# Vers une classification conceptuelle recouvrante

Guillaume Cleuziou, Bruno Crémilleux

# Vers une classification conceptuelle recouvrante

Guillaume Cleuziou*, Bruno Crémilleux**

*Université d'Orléans, INSA Centre Val de Loire, LIFO EA 4022, FR-45067 Orléans, France
guillaume.cleuziou@univ-orleans.fr,
**Université de Caen Basse-Normandie, GREYC UMR 6072, FR-14032 Caen, France
bruno.cremilleux@unicaen.fr

**Résumé.** La classification conceptuelle (non-supervisée) d'un ensemble de données décrites par un tableau binaire est une tâche difficile, réalisée généralement à l'aide d'heuristiques gloutonnes et/ou par relaxation de la contrainte de partition. Plutôt que de tolérer les recouvrements/chevauchements entre concepts par nécessité, nous étudions la possibilité de les placer au cœur du processus en considérant la recherche d'une "couverture conceptuelle" pertinente. Nous proposons, dans le formalisme de l'analyse formelle de concepts (AFC), une première définition associée à une stratégie de recherche d'une classification conceptuelle recouvrante.

## 1 Introduction et notations

Nous posons le problème du clustering conceptuel dans le formalisme de l'Analyse Formelle de Concepts (AFC) : étant donné un ensemble d'objets $G$ décrit par une relation binaire $I$ sur un ensemble d'attributs $M$, la tâche de classification conceptuelle consiste à produire de façon simultanée un ensemble de clusters (ou classes) sur les objets et une description/définition intensionnelle de ces clusters, de sorte à générer une collection de concepts formels couvrant l'ensemble $G$. Plusieurs formes de clusterings conceptuels peuvent être recherchés : des structures hiérarchiques (**?**) définissant un arbre de classification des données ou des structures "à plat" (**?**). Dans ce deuxième cas, l'espace de recherche est généralement défini par l'ensemble des classifications dont les clusters sont des concepts formels (fermés) réalisant une couverture des données. À cet espace de recherche (trop large) est adjoint une ou plusieurs contraintes structurelles ou qualitatives permettant respectivement de restreindre l'espace de solutions ou d'orienter la recherche vers une solution de "bonne qualité". Par exemple, la contrainte structurelle de partitionnement (strict) étant généralement trop forte, elle est souvent délaissée au profit d'une contrainte de tolérance sur la taille des recouvrements entre concepts (e.g. chevauchement d'au plus $n$ objets entre deux clusters/concepts).

Nous introduisons dans cette étude une nouvelle contrainte structurelle qui renverse la problématique des recouvrements entre concepts en passant de recouvrements "subis" ou "tolérés" à des recouvrements "choisis" et susceptibles de conduire à de nouvelles solutions conceptuellement pertinentes.

Dans la suite de ce résumé, $(G, M, I)$ désigne un contexte formel tel que $(g, m) \in I$ (aussi noté $gIm$) si et seulement si l'objet $g$ possède l'attribut $m$. On utilisera également les

opérateurs de dérivation usuels $A'$ et $B'$ pour tout $A \subseteq G$ et $B \subseteq M$ qui correspondent respectivement à l'intension de $A$ ($A' = \bigcup_{g \in A}\{m \in M | gIm\}$) et à l'extension de $B$ ($B' = \bigcup_{m \in B}\{g \in G | gIm\}$). Dans la suite, une classification conceptuelle désignera une collection de classes $\mathcal{C} = \{(A_1, B_1), \ldots, (A_p, B_p)\}$ vérifiant les propriétés suivantes :

$i)$ $\forall i,\ A'_i = B_i$ et $B'_i = A_i$ (les classes sont toutes des concepts formels),
$ii)$ $\forall i,\ A_i \neq \emptyset$ (les classes sont toutes non-vides),
$iii)$ $\forall i, j,\ A_i \cap A_j \notin \{A_i, A_j\}$ (il n'y a pas d'inclusion entre classes),
$iv)$ $\bigcup_{i=1}^{p} A_i = G$ (la classification réalise une couverture de $G$).

