)+NL(O)} \times (1 - \frac{L(O)}{L(O)+NL(O)})}}$
  $xv_1(P_h) \hookrightarrow N(0,1)$
- $xv_2(P_h) = \frac{NLD(P_h) - E_2}{SD2} = \frac{NLD(P_h) - D(P_h) \times \frac{NL(O)}{L(O)+NL(O)}}{\sqrt{D(P_h) \times \frac{NL(O)}{L(O)+NL(O)} \times (1 - \frac{NL(O)}{L(O)+NL(O)})}}$
  $xv_2(P_h) \hookrightarrow N(0,1)$

Si on considère H_1 l'hypothèse alternative à H_0 définie ainsi : "L'ensemble d'objets O est organisé de manière non aléatoire selon une partition P_h en z

classes telles que cette partition représente une cns valide", alors $LM(P_h)$ et $NLD(P_h)$ doivent simultanément exhiber des valeurs exceptionnellement élevées.

Nous pouvons donc maintenant bâtir un test statistique unilatéral droite pour $xv_1(P_h)$ et $xv_2(P_h)$ (nous notons ces tests T_1 et T_2). Nous considérons par la suite qu'une partition constitue une cns valide si et seulement si pour chaque test (T_1 et T_2) l'hypothèse H_1 est acceptée.

Définition 10 ***Classification non supervisée valide***
Une partition P_h est considérée comme valide avec un couple de risques du premier type (α_1,α_2) ssi :

$$xv_1(P_h) = \frac{LM(P_h) - M(P_h) \times \frac{L(O)}{L(O)+NL(O)}}{\sqrt{M(P_h) \times \frac{L(O)}{L(O)+NL(O)} \times (1 - \frac{L(O)}{L(O)+NL(O)})}}$$

$$pv_1(P_h) = 1 - \int_{-\infty}^{xv_1(P_h)} \frac{1}{\sqrt{2\Pi}} e^{\frac{-t^2}{2}} dt \quad (4.12)$$

$$pv_1(P_h) \leq \alpha_1 \Leftrightarrow xv_1(P_h) \geq F^{-1}(1-\alpha_1) \quad (4.13)$$

ET

$$xv_2(P_h) = \frac{NLD(P_h) - D(P_h) \times \frac{NL(O)}{L(O)+NL(O)}}{\sqrt{D(P_h) \times \frac{NL(O)}{L(O)+NL(O)} \times (1 - \frac{NL(O)}{L(O)+NL(O)})}}$$

$$pv_2(P_h) = 1 - \int_{-\infty}^{xv_2(P_h)} \frac{1}{\sqrt{2\Pi}} e^{\frac{-t^2}{2}} dt \quad (4.14)$$

$$pv_2(P_h) \leq \alpha_2 \Leftrightarrow xv_2(P_h) \geq F^{-1}(1-\alpha_2) \quad (4.15)$$

avec F la fonction de distribution de probabilité cumulée inverse de la loi normale centrée réduite

4.2.2.2 Méthodologie

Nous venons de présenter deux tests permettant de décider si oui ou non une partition constitue une cns valide. Cependant, il est clair que pour de nombreux jeux de données plusieurs partitions de l'ensemble des objets (O) de ces jeux de données peuvent être considérées comme des cns valides. Admettons donc que l'on dispose d'un ensemble de cns valides noté ECV. Le problème est alors de déterminer laquelle de ces cns est la cns la plus valide.

La méthodologie que nous proposons permet de résoudre ce problème dans certains cas, et permet toujours de déterminer un sous ensemble ECM de ECV ($ECM \subseteq ECV$) qui inclue toutes les cns candidates au "titre de cns la plus valide". Cette dernière situation correspond au cas pour lequel nous ne sommes pas capables de décider quelle est la cns de ECM la plus valide (si

toutefois il existe une cns plus valide que les autres) mais cependant capables de déterminer un sous ensemble de cns les plus valides. Notre méthodologie permet de plus la visualisation d'un ensemble d'informations concernant la structure de l'ensemble objets, ce qui constitue un outil utile pour l'utilisateur qui doit choisir une cns de ECM (si ECM inclue plus d'une cns).

Nous exposons cette méthodologie sous forme algorithmique (voir algorithme 3 en page 79).

REMARQUE :
Le point **4.** de l'algorithme de la méthodologie (l'extraction de ECM) peut également être réalisé par comparaison directe des couples de valeurs ($xv_1(P_i)$,$xv_2(P_i)$) (cela étant du à la relation de monotonie unissant à la fois $xv_1(P_i)$ et $pv_1(P_i)$ ainsi que $xv_2(P_i)$ et $pv_2(P_i)$). Cette comparaison mène aussi à 4 situations différentes :

- P_j est considérée comme plus valide que P_i ssi
  $(xv_1(P_i) < xv_1(P_j)$ et $xv_2(P_i) < xv_2(P_j))$ ou $(xv_1(P_i) \leq xv_1(P_j)$ et $xv_2(P_i) < xv_2(P_j))$ ou $(xv_1(P_i) < xv_1(P_j)$ et $xv_2(P_i) \leq xv_2(P_j))$
  nous notons cette relation : $P_j < b > P_i$
- P_i est considérée comme plus valide que P_j ssi
  $(xv_1(P_j) < xv_1(P_i)$ et $xv_2(P_j) < xv_2(P_i))$ ou $(xv_1(P_j) \leq xv_1(P_i)$ et $xv_2(P_j) < xv_2(P_i))$ ou $(xv_1(P_j) < xv_1(P_i)$ et $xv_2(P_j) \leq xv_2(P_i))$
  nous notons cette relation : $P_i < b > P_j$
- P_j et P_i sont considérées comme équivalentes du point de vue de la validité ssi $(xv_1(P_j) = xv_1(P_i)$ et $xv_2(P_j) = xv_2(P_i))$
  nous notons cette relation : $P_i < s > P_j$
- P_j et P_i sont considérées comme incomparables du point de vue de la validité ssi $(xv_1(P_j) < xv_1(P_i)$ et $xv_2(P_j) > xv_2(P_i))$ ou $(xv_1(P_j) > xv_1(P_i)$ et $xv_2(P_j) < xv_2(P_i))$
  nous notons cette relation : $P_i <? > P_j$

ECM est alors défini par : $ECM = \bigcup\{P_i \; telle \; que \; \nexists j \in 1..q, P_j < b > P_i\}$

Algorithme 3 Méthodologie pour la détermination de la cns la plus valide (ou de l'ensemble) des cns les plus valides

Données : $EP = \{P_1,..,P_q\}$ un ensemble de q partitions de O

1.Déterminer pour chaque partition P_i les valeurs de leurs statistiques $xv_1(P_i)$ et $xv_2(P_i)$ ainsi que les valeurs $pv_1(P_i)$ et $pv_2(P_i)$ obtenues respectivement pour les tests T_1 et T_2. Ainsi, chaque P_i ($i \in 1..q$) est caractérisée par deux couples de valeurs ($pv_1(P_i)$,$pv_2(P_i)$) et ($xv_1(P_i)$,$xv_2(P_i)$).

2. Fixer α_1 (resp. α_2) seuil sur le risque de rejeter à tort l'hypothèse H_0 (et d'accepter à tort H_1) pour les tests T_1 (resp. T_2) (ces valeurs sont fixées arbitrairement par l'utilisateur, les valeurs typiques sont 0.05, 0.025, 0.01...).

3. Créer l'ensemble de partitions ECV qui correspond à l'ensemble des cns valides (selon T_1 et T_2). ECV est formellement définit de la façon suivante :

$$ECV = \bigcup_{P_i \in EP} (P_i \; telle \; que \; pv_1(P_i) \leq \alpha_1 \; et \; pv_2(P_i) \leq \alpha_2)$$

4. Comparer la validité de deux cns P_i ($i \in 1..q$), P_j ($j \in 1..q, i \neq j$) revient à comparer leur couple de valeurs ($pv_1(P_i)$,$pv_2(P_i)$) and (pv_{1_j},$pv_2(P_j)$).
Cette comparaison mène à 4 situations differentes :

- P_i est considérée plus valide que P_j ssi
  ($pv_1(P_i) < pv_{1_j}$ et $pv_2(P_i) < pv_2(P_j)$) ou ($pv_1(P_i) \leq pv_{1_j}$ et $pv_2(P_i) < pv_2(P_j)$) ou ($pv_1(P_i) < pv_{1_j}$) et ($pv_2(P_i) \leq pv_2(P_j)$),
  nous notons cette relation : $P_i < b > P_j$
- P_j est considérée plus valide que P_i ssi
  ($pv_{1_j} < pv_1(P_i)$ et $pv_2(P_j) < pv_2(P_i)$) ou ($pv_{1_j} \leq pv_1(P_i)$ et $pv_2(P_j) < pv_2(P_i)$) ou ($pv_{1_j} < pv_1(P_i)$ et $pv_2(P_j) \leq pv_2(P_i)$),
  nous notons cette relation : $P_j < b > P_i$
- P_j et P_i sont considérées comme équivalentes du point de vue de la validité ssi $pv_{1_j} = pv_1(P_i)$ et $pv_2(P_j) = pv_2(P_i)$,
  nous notons cette relation: $P_i < s > P_j$
- P_j et P_i sont considérées comme incomparables du point de vue de la validité ssi ($pv_{1_j} < pv_1(P_i)$ et $pv_2(P_j) > pv_2(P_i)$) ou ($pv_{1_j} > pv_1(P_i)$ et $pv_2(P_j) < pv_2(P_i)$),
  nous notons cette relation : $P_i <? > P_j$

5. Extraire l'ensemble ECM des cns les plus valides, qui est formellement défini comme suit : $ECM = \bigcup\{P_i \; telle \; que \; \nexists j \in 1..q, P_j < b > P_i\}$

EXEMPLE : *Nous illustrons la méthodologie sur un exemple synthétique*[8]*. Considérons le jeu de données synthétique introduit préalablement en page 74 (voir tableau 4.3).*

Le tableau 4.4 présente les caractéristiques de chaque partition de l'ensemble des partitions possibles des objets du jeu de données (à l'exception de la partition grossière en une classe et de la partition la plus fine qui possède autant de classes que d'objets du jeu de données).

