Data mining et Statistique
Réponse aux intervenants

Philippe Besse

Je voudrais tout d'abord remercier les participants à la discussion pour leur généreuse contribution. Comme il a été signalé, les auteurs de cet article sont globalement issus de la même culture et ont un point de vue nécessairement partiel donc partial. Chaque contribution apporte alors un éclairage différent et des compléments utiles sur un sujet qui, au moins par écrit, n'a pas encore été débattu dans la communauté statistique française. En revanche, plusieurs journées et conférences autour du Data Mining ont déjà été organisées par la Société Française de Statistique. Compte tenu de la diversité du sujet, je ne répondrai qu'à quelques-uns des thèmes abordés.

Les contributions s'accordent sur de nombreux points : le rôle moteur du développement technologique, en particulier en informatique, induit la naissance d'un nouveau marché dont les produits, les logiciels, assurent la promotion des techniques d'analyse et un gain notable de convivialité, d'efficacité. La stratégie de partage de l'échantillon : apprentissage, validation, test, est communément admise. Le développement du *data mining* s'inscrit dans le contexte des sciences de l'ingénieur sous la forme d'un paradigme (G. d'Aubigny) dont on peut s'interroger si c'est une science ou une technologie (D. Zighed), une discipline ou un effet de mode (G. Saporta) ou si ce paradigme s'élèvera dans l'avenir au rang de discipline (L. Lebart). Il est donc important, par souci de cohérence, d'analyser ce paradigme du point de vue de l'utilisateur comme J.-M. Gauthier et du point de vue des méthodes plutôt que de la méthodologie (G. d'Aubigny).

Data Mining et Analyse des Données

Comme une résurgence d'un débat hexagonal déjà ancien, les critiques ou désaccords les plus marqués concernent la place de l'Analyse des Données et l'intérêt des exemples présentés. G. D'aubigny, rejoint en d'autres termes par Y. Lechevallier, s'interroge ainsi sur la *spécificité de ces études qui les différencient de l'Analyse Exploratoire des Données pratiquées depuis 25 ans. En quoi sont-elles spécifiques d'une démarche de Data Mining ?*

L. Lebart regrette que notre article n'ait pas mentionné le *caractère précurseur de nombreux travaux réalisés en France* sur le thème de *l'analyse* *exploratoire des données.* C'est certainement un manque et je le remercie de l'avoir comblé même si, nous ne pouvons que le regretter, cet aspect précurseur n'a pas été considéré dans le développement outre-atlantique du *data mining.* En particulier, l'absence de l'analyse des correspondances multiple du module SAS Enterprise Miner, comme de beaucoup de logiciels américains, est symptomatique.

Il semble simple d'éviter un faux débat en prenant soin d'expliciter le niveau auquel il se situe. Du point de vue de l'utilisateur, adopté par D. Zighed, l'analyse de données est un ensemble de techniques exploratoires multidimensionnelles. Ce fut aussi, implicitement, le point de vue des auteurs par souci de cohérence avec celui adopté pour considérer le *data mining* : un vaste ensemble d'outils regroupés autour d'une problématique. Sur le plan méthodologique, tout le monde s'accorde à dire que l'Analyse des Données, analyse secondaire (G. d'Aubigny, G. Saporta), a introduit en France une rupture dans la pratique statistique dominante de l'époque en s'intéressant aux données avant que de considérer un modèle. En ce sens, les promoteurs de l'Analyse des Données furent des précurseurs. Comme le signale G. d'Aubigny, ce même rôle est attribué outre-atlantique à l'EDA (Exploratory Data Analysis) de Tuckey mais avec des outils radicalement différents pour réfuter le cadre probabiliste : la norme L_1 dans un cas, la géométrie euclidienne et la dualité dans l'autre.

Néanmoins, cette similarité datée ne doit pas faire oublier que l'un des objectifs majeurs du *data mining* est la recherche d'un modèle prédictif parcimonieux qu'il soit gaussien, binomial ou d'une autre nature et que la procédure de choix de ce modèle joue dans ce cadre un rôle central. Sur le plan méthodologique, vouloir intégrer cet objectif à l'Analyse des Données est pour le moins contradictoire avec son "mythe fondateur". Le point de vue plus récent, qui considère les techniques factorielles (Caussinus, 1986, 1993) ou celles de classification (Celeux 1988) comme les estimations des paramètres d'un modèle, présente l'avantage d'expliciter clairement les hypothèses sous-jacentes et d'intégrer ces outils dans une démarche globale de modélisation propre à la Statistique.

