Un algorithme génétique pour la MDS

Ngouenet Roger
I.R.I.S.A - REPCO
Campus de Beaulieu
35042 RENNES CEDEX
Email: ngouenet@irisa.fr

Resume

Le problème des représentations euclidiennes par les méthodes de la MDS (multidimensional scaling) nécessite un processus de minimisation d'une fonction numérique qui n'est généralement pas différentiable. En prenant en considération certaines propriétés inhérentes à la structure de l'espace où se trouve définie une telle fonction, nous examinons son optimisation par une famille d'algorithmes stochastiques: les algorithmes génétiques (A.G.).

Mots-Clé: MDS (Multidimensional scaling), optimisation, algorithmes génétiques (A.G.) , différentiable, méthodes du gradient.

La MDS (multidimensional scaling) désigne une classe des méthodes d'exploitation des tableaux numériques et symboliques en analyse des données. L'objectif de ces méthodes est de construire des représentations géométriques, dans un espace métrique $< \Omega, d >$ ayant un nombre relativement restreint de dimensions (idéalement inférieur ou égal à 3), à partir d'une matrice $\Delta = (\delta_{ij})$ de dissimilarités entre n entités $o_1, o_2, ..., o_n$. Les distances interpoints de la configuration obtenue doivent refléter les dissimilarités initiales.

Dans l'approche exploratoire, en faisant abstraction du travail lié aux transformations de la matrice de dissimilarité, ce problème se réduit à celui de l'optimisation d'une fonction numérique (fonction coût) sur un espace de description. Une des solutions les plus fécondes a été introduite par Kruskal [6,

1964] qui formalise ce problème en termes d'un ajustement au sens des moindres carrés et considère comme fonction coût

$$STRESS(X) = \left(\frac{\sum_i \sum_j (\delta_{ij} - d_{ij}(X))^2}{\sum_i \sum_j d_{ij}(X)^2} \right)^{\frac{1}{2}}$$

où

$$d_{ij}(X) = \left(\sum_{s=1}^{k} (x_{is} - x_{js})^2 \right)^{\frac{1}{2}}$$

X et k désignent respectivement la configuration cherchée et la dimension de l'espace de représentation.

Pour la minimisation de $STRESS(X)$ Kruskal utilise un algorithme itératif de résolution basé sur le gradient. Cette approche - fonction coût de type STRESS et optimisation par la méthode du gradient - est aussitôt préconisée par plusieurs auteurs tels que Young et Seery [7, 1981] dans le logiciel KYST ou encore Takane, Young et de Leeuw [7, 1981] dans le logiciel ALSCAL.

Pourtant, comme le font remarquer Heiser et de Leeuw [2, 1989] les fonctions coût de type $STRESS$ ne sont pas différentiables aux points X pour lesquels il existe au moins un couple de points (i , j) - i distinct de j - tel que $d_{ij}{}^2(X) = 0$. C'est ainsi que Heiser et de Leeuw [4, 1986] proposent dans un nouveau logiciel (SMACOF), un algorithme de résolution basé sur le sous-gradient. De plus, l'utilisation du gradient présente des difficultés si on veut que la distance interpoint corresponde à une métrique non euclidienne (e.g. métrique L_p de Minkowski).

Nous nous proposons, par conséquent, d'adapter à la minimisation des fonctions coût de type $STRESS(X)$ une famille d'algorithmes stochastiques récentes connue sous le nom d' algorithmes génétiques (A.G.). L'avantage essentiel est la grande souplesse quant à la nature de la fonction à optimiser et la structure de l'espace de définition de la fonction. Ainsi, elles ne tiennent compte ni de la différentiabilité de la fonction ni du nombre de ses sommets.

Les algorithmes génétiques sont des techniques d'inspiration biologique dont les champs d'application couvrent plusieurs domaines aussi complexes que variés: contrôle d'un système de pipeline [3, GOL83], cadrage de radiographies médicales [5, FVG84], connexions d'un réseau de neurones [1, WBS90], systèmes de classification [8, ROB87].

Ils permettent à une population d'individus d'évoluer comme le ferait un ensemble d'êtres vivants. Dans notre cas, un individu correspond à une configuration de points. Le principe des A.G. est d'appliquer à une population des transformations syntaxiques (copies , permutations et modifications) afin de produire une population mieux adaptée. Ils s'apparentent , ainsi, à la théorie de Darwin, sur l'évolution naturelle de la vie, selon laquelle " la vie est une compétition et seuls les mieux adaptés survivent et se reproduisent " !
Dans les A.G. simples, le passage d'une génération à la suivante se fait par l'exécution d'un cycle composé des trois opérateurs génétiques standards: reproduction, croisement, et mutation.

Les résultats expérimentaux obtenus sont très interessantes.

References

[1] and C. Bogard. D. Whitley, T. Starkweather. Genetic algorithms and neural networks: optimizing connections and connectivity. *Parallel Computing*, 14(3):347–361, August 1990.

[2] D. d'Aubigny. *L'analyse multidimensionnelle des donnees de dissimilarite.* PhD thesis, Universite Joseph Fourier, Grenoble-1, January 1989.

[3] D. E. Goldberg. *Computer-aided gas pipeline operation using genetic algorithms and rule learning.* PhD thesis, University of Michigan., 1983.

[4] W. Heiser. Smacof-1. *University of Leiden*, 1986.

[5] and D. Van Gucht. J. M. Fitzpatrick, J. J. Gefenstette. Image registration by genetic search. *In Proceedings of IEEE Southeast conference.*, 1984.

[6] J. B. Kruskal. Nonmetric multidimensional scaling. *Bell Telephone system*, 4821, June 1964.

[7] S. Schiffman M. Reynolds and W. Young. Introduction to multidimensional scaling, theory, methods and applications. *Academic Press*, 1981.

[8] G. Robertson. Parallel implementation of genetic algorithms in classifier system. *Genetic algorithms and simulated annealing*, 129–140, 1987.