Si nous fixons $\alpha_1 = \alpha_2 = 0.15$ *nous obtenons alors* $ECV = \{P_6, P_{14}\}$*. La comparaison de la validité de ces 2 cns donne* $ECM = \{P_6\}$*. Remarquons que les partitions* P_6 *et* P_{14} *apparaissent clairement comme les seules cns valides et que nous pouvons aussi considérer que* P_6 *correspond à la cns la plus valide. Nous devons aussi noter que toutes les étapes de la méthodologie peuvent être résumées graphiquement comme le montre les figures 4.2 et 4.3 ; la figure 4.2 (resp. 4.3) correspond au couple* $(1 - pv_1(P_i), 1 - pv_2(P_i))$ *(resp.*$(xv_1(P_i), xv_2(P_i))$*), les lignes formées de tirets de la figure 4.2 correspondent à* $pv_1(P_i) = 1 - \alpha_1 = 0.85$ *et* $pv_2(P_i) = 1 - \alpha_2 = 0.85$*, elles délimitent la zone incluant des cns valides.*

	P_h	M	D	LM	NLD	LD	NLM	xv_1	xv_2	pv_1	pv_2
P_2	{2,3, 4},{1}	12	12	4	7	5	8	-0,298	-0,298	0,383	0,383
P_3	{1,3, 4},{2}	12	12	4	7	5	8	-0,298	-0,298	0,383	0,383
P_4	{1,2, 4},{3}	12	12	4	7	5	8	-0,298	-0,298	0,383	0,383
P_5	{1,2, 3},{4}	12	12	6	9	3	6	0,894	0,894	0,814	0,814
P_6	{1,2}, {3,4}	8	16	7	14	2	1	2,92	2,066	0,998	0,981
P_7	{1,3}, {2,4}	8	16	1	8	8	7	-1,461	-1,033	0,072	0,151
P_8	{1,4}, {2,3}	8	16	1	8	8	7	-1,461	-1,033	0,072	0,151
P_9	{1},{2},{3,4}	4	20	3	14	6	1	1,549	0,693	0,939	0,756
P_{10}	{1},{3},{2,4}	4	20	0	11	9	4	-1,549	-0,693	0,061	0,244
P_{11}	{1},{4},{2,3}	4	20	1	12	8	3	-0,516	-0,231	0,303	0,409
P_{12}	{2},{3},{1,4}	4	20	0	11	9	4	-1,549	-0,693	0,061	0,244
P_{13}	{2},{4},{1,3}	4	20	1	12	8	3	-0,516	-0,231	0,303	0,409
P_{14}	{3},{4},{1,2}	4	20	4	15	5	0	2,582	1,155	0,995	0,876

Légende : $M : M(P_h)$, $D : D(P_h)$, $LM : LM(P_h)$, $NLD : NLD(P_h)$, $NLM : NLM(P_h)$, $LD : LD(P_h)$

TAB. 4.4 –: *Partitions du jeu de données synthétique*

REMARQUES :

- La présentation graphique des résultats les résume parfaitement, et permet une bonne et rapide comparaison de la validité des cns.
- La présentation graphique des résultats permet également une visualisation du type de structure du jeu de données : cela indique quels nombres de classes doivent être considérés comme trop faibles et quels nombres

8. Attention, cet exemple vise essentiellement à illustrer les différentes étapes de la méthodologie, en effet, étant donné le faible nombre d'individus du jeu de données l'approximation normale est ici douteuse...

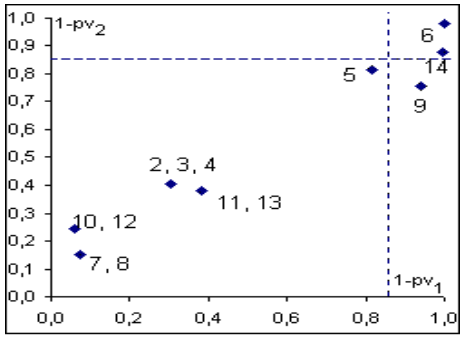

FIG. 4.2 –: *Couples* $(1 - pv_1(P_i)$,$1 - pv_2(P_i))$ *pour chaque partition*

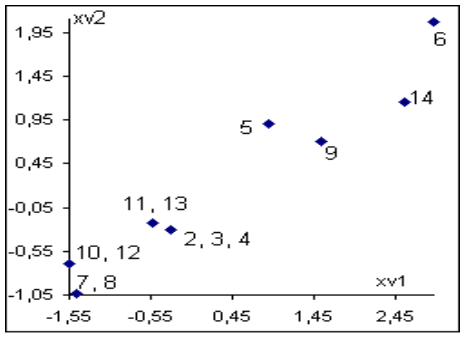

FIG. 4.3 –: *Couples* $(xv_1(P_i)$,$xv_2(P_i))$ *pour chaque partition*

de classes doivent être considérés comme trop élevés. (Cela apporte également d'autres informations que nous aborderons plus tard.)

– Nous utiliserons par la suite la représentation graphique employant les couples $(xv_1(P_i)$,$xv_2(P_i))$ (car ces deux graphiques représentent la même information et nous trouvons ce dernier type de graphique plus clair).

Avant toute autre expérimentation, nous pouvons lister certains avantages de cette méthodologie :

– présentation intuitive et graphique des résultats ;
– évaluation graphique simultanée de l'homogénéité interne des classes et de l'hétérogénéité entre classes de la cns, contrairement à la plupart des méthodologies classiques qui considèrent soit
  – une seule de ces deux notions,
  – ou, ces deux notions combinées au sein d'un unique indice,
  – ou encore ces deux notions de manière séparée.

- permet la comparaison de cns indépendamment de :
  - leurs nombres de classes (pas de tendance associée au nombre de classes),
  - l'algorithme (et les paramètres de l'algorithme) mis en œuvre pour les mettre à jour...
- une seule passe sur les données et coût calculatoire relativement faible (particulièrement dans le cas des données catégorielles) ;
- permet une caractérisation de la structure des données (ce point est complété ultérieurement).

4.2.2.3 Expérimentations

Nous avons présenté notre méthodologie sur un exemple synthétique, nous poursuivons maintenant les expérimentations et décrivons les résultats de ces dernières.

4.2.2.4 Expérimentations sur le jeu de données Small Soybean Disease

Nous utilisons tout d'abord le jeu de données "small soybean diseases" (provenant de la collection de l'Université de Californie à Irvine (UCI) [MM96]) classiquement adopté dans la littérature traitant de la cns. Ce jeu de données possèdent 47 objets (des graines de soja) décrits par 35 variables catégorielles ; de plus, à chaque objet est associé une des 4 étiquettes suivantes (qui correspondent à des pathologies) : diaporthe-stem-canker(notée D1), charcoal-rot (notée D2), rhizoctonia-root-rot(notée D3), phytophthora-rot(notée D4). Exceptée D4 qui est associée à 17 objets, toutes les autres pathologies sont associées à 10 objets chacune (voir page 217 pour de plus amples informations). Les expérimentations menées visent à montrer la puissance de la méthodologie présentée pour déterminer et comparer la validité de cns ainsi que pour déterminer la structure du jeu de données.

Description Nous avons mené deux expérimentations différentes :

- **Experience #1.** Nous avons utilisé KEROUAC (méthode de cns pour données catégorielles présentée dans [JN03c] [JN03a] et au chapitre 3) avec des paramètrages différents (des valeurs différentes pour le facteur de granularité) de manière à générer des cns possédant des nombres de classes différents (les paramètres ont été fixés de manière à obtenir des cns en 2, 3, 4, 5, 6, 7, 8, 9, 10 classes). (Cette méthode utilise une variation du Nouveau Critère de Condorcet [Mic97] et essaie donc de trouver une cns minimisant simultanément le nombre de dissimilarités entre objets de même classe et le nombre de similarités entre objets appartenant à des classes différentes). Les résultats sont exposés dans le tableau 4.5 (page 83). Ce tableau donne les informations suivantes :
  - le nombre de classes (#Cl.) de chaque partition (cns),

- la valeur à laquelle a été fixée le facteur de granularité (α) afin d'obtenir la partition et ce pour chaque partition,
- les valeurs de xv_1 et xv_2 pour chaque partition,
- la valeur du critère à minimiser sous-jacent à la méthode K-Modes (QKM) pour chaque partition,
- la valeur du critère à minimiser sous-jacent à la méthode KEROUAC (NCC) pour chaque partition,
- le taux de correction (T.C.) de chaque partition par rapport au concept "pathologie".

n°	#Cl.	α	xv_1	xv_2	QKM	NCC	T.C.
1a	2	1.5	18.02	24.98	310	13991	57.45%
2a	3	2	28.39	23.57	236	17009	78.72%
3a	4	3	31.41	18.17	199	19739	100%
4a	5	3.5	31.24	17.36	188	20055	100%
5a	6	4	29.86	13.74	173	21429	100%
6a	7	5	28.78	12.04	154	22055	100%
7a	8	5.1	27.43	10.41	141	22640	100%
8a	9	5.15	26.04	8.85	132	23206	100%
9a	10	5.3	25.86	8.66	128	23277	100%

TAB. 4.5 –: *Résultats de l'Expérience #1*

- **Experience #2.** Nous avons utilisé la méthode des K-Modes [Hua97] (qui est une extension des K-Means pour les données catégorielles) afin de réaliser 9 séries de cns. Chaque série correspond à 10 cns possédant un même nombre fixé de classes. (Nous avons réalisé des séries de cns avec les mêmes valeurs de paramètres car cette méthode est fortement sensible à l'initialisation contrairement à KEROUAC qui fournit toujours le même résultat pour une valeur donnée du facteur de granularité). Le nombre de classes (#Cl.) a été fixé respectivement à 2, 3, 4, 5, 6, 7, 8, 9, 10. Le tableau 4.6 (page 84) récapitule les résultats pour la "meilleure" partition de chaque série. (On désigne par "meilleure" partition, la partition possédant la valeur la plus faible pour QKM, c'est à dire la meilleure partition au sens du critère sous-jacent à la méthode des K-Modes). Le tableau 4.7 récapitule l'ensemble des résultats pour cette expérience en se contentant toutefois de ne décrire qu'une seule fois une même partition même si plusieurs processus de cns ont mené à une même partition. Ces tableaux donnent les informations suivantes pour chaque cns :
  - le nombre de classes (#Cl.) de chaque partition (cns),
  - les valeurs de xv_1 et xv_2 pour chaque partition,
  - la valeur du critère à minimiser sous-jacent à la méthode K-Modes (QKM) pour chaque partition,

- la valeur du critère à minimiser sous-jacent à la méthode KEROUAC (NCC) pour chaque partition,
- le taux de correction (T.C.) de chaque partition par rapport au concept "pathologie".

n°	#Cl.	xv_1	xv_2	QKM	NCC	T.C.
1x	2	18.96	21.22	308	15563	57.45%
2x	3	28.39	23.57	236	17009	78.72%
3x	4	31.41	18.17	199	19739	100%
4x	5	29.47	14	177	21265	100%
5x	6	28.58	12.6	165	21776	100%
6x	7	26.17	10.42	149	22486	97.87%
7x	8	25.78	9.57	141	22843	100%
8x	9	25.94	9.42	130	22932	100%
9x	10	24.28	8.13	126	23367	97.87%

TAB. 4.6 –: *Meilleurs résultats de l'expérience #2*

Les résultats concernant les valeurs de xv_1, xv_2, T.C., #Cl., QKM, NCC pour chaque cns (partition) de la première expérimentation ainsi que pour chaque cns de la seconde expérimentation) sont graphiquement présentés sur les figures 4.4 (page 86) et 4.5 (page 87). Ces figures reprennent l'ensemble des informations données dans les tableaux 4.5, 4.6 et 4.7.

