JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

LUDOVIC LEBART

Discussion et commentaires. Data mining et statistique

Journal de la société française de statistique, tome 142, n° 1 (2001), p. 73-76

<http://www.numdam.org/item?id=JSFS_2001__142_1_73_0>

DISCUSSION ET COMMENTAIRES

Data Mining et Statistique

Ludovic LEBART *

L'article de Philippe Besse, Caroline Le Gall, Nathalie Raimbault et Sophie Sarpy présente le point de vue d'un(e) statisticien(ne) sur les interactions et les éventuels recouvrements entre Data Mining et Statistique. D'emblée, je trouve cet article bien fait et utile, j'adhère à beaucoup de ses réflexions, et je pense que sa publication est opportune et correspond certainement à une attente des lecteurs du Journal de la SFdS.

Je commencerai cependant l'exercice de la discussion en exprimant quelques regrets : que pour un des premiers articles sur le data mining publié dans un journal représentatif de la communauté des statisticiens francophones ne soient pas mentionné le caractère précurseur de nombreux travaux réalisés en France au cours des trente dernières années sur le thème de l'analyse exploratoire automatique des données. Ces travaux furent réalisés à la suite des idées et impulsions de Jean-Paul Benzécri, qui voyait en l'ordinateur un «novius organum» capable d'extraire «de la gangue des données le pur diamant de la véridique nature», (formule ayant donné lieu à maintes polémiques en son temps) et affirmait : «Ce dont nous avons besoin, c'est d'une méthode rigoureuse qui extraie des structures à partir des données» (deuxième principe de l'analyse des données, in : L'analyse des données, Tome II A, chapitre 1). Il faut lire la contribution de J.P. Benzécri à l'Organum de l'Encyclopaedia Universalis («La place de l'a priori», 1974) et les premiers chapitres des tomes 1 et 2 de son traité d'analyse des données (Dunod, 1973) pour se convaincre de son indéniable rôle de précurseur en la matière. Pourquoi ce courant - très diversifié par la suite - a eu du mal à s'exporter, et pourquoi, malgré son caractère stimulant et fécondant dans la communauté des statisticiens français, il n'a pas eu le statut académique attendu,... voilà l'objet d'un autre article. D'ailleurs certaines observations de Philippe Besse et de ses collaborateurs mettant en jeu la statistique, le data mining et les institutions universitaires pourraient également contribuer à un tel débat. J'approuve en effet sans réserve leurs remarques, à l'appui du rapport sur la statistique de l'Académie des Sciences, sur ce qu'ils désignent par le malaise de la statistique en France (fin du paragraphe 2.5).

* ENST, 46 rue Barrault, 75013, Paris
lebart@enst.fr.

Le problème de la stratégie

Les revues de statistique fourmillent de méthodes nouvelles, ou de variantes des méthodes actuelles, adaptées à des situations particulières : combien de variantes de la régression, des méthodes factorielles, des méthodes de classification ? Quelques centaines au bas mot. Comment un utilisateur peut-il s'y retrouver ? A quel saint se vouer ? Mais surtout où trouver le logiciel idoine ? Et, plus que tout : le détour et l'investissement méthodologique en valent-ils la peine ? L'achat d'un logiciel spécialisé (si celui-ci existe) est-elle justifié ? Quoi de plus chronophage, en effet, qu'une prise de contact avec un nouveau logiciel !

Je crois que nous, statisticiens, ne remplissons pas toujours notre contrat moral de ce point de vue vis-à-vis du public des utilisateurs, faute de moyen et de temps, et peut-être aussi faute de valorisation de cette partie de notre activité. Nous fournissons une flore de méthodes, mais pas de stratégie ni d'outils polyvalents permettant à un utilisateur non-statisticien de choisir la bonne méthode. Une niche est alors laissée libre pour des logiciels généralistes opérant une sélection des méthodes les plus courantes et les plus versatiles. Voilà une des fonctions que remplissent les logiciels de data mining, fonction qui correspond à un besoin et à une demande économique parfaitement justifiée dans l'état actuel des disciplines et de leurs développements.

Le problème des logiciels, la rigidité, la déqualification

Je reprends ici, en d'autres termes et d'un point de vue légèrement différent, les excellentes remarques du paragraphe « automatisation » (section 2.5 de l'article).

Cette « stratégie incorporée » et cette relative facilité d'utilisation peuvent aussi s'accompagner - implicitement - d'un certain déni de la culture statistique, de ses cadres conceptuels. En bref, il serait possible de se passer des services des statisticiens, un peu (un peu seulement) comme les traitements de textes, tableurs, logiciels graphiques et de présentation ont pratiquement fait disparaître les dactylos, les assistants de recherche, les dessinateurs. Ce n'est pas être corporatiste que de dire qu'il s'agirait là d'une dangereuse déviation. La réalité ne cesse pas d'être complexe parce qu'on utilise, pour l'analyser, un logiciel convivial. Explorer un gigantesque domaine vierge de toute analyse avec un logiciel puissant est une opération dans laquelle l'utilisation d'un logiciel de data mining sera éminemment profitable lors d'une première approche... mais utiliser les mêmes outils sur les plates-bandes délicates de certains plans d'expérience, sur de petits échantillons laborieusement constitués, des données censurées, des processus modélisables, c'est s'exposer à des risques ou à des déceptions.