# 2 Couverture conceptuelle

Les propriétés précédentes sont certes contraignantes mais insuffisamment pour espérer extraire des classifications conceptuelles intéressantes d'un espace de recherche très vaste. En effet aucune contrainte relative (entre concepts), autre que l'inclusion, n'étant spécifiée, plusieurs concepts ne se différenciant que par un seul objet pourront par exemple apparaître dans une solution sans apporter d'intérêt particulier à sa sémantique mais tout en augmentant inutilement la taille de cette solution et par la même la taille de l'espace de recherche.

Nous introduisons une nouvelle contrainte structurelle visant à modéliser conceptuellement les recouvrements/chevauchements entre classes. Nous considérons que pour être pertinent, un recouvrement entre deux ou plusieurs concepts formels doit pouvoir s'expliquer conceptuellement par les attributs des concepts concernés, et uniquement ceux-ci. Ainsi, considérant un objet $g$, situé par exemple à l'intersection de deux concepts formels $A_1$ et $A_2$ alors cet objet possède au minimum tous les attributs de $A'_1 \cup A'_2$ ; si $g$ possède un attribut supplémentaire $m$, celui-ci n'étant pas caractéristique des concepts initiaux ($m \notin A'_1 \cup A'_2$) il ne devrait pas non plus être caractéristique d'un objet de l'intersection, autrement dit il doit exister des objets dans les concepts initiaux qui possèdent également cet attribut. Cette nouvelle contrainte structurelle est formalisée par la propriété $(v)$ suivante

$$v)\quad \forall m \in M,\ \forall \mathcal{S} \subset \mathcal{C},\ \forall g \in \bigcap_{\mathcal{S}} A_i,\quad gIm \Rightarrow \forall A_i \in \mathcal{S},\ \exists g_i \in A_i \setminus \bigcap_{\mathcal{S}} A_i,\ \text{t.q. } g_i I m$$

obligeant chaque attribut $m$ d'un objet $g$ situé à l'intersection d'une famille $\mathcal{S}$ de concepts, à être possédé par au moins un objet de chaque concept (en dehors de l'intersection).

**Définition 2.1** *Soit $\mathcal{C} = \{(A_1, B_1), \ldots, (A_p, B_p)\}$ un ensemble de $p$ concepts associé à un contexte formel $(G, M, I)$, $\mathcal{C}$ est une* **couverture conceptuelle** *si et seulement si $\mathcal{C}$ satisfait aux propriétés $i)$, $ii)$, $iii)$, $iv)$ et $v)$ définies précédemment.*

Étant donné un ensemble de $p$ concepts $\mathcal{C}$, vérifier que $\mathcal{C}$ satisfait aux exigences d'une couverture conceptuelle est a priori coûteux. Les propriétés $i)$ à $iv)$ peuvent être vérifiées simplement en observant chaque concept (pour les propriétés $i)$, $ii)$ et $iv)$) et chaque paire de concepts (pour la propriété $iii)$) ; en revanche la propriété $v)$ que nous avons introduite nécessiterait a priori de considérer l'ensemble des $2^p$ sous-familles de concepts. Cependant, on peut montrer que cette contrainte présente une bonne propriété de monotonie permettant de vérifier la satisfaction en ne considérant que les paires de concepts.

**Proposition 2.1** *Soient un contexte formel $(G, M, I)$ et $\mathcal{S}$ une famille d'au moins $k+1$ concepts formels extraite de ce contexte. Si $\mathcal{S}$ satisfait la propriété $v$) pour tout recouvrement de $k$ de ses concepts alors $v$) est également satisfaite pour tout recouvrement de $k+1$ concepts de $\mathcal{S}$.*

Sur la base Iris (**?**) discrétisée de manière à décrire les 150 objets iris selon 8 attributs binaires, nous avons généré l'ensemble des solutions possibles en considérant :

— les classifications conceptuelles non-contraintes (propriétés $i$) à $iv$))
— les partitions conceptuelles (idem en interdisant tout recouvrement entre concepts)
— les couvertures conceptuelles (cf. définition **??**)

La Figure **??** (graphique de gauche) présente la distribution des solutions relativement au nombre de concepts qu'elles contiennent. On note alors que le filtre opéré par la contrainte structurelle $v$) permet de réduire significativement l'espace des solutions puisque seulement un tiers des classifications satisfont les propriétés d'une couverture conceptuelle [1]. Le graphique de droite fournit une analyse qualitative de ces trois différents ensembles de solutions par rapport à la classification de référence à l'aide de la mesure F-Bcubed (**?**) permettant d'évaluer la correspondance entre deux classifications (strictes ou recouvrantes). Il est alors très intéressant d'observer la qualité du filtre opéré par la contrainte $v$), en effet les couvertures conceptuelles apparaissent en moyenne d'avantage en correspondance avec la classification de référence que les classifications non-contraintes, mais également par rapport aux partitions conceptuelles.

