Génération de partition minimisant l'information partagée

Maurin NADAL

22 septembre 2011

1 Heuristique d'exploration par modifications

Le principe de cette heuristique est de partir d'un jeu de paramètre de base et de lui appliquer des modifications successives en vue de l'améliorer progressivement.

L'objectif principal sera de réalisé le plus grand nombre possible de modifications en parallèle. Il faudra pour cela évaluer plusieurs jeux de paramètres qui seront porteur de telle ou telle modification. La manière la plus simple de réaliser cela est d'appliquer une unique modification à chaque individu (ie jeu de paramètres) puis de comparer le score obtenu avec celui obtenu sans la modification. Cependant, cette manière de faire n'apporte aucune information sur les croisements entre différentes modification. De plus, il faut nécessairement évaluer 1 individus pour chaque modifications réalisées.

L'objectif de cette heuristique est donc d'appliquer plusieurs modifications à chaque individus en appliquant une répartition permettant de minimiser l'impact d'une modification sur le score d'une autre.

Pour cela, on réalise des bi-partitions de l'ensemble de nos individus. Si n est le nombre d'individu, chacune des deux parties de la bi-partition sera de cardinalité $\frac{n}{2}$. Par exemple, une bi-partition possible pour un ensemble à 4 individus serait : $\{1, 0, 1, 0\}$ et son complémentaire : $\{0, 1, 0, 1\}$. La modification qui serait associé à cette bi-partition serait donc appliquées aux individus 1 et 3, les 2 et 4 n'étant pas affectés.

De plus, on cherche à rendre constante la cardinalité de l'intersection de 2 bi-partition différentes. Ces intersections doivent être de taille $\frac{n}{4}$. Cela permet que la propriété suivante soit vérifiés, quelque soit 2 modifications A et B différentes, il y a autant d'individu non modifiés, que d'individus ayant que la modification A, que d'individus ayant la modification B et que d'individus ayant subi les 2 modifications.

Un corollaire de cette propriété est que lorsque l'on évalue le score de la modification A, la moitié des individus considérés ont subi la modification B et l'autre non. Il en va de même pour les individus témoins (ceux n'ayant pas subi la modification A). Cela permet de neutraliser l'impact de la modification B sur

le score de la A.

Quand on considère des bi-partitions à 4 individus, voilà celles qui peuvent être envisagées :

- 1100
- 1010
- 1001

Par convention, on considère que le premier élément est toujours sélectionné. Cette convention sera expliquée plus en détail dans la partie suivante. De plus, quand on considère le problème à 4 individus, la solution est triviale (toutes les bi-partitions respectant la convention ont une intersection de 1, donc toutes peuvent être sélectionnées simultanément).

Une fois une modification évaluée, il faut décider de son intérêt. Elle est alors soit "oubliée", soit appliquée à l'ensemble des individus (elle fait alors partie du jeu de paramètres de base). Une nouvelle modification pourra alors être testée en utilisant la répartition qui vient d'être libérée.

Le principal problème de cette heuristique est donc de générer les différentes répartitions des modifications sur les individus.

Introduction du problème

Le problème considéré est le suivant :

À partir d'un ensemble fini de taille n, on cherche à générer un maximum de bi-partitions équilibrées ($\frac{n}{2}$ éléments par partition) de cet ensemble en minimisant le nombre d'éléments communs à 2 parties de 2 partitions différents à $\frac{n}{4}$.

Exemple :
Soit $E = \{a, b, c, d, e, f, g, h\}$
et A, B et C trois bipartitions.
Avec $A = \{A_1, A_2\}$ ainsi répartis : $A_1 = \{a, b, c, d\} et A_2 = \{e, f, g, h\}$.

Les bi-partitions couvrent toujours entièrement E, il n'est donc finalement pas nécessaire d'exprimer la 2e moitiée de la partition, car elle correspond au complémentaire de la première part.

On a donc $B_1 = \{a, b, e, f\}$ et $C_1 = \{a, b, f, g\}$.

Par définition, aucun élément n'est commun entre une moitié de bi-partition et son complétementaire. Les cas à étudier seront donc les intersections entre par exemple A_1 et B_1, B_2, C_1, C_2. Dans le cas présents, les intersections contiennents toujours 2 éléments.

Il est possible de ne s'intéresser que aux intersections entre A_1 et B_1, C_1. En effet, le passage au complémentaire permet de connaître immédiatement la taille de $A_1 \cap B_2$ à partir $A_1 \cap B_1$ à l'aide de la relation suivante : $|A_1 \cap B_2| =$

$\frac{|E|}{2} - |A_1 \cap B_1|$. Cette règle permet de fixer le nombre d'élément commun souhaité. En effet, pour minimiser cette intersection, il faut prendre $\frac{|E|}{4}$, sinon la relation présentée précédement implique que l'intersection au complémentaire sera trop grande.