D'ailleurs G. Saporta ne s'y trompe pas en adoptant ce point de vue plus large. Il s'interroge plutôt sur ce qui différencie les exemples présentés d'une *analyse statistique classique.* Il y répond en rappelant qu'en *data mining le modèle provient des données et n'est pas choisi a priori.* Je partagerais ce point de vue à condition, comme le suggère L. Lebart, que la *Statistique accorde l'hospitalité*[1] aux autres techniques issues de l'Intelligence Artificielle. Ceci suppose de reconnaître honnêtement leur apport et leur intérêt et parfois d'accepter de prendre en compte des résultats heuristiques avant que l'étude mathématique n'en soit développée.

[1]Reconnaissons à ce sujet aux concepteurs et développeurs du logiciel SPAD leur intérêt très précoce pour les réseaux de neurones.

Une approche pragmatique

Revenons à des aspects plus concrets que méthodologiques dans l'esprit de l'article initial. La discussion suggère en effet quelques remarques complémentaires.

Par souci de concision, les jeux de données ont été choisis de taille relativement modeste et nous avons volontairement omis toute la gestion préalable des données et la constitution des *datamarts* par la fusion (D. Zighed) de sources hétérogènes. Certaines spécificités du *data mining* se trouvent ainsi éludées contrairement à l'exemple *industriel* décrit par C. Derquenne et G. Hébrail que nous remercions de nous avoir complétés sur ce point.

L'étude des données bancaires est un exemple typique de Gestion de la Relation Client (GRC) qui, pour l'essentiel, assure la promotion commerciale du *data mining* à travers l'utilisation des logiciels généralistes. Ces derniers ne savent pas, ou pas encore, s'adapter aux spécificités structurelles des données (C. Derquenne et G. Hébrail) mais ils sont efficaces tout en requérant beaucoup de prudence. Je voudrais à ce sujet rassurer L. Lebart dont nous partageons le point de vue sur l'analyse discriminante (au sens de Fisher) et ses qualités de robustesse. Notre texte est malencontreusement expéditif mais cette technique a été soigneusement testée dans toute ses versions (paramétriques linéaire ou quadratique, non paramétriques, k plus proches voisins) avant d'être écartée car mal adaptée, sur cet exemple, à un problème de discrimination nettement non linéaire.

La présentation des exemples se focalisent sur les problèmes de choix de modèles et de choix de méthodes et sur les estimations des taux d'erreur afférents sur des échantillons tests. La recherche d'une défaillance dans un procédé industriel est spécifique dans le sens où c'est le modèle choisi qui apporte la réponse plus que l'estimation des paramètres. Cet exemple est, sous une forme simplifiée, un type de problème hautement multidimensionnel, c'est-à-dire dans lequel le nombre de variables (les paramètres du procédé) pourrait être beaucoup plus grand que le nombre d'observations (les lots). La détection de pompage piloté, qui vise à discriminer des portions de courbes, rentre également dans cette catégorie mais, encore une fois, nous avons éludé la description des prémisses qui ont permis d'identifier les caractéristiques (fréquence, déphasage, seuillage) les plus discriminantes des signaux.

Les exemples aéronautiques ont été choisis pour montrer combien les interactions sont nécessaires entre disciplines. Les modèles statistiques usuels se montrent incapables, sur l'exemple du pilote automatique, de prendre en compte efficacement les non-linéarités contrairement aux réseaux de neurones. En revanche, l'expertise statistique redevient incontournable pour aborder le problème de la certification des réseaux ainsi estimés. J.-M. Gauthier remarque avec raison que des compétences en Statistique et en Informatique ne suffisent pas à faire de bons modèles, il faut y associer les compétences du *métier*. Ce n'est pas toujours explicite dans notre texte mais cette pluridisciplinarité est d'une évidente nécessité. Sa prise en compte dans nos exemples est implicitement indiquée par les appartenances des cosignataires. Enfin, l'effort pour faire intervenir les questions de causalité a été constant dans nos traitements. Cela est plus naturel et plus aisé dans les problèmes relevant de la physique que dans le domaine du marketing.

D'autres développements

Nous avons donc adopté, dans cet article, un point de vue très pragmatique cherchant à tirer quelques enseignements à partir de l'expérience acquise sur des exemples limités. C'est cohérent avec le contexte d'émergence du *data mining* mais cela restreint la généralité de la présentation. Certaines négligences ont déjà été évoquées ci-dessus : la fouille de données hétérogènes et complexes (textes, sons, images) peut devenir l'un *des enjeux majeurs du data mining* (D. Zighed). D'autres techniques nécessiteraient des développements plus spécifiques : l'exploration graphique de grandes bases de données telles quelles sont évoquées par (Ladiray, 2001) ou encore les méthodes de Support Vector Machine (Vapnik, 1999).

