JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

JEAN-MICHEL GAUTIER

Discussion et commentaires. Data mining et statistique

Journal de la société française de statistique, tome 142, n° 1 (2001), p. 67-72

<http://www.numdam.org/item?id=JSFS_2001__142_1_67_0>

DISCUSSION ET COMMENTAIRES

Data Mining et Statistique

Jean-Michel GAUTIER *

Aborder une discussion sur un thème aussi débattu que celui choisi par les auteurs peut se faire selon différents registres, du débat technique au débat d'idées, en passant par le débat de discipline (territoire, légitimité). J'ai préféré utiliser le mode testimonial pour introduire, à travers le point de vue d'un praticien du Data Mining en Marketing, une vision, peut-être sectorielle, mais résolument interdisciplinaire.

Durant 20 ans d'une carrière consacrée à l'étude des comportements des consommateurs pour l'optimisation des politiques marketing et commerciales des entreprises, comme enseignant, chercheur, mais aussi directement au service des entreprises, je me suis souvent demandé comment classer mon expertise.

À la fois statisticien et psychologue par formation, homme de marketing par destination et informaticien par nécessité, je n'ai jamais été capable de donner un nom à cette spécialité :

- S'agissait-il principalement de bâtir des démarches méthodologiques d'investigation et de modélisation statistique au service du marketing ?
- Ou bien de s'assurer que les démarches en question étaient utilisables au quotidien dans les entreprises avec les systèmes informatiques, les données et les logiciels disponibles ?
- S'agissait-il le cas échéant d'imaginer de nouveaux systèmes de collecte ou d'enrichissement de l'information client en amont de la modélisation ?
- Ou encore de s'assurer de la validité des démarches de collecte et de modélisation de l'information par rapport à une expertise « métier » des comportements du consommateur ?

Probablement tout à la fois, au vu de toutes les applications que j'ai pu mener au sein de nombreuses entreprises au service du Marketing Direct et plus généralement de la Gestion de la Relation Client.

Lorsque j'ai vu apparaître pour la première fois le terme Data Mining, j'ai pensé qu'il décrivait particulièrement bien mon activité : extraire et filtrer une connaissance client directement opérationnelle (prédictive !) au

* Département Systèmes d'Information et d'Aide à la Décision Groupe HEC, 78351 JOUY EN JOSAS CEDEX ; Email : gautier@hec.fr gautier.jm@wanadoo.fr

service des actions marketing des entreprises, sur le meilleur terrain (système d'information) et avec les outils industriels les plus perfectionnés disponibles.

J'ai commencé alors à suivre les travaux et les produits (logiciels) regroupés sous cette appellation :

- En 1995, premiers logiciels avec quelques algorithmes proposant des arbres de segmentation sur des critères « exotiques » (du point de vue d'un statisticien), sans la moindre considération inférentielle.
- De 1996 à 1998, développement de logiciels rendant plus faciles des enchaînements méthodologiques, déjà couramment pratiqués par les statisticiens, en matière de modélisation et de classification.
- Depuis 1999 des travaux plus axés sur la recherche de nouveaux algorithmes permettant d'explorer des modélisations jusqu'alors impossibles, faute d'outils de résolution des modèles proposés.

Après quelques débats stériles entre statisticiens et informaticiens sur l'existence d'une dimension inférentielle en Data Mining (un remix d'une querelle identique déjà vécue sur les réseaux de neurones), tout le monde s'accorde aujourd'hui sur une prise en compte minimale à travers les concepts de sélection de prédicteurs, de modèles parcimonieux et de validation / test évoqués par les auteurs.

Globalement, le courant Data Mining pouvait être vu comme un souci d'industrialisation et d'opérationnalisation de démarches de structuration de connaissance et de modélisation décisionnelle.

Si les statisticiens, premiers utilisateurs des outils produits par le Data Mining, débattent aujourd'hui encore de l'existence même d'une discipline « Data Mining », c'est que, tout à la fois, dans la plupart des cas, l'objectif est identique à celui de la modélisation « descriptive » ou « explicative » effectuée par les méthodes statistiques classiques et, en même temps, à travers la facilitation des traitements informatiques, il introduit un nouveau débat : peut-on modéliser sans être statisticien ? ou « des dangers des outils de Data Mining placés entre n'importe quelles mains », comme le soulignent les auteurs.

De mon point de vue de praticien, c'est en réalité un faux débat, car il s'agît plutôt d'identifier les qualités qui font de quelqu'un un « bon » modélisateur dans un domaine donné.

Avant de revenir sur cette question fondamentale, je voudrais tenter de mieux cerner l'apport du Data Mining par rapport à une situation antérieure où cette approche n'avait pas cours.