Enfin, il est possible de fixer comme règle que le premier élément de E appartiendra toujours à la première partie de la bi-partition. Si ce n'est pas le cas, cet élément appartient au complémentaire, et il suffit donc d'inverser les 2 partie de cette bi-partition pour respecter cette règle.

2 Problème équivalent : Ensemble indépendant maximum ou Clique maximale

Ce problème peut être ramené au problème classique de l'ensemble indépendant maximum ou à celui de sous clique maximale.

En effet, en se plaçant dans le cadre de l'exemple fourni en introduction, on considère le graphe complet $G = \{V_G, E_G\}$ avec
V_G = l'ensemble des sous-ensembles de E contenant a et de cardinalité 4 (soit $|E|/2$).
E_G = l'ensemble des arêtes étiquettées par la cardinalité de l'intersection des deux sommets.

Il est possible de considérer deux graphes induits de G. Tout d'abord, en ne conservant que les arêtes ayant 2 pour étiquette. Dans ce cas, le problème initial revient à recherche la plus grande clique de ce graphe. En effet, cela fournira une liste de moitié de bi-partition ayant toutes que 2 éléments en communs, cela sera donc aussi le cas pour les complémentaires.

Le deuxième graphe induit intéressant correspond au problème dual. On ne conserve que les arêtes ayant une étiquette différente de 2 (donc 1 ou 3) et on recherche le plus grand ensemble indépendant de ce graphe.

3 Extension possible du problème

Plusieurs extensions à ce problème devront être envisagées. Tout d'abord, il est possible de considérer des bi-partitions non-équilibrées. Cela pose deux problèmes. Tout d'abord, pour la partition plus petite, soit P_1 cette partie, si on imagine que $|P_1| = 3$ et qu'il existe une bi-partition A telle que $A_1 \cap P_1 = 2$. Alors l'information partagée par P_1 avec A_1 représentera 2/3 de l'information contenue dans P_1. Le problème symétrique se pose pour P_2 qui sera de cardinalité 5 et qui contiendra donc trop d'élément présents dans les autres bi-partitions.

Une autre extension consiste a ne pas réaliser des partitions couvrant complètement l'ensemble initial. Cela permettrait de diminuer fortement l'information partagées

entre les différentes parties de ces bi-partitions, mais limiterait le nombre de partition qu'il est possible de créer.

4 Théorème

Théorème 1. *Soit un ensemble E de taille 2^n. Alors, le nombre maximal de bi-partition équilibrée n'ayant entre elles que 2^{n-2} éléments en commun est $2^n - 1$.*

Démonstration. On considérera dans le cadre de cette preuve que les ensembles sont représentés par un tableau de 2^n bits. Si l'élément est présent, le bit associé est a 1, si il est absent, le bit est à 0. Cela permet d'envisager facilement des permutations permettant de remplacer un sommet par un autre.

Lemme 1.1. *Le graphe des bi-partitions ayant une intersection de 2^{n-2} éléments est sommet-transitif.*

Démonstration. Toute permutation des éléments dans l'ensemble de base correspond à un auto-morphisme sur les sommets du graphe. De plus, tous les sommets ayant un même nombre d'éléments, il est possible de construire une permutation qui permet de passer d'un sommet à un autre. La permutation la plus simple est de conserver l'ordonnancement des éléments, ie. soit 2 sommets A et B, la permutation échange le premier élément de A avec le premier de B, le deuxième de A avec le deuxième de B et ainsi de suite.

Donc, $\forall A, B \in V \exists P$ telle que $P(A) = B$.

Ce graphe est donc bien sommet transitif.

□

Lemme 1.2. *Le graphe des bi-partitions ayant une intersection de 2^{n-2} éléments est arête-transitif.*

Démonstration. Soit u_1 et u_2 deux sommets quelconques du graphes relié par une arête. Ils ont donc 2^{n-4} éléments en communs. Soit v_1 et v_2 deux autres sommets du graphes, eux aussi reliés par une arête.
Il existe au moins une permutation P telle que $P(u_1) = v_1$ et $P(u_2) = v_2$.
Cette permutation se construit de la manière suivante :

1. Les 2^{n-2} éléments en communs entre u_1 et u_2 sont envoyés vers l'intersection de v_1 et v_2.
2. Les 2^{n-2} éléments n'appartenant que a u_1 sont envoyés sur les éléments n'appartenant que à v_1.
3. Les 2^{n-2} éléments n'appartenant que à u_2 sont envoyés sur ceux n'appartenant que à v_2.
4. Les 2^{n-2} éléments n'appartenant ni à u_1 ni à u_2 sont envoyés sur ceux n'appartenant ni à v_1 ni à v_2.

