Une approche algébrique au problème du consensus de partitions

Frédéric Dumonceaux, Guillaume Raschia, Marc Gelgon

Laboratoire d'Informatique Nantes Atlantique
(LINA, UMR CNRS 6241)
Université de Nantes
Rue Christian Pauc, La Chantrerie
44306 Nantes cedex 3, France
prenom.nom@univ-nantes.fr

Résumé. En classification non-supervisée, le consensus de partitions a pour objectif de produire une partition unique, représentant le consensus, à partir d'un ensemble de partitions où chacune est engendrée indépendamment des autres, voire avec des méthodologies différentes. En complément des techniques ayant leur qualité propre en terme de robustesse ou de passage à l'échelle, nous apportons un point de vue original sur le consensus de partitions, c'est-à-dire, par le biais de définitions algébriques qui permettent d'établir la nature des déductions pouvant être réalisées dans une approche systématique (p.ex. un système à base de connaissances). Nous fondons notre approche sur le treillis des partitions pour lequel nous montrons comment peuvent être adjoint des opérateurs dans le but de formuler une expression caractérisant le consensus à partir d'un ensemble de partitions.

1 Introduction

La classification non supervisée (clustering) de données constitue une tâche fondamentale et classique de structuration, pour l'analyse exploratoire de jeux de données. Elle a été l'objet d'un nombre considérable de travaux depuis des décennies, à la fois comme objet général d'analyse de données ou dans des contextes plus appliqués (bioinformatique Hu et Yoo (2004), segmentation d'images Hong et al. (2008), etc ...). Toutefois, la nature même du problème ne fournit généralement pas de vérité-terrain simple, tant sur la composition des classes que leur nombre. Les algorithmes proposés dans la littérature optimisent des critères très divers et font des hypothèses elle aussi diverses sur les propriétés de cohérence intra-classe.

Depuis une dizaine d'années, un axe de recherche, le *clustering d'ensemble*, s'est développé, élaborant des critères et des méthodes pour aggréger différentes partitions d'un même jeu de données, construites par des critères antagonistes.

Plusieurs méthodes sont alors envisageables pour pallier l'incertitude quant à la plausibilité du résultat :

- l'entremise de connaissances obtenues par le biais d'un « expert », ou inhérentes au domaine dont est issu le jeu de données, permet de formuler des hypothèses sur le mécanisme générateur sous-jacent et donc sur la nature des regroupements les plus vraisemblables ;
- la recherche d'un compromis explicite entre les différents résultats, en les combinant et suivant généralement un critère à optimiser qui définit les propriétés attendues dans le clustering devant réaliser le consensus.

Il existe deux grandes familles complémentaires d'approches, pour traiter le problème de la combinaison de clustering. La première approche permet d'établir une étude des mécanismes de génération des partitions (*p.ex.* Von Der Gablentz et al. (2000); Dudoit et Fridlyand (2003)). Celle-ci tend à améliorer la fiabilité du résultat dès lors que des variations sont appliquées dans le protocole expérimental, où différents algorithmes avec des paramétrages différents peuvent être utilisés. Différentes projections peuvent être également utilisées dans l'espace des attributs pour engendrer chaque partition. La seconde approche procède par la définition d'une fonction de consensus qui correspond à un critère à optimiser dépendant du choix d'un modèle de représentation des partitions (*i.e.* hypergraphe des clusters, matrice d'associations) et d'une mesure de distance adaptée à cette dernière (*p.ex.* Strehl et Ghosh (2003); Topchy et al. (2005)).

Problématique et enjeux Il n'existe pas à ce jour, à notre connaissance, de théorie qui valide un ensemble de propriétés axiomatiques devant être implicitement vérifiées par chaque méthode en particulier, au-delà de la taxonomie propre à la discipline et distinguant les fonctions de consensus et les méthodes génératives. Plus généralement, chaque méthode fonde son propre compromis sur une base opérationnelle qui ne permet pas de *justifier* l'ensemble des choix appliqués pendant la construction du résultat final.

Par ailleurs, chaque résultat de clustering est naturellement définissable comme la donnée d'une partition P, définie sur un ensemble d'objets Ω, figurant un jeu de données quelconque. Construire un résultat satisfaisant vis-à-vis des partitions originales représente en soi une gageure à cause de la nature combinatoire des partitions. De plus, la nature exploratoire de la tâche rend délicate la définition de propriétés permettant de construire inductivement un résultat sans introduire de biais dans le raisonnement.

Une présentation algébrique peut alors contribuer à la compréhension des mécanismes intervenant lors de la combinaisons de partitions de sorte à définir effectivement la nature du raisonnement applicable sur celles-ci et en particulier comment établir qu'une partition soit la représentation légitime du consensus de plusieurs autres partitions.