Analyse des Résultats La question qui se pose après avoir réalisé ces expérimentations est de déterminer laquelle de toutes les cns obtenues est la plus valide. Nous procédons ici à une analyse segmentée en 3 points :

- nous analysons tout d'abord les résultats de l'expérience 1, le problème est alors de déterminer laquelle des cns obtenues par l'intermédiaire de la méthode KEROUAC peut être considérée comme la plus valide ;
- puis, nous nous penchons sur le même problème pour l'expérience 2 (laquelle des cns obtenues par l'intermédiaire de la méthode K-Modes peut être considérée comme la plus valide ?) ;
- le troisième problème abordé est celui de la détermination de la meilleure des cns, que celles ci soient obtenues grâce aux K-Modes ou à KEROUAC.

Le choix de cette segmentation de l'analyse en 3 points est ici motivé par le désir d'exposer les divers intérêts et avantages de notre méthode dans différentes situations d'évaluation/comparaison de la validité de cns :

- lorsque l'ensemble des cns est obtenue par utilisation d'une unique méthode (points 1 et 2);
- la comparaison de nos critères et méthodologie avec l'utilisation de critères internes au sein d'un mode d'évaluation relatif, et ce, que le critère

n°	#Cl.	xv_1	xv_2	QKM	NCC	T.C.	n°	#Cl.	xv_1	xv_2	QKM	NCC	T.C.
1b	2	18,02	24,98	310	13991	57,45%	6c	7	26,95	11,07	155	22278	100%
1c	2	20,81	21,53	309	15983	57,45%	6d	7	29,51	15,68	180	20522	100%
1d	2	18,14	21,82	311	15083	57,45%	6e	7	26,6	10,92	157	22302	100%
1e	2	18,96	21,22	308	15563	57,45%	6f	7	25,35	10,14	156	22521	100%
1f	2	18,13	23,84	310	14375	57,45%	6g	7	25,92	10,4	157	22464	100%
2a	3	16,6	21,92	297	14563	57,45%	6h	7	27,54	11,69	158	22063	100%
2b	3	22,09	15,32	270	19221	76,6%	6i	7	28,2	14,05	168	21061	100%
2c	3	28,39	23,57	236	17009	78,72%	6j	7	27,49	11,88	161	21970	100%
2d	3	21,9	15,32	271	19173	78,72%	7a	8	25,83	10,96	155	22183	97,87%
3a	4	26,74	21,15	226	17624	78,72%	7b	8	26,6	11,25	158	22148	97,87%
3b	4	19,59	11,05	246	20969	78,72%	7c	8	26,12	10,32	152	22528	100%
3c	4	26,98	21,38	223	17581	78,72%	7d	8	25,74	9,6	144	22826	100%
3d	4	26,96	21,36	224	17583	78,72%	7e	8	25,78	9,57	141	22843	100%
3e	4	28,5	16,08	202	20183	89,36%	7f	8	25,66	10,22	150	22520	100%
3f	4	28,5	16,08	202	20183	89,36%	7g	8	27,63	13,57	162	21200	100%
3g	4	29,84	16,84	199	20065	95,74%	7h	8	26,72	11,3	154	22140	100%
3h	4	29,98	16,91	199	20053	95,74%	7i	8	26,16	10,46	149	22467	100%
3i	4	31,41	18,17	199	19739	100%	7j	8	26,83	10,43	144	22560	100%
4a	5	25,66	19,47	210	18101	78,72%	8a	9	24,98	9,44	145	22815	97,87%
4b	5	26,48	20,7	219	17756	78,72%	8b	9	26,23	9,53	131	22914	100%
4c	5	28,12	14,86	190	20676	95,74%	8c	9	25,46	11,49	156	21871	100%
4d	5	29,85	16,24	190	20325	100%	8d	9	27,43	10,93	145	22402	100%
4e	5	29,47	14	177	21265	100%	8e	9	27	10,8	146	22411	100%
4f	5	29,99	16,32	188	20313	100%	8f	9	25,94	9,42	130	22932	100%
4g	5	30,29	16,44	184	20306	100%	8g	9	23,99	8,44	143	23187	100%
4h	5	28,67	13,62	181	21327	100%	8h	9	25,03	9,57	142	22754	100%
5a	6	26,69	13,22	174	21219	95,74%	8i	9	25	9,8	150	22640	100%
5b	6	27,4	13,14	175	21368	95,74%	8j	9	26,19	9,98	137	22699	100%
5c	6	28,12	13,85	177	21143	97,87%	9a	10	24,28	8,13	126	23367	97,87%
5d	6	27,44	12,1	169	21858	97,87%	9b	10	24,53	8,3	132	23313	97,87%
5e	6	28,12	12,61	165	21713	97,87%	9c	10	24,6	8,32	129	23309	100%
5f	6	28,54	12,93	175	21624	97,87%	9d	10	23,74	8,7	139	23028	100%
5g	6	28,58	12,6	165	21776	100%	9e	10	24,32	8,47	132	23207	100%
5h	6	30,04	16,05	177	20440	100%	9f	10	24,5	8,37	127	23275	100%
5i	6	29,81	15,54	177	20629	100%	9g	10	24,35	8,16	127	23363	100%
5j	6	29,57	15,8	182	20480	100%	9h	10	25,06	8,65	127	23203	100%
6a	7	26,17	10,42	149	22486	97,87%	9i	10	25,57	9,74	135	22739	100%
6b	7	27,13	10,97	152	22344	100%	9j	10	22,77	7,35	133	23590	100%

TAB. 4.7 –: *Résultats de l'expérience #2*

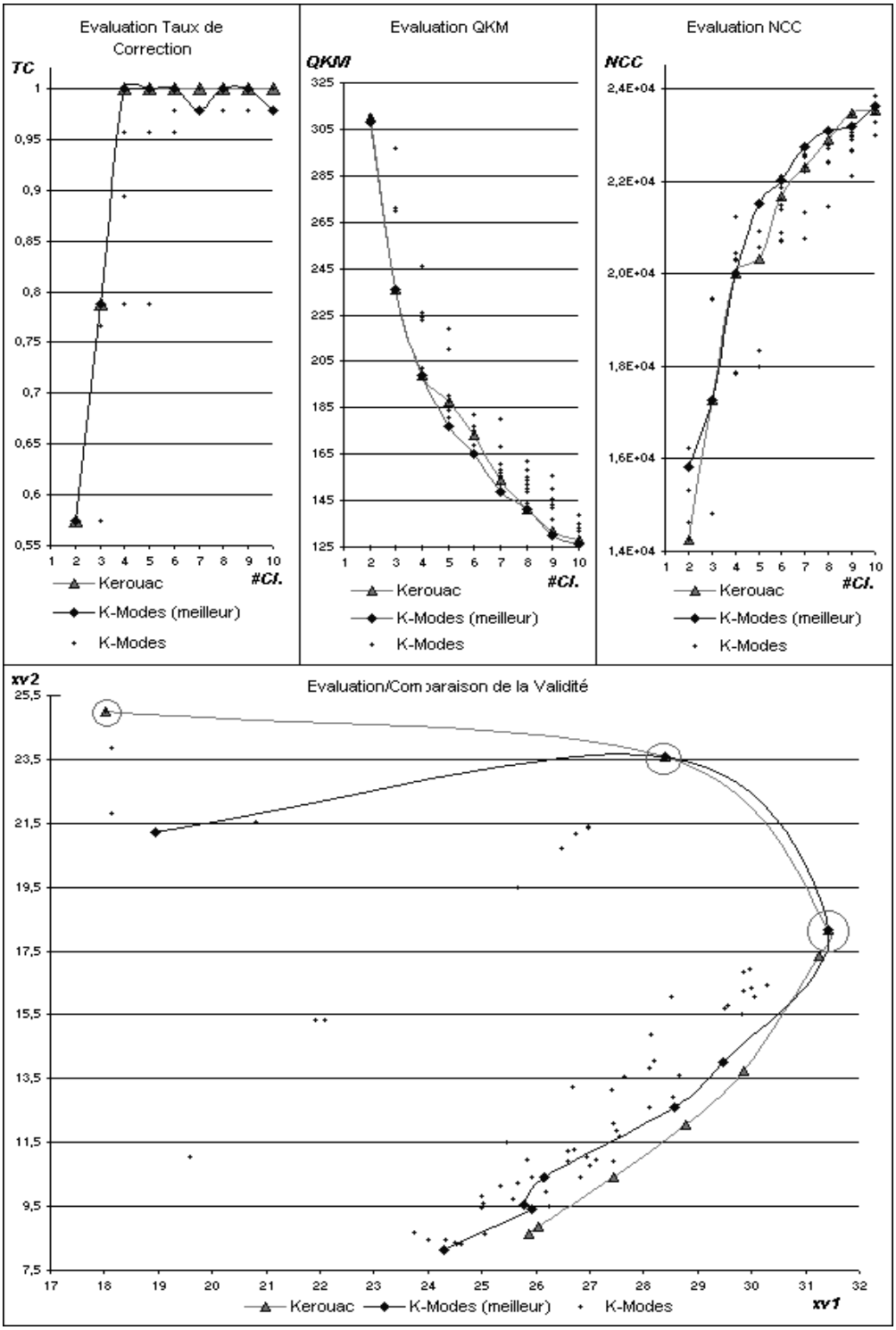

FIG. 4.4 –: *Eléments pour l'évaluation de la validité des cns sur le jeu de données Soybean Disease*

