Conférences invitées

Visualisation des données textuelles et inférence statistique

Ludovic Lebart,
CNRS, GET-ENST, 46 rue Barrault, 75013, Paris.
lebart@enst.fr
http://egsh.enst.fr/lebart/

Les outils de visualisation jouent un rôle important dans l'acquisition des connaissances à partir de données complexes, notamment celles qui se présentent sous la forme de textes. *Voir* et *comprendre* sont des verbes qui sont parfois synonymes, et l'organisation spatiale et géométrique des cartes, images ou graphes favorise l'assimilation, la mémorisation, et stimule l'imagination et la pensée.

Les principaux outils de visualisation de données multidimensionnelles sont d'une part les analyses en axes principaux, qui portent des noms divers mais qui sont presque toutes fondées sur la décomposition aux valeurs singulières, et d'autre part les méthodes de classification, incluant les cartes auto-organisées de Kohonen. Ces deux familles d'outils sont d'ailleurs complémentaires, et leur usage simultané s'impose lorsqu'il s'agit de traiter de très grands ensembles de données.

Comment valider les visualisations obtenues ? Ce sont les méthodes de simulation en général, et plus particulier les méthodes de re-échantillonnage et de bootstrap qui sont les plus opératoires actuellement.

Dans le domaine textuel, les difficultés viennent du fait que les unités statistiques sont diverses (caractères, mots, segments, phrases, unités de contexte, paragraphes, messages, rapports, chapitres, livres, locuteurs, etc.) et modulables en fonction de seuils de fréquences ou de prétraitement (analyseurs syntaxiques, segments) mais la souplesse du bootstrap permet de s'adapter à des situations variées.

On peut donc savoir si un diagramme représentant des corrélations ou des cooccurrences de mots n'est guère qu'un motif provenant d'un kaléidoscope aléatoire, ou au contraire l'image d'une structure réelle (ou plus vraisemblablement : on peut savoir quelle partie du diagramme mérite d'être interprétée).

Les zones de confiance statistiques que l'on peut tracer autour des points (cas des visualisations par axes principaux) ne seront utilisables en pratique que dans un environnement interactif. Il en est de même pour la mise en oeuvre des méthodes de validation externe (à partir de méta-données) et pour l'utilisation combinée des analyses en axes principaux et des méthodes de classification.