En économie et plus particulièrement dans la théorie du choix social, un consensus est considéré comme la réalisation d'un arrangement ou d'une combinaison à partir d'un ensemble de profils formulés par des individus de sorte à obtenir un profil commun, avec la contrainte que ce dernier représente fidèlement les préférences initiales du plus grand nombre possible d'individus. Dans ce cas, il est alors possible de formuler l'analogie *associant* un résultat de clustering au profil propre à un individu. Par conséquent, il est aisé d'interpréter l'agrégation de deux objets $a, b \in \Omega$ comme une relation de nature préférentielle résultant des paramètres propres à l'algorithme de clustering et du critère statistique qu'il a employé, puis de formuler le problème de satisfaction par la recherche d'une partition qui soit maximalement *compatible* avec l'ensemble de départ (*i.e.* typiquement, par l'usage d'une relation d'ordre partielle).

L'intérêt de ce paradigme est qu'il permet de contraindre les procédures de décisions qui vont régir explicitement comment les profils peuvent se combiner entre eux et ainsi d'imposer des *modalités* propres aux usages espérés dans un contexte applicatif.

Un exemple immédiat est de considérer les *interactions* attendues entre des partitions dans un *système de recommandations* : ceux-ci ont pour but de mettre en relation un individu α dont le profil de préférences P_α est proche de celui d'un autre individu β. La logique sous-jacente est alors très similaire à celle d'un système multiagent où les individus ont la faculté de raisonner (*i.e.* réaliser des déductions) depuis la connaissance induite par leurs propres préférences *et* de celles exprimées par d'autres individus afin de modifier leur propre comportement. La relation entre deux individus α et β va donc dépendre de la *permissivité* de la relation définissant la proximité entre P_α et P_β et décidant quand un individu peut avoir accès aux préférences de l'autre, et réciproquement.

Cependant, le but de cette article n'est pas de proposer un système logique modélisant l'ensemble des cadres applicatifs susceptibles d'inclure des partitions mais de présenter la construction d'un calcul algébrique basé sur *le treillis des partitions* où l'analogie précédente va permettre de qualifier assez simplement le pendant algébrique de la définition du consensus. Cette approche sera mis en perspective avec celle présentée dans Barthelemy et Leclerc (1995).

Contributions

- Modélisation d'un nouvel opérateur algébrique sur le treillis des partitions ;
- Définition d'une fonction de consensus exploitant cette construction ;

Cet article est une version condensée du rapport de recherche disponible dans Dumonceaux et al. (2013).

Dans la prochaine section, on présentera succinctement l'algèbre permettant de manipuler des partitions et la nature du raisonnement applicable entre celles-ci et comment l'exploiter par le biais d'opérateurs adjoints au treillis.

2 Consensus de partitions

Un treillis est un ensemble partiellement ordonnée $(P, \leq)$ dont les éléments sont soit mutuellement comparables par sa relation (réflexive, antisymétrique et transitive), ou pour lesquels il existe deux bornes accessibles par cette même relation et symbolisées par les opérateurs $\vee$ et $\wedge$ (idempotent, commutatif et associatif). La relation d'ordre partielle ($\leq$) permet de distinguer les éléments vérifiant des propriétés particulières et se confond avec la relation de déduction ($\vdash$) qui lie une proposition logique prise pour hypothèse et sa conséquence. Les éléments d'un treillis sont alors une représentation concrète des énoncés d'un système logique.

Soit $(\Pi_\Omega, \vee, \wedge)$ le treillis des partitions, où Ω est l'*ensemble support* pour lequel chaque partition $P \in \Pi_\Omega$ est le résumé d'une séquence d'opérations d'agrégation dont les clusters $c \subseteq \Omega$ sont les représentations idoines et mutuellement disjoints. La relation de raffinement entre deux partitions $P \leq Q$ est alors vérifiée dès lors que chacun des clusters de P est inclus dans un cluster de Q. Par exemple, sachant $P = 12|3|45|6$ et $Q = 123|456$, P est alors *plus fine* que Q. En particulier, imaginons que P soit un résultat d'expérience ou l'énonciation

d'une hypothèse dans un problème quelconque, on écrit alors que de P, on déduit Q puisque on remarque aisément que toutes les paires d'éléments dans $\Omega \times \Omega$ sont *préservées* dans Q [1]

Dans Π_Ω, la partition $P \wedge Q$ est telle que les clusters sont *simultanément* inclus dans P et dans Q, chacun des clusters de P et de Q sont simultanément inclus dans $P \vee Q$. Par exemple, étant donné $P = 12|345|67$ et $Q = 123|45|67$, alors $P \wedge Q = 12|3|45|67$ et $P \vee Q = 12345|67$. Le point de vue logique adopté ici considère P et Q comme des hypothèses, autrement dit P, Q sont les *prémisses* du raisonnement pour lequel on peut appliquer les règles d'introductions classiques des connecteurs logiques, et ainsi calculer les propositions correspondantes.