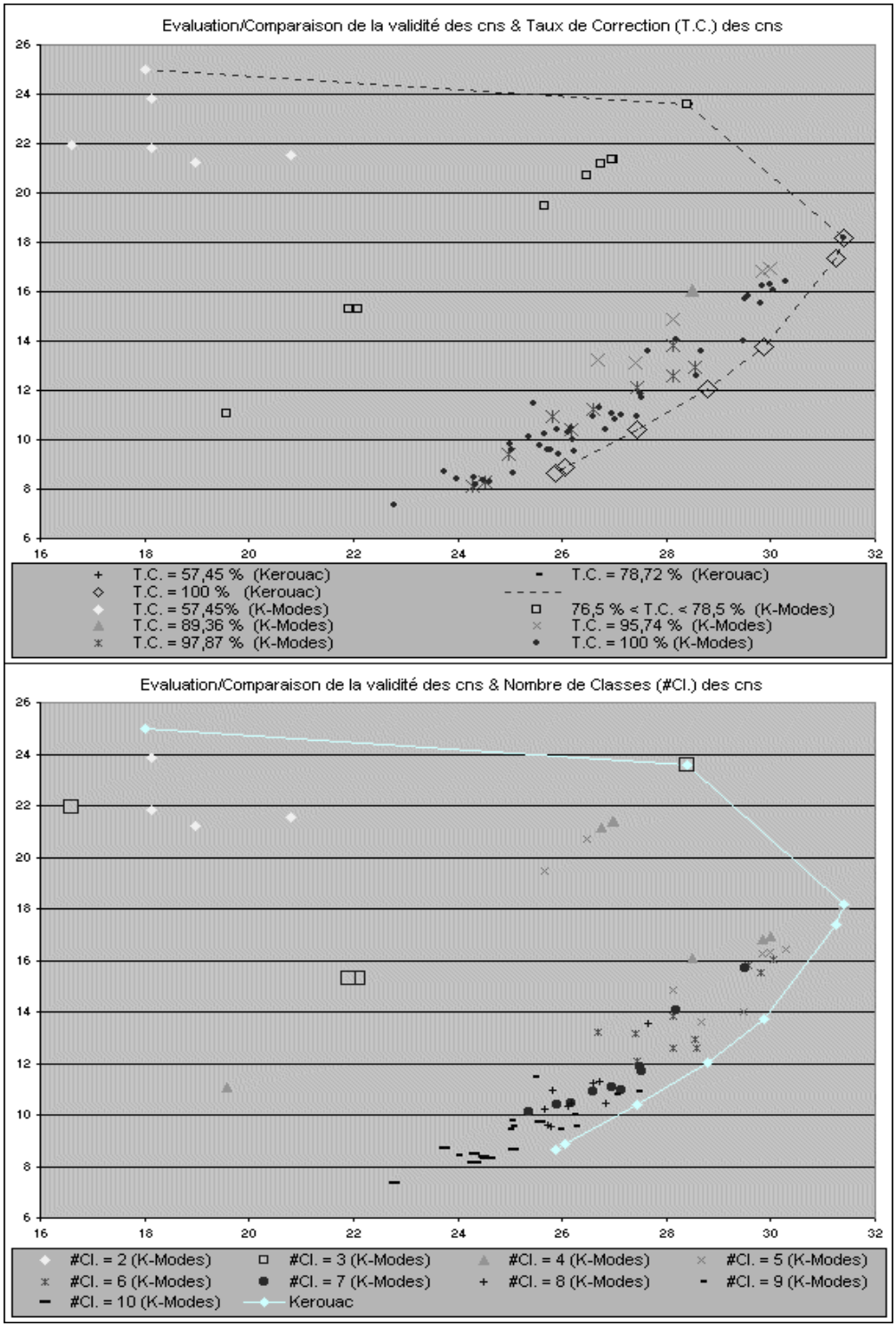

FIG. 4.5 –: *Eléments pour l'évaluation de la validité des cns sur le jeu de données Soybean Disease*

possède (QKM) ou non (NCC) une tendance selon le nombre de classes (points 1 et 2) ;
- la comparaison de nos critères et méthodologie avec l'utilisation d'un critère externe (T.C.) (points 1, 2 et 3) ;
- l'évaluation/comparaison de la validité de cns obtenues par des méthodes différentes (point 3).

Analyse de l'expérience #1 Pour déterminer quelle cns peut être considérée comme la plus valide nous pouvons utiliser plusieurs modes d'évaluation :

Mode d'évaluation relatif avec utilisation d'un critère interne : Le premier problème est le choix du critère à employer (le mode d'évaluation relatif a été choisi car son coût calculatoire est relativement faible et proche de celui de notre méthode). Nous choisissons ici les critères NCC et QKM car ils correspondent aux critères des méthodes de cns employées dans les expériences 1 et 2. Si l'on considère le critère :

- **NCC** : nous pouvons observer que pour cette expérience la valeur de ce critère croît avec le nombre de classes, or ce critère est indépendant du nombre de classes et doit être minimisé, on peut donc en conclure que *la cns 1a (en deux classes) apparaît comme la plus valide* (en fait une cns en une classe mènerait à une valeur du NCC plus faible encore, ce qui semble vouloir signifier qu'il n'existe pas de structure dans les données)
- **QKM** : ce critère doit être minimisé et est dépendant du nombre de classes de la cns (il tend à décroître avec un nombre croissant de classes). On observe ici que cette tendance à la décroissance est bien respectée. La méthodologie d'évaluation veut que l'on cherche alors un changement local marquant dans la courbe, or ce changement n'apparaît pas vraiment, *on ne peut donc rien conclure* de satisfaisant avec cette méthodologie quant à la cns qui doit être la plus valide.

Mode d'évaluation relatif avec utilisation d'un critère externe : Nous utiliserons ici une version ultra simpliste des critères externes basés sur la théorie de l'information : le taux de correction de la cns par rapport à une structure prédéfinie, cette structure étant la partition des objets impliquée par le concept comestibilité, le critère est donc le critère de correction par rapport au concept comestibilité. On observe que la correction croît simultanément avec le nombre de classes (entre 2 et 4 classes) puis se stabilise à 1 (100%) pour un nombre de classes supérieur ou égal à 4. On peut en conclure que *la cns la plus valide est donc celle en 4 classes* puisqu'elle traduit parfaitement le concept comestibilité avec un nombre de classes le plus restreint possible.

Notre méthodologie d'évaluation : Remarquons tout d'abord, que toutes les partitions présentées correspondent à des cns valides selon la définition 10 avec $\alpha_1 = \alpha_2 = 0.001$ (conséquemment aucune ligne n'est tracée sur le

graphique pour délimiter la zone incluant les cns valides). L'ensemble des meilleures cns est ici constitué par les cns 1a, 2a, 3a (i.e. en 2, 3 et 4 classes, voir tableau 4.5).
Les analyses des figures 4.4, 4.5 révèlent également :

- La structure des données par rapport à l'hétérogénéité entre classes : les valeurs de xv_2 sont organisées selon le nombre de classes de la cns. En fait, il apparaît que plus une cns possède de classes, moins significative est sa valeur pour xv_2.

  Une analyse plus précise de ces valeurs montre que :

  - il n'y a pas de forte différence entre les cns en 1a et 2a (i.e en 2 et 3 classes) de ce point de vue (par opposition au point de vue de l'homogénéité des classes (i.e. des valeurs de xv_1))
  - la décroissance en significativité des valeurs xv_2 ralentit lorsque le nombre de classes est élevé.

  Nous pouvons, à partir de ces observations, conclure que l'hétérogénéité entre classes caractérise évidemment la structure du jeu de données, mais que cet aspect n'est pas très fort puisque la relation entre le nombre de classes et xv_2 est monotone (la monotonicité ne traduit pas une caractéristique structurelle spéciale pour les données car on peut considérer normal que les cns possédant un nombre de classe élevé possèdent également une hétérogénéité entre classes moins significative). De surcroît, dans la mesure où la valeur pour xv_1 est beaucoup plus significative pour la cns en 3 classes que pour la cns en 2 classes, (et que par ailleurs leurs valeurs pour xv_2 sont quasi équivalentes) nous pouvons conclure que la cns en 3 classes est plus valide que la cns en 2 classes.
- La structure des données par rapport à l'homogénéité des classes : les valeurs de xv_1 sont organisées de manière non monotone par rapport au nombre de classes des cns. En effet, les valeurs de xv_1 croissent dans un premier temps pour des nombres de classes croissants (jusqu'à #Cl.=4) puis décroissent. Nous pouvons ainsi conclure de ces observations que l'homogénéité entre classes caractérise évidemment la structure des données, et que cet aspect est assez fort puisque la relation entre le nombre de classes et xv_1 est non-monotone (cette non-monotonicité est effectivement intéressante car elle s'oppose à la corrélation naturelle entre nombre de classes et significativité de l'homogénéité entre classes).
- Ces deux premiers points indiquent que la cns la plus valide est soit la cns 2a ou 3a (i.e. la cns en 3 classes ou en 4 classes). Etant donné que nous avons observé que l'aspect homogénéité des classes caractérise plus fortement la structure des données, nous pouvons conclure que *la cns 2a est la plus valide*. Enfin, si la cns 4a (en 5 classes) n'a pas été sélectionnée parmi l'ensemble des cns les plus valides, nous pourrions toutefois réviser ce choix car ses valeurs pour xv_1 et xv_2 sont proches de celles de la cns en 4 classes. (En définitive, si l'on recherchait une cns valide avec le

souhait qu'elle inclue plus de 4 classes, choisir cette cns ne constituerait pas une mauvaise idée.)

Analyse de l'Expérience #2 Nous analysons maintenant les résultats obtenus lors de l'expérience #2 (cns obtenues avec la méthode des k-modes). Afin de déterminer laquelle des cns peut être considérée comme la plus valide, nous pouvons utiliser les modes d'évaluation considérés pour l'analyse des résultats de l'expérience #1 :

Mode d'évaluation relatif avec utilisation d'un critère interne :
- **NCC** : les observations et donc les conclusions sont les mêmes que pour l'expérience #1, on conclut donc que *la cns 1b (en 2 classes) apparaît comme la plus valide*
- **QKM** : mêmes observations et donc conclusions que pour l'expérience #1, *on ne peut donc rien conclure* de satisfaisant avec cette méthodologie.

Mode d'évaluation relatif avec utilisation d'un critère externe : observations quasi-similaires à celles de l'expérience #1 et donc mêmes conclusions, *on conclut que la cns la plus valide est donc la cns 3i (en 4 classes)*.

Notre méthodologie d'évaluation : Tout d'abord, toutes les partitions obtenues correspondent à des cns valides selon la définition 10 avec $\alpha_1 = \alpha_2 = 0.001$. L'ensemble des meilleures cns est ici constitué par les cns 1b, 2c, 3i (en 2, 3 et 4 classes ; entourées sur la figure 4.4).
Si nous considérions chaque série de cns (correspondant à un même nombre de classes), sélectionnions parmi ses cns la cns la plus valide, et tracions la courbe joignant les cns sélectionnées selon l'ordre sur le nombre de classes des cns sélectionnées, nous obtiendrions alors une courbe ressemblant à celle de l'expérience #1. (Notons bien qu'il s'agit d'une courbe non tracée et qu'il ne s'agit pas de la courbe nommée K-Modes (meilleur) qui joint les meilleures cns de chaque série au sens du critère QKM.) De plus, les cns 1b, 2c, 3i correspondent respectivement aux cns 1a, 2a, 3a de l'expérience #1. Cela permet les mêmes conclusions que pour l'expérience #1 :
- l'hétérogénéité entre classes caractérise évidemment la structure du jeu de données, mais cet aspect n'est pas très fort puisque la relation entre le nombre de classes et xv_2 est monotone, alors que pour l'homogénénité interne des classes, la relation de monotonie est brisée ce qui montre qu'elle caractérise fortement la structure des données.
- l'ensemble de ces points mènerait à la sélection de la cns en *4 classes comme cns la plus valide*. (notons que cette cns est identique à la cns en 4 classes obtenue par KEROUAC).

