Classification hiérarchique d'une matrice de distance avec contrainte d'adjacence

A. Dehman (a), P. Neuvial (a), G. Rigaill (c), M. Koskas (b) and C. Ambroise (a)

(a) Laboratoire de Mathématiques et Modélisation d'Evry, CNRS UMR 8071/Université d'Evry val d'Essonne/ENSIIE/USC INRA, 23 boulevard de France, 91000 Evry

{alia.dehman,pierre.neuvial,christophe.amrboise}@genopole.cnrs.fr

(b) Université d'Evry et Institut de Sciences des Plantes de Paris-Saclay

guillem.rigaill@univ-evry.fr

(c) INRA MIA/AgroParisTech, CNRS UMR 518

michel.koskas@agroparistech.fr

Mots clefs : Classification ascendante hiérarchique, classification sous contraintes spatiales, méthode de Ward, études d'association génome entier, déséquilibre de liaison.

Contexte: détection de blocs de déséquilibre de liaison dans les études d'association

Les études d'association génome entier (GWAS pour Genome-Wide Association Studies) visent à identifier des marqueurs génétiques associés à un trait phénotypique, par exemple une maladie. Les marqueurs génétiques étudiés sont généralement des polymorphismes d'un seul nucléotide (SNP pour Single Nucleotide Polymorphism). Les expériences de puces à ADN ou de séquençage permettent de mesurer le génotype d'un très grand nombre ($p \sim 10^5 - 10^6$) de SNP chez un grand nombre d'individus ($n \sim 10^2 - 10^4$). Ces p variables ont une structure de dépendance par blocs le long du génome, liée au phénomène de déséquilibre de liaison (DL) dû à la recombinaison génétique.

Nous avons récemment proposé une méthode permettant l'identification de blocs de LD associés à un phénotype d'intérêt [1]. Cette méthode repose sur une première étape de classification ascendante hiérarchique avec contrainte d'adjacence sur la base d'une similarité entre SNP induite par le LD. Une limitation pratique de cette méthode est que l'algorithme de classification est intrinsèquement quadratique en p, à la fois en temps et en espace. Cette complexité rend difficile, voire impossible, le traitement de problèmes où $p \sim 10^5 - 10^6$.

Idée: exploiter la structure en bande de la matrice des distances

Nous proposons d'exploiter une information biologique supplémentaire: le fait que la taille maximale h des blocs de DL est généralement inférieure à p de plusieurs ordres de grandeurs. Cette propriété biologique est illustrée sur la Figure 1, qui illustre la structure bloc-diagonale du DL entre les 50 premiers SNP du chromosome 22 dans le cas d'une étude sur le VIH [2].

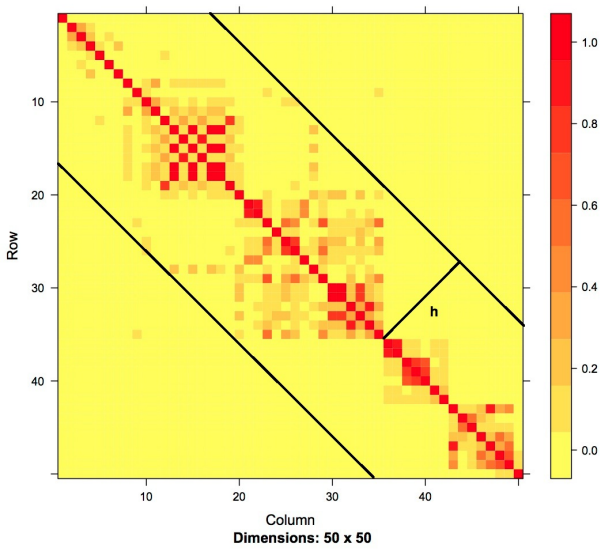

Figure 1: Structure bloc-diagonale du DL entre les 50 premiers SNP du chromosome 22 dans une population humaine [2].

Contribution: un algorithme de classification sous-quadratique

La prise en compte de cette information biologique permet d'obtenir un algorithme approché dont la complexité est sous-quadratique. Cet algorithme est donc applicable à des données où p est "grand".
L'algorithme proposé prend entrée la matrice (creuse) des distances entre tous les couples de variables dont les indices sont distants de moins de h, où h est fixé à l'avance. Grâce (i) au pré-calcul de certaines sommes cumulatives des distances, et (ii) à une structure de tas binaire pour le stockage des fusions successives entre classes, l'algorithme proposé a une complexité $O(p(h + \log(p)))$ en temps et $O(ph)$ en espace.

Nous proposons une implémentation en `R` de cet algorithme, faisant appel à du `C++`. Nos expériences numériques réalisées à partir de données réelles montrent que cet algorithme fournit une très bonne approximation de la solution obtenue sans la contrainte de bande sur-diagonale.

Références

[1] Dehman, A. Ambroise, C. and Neuvial, P. (2015). Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics* 16:148.

[2] Dalmasso, C *et al.* (2008) Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS Genome Wide Association 01 study. *PloS One* 3(12):3907.