DISCUSSION ET COMMENTAIRES

Data Mining et Statistique

Gilbert SAPORTA *

L'article de P.Besse *et al.* vient à point nommé pour alimenter le débat sur la nature du Data Mining : est-ce une nouvelle discipline ? un simple effet de mode ? de la statistique sous un autre nom ? Voici les réflexions que m'inspire cette contribution.

À la fréquentation des forums, salons, en regardant offres d'emploi et organigrammes de sociétés, on s'aperçoit qu'il est de plus en plus fréquent d'entendre parler de Data Mining que de Statistique : des services statistiques de grandes banques sont devenus des départements de Data Mining, il semble plus valorisant de dire que l'on utilise des outils de Data Mining qu'un logiciel statistique d'analyse discriminante, etc. Indépendamment du succès habituel en France des termes anglo-américains importés qui donnent un air de modernité à ceux qui les emploient, on peut se demander si ce n'est pas dû au fait que le mot statistique fait peur, du moins une certaine statistique trop académique et coupée des réalités. En sens inverse le Data Mining irrite nombre de statisticiens pour des raisons diverses tenant à son caractère exploratoire et non lié à une théorie (j'y reviendrai plus loin) ainsi qu'à la peur qu'évoque l'utilisation d'outils automatiques de détection de structures cachées. De telles appréhensions sont irrationnelles et renvoient à des débats déjà anciens lorsque les premiers logiciels de statistique sont apparus : que n'a-t-on pas entendu alors ! le métier de statisticien était condamné puisque tout un chacun pouvait faire de la statistique sans formation...

Le temps est passé et on s'est aperçu au contraire que l'on avait besoin de plus en plus de spécialistes formés à la statistique, mais maîtrisant en outre l'informatique. Notre secteur s'est développé, il suffit de compter le nombre de formations de statisticiens professionnels existant maintenant dont beaucoup se sont créées ces dernières années : 11 départements STID d'IUT, 4 licences professionnelles, 7 IUP, 16 DESS, 3 grandes écoles, sont recensés sur le site internet de la SFdS.

Les logiciels de Data Mining apportent un nouveau confort et des possibilités accrues, en particulier de comparaison de modèles. L'évolution vers plus de convivialité et de puissance ne peut en aucun cas être un inconvénient quand elle contribue à diminuer certaines tâches fastidieuses.

---

* Chaire de Statistique Appliquée Conservatoire National des Arts et Métiers. 292 rue Saint Martin, 75141 Paris cedex 03
e-mail : saporta@cnam.fr

Venons-en aux objectifs et aux méthodes : deux des définitions les plus couramment acceptées font référence à la découverte de faits cachés, inattendus et économiquement exploitables, enfouis sous les montagnes de données que sont les gigabases qui sont souvent alimentées automatiquement. « *Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* » (U.M. Fayyad). « *I shall define Data Mining as the discovery of interesting, unexpected, or valuable structures in large data sets* » (D.J. Hand). La métaphore du Data Mining est alors claire : avec les bons outils on va découvrir les pépites. Cette ambition n'est pas nouvelle et une large partie de la statistique la fait sienne : c'était en particulier l'ambition des fondateurs de l'analyse exploratoire des données, et la citation suivante de J.P. Benzécri en 1973 peut être reprise : « *L'analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature* ». Comme le rappelait en 1997 J. Kettenring, alors président de l'American Statistical Association : « *Statistics is the science of learning from data. Statistics is essential for the proper running of government, central to decision making in industry, and a core component of modern educational curricula at all level.* »

Cependant on s'aperçoit bien vite que les principales applications du Data Mining concernent plus la recherche de modèles que de structures ou « patterns » comme les définit D. Hand. Un modèle est une représentation simplifiée des relations existant entre variables dans un but de synthèse et de prévision, tandis que pour la découverte de « patterns » « *the aim is not to build an overall global descriptive model, but is rather to detect peculiarities, anomalies, or simply unusual or interesting patterns in the data.* »

Modéliser a toujours été une activité essentielle des statisticiens et on peut se demander en effet ce qu'il y a de nouveau avec le Data Mining. En quoi les exemples donnés dans l'article de P. Besse *et al.* diffèrent-ils d'une analyse statistique classique, d'autant qu'il s'agit de petits jeux de données ?

La réponse est qu'en Data Mining le modèle provient des données et n'est pas choisi *a priori*.

Dans la pratique statistique habituelle, le modèle découle d'une théorie (économique, physique, biologique..) et le but est d'estimer et de tester les paramètres du modèle. En Data Mining le modèle final vient après une exploration combinatoire d'un grand nombre de modèles : pour un problème de discrimination on comparera les performances de la fonction de Fisher, de la régression logistique, de réseaux de neurones, d'arbres de décision, etc., pour retenir le modèle le plus efficace. De plus il s'agit d'une *analyse secondaire* de bases de données recueillies à d'autres fins que statistiques (transactions, fichiers clients) : les modes usuels d'obtention des données comme les plans d'expériences ou les sondages sont en général absents.

On comprend alors les réticences des théoriciens et des spécialistes d'un domaine devant une telle attitude pragmatique ou empirique : le Data Mining ne vise pas à la connaissance scientifique mais à l'action : il est né de la préoccupation de rentabiliser les bases de données et trouve actuellement une de ses principales applications en entreprise avec le développement de

la Gestion de la Relation Client (en anglais CRM ou Customer Relationship Management) qui remplace le marketing produit par le marketing client. Il s'agit de comprendre et d'anticiper le comportement de clients tous différents pour les fidéliser, leur proposer des produits et services personnalisés, etc.