Analyse combinée des Expériences #1 et #2 L'analyse simultanée des résultats des 2 expériences peut permettre de choisir la cns la plus valide parmi

un ensemble de cns résultant de 2 méthodologies différentes. Afin de déterminer laquelle des cns peut être considérée comme la plus valide nous pouvons utiliser les modes d'évaluation considérés pour les expériences #1 et #2 :

Mode d'évaluation relatif avec utilisation d'un critère interne :

- **NCC** : la cns possédant la plus faible valeur pour le critère NCC est la cns en 2 classes obtenue en utilisant KEROUAC, on conclut donc que *la cns 1a (en 2 classes) obtenue avec* KEROUAC *ou encore la cns 1b (qui est la même mais obtenue cette fois avec les K-Modes) apparaît comme la plus valide*
- **QKM** : mêmes observations et donc conclusions que pour les expériences #1 et #2, *on ne peut donc rien conclure* de satisfaisant avec cette méthodologie.

Mode d'évaluation relatif avec utilisation d'un critère externe : observations similaires à celles des expériences #1 et #2, *on conclue que la cns la plus valide est donc celle en 4 classes obtenue avec* KEROUAC *(cns 3a) qui est également l'une des cns obtenues avec les K-Modes (cns 3i).*

Notre méthodologie d'évaluation : On peut reprendre les observations des deux expériences précédentes, celles ci mènent à la conclusion que *la cns la plus valide est celle en 4 classes obtenue avec* KEROUAC *(cns 3a) qui est également l'une des cns obtenues avec les K-Modes (cns 3i).*

Récapitulatif, Confrontation des Analyses des Résultats Les résultats des diverses analyses sont regroupés dans le tableau 4.8. Il apparaît clairement que les conclusions apportées par notre méthode sont en adéquation totale avec celles obtenus par l'utilisation d'un critère externe, cela semble donc signifier que les résultats de notre méthode sont excellents, contrairement à ceux obtenus par les critères internes. En effet, pour le critère QKM (qui correspond à une version spéciale du critère SSE), l'utilisation du mode d'évaluation relatif ne mène à aucun résultat car l'interprétation graphique des résultats est impossible. Les seules informations que l'on peut tirer de ce critère concerneraient la recherche de la cns la plus valide pour un nombre de classes fixé. (Notons d'ailleurs que l'utilisation de notre méthodologie pour ce type de recherche mènerait dans la majorité des cas à l'obtention de résultats similaires à ceux obtenus avec ce critère). Concernant, l'utilisation du critère NCC les résultats fournis ne semblent pas vraiment valables puisqu'ils contredisent les connaissances que l'on a sur les données (les différentes pathologies associées aux graines), et de plus, ils tendent à signifier l'absence de toute structure dans les données. Enfin, les figures 4.4 et 4.5 permettent de conclure, que la cns la plus valide peut être obtenue aussi bien avec les K-Modes qu'avec KEROUAC, mais que par contre, à nombre de classes égales, les cns obtenues par KEROUAC sont le plus souvent plus valides que celles obtenues par les K-Modes, ce dernier élément est intéressant pour la comparaison globale de la

validité des structures fournies par deux méthodes différentes pour un même jeu de données.

Expérience	Mode d'Evaluation Comparaison de la validité	cns la plus valide	#Cl.	Méthode de cns utilisée
Expérience #1	Mode Relatif + NCC	1a	2	KEROUAC
Expérience #1	Mode Relatif + QKM	aucune		
Expérience #1	Critère Externe (T.C.)	4a	4	KEROUAC
Expérience #1	Notre Méthodologie	4a	4	KEROUAC
Expérience #2	Mode Relatif + NCC	1b	2	K-Modes
Expérience #2	Mode Relatif + QKM	aucune		
Expérience #2	Critère Externe (T.C.)	3i	4	K-Modes
Expérience #1	Notre Méthodologie	3i	4	K-Modes
Expérience #1 & #2	Mode Relatif + NCC	1a ou 1b	2	KEROUAC ou K-Modes
Expérience #1 & #2	Mode Relatif + QKM	aucune		
Expérience #1 & #2	Critère Externe (T.C.)	3a ou 3i	4	KEROUAC ou K-Modes
Expérience #1 & #2	Notre Méthodologie	3a ou 3i	4	KEROUAC ou K-Modes

TAB. 4.8 –: *Récapitulatif des Analyses des Résultats*

4.2.3 Expériences sur le jeu de données Mushrooms

Nous avons ensuite utilisé les données "Mushrooms" (provenant également de la collection de l'UCI [MM96]) un autre jeu de données classiquement utilisé. Ce jeu de données inclut les descriptions d'échantillons de 23 espèces de champignons des familles Agaricus et Lepiota. Il inclut 8124 champignons différents décrits par 22 variables catégorielles. Chaque champignon est de plus identifié comme comestible ou vénéneux. (voir page 217 pour de plus amples informations sur ce jeu de données)

4.2.3.1 Description

Nous avons mené les mêmes expériences que celles précédemment décrites pour le jeu de données Soybean Disease, mais en employant des paramètres différents pour KEROUAC et pour les K-modes de manière à obtenir des cns en 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, et 25 classes. Les résultats sont présentés sur les figures 4.6 et 4.7 (toutes les partitions obtenues correspondant à des cns valides selon la définition 10 avec $\alpha_1 = \alpha_2 = 0.001$; conséquemment, aucune ligne n'est tracée pour délimiter la zone incluant les cns valides).

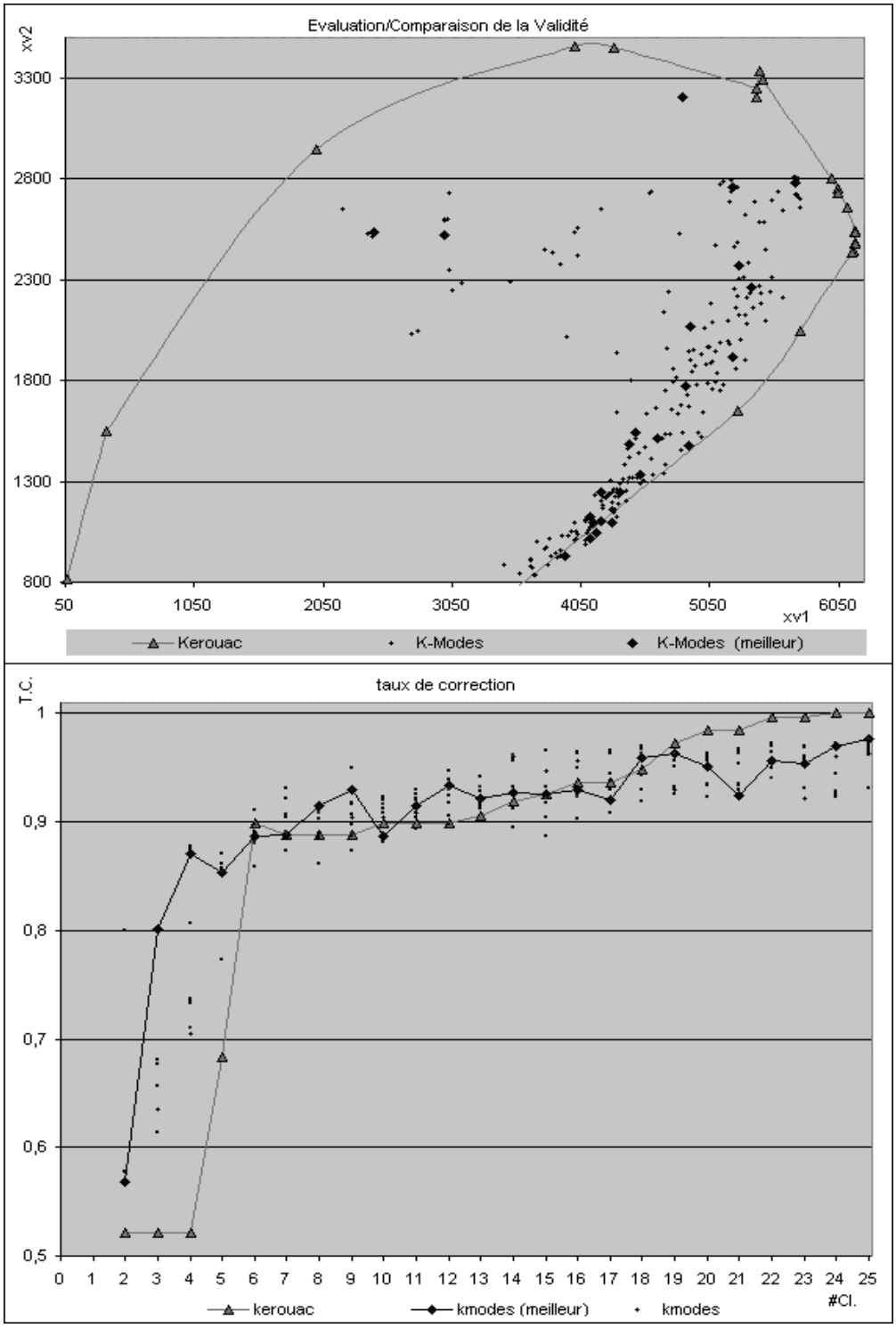

FIG. 4.6 –: *Divers éléments pour l'évaluation de la validité des cns sur le jeu de données Mushrooms*