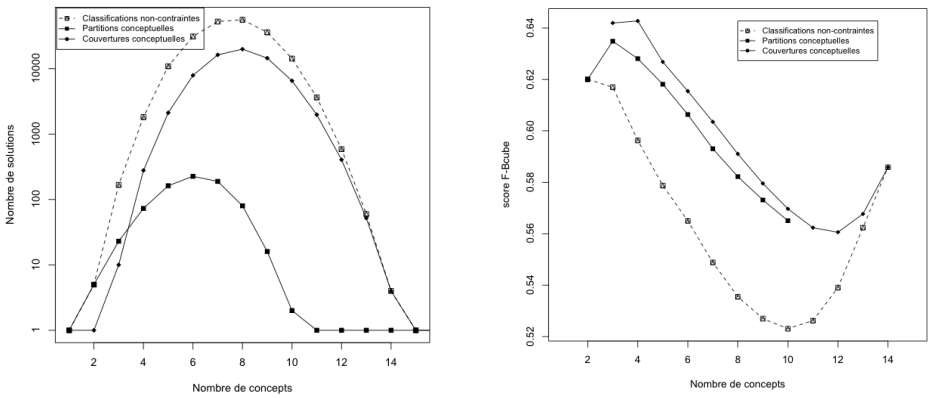


FIG. 1 – *Taille de l'espace de recherche et évaluation externe des classifications sur la base Iris discrétisée.*

# 3 Stratégie de recherche et conclusion

L'exploration de l'espace des couvertures conceptuelles reste un problème difficile au même titre que la recherche d'une partition conceptuelle. Une manière de procéder consisterait

1. On compte au total 69,882 couvertures conceptuelles contre 207,259 solutions sans la contrainte $v$) et 767 partitions conceptuelles.

à explorer l'espace des classifications conceptuelles (non contraintes) par fusions successives de concepts en partant d'une solution particulière dite "base minimale", définie par l'ensemble des concepts formels obtenus par les extensions élémentaires sur chaque objet de $G$. Partant de cette base minimale, nous proposons une approche hiérarchique bottom-up qui consiste à chaque étape de fusion à (1) favoriser l'émergence d'une couverture conceptuelle (ou à conserver ce statut lorsque la solution en cours correspond déjà à une couverture conceptuelle) et (2) préférer la génération de concepts plus spécifiques.

Cette stratégie, que nous ne détaillerons pas d'avantage dans ce résumé, n'offre pas en théorie l'assurance d'atteindre une couverture conceptuelle ; en pratique néanmoins, on observe sur diverses expérimentations que cette approche permet effectivement d'extraire un ensemble de couvertures conceptuelles.

Nous avons proposé dans ce résumé une nouvelle approche pour la problématique de classification conceptuelle. Partant de l'hypothèse que les concepts formels représentent effectivement de bons candidats pour la tâche de clustering, nous avons introduit et étudié les propriétés d'une contrainte structurelle régissant la coexistence d'un ensemble de concepts au sein d'une classification, en observant les propriétés conceptuelles de leurs chevauchements. Conforté par l'analyse préliminaire à la fois quantitative et qualitative des nouvelles classifications induites par cette contrainte nous avons posé les premières bases d'une méthodologie d'exploration de l'espace des couvertures conceptuelles, ainsi définies.

# Références

Amigó, E., J. Gonzalo, J. Artiles, et F. Verdejo (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr. 12*(5), 613.

D.J. Newman, S. Hettich, C. B. et C. Merz (1998). UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine learning 2*(2), 139–172.

Michalski, R. S. et R. E. Stepp (1983). Learning from observation : Conceptual clustering. In *Machine learning*, pp. 331–363. Springer.