Le principal constat est que le Data Mining s'est traduit par une amélioration des performances et de la facilité d'usage des logiciels mis à disposition des chercheurs et des praticiens de la modélisation. A travers des recherches sur le lien entre les modèles d'organisation de données et l'efficacité des outils, ainsi que sur de nouveaux algorithmes de construction de modèles, le Data Mining a doté les chercheurs/modélisateurs de meilleurs outils.

Du fait même que le Data Mining a pour vocation d'élargir les cadres traditionnels de la modélisation en rendant les méthodologies plus faciles à mettre en oeuvre et en permettant de traiter de plus gros volumes de données, les chercheurs se sont également intéressés aux processus de regroupement de l'information à partir de sources hétérogènes, aboutissant à la mise au point de nouveaux outils comme les E.T.L. (logiciels d'extraction, de transformation et de chargement), ou les outils de validation / cleaning de l'information, de nouveaux concepts comme celui de DataWarehouse (entrepôt de données, souvent hébergé par un SGBD relationnel, qui concentre toute l'information brute plus ou moins normalisée et en assure l'historisation et la pérennisation), ou celui de Datamart (Base de donnée dédiée à un métier et optimisée pour un type d'usage, extraite du DataWarehouse).

Vu par un praticien de la modélisation opérationnelle en Marketing, le Data Mining a eu pour résultat de traiter des problématiques qui intéressaient peu les chercheurs en statistique, telles que :

- la constitution et la structuration de systèmes d'information servant de base aux outils de modélisation des connaissances et de modélisation décisionnelle,
- l'optimisation de ces systèmes par rapport aux outils de Data Mining,
- la convivialité des enchaînements méthodologiques dans une démarche de modélisation et l'application opérationnelle des résultats des modèles,
- mais également, la mise à disposition d'algorithmes et logiciels proposant des méthodes nouvelles de modélisation des connaissances, même si elles ne se placent pas toujours dans un contexte inférentiel.

En énonçant ces constats, j'en viens à la conclusion que le Data Mining n'est finalement pas un concept apte à décrire complètement mon travail de modélisateur. Il déborde à la fois le champ de mon expertise par la diversité de ses applications, et en même temps ne la contient pas. Mais il n'est pas non plus une simple mise en règles et en logiciels de mes pratiques antérieures, et les nouveaux outils et modes de raisonnement sur l'information qu'il apporte ont fait progresser mon expertise.

Finalement, à travers les interrogations des auteurs sur la qualité des modèles réalisés par les utilisateurs des outils de Data Mining, on identifie une question implicite sur les compétences nécessaires à l'élaboration de « bons » modèles. Serait-il possible qu'avec ces nouveaux outils le mythe du presse bouton, du tout automatique devienne réalité ? Après tout dans d'autres domaines d'usages quotidiens (télévision, téléphonie, voiture et même micro-ordinateur) l'utilisateur a bien peu de connaissances et de compétences sur le fonctionnement des produits qu'il utilise, ce qui ne l'empêche pas de les exploiter intelligemment à son profit. C'est en tout cas ce que certains commerciaux incompétents (ou sans éthique ?) voudraient nous faire croire à propos de ces nouveaux logiciels qu'ils vendent à prix d'or, sous couvert de « redonner le pouvoir » à l'utilisateur final. Y aurait-il, derrière cet argument qui semble porter ses fruits, des frustrations irrésolues, des « abus de pouvoir » des informaticiens ou des statisticiens par rapport à l'homme de métier

(dans mon domaine, le responsable marketing), ou plus simplement des incompréhensions ? Certainement, et de façon très diverse d'une entreprise à l'autre ! Suffisamment en tout cas pour que les outils de Data Mining puissent être utilisés (et vendus) aussi bien à des informaticiens comme contre pouvoir vis-à-vis des directions des études, à des responsables marketing pour les rendre autonomes vis-à-vis des études et de l'informatique, et à des statisticiens pour les affranchir des contraintes informatiques.

Ces pratiques et les discours qui les sous tendent sont génératrices de conflits de territoires / disciplines particulièrement stériles, dont même les auteurs de l'article sont victimes. Car, si l'on peut s'accorder avec eux sur le fait que l'utilisation d'outils, si conviviaux soient-ils, ne dispense pas les utilisateurs de solides connaissances en statistique (et d'ailleurs en informatique également !), fait-on vraiment de « bons » modèles lorsque l'on est « seulement » statisticien et informaticien ?