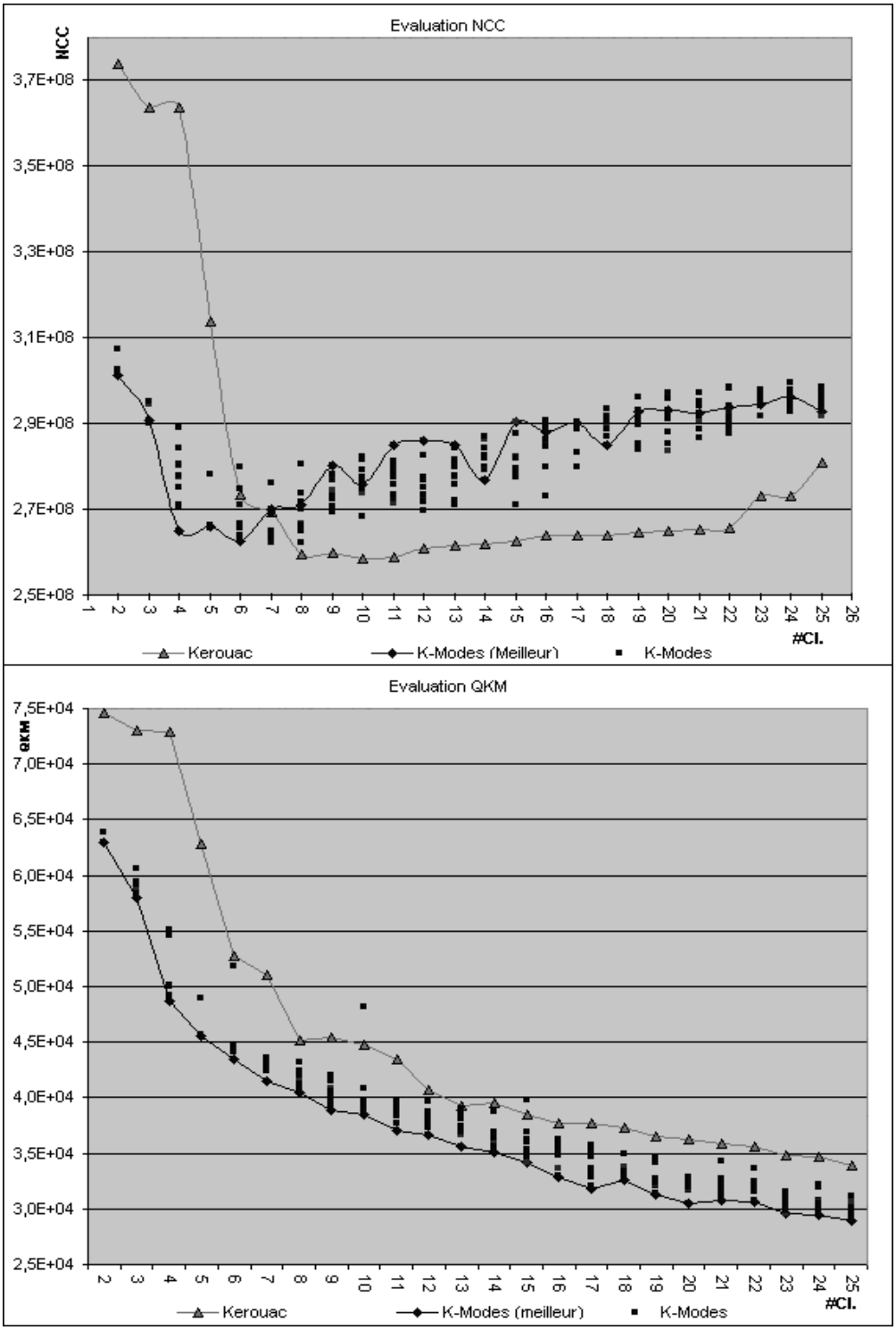

FIG. 4.7 –: *Divers éléments pour l'évaluation de la validité des cns sur le jeu de données Mushrooms*

4.2.3.2 Analyse des Résultats

Nous ne fournissons pas ici une analyse complète des résultats. Les cns les plus valides sont celles possédant 6, 10, 15 et 20 classes avec KEROUAC qui correspondent également à des cns obtenues en utilisant les k-modes. Une analyse complète mènerait à sélectionner la cns en 20 classes qui possède un taux de correction par rapport au concept comestibilité de 98.42 %. Pour cela nous considérons que, comme pour le jeu de données Small Soybean Diseases, les données sont fortement caractérisées par l'aspect homogénéité interne des classes et plus faiblement par l'aspect hétérogénéité entre classes. Cela n'apparaît pas de manière totalement évidente sur le graphique car pour ces deux aspects la relation de monotonie est brisée mais il nous semble cependant que cela est plus marquant pour l'aspect homogénéité interne que pour l'aspect hétérogénéité entre classes dans la mesure où, pour ce dernier aspect, la cassure intervient pour un nombre de classes égal à 5 qui nous semble très faible.

Notons également que quelle que soit la méthode utilisée (K-Modes ou KEROUAC) notre méthodologie mène au choix de la même cns. Concernant les cns parfaitement correcte du point de vue de la correction par rapport au concept comestibilité, la cns en 24 classes obtenue par KEROUAC est celle possédant le moins de classes, et nous pouvons remarquer que son niveau de validité est très proche de celui de la cns la plus valide. Enfin une comparaison des deux méthodes par l'intermédiaire de notre méthodologie mènerait à la conclusion que sur ce jeu de données, KEROUAC semble fournir des résultats plus valides.

L'utilisation du critère NCC mènerait au choix de la cns en 10 classes obtenue par KEROUAC (si l'on ne considérait que les cns obtenues par les K-Modes, on choisirait alors la "meilleure" cns en 6 classes). Ainsi, l'utilisation du critère NCC peut permettre de déterminer une cns plus valide et ce quelle que soit la méthode utilisée (il est toutefois clair que la méthode KEROUAC verra ses résultats privilégiés), cependant ce choix de cns semble peu conforme aux connaissances que l'on possède (concept comestibilité).

L'utilisation du critère QKM ne permet pas le choix d'une cns apparaissant plus valide que les autres si l'on ne sépare pas les résultats des 2 méthodes employées. Ainsi, pour les cns obtenues par KEROUAC, on observe un "coude" dans la représentation graphique pour un nombre de classes valant 8, on choisirait alors la cns en 8 classes. Par contre pour les cns obtenues grâce aux K-Modes repérer le coude semble très hasardeux, la prise de décision s'avère ici très difficile... Comme les cns obtenues par les K-Modes sont les plus valides au sens du critère QKM on ne pourra déterminer la cns la plus valide parmi l'ensemble de toutes les cns (la seule indication est que cette dernière a été obtenue grâce aux K-Modes, ce qui peut apparaître normal vu la relation unissant le critère QKM et la méthode des K-Modes).

Ainsi, l'utilisation du critère QKM peut permettre de déterminer une cns plus valide dans le cas des cns obtenues par la méthode KEROUAC mais ne semble ni permettre la comparaison des cns obtenues par des méthodes dif-

férentes ni la détermination de la cns la plus valide pour les résultats obtenus pour les K-Modes... De plus le choix opéré pour les cns issues de l'utilisation de KEROUAC ne semble pas des plus judicieux.

De manière globale il nous semble donc que l'application de notre méthodologie soit le meilleur choix que l'on puisse faire pour ce jeu de données.

Les expérimentations menées sur les jeux de données Mushrooms et Soybean Diseases indiquent clairement que la méthodologie que nous proposons permet toujours de déterminer une cns apparaissant comme la plus valide, et que les résultats fournis sont en presque parfaite adéquation avec les connaissances dont on dispose sur les données. De plus, aucun surcoût calculatoire n'est impliqué...

4.2.4 Résumé et Informations Supplémentaires

Nous venons de présenter deux critères et une méthodologie pour l'évaluation et la comparaison de la validité de cns. Ses principaux attraits sont :

- une évaluation simultanée et séparée de l'homogénéité des classes et de l'hétérogénéité entre classes contrairement aux approaches classiques qui considèrent soient une seule de ces deux notions, soit une évaluation de ces deux notions au sein d'un critère les combinant ;
- sa capacité à traiter des cns obtenues par différentes méthodes ;
- sa capacité à traiter des cns ayant des nombres de classes différents ;
- la visualisation relativement clair de ses résultats ;
- elle permet la caractérisation visuelle de la structure sous-jacente aux données contrairement aux méthodes existantes (point développé plus tard) ;
- son coût calculatoire associé relativement faible et la non-utilisation de la méthode de Monte-Carlo ;
- sa capacité à traiter des types de données différents et hétérogènes (en utilisant différents types de fonctions $Lien$).

La caractérisation visuelle de la structure sous-jacente aux données est en effet possible par l'intermédiaire de la visualisation de la validité de différentes cns possédant des nombres de classes différents. Par exemple, si nous utilisons la méthode KEROUAC, nous pouvons tracer la courbe joignant les cns selon l'ordre croissant sur le nombre de classes (voir les figures 4.4, 4.5, 4.6), et nous pouvons ensuite associer cette courbe à l'un des 3 types de structures suivants (voir figure 4.8[9]) :

- Type A : Structures impliquant une cns plus valide que toute les autres, i.e. il existe une cns possédant l'homogénéité des classes la plus significative et l'hétérogénéité entre classes la plus significative. Ce type de

9. sur ces figures les portions de courbes incluant les cns les plus valides sont comprises entre les flèches $\downarrow$ et $\leftarrow$

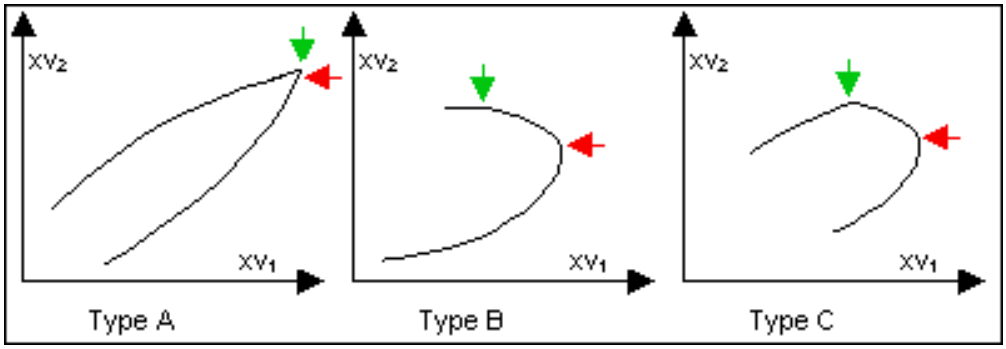

FIG. 4.8 –: *Différents Types de Structures*

structure est à la fois fortement caractérisé par l'homogénéité des classes et l'hétérogénéité entre classes.

- Type B : Structures impliquant un ensemble de cns considérées comme les plus valides et caractérisées fortement par l'homogénéité des classes, i.e. il existe plusieurs cns différentes pouvant être considérées comme les plus valides et caractérisées essentiellement par l'homogénéité des classes. (Dans ce cas, nous pensons que la cns la plus valide est celle exhibant l'homogénéité des classes la plus significative)
- Type C : Structures impliquant un ensemble de cns considérées comme les plus valides et caractérisées fortement à la fois par l'homogénéité des classes et l'hétérogénéité entre classes. (Dans ce cas, nous pensons que la cns pouvant être considérée comme la cns la plus valide est soit celle exhibant l'hétérogénéité entre classes la plus significative, soit celle exhibant l'homogénéité des classes la plus significative.).

Un ensemble de tests réalisé sur 15 jeux de données issus de la collection de l'université de Californie à Irvine [MM96]) [10] illustre ces assertions.

Les résultats de ces tests sont présentés sur les figures 4.9, 4.10, 4.11, 4.12, 4.13 (la forme générale de la courbe permettant la caractérisation de la structure du jeux de données est dessinée en rouge sur ces graphiques).