La contribution de R. De Veaux entre dans cette catégorie. Seul étranger, il n'est pas concerné par le débat sur les origines évoqué plus haut et centre son intervention sur l'avenir et la description d'algorithmes émergeants : *bagging*, *boosting*, qui n'ont été qu'évoqués dans le corps de l'article. Imaginés au sein de la communauté de l'apprentissage machine qui a illustré, de façon heuristique, leurs performances sur de nombreux exemples réels et simulés, ces algorithmes intéressent actuellement les statisticiens qui cherchent, sinon à prouver, du moins à expliquer leurs performances. Schématiquement, celles-ci sont liées à la résistance au sur-ajustement (resp. sur-apprentissage) qui autorise une amélioration de l'adéquation du modèle moyen ou du comité de modèles sans dégrader la capacité de prédiction (resp. de généralisation).

L'algorithme de *boosting* (Freund et Schapire, 1996), ou plutôt la version de Friedman et col. (2000)[2], ainsi que les forêts aléatoires (Breiman 2001)[3] ont été appliquées pour constituer un comité d'arbres de décision estimés pour prévoir la possession de la carte Visa Premier (données du premier exemple traitée). Le principe des forêts aléatoires est d'introduire une double randomisation : bootstrap des observations pour l'estimation de chaque arbre de la forêt (comme en *bagging*) et tirage aléatoire d'un sous-ensemble des variables explicatives considérées à chaque nœud d'un arbre. Trente échantillons tests ont successivement été tirés afin d'observer les distributions des taux de mauvais classement, distributions qui ont été comparées à celles obtenues par les méthodes classiques (arbre de décision,

[2]J. Friedman en fournit le programme : `www-stat.stanford.edu/ jhf/MART.html`.

[3]Le programme se trouve sur la page de L. Breiman : `http ://www.stat.Berkeley.edu/users/breiman/`.

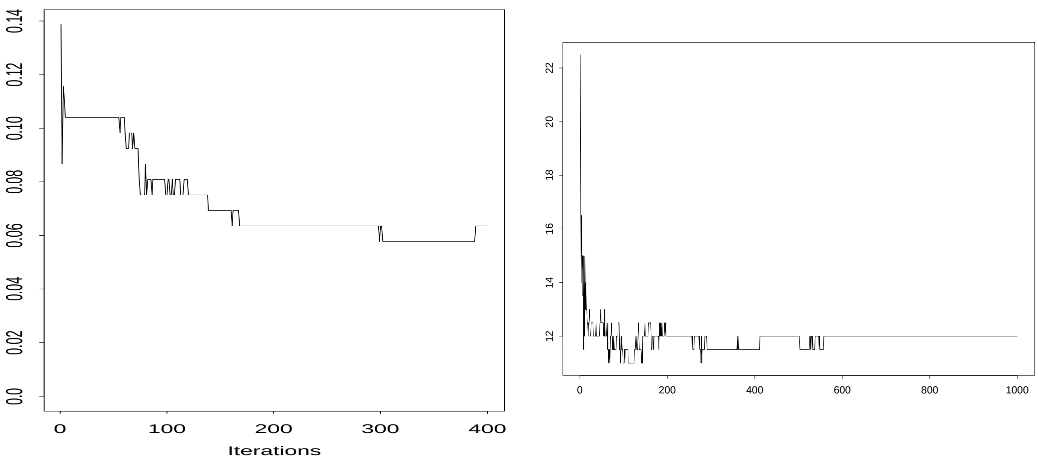

FIG. 1 – *Algorithmes AdaBoost et Random forests. Évolution, pour un échantillon test, du taux de mal classés en fonction du nombre d'arbres intervenant dans la combinaison de modèles.*

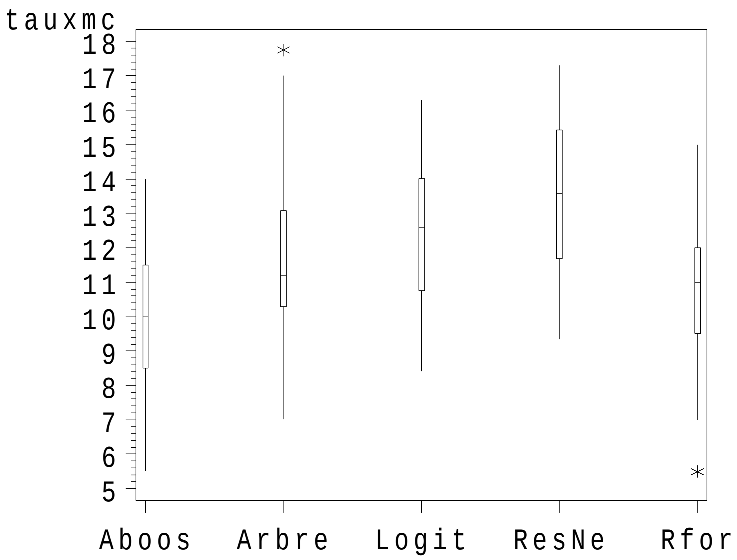

FIG. 2 – *Diagrammes boîtes des taux d'erreurs observés sur 30 échantillons tests et pour chaque méthode.*