L'agencement interne de ces différentes parties de la permutation n'est pas important.

Donc, par contruction, cette permutation existe effectivement. Le graphe est donc bien arête-transitif. □

Lemme 1.3. *Au moins une des cliques maximales contient le sommet u_n contenant uniquement les 2^{n-1} premiers éléments de l'ensemble (soit 2^{n-1} uns suivi de 2^{n-1} zéros dans la représentation sous la forme d'un tableau de bits). On appellera cette clique C_n.*

Démonstration. Le graphe considéré est arête-transitif, il est donc aussi sommet-transitif. Soit C la clique maximale, on considère une permutation P qui envoi un des sommet de C sur u_n.
La clique maximale $C_n = P(C)$ existe donc bien et respecte la condition demandée par le lemme.

□

Lemme 1.4. *Sauf le sommet u_n, tous les sommets de C_n sont équilibrés, ie. ils contiennent autant d'éléments appartenant à la première moitié de l'ensemble qu'à la deuxième moitié (soit 2^{n-2} éléments de chaque moitiés).*

Démonstration. On sait que chaque sommet de C_n à une intersection de 2^{n-2} avec chaque de ses voisins. De ce fait, chaque sommet de C_n à une intersection de 2^{n-2} sommets avec u_n. Comme u_n ne contient que les 2^{n-1} premiers éléments de l'ensemble complet, on en déduit que tous ses voisins contiennent uniquement 2^{n-2} éléments appartenant à cette première moitié. Comme chaque sommet contient 2^{n-1} éléments, on en déduit qu'ils contiennent de même 2^{n-2} éléments appartenant à la deuxième moitié de l'ensemble. □

A partir de ce point, il est possible de s'intéresser aux moitiés de chaque partition. En effet, dans la clique C_n, hormis u_n, chaque moitié de sommet contient 2^{n-2} éléments pris dans un ensemble de 2^{n-1} éléments.

Définition 1. *Une **clique parfaitement équilibrée** est une clique dont tous les sommets sont parfaitement équilibrés.*

Définition 2. *Un **sommet parfaitement équilibré** est un sommet dont l'ensemble associé vérifie la propriété suivante :*

- *Chaque intervalle de la forme $[k * 2^i, (k+1) * 2^i]$ avec $i \in [\![1, log_2(|E|)]\!]$ contient 0 ou 2^j éléments avec $j \in [\![0, i]\!]$.*

Lemme 1.5. *Quelque soit C une clique de G, il existe une permutation P telle que $P(C)$ soit une clique parfaitement équilibrée.*

Démonstration. Cette preuve va reposer sur la construction d'une telle permutation.

La première étape va consister à créer un sommet de la forme 2^{n-1} uns suivis de 2^{n-1} zéros. Pour cela, on sélectionne un sommet de la clique, on construit progressivement P en permutant chaque "1" de la 2^e moitié avec un zéro de la

première.

On obtient alors une clique pour laquelle hormis le sommet sélectionné qui est maintenant identique à u_n, tous les sommets contiennent autant de 1 dans la première moitié que dans la deuxième (soit 2^{n-2} 1 dans chaque moitié).

Nous allons maintenant réalisé des permutations qui permettent de conserver u_n, ie qui ne permutent que des éléments à l'intérieur de chaque moitié (voir 1.4).

Notre objectif va maintenant être de faire apparaître le sommet suivant : 1...10...01...10...0. Pour cela, nous traitons tout d'abord la première moitié. Nous sélectionnons un sommet de la clique différent de u_n et nous permutons tous les 1 de sont 2^e quart avec les 0 de son 1^{er} quart. Ensuite nous réalisons la même opération avec la 2^e moitié. Nous appellerons par la suite ce sommet u_1.

Hormis u_n, que nous ne prendrons maintenant plus en compte (tout en s'arrangeant pour que les permutations effectuées par la suite le laissent invariant), Tous les sommets ont dans chaque quart, soit $0, 2^{n-3}$ ou 2^{n-2} 1 dans chaque quart. Un quart contenant 0 éléments est alors toujours associé à un quart en contenant 2^{n-2} afin que la moitié correspondant à ces 2 quarts contiennent 2^{n-2} éléments.