On peut ainsi remarquer que :

- le jeu de données BREAST possède une structure de type A ;
- les jeux de données CANCER (figure 4.9), GERMAN (figure 4.13), Mushrooms (figure 4.6) possèdent une structure de type C ;
- les jeux de données restant (figures 4.9, 4.10, 4.11, 4.12, 4.13), et le jeu de données Soybean Disease (figure 4.4) possèdent une structure de type B.

10. Ces jeux de données sont les jeux : CANCER, HOUSE-VOTES84 (noté HVOTES), CONTRACEPTION (noté CONTRA.), SPAM, MONKS 3, CAR, NURSERY, FLAGS, ION, WINE, PIMA, BREAST CANCER (noté BREAST), SICK, GERMAN, VEHICLE. Ils sont présentés en annexe (voir page 217). De plus, toutes les variables numériques ont subi un processus de discrétisation supervisée suivant la méthode FUSINTER [ZRR98].

Pour conclure, nous désirons en premier lieu insister sur la simplicité d'utilisation de cette méthode et sur les informations intéressantes qu'elle peut fournir à l'utilisateur grâce à la représentation graphique.

Nous envisageons de poursuivre ces travaux par une étude expérimentales extensive de cette méthodologie (incluant d'autres méthodes de cns, des jeux de données hétérogènes, et la comparaison avec d'autres critères d'évaluation de la validité), nous montrons dans le chapitre suivant que ces résultats nous ont permis de développer deux méthodes de sélection de variables efficaces et rapides dédiées respectivement à l'apprentissage supervisé et à l'apprentissage non supervisé.

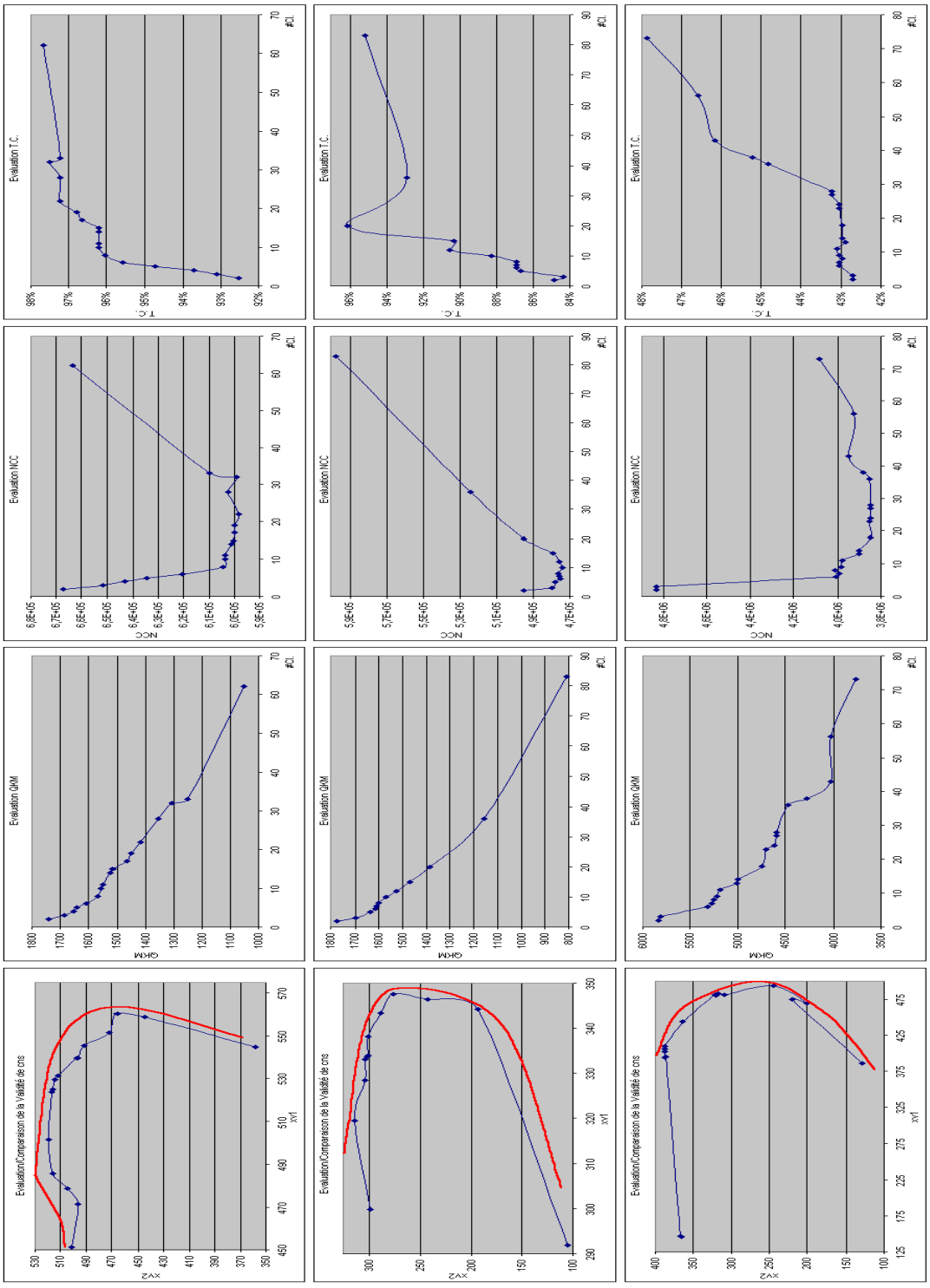

FIG. 4.9 –: *Représentations graphiques pour la détermination des structures des jeux de données : CANCER, HVOTES, CONTRA.*

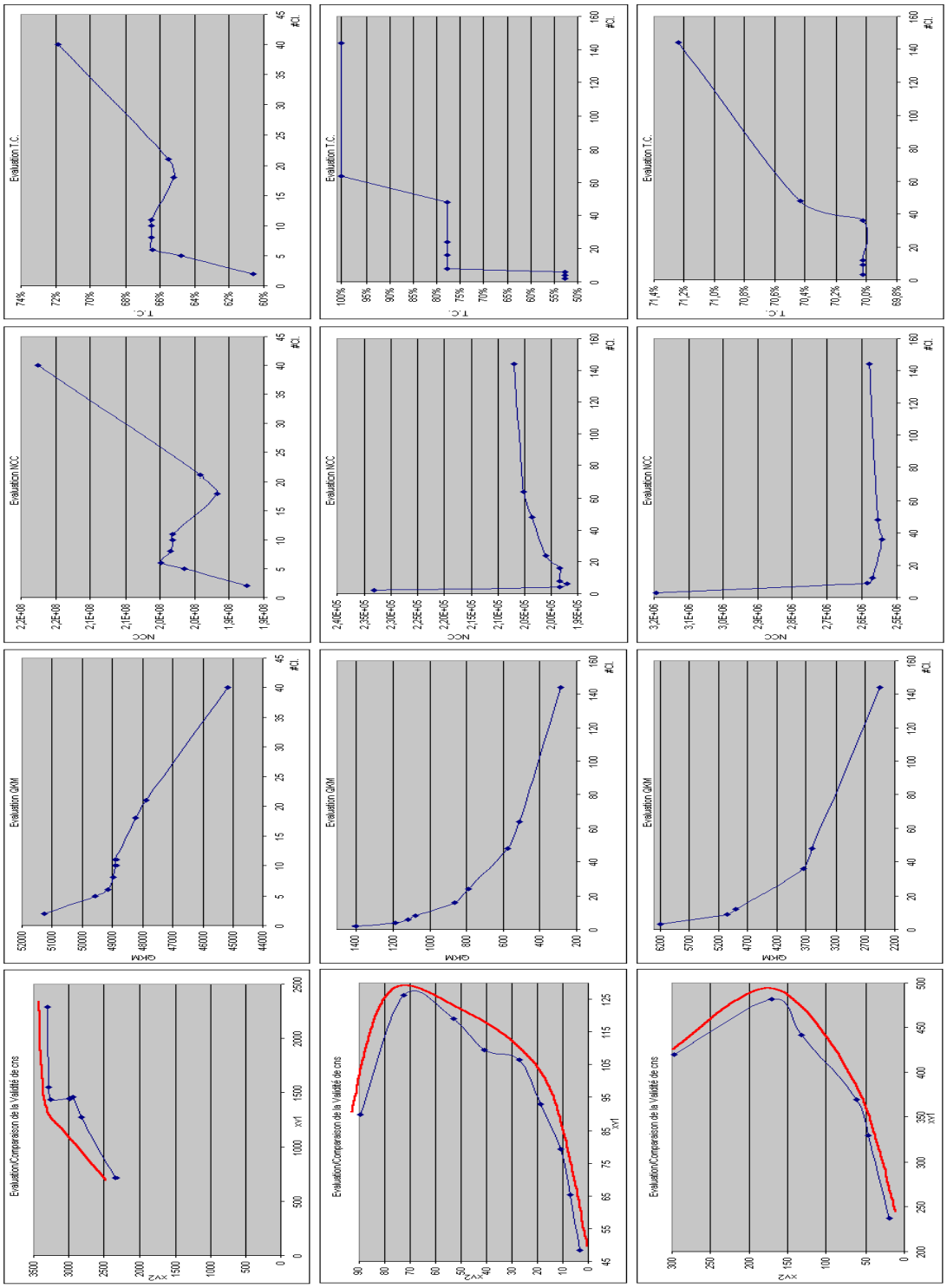

FIG. 4.10 –: *Représentations graphiques pour la détermination des structures des jeux de données : SPAM, MONKS 3, CAR*