On peut distinguer le « sur mesure » des chercheurs en statistique élaborant ou créant leurs propres programmes (université, écoles, centres de recherche),

le « prêt-à-porter » des grands logiciels statistiques présentant de multiples options (groupe industriels, agronomie, biopharmacie, grandes administrations), et le prêt à porter grand public (petites entreprises, services autonomes des grandes entreprises ne comportant pas de statisticiens, consultants isolés) auxquels seraient plus particulièrement dédiés les logiciels les plus courants de data mining.

Le carcan d'un logiciel optimisé, mais simplifié, a pour contrepartie de rendre l'utilisateur prisonnier d'une panoplie restreinte. Les auteurs de l'article qui est l'objet de notre discussion sont parfaitement conscients de cette faiblesse, mais n'y échappent pas. Il y est dit, section 3.3 : « L'analyse discriminante et ses variantes n'ont pas donné de bons résultats ; absente de la version basique de SEM (2001), elle est laissée de côté ». Indépendamment de questions de terminologie (certains auteurs considèrent que la régression logistique et le perceptron multicouche sont aussi des méthodes d'analyse discriminante, qui coïncident d'ailleurs pratiquement avec l'analyse discriminante de Fisher dans certaines configurations) peut-on, dans une analyse comparative, laisser de côté l'une des méthodes statistiques les plus utilisées (*cf.* l'article de Gnanadesikan, Discriminant Analysis and Clustering, dans Statistical Science, n° 4, 1989) et fournissant des résultats dont la robustesse est rarement démentie (*cf.* par exemple, Bardos, Analyse discriminante, Dunod, 2001) ? Peut-on se contenter de dire qu'elle a fourni de mauvais résultats, alors qu'il s'agirait pratiquement d'un contre exemple peu courant pour ce « Jack of all trades » qui est quasiment une norme et un repère dans les problèmes de discrimination, de reconnaissance des formes ? La compétence statistique reconnue des auteurs ne fait que mettre en relief le caractère insidieux de ce type de biais logiciel. Une application statistique se fait toujours avec des contraintes de temps et de moyens, et le choix des outils est rarement sans influence sur les résultats.

Recherche en data mining et pratique du data mining

Je crois par ailleurs qu'il faut bien distinguer la recherche en data mining et la pratique du data mining. Les statisticiens n'aiment pas être jugés à l'aune des usages inconsidérés de leur discipline, accordons la même faveur aux « data miners ». Cela dit, la pratique du data mining (le contenu des logiciels les plus diffusés) fait beaucoup penser à une importation sous un autre nom des techniques d'analyse des données (incluant un branchement sur les bases de données qui est maintenant indispensable), avec de nouveau les erreurs de jeunesse contre lesquelles nous nous sommes battus au cours des décennies précédentes et dont certaines sont d'ailleurs relevées par les auteurs de l'article (pratiquement pas d'épreuves de validité pour la partie exploratoire, monomanie pour la méthode à la mode, peu d'attention portée à la sémantique ni à la qualité des données de base).

En revanche, la recherche en data mining se situe dans un paradigme qui se révèle très fécond et stimulant qui est celui de la « découverte de connaissance

dans les bases de données » (KDD). Comme on le sait depuis Thomas Kuhn, les travaux autour d'un paradigme nouveau ont tendance à s'ériger en discipline autonome et à bien marquer leurs frontières, phénomène bien naturel déjà observé pour le connexionisme. L'idée d'extraire des connaissances (et non plus seulement des patterns ou des structures) est stimulante, comme l'idée de traiter des connaissances au même titre que des données en analyse des données symboliques. Je ne crois pas personnellement que des résultats spectaculaires aient déjà été obtenus, mais certaines avancées sont incontestables, d'autres imminentes.

La statistique n'est pas essoufflée. Au contraire, son domaine est devenu tellement vaste qu'elle a du mal à contrôler ses frontières, surtout quand ses voisins sont dynamiques, voire enthousiastes. Il faut espérer que des confrontations de cette nature renouvellent nos problématiques, nous incitent à une critique interne sans complaisance, nous enrichissent sans frilosité aucune ! Les réseaux de neurones sont en grande partie devenus une branche de la statistique, malgré les redondances que l'on sait en matière de concepts et de notations, tout simplement parce que l'on ne peut pas calculer une erreur de prévision sans probabilité.

On peut dire, en lisant les applications variées et les commentaires sereins de l'article de Philippe Besse et ses collaborateurs, que le data mining, qui fait largement appel aux outils statistiques, appartient déjà à notre nébuleuse. La statistique n'étant pas un lieu de pouvoir (surtout en France !) cet appartenance n'est pas l'impérialisme d'une discipline, mais l'hospitalité d'une culture.