Cela se justifie de la manière suivante :
On sait que l'intersection de deux sommets de la clique est de taille 2^{n-2} (cf. définition du graphe). En particulier l'intersection de tous les autres sommets avec u_1. On en déduit que chaque sommet contient 2^{n-2} éléments dans l'union de son premier et son troisième quart. Chaque sommet contenant 2^{n-1} éléments, on en déduit que l'union du deuxième et le quatrième quart contiennent aussi 2^{n-2} éléments.

Donc hormis u_n et u_1, tous les sommets vérifient le système suivant (avec a, b, c et d représentant respectivement le nombre d'éléments du premier, deuxième, troisième et quatrième quart) :

$$a + b = c + d \text{ et } a + c = b + d$$

On en déduit que $a = d$ et $b = c$ (par substitution).
Deux cas sont alors possible, soit $a = d = 2^{n-2}$ et $b = c = 0$ ou $a = b = c = d = 2^{n-3}$. Le cas $b = c = 2^{n-2}$ est impossible car il a été défini au début de ce document que l'on prenait toujours le premier élément de l'ensemble (afin de se démarquer des complémentaires qui ne l'ont jamais).

Les étapes suivantes se font exactement de la même manière en découpant les quarts en 8e, puis en 16e, jusqu'a arriver à des parties de 2 éléments qui sont soit égales à $00, 01, 10 ou 11$, tout en conservant la règle du premier élément à 1. Les permutations sont toujours restreintes alors à l'intervalle dans lequel on travail afin de ne pas détruire le travail réalisé dans les étages supérieurs.

La permutation ainsi construite permet alors de passer de la clique initiale à une clique parfaitement équilibrée.

□

Lemme 1.6. *Il existe des cliques parfaitement équilibrées de tailles* $2^n - 1$.

Démonstration. Il est possible de construire récursivement une telle clique à partir d'une clique issue d'un ensemble de taille 2^{n-1}.

Supposons qu'il existe une clique parfaitement équilibrée de taille $2^{n-1} - 1$ associée à un ensemble de taille 2^{n-1}.
Dans ce cas, tout sommet de cette clique à une intersection de 2^{n-3} avec tous les autres sommets de cette clique et avec tous leurs complémentaires.

La construction de la nouvelle clique va se faire moitié par moitié. Tout d'abord, pour la première moitié, on prend chaque sommet de la clique précédente 2 fois. On associe ensuite à chaque doublon le même sommet pour le premier élément et son complémentaire pour le deuxième.

Ainsi, les intersections entre des sommets ayant une première moitié différente est égal à $2^{n-3} * 2 = 2^{n-2}$. En effet, toutes les premières moitiés différentes ont une intersection de 2^{n-3} et il en va de même pour la deuxième moitié.

Pour deux sommets partageant la même première moitié, elles ont une intersection de 2^{n-2} puisqu'elles sont identiques et de taille 2^{n-2}. L'intersection de leurs deuxièmes moitiés est vide puisque ces dernières sont complémentaires.

On a donc ainsi construit $(2^{n-1} - 1) * 2 = 2^n - 2$ sommets ainsi, qui ont tous deux à deux une intersection de taille 2^{n-2}

Toutes les sommets contiennent exactement 2^{n-2} éléments dans leur première moitié. De ce fait, on peut ajouter le sommet 1...10...0 composés de 2^{n-1} 1 suivis de 2^{n-1} 0. Donc ce sommet aura une intersection de 2^{n-2} avec tous les autres sommets, il peut donc être ajouté à la clique.

La composition d'un sommet à partir de deux moitiés parfaitement équilibrées forme un sommet parfaitement équilibré. De plus, le dernier sommet que nous ajoutons est lui aussi parfaitement équilibrée, donc nous obtenons bien une clique parfaitement équilibrée de taille $2^n - 1$.

L'initialisation de cette récurrence se fait à partir de la clique parfaitement équilibrée associée à un ensemble de taille 2. Le seul sommet de cette clique est 10, c'est aussi la seule bipartition réalisable (son complémentaire 01 est la seule autre possibilité de choisir un élément parmi 2).

On a donc bien pour un ensemble de taille 2^n une clique parfaitement équilibrée de taille $2^n - 1$.

□

Lemme 1.7. *La répartitions des éléments de l'ensemble dans cette clique est homogène, ie. chaque élément (hormis le premier) est contenu dans un même nombre de sommet, plus exactement dans* $2^{n-1} - 1$ *sommets.*

Démonstration. On suppose que dans la clique au rang $n-1$, les éléments sont homogènement répartis dans les différents sommets. C'est à dire que le premier élément appartient à tous les sommets (par convention) et tous les autres appartiennent à 2^{n-2} sommets.

Nous avons vu dans le lemme précédent que la clique de rang suivant se construisait moitié par moitié. Nous allons donc dans un premier temps étudier les éléments appartenant à la première moitié.