Cependant, le treillis des partitions n'étant pas *distributif*, il est impossible de décomposer sous une forme *invariable* et *minimale* une partition comme la composition de plus petite partitions *atomiques* et figurant une unique association entre deux éléments pris dans Ω (*cf.* Birkhoff (1937)). La conséquence immédiate est qu'on ne peut déduire la suite *exacte* des agrégations d'une partition particulière et par induction, faire appel de multiples fois à $(\vee)$ revient à inclure lors de la combinaison de partitions, un nombre croissant d'associations entre des éléments qui n'ont pas été formulées explicitement.

Une autre conséquence est l'impossibilité de définir des opérateurs adjoints à la structure comme c'est l'usage pour le treillis des parties d'un ensemble 2^Ω avec les opérateurs d'implication $(\rightarrow)$ et de différence $(-)$. Ces opérateurs étant définis comme des connections de galois, requérant la propriété de distributivité, on choisit alors d'amender cette définition. Par exemple, $Q \rightarrow P$ est définit par $\bigvee\{R \mid R \wedge Q \leq P\}$ dont le résultat R figure la plus grande partition dont les parties communes avec Q sont compatibles avec P et on impose que le résultat soit choisi parmi 2^P, et ainsi on a toujours $Q \rightarrow P \leq P$ quelque soit P. L'usage de cet opérateur va permettre de *minorer* l'importance des petits clusters déja inclus dans ceux d'une autre partition.

Arrow (1951) établit un ensemble de trois propriétés axiomatiques qui devrait être vérifié par toute méthode de consensus, et démontre dans la foulée que ses critères ne peuvent être satisfait simultanément, ce fait est communément appelé *paradoxe de Arrow*, dont le précurseur fut Condorcet. Une solution communément admise est d'envisager la relaxation de l'une des propriétés voire d'enfreindre une ou plusieurs de ces contraintes. Ces critères sont les suivants :

- L'indifférence face aux alternatives non-pertinentes : deux éléments $x, y \in \Omega$ seront agrégés dans le consensus indépendamment des éléments dans $\Omega - \{x, y\}$ et leur agrégation dépend uniquement de la position exprimée sur *ce* couple par l'ensemble des partitions dans $\mathcal{P}$;
- Optimum de Pareto : si toutes les partitions sont *unanimement* d'accord sur l'agrégation de deux éléments $x, y \in \Omega$, alors ils sont également agrégées au niveau de la partition résultant du consensus ;
- Non-didactorial : il ne doit pas exister de sous-ensemble $\mathcal{P}' \subset \mathcal{P}$ tel que le consensus soit obtenu à l'unanimité sur $\mathcal{P}'$, rejetant les préférences induites par les partitions dans $\mathcal{P} - \mathcal{P}'$.

On définit une fonction de consensus par $f : \Pi_\Omega^n \rightarrow \Pi_\Omega$. Une fonction qui satisfait de manière évidente à chacun de ses trois critères est celle requérant l'unanimité :

$$u(\mathcal{P}) =_{def} \bigwedge_{P_i \in \mathcal{P}} P_i$$

1. En particulier, soit une bijection $\sigma : \Omega \rightarrow \Omega$, telle que P_σ représente la partition résultante, alors $P \leq Q \Rightarrow P_\sigma \leq Q_\sigma$ et la relation d'ordre dans le treillis est indépendante de l'étiquetage employé pour l'ensemble support.

mais celle-ci est largement susceptible de renvoyer des résultats peu probants car ne sauvegardant pas suffisamment d'associations entre éléments. Il faut donc assouplir la contrainte. La règle suivante agrège l'ensemble des consensus *unanimement* obtenus par l'ensemble de toutes les *majorités* formées sur $\mathcal{P}$:

$$m(\mathcal{P}) =_{def} \bigvee \bigwedge_{\{S \in 2^{\mathcal{P}} || S| \geq |\mathcal{P}|/2\}} S$$

Dans notre méthode *alg(.)*, $\mathcal{P}'$ conserve pour chaque partition, chacun des clusters qui n'est *strictement* pas inclus dans un cluster d'une autre partition de l'ensemble. Si l'on se représente l'ensemble des clusters de $\mathcal{P}$ comme un hypergraphe $\mathcal{H}$, alors l'hypergraphe $\mathcal{H}'$ résultant de $\mathcal{P}'$ forme une famille de Sperner telle que $\forall e, f \in \mathcal{H}', e \not\subset f$ et $f \not\subset e$. Les associations préservées dans le consensus est alors le résultat de l'intersection entre chaque paire d'hyperarêtes dans $\mathcal{H}'$:

$$\begin{array}{rl} alg(\mathcal{P}) & =_{def} \bigvee_{\{(x,y) \in 2^{\mathcal{P}'}\}} (x \wedge y) \\ \text{où } \mathcal{P}' & = \{\bigwedge_{\{y_i \in P - x\}} (y_i \rightarrow x) \mid x \in \mathcal{P}\} \end{array}$$