À mon sens la réponse est non, et il ne s'agit pas de la reprise du vieux débat sur théorie et application. Le meilleur des statisticiens - informaticiens, spécialiste de la modélisation appliquée, peut être particulièrement inefficace sur un domaine qu'il ne connaît pas. Cette remarque va bien au-delà du traditionnel : « On va plus vite pour identifier les bons modèles quand on connaît bien le sujet sur lequel on travaille ». Le fond du problème réside dans la stabilité temporelle des modèles réalisés. En l'absence d'expertise « métier » sur les comportements modélisés, les modèles vont faire intervenir sans discernement les causes possibles du phénomène et des variables collatérales. Or, en Marketing, on observe une grande variabilité dans le temps des comportements des consommateurs, liée à court terme aux évolutions de la conjoncture, de l'actualité, des modes, des innovations, des marchés, et à moyen terme aux mutations de la société. Il en résulte une instabilité temporelle forte des modèles de comportement qui sont la base des outils de Gestion de la Relation Client. Le point clé est que le niveau de cette instabilité est fortement accru dès que les modèles n'identifient pas correctement les facteurs structurant des comportements, et intègrent des mesures collatérales, plutôt que les causes présumées. Paradoxalement, le meilleur modèle « prédictif » d'un comportement d'une population à un moment donné, n'est pas forcément celui qui présente la meilleure stabilité dans le temps. Or, les modèles utilisés en marketing opérationnels sont mis en œuvre avec un décalage de plusieurs mois par rapport aux échantillons ayant servis à leur construction et sont utilisés pendant plusieurs mois avant d'être reétalonnés ou reconstruits. C'est donc à l'aune de l'efficacité à terme que l'on devrait choisir un modèle en Marketing, ce qui est bien sûr paradoxal. Se rapprocher le plus possible de la prise en compte dans le modèle des causes présumées des comportements étudiés, reste aujourd'hui la méthode la plus efficace pour améliorer la pérennité des modèles construits. D'une part, cette compétence est rarement la spécialité des statisticiens - informaticiens chargés de l'élaboration des modèles, mais c'est plutôt celle des « hommes de métier » ; d'autre part, et c'est très symptomatique, si l'accent est souvent mis sur l'identification des causes dans la recherche fondamentale en Marketing, ce concept est quasi absent de toute la littérature concernant le Data Mining

et les modélisations des comportements clients pour la Gestion de la Relation Client. Il n'en n'est d'ailleurs pas question dans l'article débattu ici.

Pour en revenir à l'origine du débat, il me semble tout aussi dangereux pour un statisticien - informaticien de réaliser un modèle techniquement correct qui n'est pas validé par une connaissance « métier » des comportements observés, que pour un utilisateur « métier » de profiter d'outils de Data Mining pour fabriquer des modèles qui ne sont pas techniquement validé, ou qui vont se révéler inapplicable en raison de contraintes techniques ou de performance.

Il ressort de cette discussion, que les compétences nécessaires à un bon exercice de la modélisation des comportements clients sont de trois ordres :

- une compétence « métier », nécessaire à la compréhension des phénomènes étudiés, et à la définition des contenus des systèmes d'information,
- une compétence « informatique - data mining », nécessaire à la manipulation de l'information, à la structuration des systèmes d'information, à la mise en œuvre des modèles,
- une compétence « statistique », nécessaire à l'élaboration et à la validation des modèles.

Le mariage des compétences peut relever de l'interdisciplinarité, dans des équipes mixtes, sur des gros projets, comme de la polyvalence, plus efficace sur des petits projets.

Cette polyvalence commence d'ailleurs à s'exprimer dans les parcours éducatifs et professionnels tant individuels que collectifs, avec des profils de :

- Statisticiens/informaticiens
- Informaticiens/statisticiens,
- Statisticiens/marketing,
- Informaticiens/marketing.

Si aujourd'hui la mise en œuvre des outils du Data Mining requiert encore une compétence informatique forte, et des notions sur la performance des algorithmes utilisés, on peut probablement envisager une transparence à terme des contraintes informatiques et de la mécanique des outils de Data Mining vis-à-vis de la démarche de modélisation. Il sera plus difficile de se passer de la compétence statistique dont la mise en œuvre reste très complexe, et difficilement modélisable de façon satisfaisante. Quant à la compétence « métier », c'est probablement celle qui peut le plus difficilement être incluse dans les outils.

En conclusion de cette discussion, je dirais que :

1. Le Data Mining existe bien comme discipline prenant en charge des aspects de la modélisation des connaissance que la statistique académique a délaissé, car relevant plus de la mise en œuvre ou des bonnes pratiques,
2. Le Data Mining comme la statistique sont des disciplines qui contribuent à la réalisation de modèles de connaissances dans différents domaines, mais ne peuvent prétendre, ni seules ni ensemble, suffire à la réalisation de ces modèles,

3. En matière d'application, et en particulier dans les sciences humaines comme le Marketing, il est fondamental de réintégrer la notion de causalité dans le discours du Data Mining et de la statistique sur la modélisation.