Le principe de construction de la première moitié est d'utiliser deux fois chaque sommet de la clique de rang précédent. Le premier élément sera donc bel et bien dans tous les sommets. Pour tous les autres, ils seront présent deux fois plus de fois que dans la clique de rang précédent, soit $(2^{n-2}-1)*2 = 2^{n-1}-2$.

De plus le dernier sommet ajouté, qui contient uniquement les 2^{n-1} premiers éléments ajoute 1 sélections à chaque éléments de la première moitié. Donc l'ensemble de ces éléments sont sélectionnés $2^{n-1}-2+1 = 2^{n-1}-1$ fois.

La propriété est donc vérifié pour la première moitié. Nous allons maintenant nous intéresser à la seconde.

Cette deuxième moitié est construite en prenant chaque sommet de la clique de rang inférieur une fois, et en prenant son complémentaire ensuite. Ces deux motifs associés contiennent exactement une fois chaque élément (ie $\forall A, A \cup \overline{A} = 1...1$). $2^{n-1}-1$ tels groupes sont ainsi formés, donc chaque élément est sélectionné $2^{n-1}-1$ fois. Le dernier sommet ajouté à la clique (1...10...0) ne contient aucun élément de la deuxième moitié.

Donc, chaque élément hormis le premier est bien sélectionné $2^{n-1}-1$ fois au rang n (sous réserve de l'hypothèse de récurrence.

L'initialisation est très simple à faire au rang 2, la clique est alors la suivante :
- 1100
- 1010
- 1001

Chaque élément hormis le premier est bien sélectionné une seule fois et $2^{2-1}-1 = 2^1-1 = 1$.

□

Lemme 1.8. *Les cliques construites dans le lemme 1.6 sont maximales.*

Démonstration. Nous avons vu dans le lemme précédent que la répartition des éléments dans les différents sommets des cliques étaient homogènes. Nous allons montrer grâce à cela qu'il n'est alors plus possible d'ajouter de nouveau sommet à ces cliques.

Tout d'abord, rappelons que chaque sommet de la clique doit avoir une intersection de 2^{n-2} éléments avec chaque autre sommet au rang n. Si nous réalisons la somme des cardinalités des intersections entre un nouveau sommet et tous

ceux déjà présent dans la clique, nous obtenons donc : $(2^n - 1) * 2^{n-2}$. Les intersections liés au premier éléments ne sont pas pris en compte dans ce calcul afin de simplifier le raisonnement. Cependant, comme tous les sommets contiennent cet élément, cela n'a pas d'influence sur la valeur de la preuve.

D'un autre côté, le nouveau sommet sera construit à partir du premier élément auquel on ajoutera $2^{n-1} - 1$ autres éléments. Chacun de ces autres éléments va forcément "intersecter" $2^{n-1} - 1$ autres sommets, puisqu'on a vu dans le lemme précédent que chaque élément était déjà sélectionné $2^{n-1} - 1$ fois. Il est donc possible de calculer le nombre de ces intersections à l'aide du calcul suivant : $(2^{n-1} - 1) * (2^{n-1} - 1)$.

En posant l'égalité de ces 2 nombres d'intersections :
$(2^n - 1) * 2^{n-2} = (2^{n-1} - 1)^2$
$\Leftrightarrow 2^{2n-2} - 2^{n-2} = 2^{2n-2} - 2^n + 1$
$\Leftrightarrow 2^n - 2^{n-2} = 1$
$\Leftrightarrow 2^{n-2} * (4 - 1) = 1$
$\Leftrightarrow 2^{n-2} = \frac{1}{3}$

Cette équation n'a pas de solution pour n entier. On en déduit donc qu'il n'est pas possible d'ajouter un sommet à cette clique, elle est donc maximale.

□

L'objectif est maintenant de montrer que la clique équilibrée maximum est de même taille que celle que l'on vient de construire.

Pour cela, j'envisage plusieurs approches possibles :
- Soit de montrer que la clique ainsi construite est maximale, et ensuite que toutes les cliques parfaitement équilibrées maximales sont de même taille.
- Soit de montrer que l'on construit ainsi toutes les cliques parfaitement équilibrées et qu'elles sont donc de ce fait toutes maximales et de même taille.
- Soit de trouver des propriétés sur les intersections dans les cliques parfaitement équilibrées, cela nous permettrait peut être d'aboutir à la construction d'une permutation nous permettant de passer de la clique maximum à la clique que l'on vient de construire.

Lemme 1.9. *Toute clique parfaitement équilibrée maximale est de taille* $2^n - 1$.

□