Considérant l'ensemble de partitions $\mathcal{P} = \{P_1, P_2, P_3, P_4, P_5\}$ avec $P_1 = 147|2|356$, $P_2 = 1234|57|6$, $P_3 = 126|3|47|5$, $P_4 = 123|4567$, $P_5 = 124|35|67$ et le multiensemble $\mathcal{P}_m = \mathcal{P} \cup \{P_1\}$.

Jeu de partitions	$u(.)$	$m(.)$	$alg(.)$
$\mathcal{P}$	$\perp$	$12\|3\|4\|5\|6\|7$	$12\|3\|47\|56$
$\mathcal{P}_m$	$\perp$	$1247\|3\|5\|6$	$1247\|356$

FIG. 1 – *Comparaison des résultats obtenues pour chaque méthode et sur les deux versions du jeu.*

Concernant le premier jeu, Fig. 1 montre que notre méthode est compatible (*i.e.* par la relation d'ordre) avec un nombre inférieur de partitions mais en revanche, elle intègre un plus grand nombre d'associations qui demeurent compatibles avec celles déjà incluses. La même chose est observable en dupliquant la première partition, cependant cela est dû en grande partie au fait qu'aucun cluster de P_1 ne sera filtré. P_1 devient librement combinable avec les clusters préservés dans les autres partitions et on obtient $alg(\mathcal{P}_m) = p_1 \vee alg(\mathcal{P})$. Par ailleurs, ce comportement peut devenir préjudiciable si le nombre de partitions identiques s'accroît.

3 Conclusion et Perspectives

Nous avons proposé une approche constructiviste dans le cadre de la définition réalisant partiellement l'adjonction avec les opérateurs traditionnelles du treillis et pour lesquelles, il n'existe pas de cadre formel décrivant l'interprétation de la dualité les liant. Par ailleurs, nous avons proposé la modélisation du problème de consensus entre des partitions par l'usage de l'un de nos opérateurs.

Identifier explicitement les contextes hypothétiques dans lequel une partition peut-être *vraie*, semble une voie prometteuse. En effet, la construction de celles-ci procède invariablement d'une étape d'uniformisation et qui résulte en une perte d'information, susceptible de

préciser le contexte initial. Un procédé abductif serait donc en mesure de chercher des corrélations entre une partition vis-à-vis des contextes validant ou réfutant une autre partition afin d'affiner le processus propre à la recherche d'une consensus à partir de partitions.

Références

Arrow, K. J. (1951). *Social Choice and Individual Values*. New York, NY : John Wiley and Sons.

Barthelemy, J.-P. et B. Leclerc (1995). The median procedure for partition. *AMS DIMACS Series in Discrete Math. 19*, 3–34.

Birkhoff, G. (1937). Rings of sets. *Duke Mathematical Journal 3*, 443–454.

Dudoit, S. et J. Fridlyand (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics 19*(9), 1090–1099.

Dumonceaux, F., G. Raschia, et M. Gelgon (2013). Consensus entre partitions : un point de vue algébrique dans le treillis des partitions. Technical report.

Hong, Y., S. Kwong, Y. Chang, et Q. Ren (2008). Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm. *Pattern Recognition 41*(9), 2742–2756.

Hu, X. et I. Yoo (2004). Cluster ensemble and its applications in gene expression analysis. In *Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29*, pp. 297–302. Australian Computer Society, Inc.

Strehl, A. et J. Ghosh (2003). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research 3*, 583–617.

Topchy, A., A. K. Jain, et W. Punch (2005). Clustering ensembles : Models of consensus and weak partitions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 27*(12), 1866–1881.

Von Der Gablentz, W., M. Koppen, et E. Dimitriadou (2000). Robust clustering by evolutionary computation. In *Proceedings of the 5th Online World Conference on Soft Computing in Industrial Applications (CDROM)*. Citeseer.

Summary

In clustering, consensus clustering aims at providing a single partition fitting a consensus from a set of independently generated. Common procedures, which are mainly statistical and graph-based, are recognized for their robustness and ability to scale-up. In this paper, we provide a complementary and original viewpoint over consensus clustering, by means of algebraic definitions which allow to ascertain the nature of available inferences in a systematic approach (e.g. a knowledge base). We found our approach on the lattice of partitions, for which we shall disclose how some operators can be added with the aim to express a formula representing the consensus.