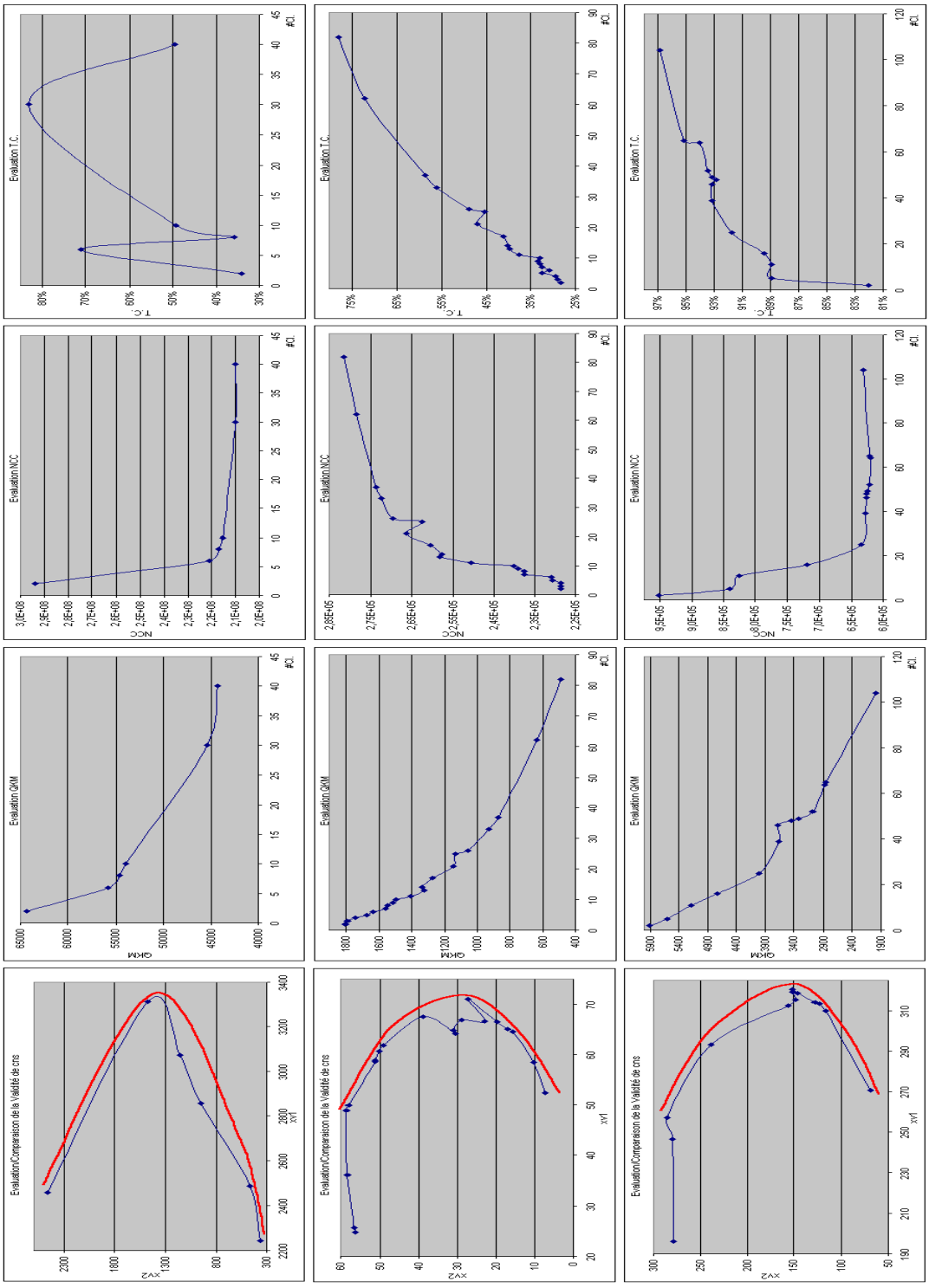

FIG. 4.11 –: *Représentations graphiques pour la détermination des structures des jeux de données : NURSERY, FLAGS, ION*

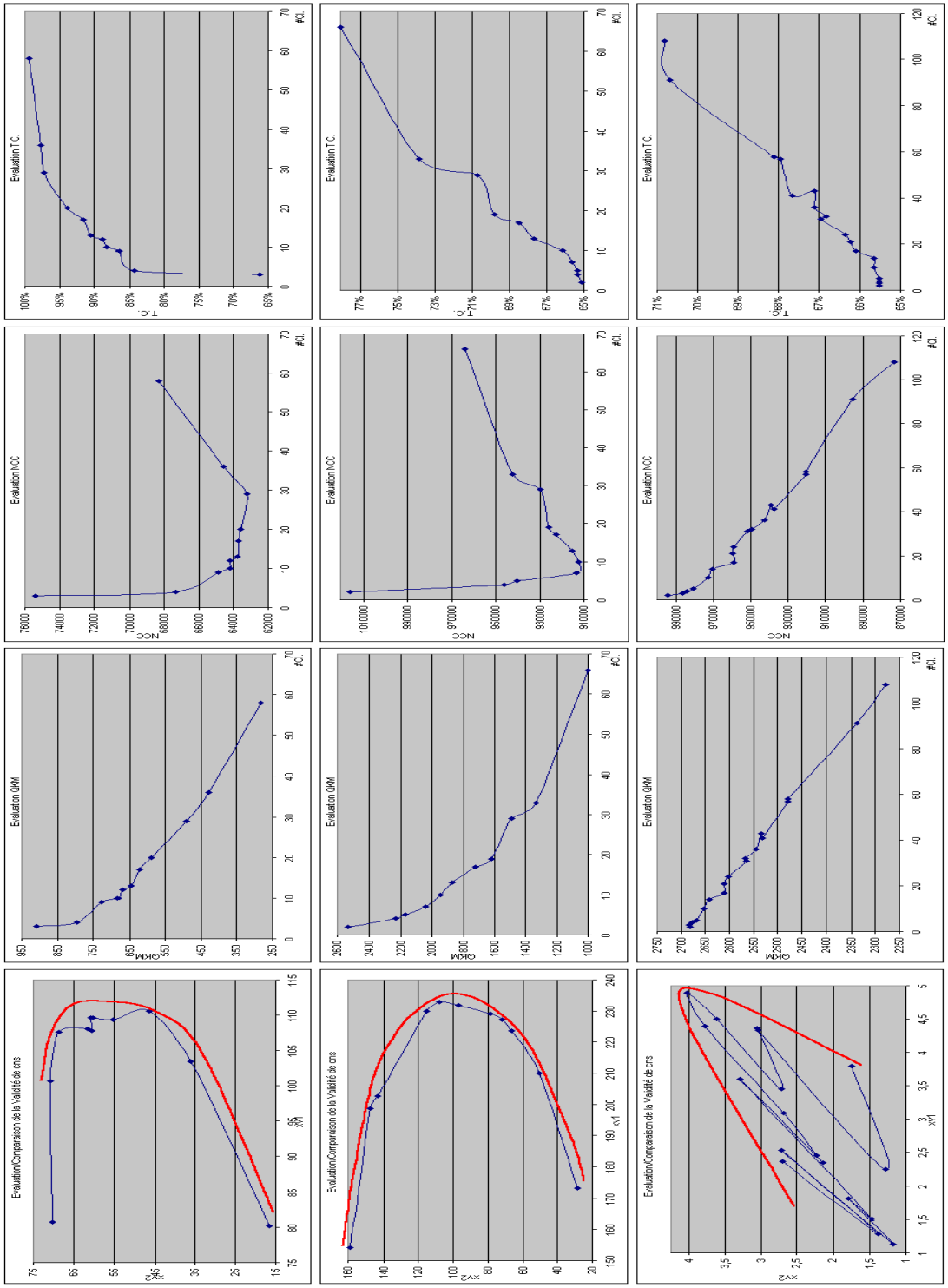

FIG. 4.12 –: *Représentations graphiques pour la détermination des structures des jeux de données : WINE, PIMA, BREAST*

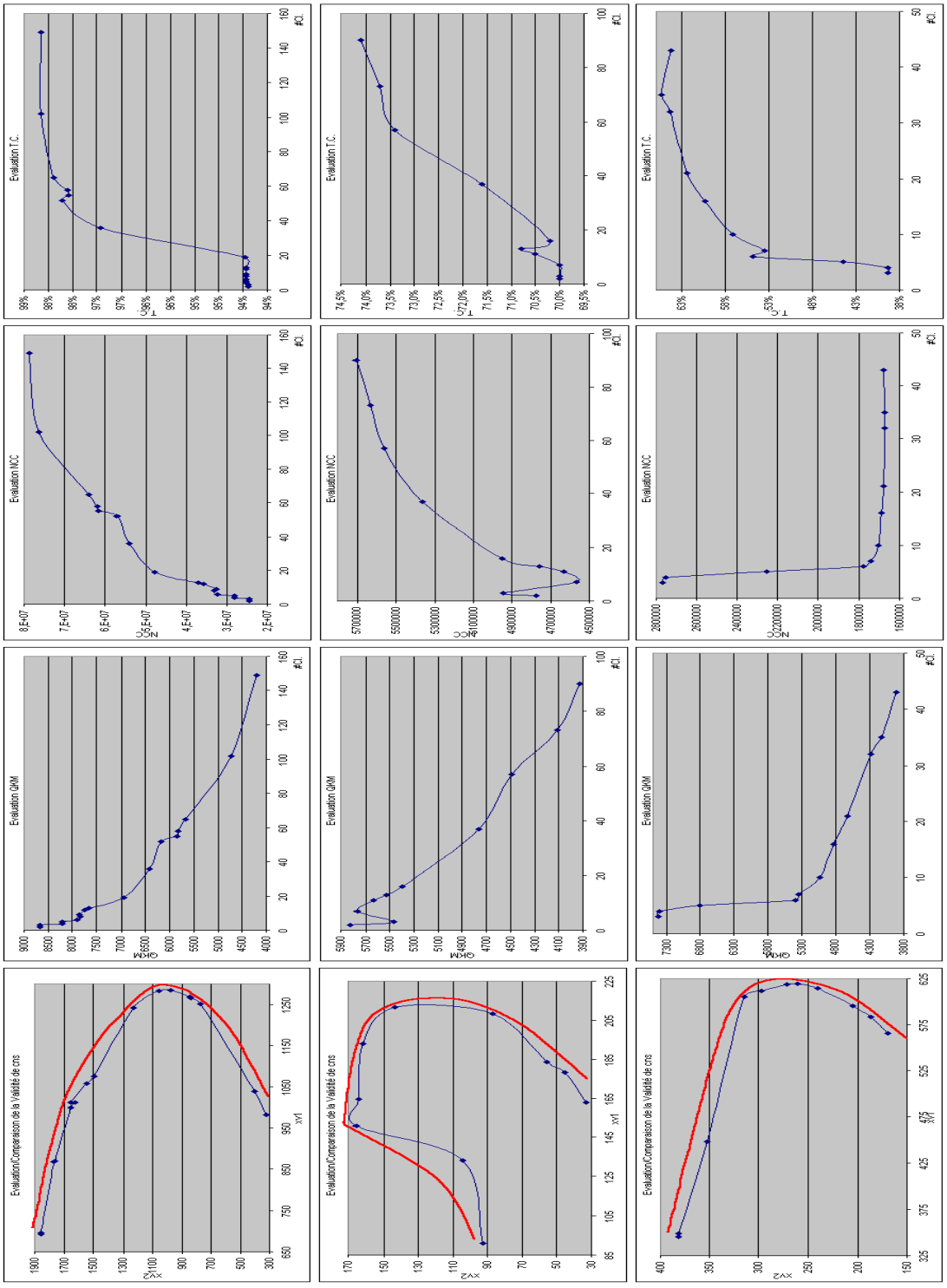

FIG. 4.13 –: *Représentations graphiques pour la détermination des structures des jeux de données : SICK, GERMAN, VEHICLE*