Partition consensus

Alain Guénoche, Laurence Reboul

IML - CNRS ; 163 Av. de Luminy, 13288 Marseille cedex 9
guenoche,reboul@iml.univ-mrs.fr

Nombreuses sont les situations où l'on dispose d'un ensemble de partitions, dit *profil*, sur un même ensemble X. L'objectif est de construire une partition *consensus* qui possède un nombre maximum de paires réunies et de paires séparées dans le profil. Ceci est équivalent au problème de la partition médiane Régnier (1965), Barthélemy & Leclerc (1995), dont la somme des distances aux partitions du profil est minimum.

1 Un problème d'optimisation

Soit X un ensemble fini de n éléments, $\mathcal{P}$ l'ensemble de toutes les partitions de X et $\Pi \subset \mathcal{P}$ un *profil* de m partitions (classes disjointes, non vides, et dont l'union est égale à X). Pour une partition donnée $P \in \mathcal{P}$, tout élément $x_i \in X$ appartient à la classe notée $P(i)$. Les partitions sur X étant des relations d'équivalence sur les paires (d'éléments de X), on mesure l'écart entre deux partitions P et Q par la distance de la différence symétrique Δ entre ces relations et leur similitude par $S(P,Q) = \frac{n(n-1)}{2} - |\Delta(P,Q)|$. Le score d'une partition P relativement à un profil $\Pi = (P_1, \ldots, P_m)$ est défini par la somme des similitudes de P relativement à chacune des partitions du profil.

Etant donné un profil $\Pi = \{P_1, \ldots, P_m\}$, soit T_{ij} le nombre de partitions dans lesquelles deux éléments donnés x_i et x_j sont réunis et $R(P)$ l'ensemble des paires réunies dans P. Maximiser $S_\Pi(P)$ est équivalent à maximiser la quantité :

$$S'_\Pi(P) = \sum_{(i,j)\in R(P)} \Big(T_{ij} - \frac{m}{2}\Big). \tag{1}$$

Pour une partition P, une paire éléments réunis a une contribution positive (resp. négative) au critère quand ces deux éléments sont réunis dans plus (resp. moins) de la moitié des partitions de Π.

Soit $\mathbf{K}_n$ le graphe complet sur X dont les arêtes sont pondérées par $w(i,j) = T_{ij} - m/2$. Maximiser S'_Π revient à construire un ensemble de cliques disjointes qui soit de poids maximum. C'est une extension aux graphes pondérés du problème de Zahn (1971), bien connu pour être NP-difficile.

2 Méthode exacte et méthode approchée

Comme il est déjà signalé dans Régnier (1965), le problème de la partition consensus (ou centrale) est un problème d'optimisation discrète que l'on peut résoudre par un programme linéaire en nombres entiers à $n(n-1)/2$ inconnues et $O(n^3)$ contraintes. Le critère S'_Π s'écrit

$$S'_\Pi(\alpha) = \sum_{i<j} \alpha_{ij} w(i,j) \tag{2}$$

avec la contrainte que α est une relation d'équivalence sur X. Le problème d'optimisation revient donc à trouver α maximisant S'_Π sous les contraintes :

$$\begin{cases} \forall (i,j), \alpha_{ij} \in \{0,1\} \\ \forall (i,j,k), \alpha_{ij} + \alpha_{jk} - \alpha_{ik} \leq 1 \end{cases}$$

De nombreuses méthodes approchées ont été envisagées, à commencer par la *Méthode des transferts* proposée par Régnier. Nous avons défini et testé une méthode de Fusion-Transfert (FT) de complexité polynomiale $O(mn^2) + O(n^3)$.

Dans la partie Fusion, on part de la partition atomique P_0 et, à chaque étape, on réunit les deux classes qui maximisent la valeur de la partition résultante ; le processus s'arrête quand aucune fusion ne permet plus d'accroître le critère. Dans la partie transfert, on commence par calculer le poids de chaque élément comme sa contribution au critère. Ensuite on transfère les éléments dont la contribution est négative, et on les affecte soit à une autre classe à laquelle ils contribuent positivement (s'il en est), soit à une classe supplémentaire. Dans ce cas, ces éléments devenus singletons ont une contribution nulle au score, ce qui fait que le critère augmente.

Afin de comparer FT à la solution optimale, nous avons fait des simulations : on part d'une partition de X à n éléments en p classes équilibrées, c'est la partition initiale du profil. Ensuite, on génère $m-1$ partitions en appliquant à la partition initiale t transferts aléatoires, dans lesquels un élément pris au hasard est affecté à une classe de la partition en cours, ou à une nouvelle classe. Ainsi les partitions obtenues n'ont généralement pas le même nombre de classes.

A valeurs fixées pour n et m, suivant la valeur de t on peut obtenir des profils homogènes dans lesquels la partition initiale est la partition consensus ou des profils très dispersés pour lesquels la partition atomique est le plus souvent la partition consensus. Nous allons nous restreindre aux cas difficiles, avec $p = n/10$, $m = n$ et $t = n/2$. Sur 100 profils tirés au hasard suivant ces paramètres, nous mesurons :

- le score de la meilleure partition du profil ($S_{Best\Pi}$)
- le score de la partition optimale (S_{Opt}),
- le score de la partition calculée par FT (S_{FT}), et
- le pourcentage de problèmes pour lesquels FT a trouvé l'optimum ($\%_{FT}$).

qui prouvent l'efficacité de l'heuristique.

$n = m$	$S_{Best\Pi}$	S_{Opt}	S_{FT}	$\%_{FT}$
20	102.4	171.8	171.7	96
30	-88.2	193.9	193.8	93
50	-1059.3	166.6	166.0	86
100	-7348.6	68.0	67.9	96

Références

Barthélemy J.P., Leclerc, B. (1995) The median procedure for partitions, *DIMACs series in Discrete Mathematics and Thoretical Computer Science*, 19, 3-34

Régnier S. (1965) Sur quelques aspects mathématiques des problèmes de classification automatique, *Mathématiques et Sciences humaines*, 82, 1983, 13-29, reprint of *I.C.C. bulletin*, 4, 1965, 175-191.

Zahn C.T. (1971) Graph-theoretical methods for detecting and describing gelstalt clusters, *IEEE Trans. on Computers*, 20.