C'est sans doute une des raisons qui font que les Instituts Nationaux de Statistique ignorent le concept de Data Mining (mais pas certaines de ses méthodes). On y cherche en vain des publications avec ce mot-clé alors que ces Instituts produisent des montagnes de données. Il est clair que l'exploration sans *a priori* de bases de données est largement étrangère à la culture des statisticiens officiels et des économètres pour qui le modèle doit provenir d'une théorie économique et non de données somme toute particulières. Notons cependant qu'Eurostat (l'Office Statistique des Communautés Européennes) finance actuellement plusieurs projets de recherche en Data Mining pour les Instituts Nationaux de Statistique : KESO (Knowledge Extraction for Statistical Offices). SPIN (Spatial Mining for Data of Public Interest).

Que peut-on trouver et prouver réellement avec le Data Mining ?

L'utilisation d'une approche combinatoire entraîne le risque de trouver fatalement des relations inattendues mais non signifiantes. C'est alors qu'entre en jeu une nouvelle manière de faire de l'inférence pour de très grandes bases de données. Il est d'une part bien connu que les tests d'hypothèses traditionnels perdent leur sens car la plupart des hypothèses sont rejetées. Ainsi l'absence de corrélation sera rejetée dès que r dépasse 0.01 lorsque l'on dispose de plus de 400 000 observations. Une telle valeur de r est évidemment sans intérêt pratique. D'autre part la validité prédictive d'un modèle ne peut se juger seulement par son ajustement aux données qui ont permis de l'estimer, et les pénalisations du nombre de paramètres de type Akaïke ou Schwartz ne sont pas suffisantes pour choisir un modèle, d'autant plus qu'il faut spécifier des hypothèses de distribution pas toujours réalistes.

Pour les très grands jeux de données, la pratique de la séparation en trois ensembles (apprentissage, test, validation) permet de répondre à la question : on estime chaque modèle candidat sur l'ensemble d'apprentissage, on choisit le meilleur sur l'ensemble test, et on calcule sa précision sur l'ensemble de validation. Les travaux de V. Vapnik, qui sont des travaux de statistique mathématique, donnent des résultats fondamentaux liant l'erreur de généralisation à la «dimension» du modèle et à la taille des ensembles. Ils sont également à la base de nouvelles techniques d'apprentissage comme les SVM (support vector machines).

La reproductibilité de certains phénomènes sur l'ensemble de test ne doit pas pourtant leurrer le praticien : ce n'est pas pour autant que l'on peut agir et on retrouve ici la distinction corrélation-causalité. Il en va ainsi de l'exemple souvent cité pour illustrer la méthode des règles d'association où l'on découvre que dans les achats de supermarché il y a une association significative entre les achats de couches et de bière : il est vraisemblable qu'une promotion sur un des produits sera sans effet sur les ventes de l'autre.

Il faut donc se garder de certaines illusions : découvrir des structures «inattendues» est également une idée trompeuse : on a d'autant plus de

chances de trouver quelque chose d'intéressant que l'on connaît mieux ses données, et l'expertise et l'intervention du spécialiste seront presque toujours nécessaires.

En conclusion, je partage le point de vue des auteurs et celui de J. Friedman : les statisticiens, qui ont laissé se développer en dehors d'eux de nombreux champs faute d'écoute suffisante et d'efforts en direction des utilisateurs, doivent participer au développement du Data Mining, qui n'est pas une mode et va certainement encore se développer. Leur formation au risque et à l'aléatoire leur donne le recul nécessaire pour l'interprétation des résultats et des méthodes. Le Data Mining doit nous rappeler que la vocation première de la Statistique est d'analyser des données. Si les statisticiens s'en désintéressent, d'autres prendront la place qu'ils auront laissée, et la statistique s'en trouvera marginalisée. Le récent rapport de l'Académie des Sciences préconise dans sa recommandation n° 6 : « *les laboratoires de statistique doivent pouvoir recruter des enseignants et chercheurs en informatique, et être dotés des moyens de calcul adéquats, pour mener des recherches coordonnées en statistique et informatique, orientées par exemple vers le « Data Mining» et les domaines connexes* ». Mais le mouvement doit concerner non seulement les statisticiens qui doivent se rapprocher des informaticiens (beaucoup l'ont déjà compris, voir les nouveaux diplômes) mais aussi les informaticiens qui ont trop souvent diminué la place des mathématiques et en particulier de la statistique dans les cursus universitaires. Comme le disait B. Efron : « *Statistics has been the most successful information science. Those who ignore Statistics are condemned to reinvent it.* ».

RÉFÉRENCES

Académie des Sciences (2000). *Rapport sur la science et la technologie n° 8, La statistique*, Paris.

FAYYAD U.M., PIATETSKY-SHAPIRO G., P. Smyth and R. Uthurusamy (eds.) (1996), *Advances in Knowledge Discovery and Data Mining*. Menlo Park, California : AAAI Press.

FRIEDMAN J. (1997), Data Mining and statistics, what's the connection ?
http ://www-stat.stanford.edu/ jhf/ftp/dm-stat.ps

FRIEDMAN J. (1999), The role of Statistics in Data Revolution. *ISI, Helsinki*. http ://www.stat.fi/isi99/index.html

HAND D.J. (1999), Why data mining is more than statistics write large, Bulletin of the International Statistical Institute, 52nd Session, Helsinki, Vol. 1. 433-436. http ://www.stat.fi/isi99/index.html

HAND D.J. (2000), Methodological issues in data mining. *Compstat 2000 : Proceedings in Computational Statistics*, ed. J.G. Bethlehem and P.G.M. van der Heijden, Physica-Verlag, 77-85.

SAPORTA G. (2000), Data Mining and Official Statistics. *Quinta Conferenza Nationale di Statistica*, ISTAT, Roma.

VAPNIK V. (1998), *Statistical learning theory*, Wiley.