TAB. 1 – *Moyennes des taux d'erreurs de classement calculés sur 30 échantillons test pour chaque modèle de prédiction*

Méthode	Adaboost	Arbre	Régression	Perceptron	Forêt
Moyenne	9.7	11.8	12.5	13.4	10.6
Écart-type	2.0	2.3	2.0	2.3	2.2

régression logistique et réseaux de neurones) au moyen de la procédure décrite en section 3.4 de l'article. La figure 1 montre les évolutions du taux de mal classés sur l'échantillon d'apprentissage en fonction du nombre d'arbres estimés pour un exemple de tirage. Malgré la complexité des combinaisons de modèles finalement obtenues, le taux atteint une limite, il n'y a pas sur-apprentissage. Ces algorithmes fournissent des résultats qui, en moyenne, se montrent sensiblement plus performants (cf. figure 2 et tableau 1) sur un échantillon test. Les écarts-types dépendant de la taille de l'échantillon test y sont relativement stables. Les moyennes montrent, sur cet exemple, que le *boosting* prédit un peu mieux que les forêts aléatoires. Cela est cohérent avec les nombreuxses études publiées récemment.

Bien sûr, comme le signale R. De Veaux, ce qui est gagné en prédictibilité est perdu en interprétabilité par rapport à un modèle classique et les aspects de causalité (J. -M. Gauthier) sont plus difficiles à prendre en compte. Néanmoins le gain réalisé est souvent étonnant et n'a pas encore trouvé de justification théorique satisfaisante. Le débat reste ouvert dans les deux revues phares des disciplines concernées : *Machine learning* d'une part et *The Annals of Statistics* de l'autre. L'une des avancées encore en gestation concernant ces algorithmes, et plus particulièrement les forêts aléatoires, est la prise en compte des problèmes posés par les données hautement multidimensionnelles tels qu'ils se posent par exemple avec l'analyse des biopuces en génomique.

Conclusion

Il serait vain de vouloir conclure définitivement sur un thème en pleine ébullition ; le débat continue. La fouille de données ne constitue pas une discipline autonome. À l'écoute des besoins d'une discipline d'application de sa problématique et compte tenu des données fournies, le prospecteur emprunte à l'Informatique ou à la Statistique les outils qui lui sont nécessaires, et sans qu'un ne soit *a priori* meilleur qu'un autre, pour rendre ses *connaissances opérationnelles* (J.-M. Gauthier, D. Zhighed) en vue d'une *action* (G. Saporta). Il est de la responsabilité du statisticien d'entrer dans le débat, d'y être à l'écoute, en s'attachant à résoudre les problèmes qui lui sont posés par les outils les plus adaptés et pas seulement par ceux dont il a déjà l'habitude ou l'expertise mathématique. Il est de la responsabilité de l'universitaire de former des étudiants susceptibles de répondre à ces besoins une fois qu'ils ont été identifiés, de leur transmettre *la vraie richesse qui réside dans la compétence actualisable du prospecteur* (G. d'Aubigny). Je ne pense pas qu'il y ait, comme le craignent C. Derquenne et G. Hébrail, une *pénurie de main d'œuvre statisticienne* mais plutôt une identification encore inadaptée de certains besoins et donc de profils des métiers de l'industrie. En revanche, le tertiaire ayant identifié ses besoins en gestion et traitement des données depuis plusieurs années, l'Université y répond déjà par des formations diversifiées[4] (G. Saporta) : IUT STID, licences professionnelles, IUP, DESS, Magistère, grandes écoles.

Références

Breiman, L. (2001). Random forests random features. *Machine Learning, à paraître.*

Caussinus, H. (1986). Models and uses of principal component analysis. In J. de Leeuw et al. (Eds.), *Multidimensional Data Analysis*, pp. 149–170. DSWO Press.

Caussinus, H. (1993). Modèles probabilistes et analyse des données multidimensionnelles. *Journal de la Société de Statistique de Paris 134*(2), 15–32.

Celeux, G. (1988). Classification et modèles. *Revue de Statistique Appliquée 36*(4), 43–57.

Freund, Y. and R. Schapire (1996). Experiments with a new boosting algorithm. In *Machine Learning : proceedings of the Thirteenth International COnference*, pp. 148–156. Morgan Kaufman. San Francisco.

Friedman, J. H., H. Hastie, and R. Tibshirani (2000). Additive logistic regression : a statistical view of boosting. *The Annals of Statistics 28*, 337–407.

Ladiray, D. (2000). Graphiquez vos données, commentaires sur l'article de J.-P. Valois : Approche graphique en analyse des données. *Journal de la Société Française de Statistique 141*, 61–67.

Vapnik, V. (1999). *Statistical learning theory.* Wiley Inter science.

[4]Consulter le site de la SFdS : `www.sfds.asso.fr`