JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

CHRISTIAN DERQUENNE
GEORGES HÉBRAIL

Discussion et commentaires. Data mining et statistique

Journal de la société française de statistique, tome 142, n° 1 (2001), p. 59-65

<http://www.numdam.org/item?id=JSFS_2001__142_1_59_0>

DISCUSSION ET COMMENTAIRES

Data Mining et Statistique

Christian DERQUENNE * et Georges HÉBRAIL *

Nous présentons ici le point de vue de l'utilisateur des techniques de data mining dans les entreprises. Notre contribution à la discussion comprend les parties suivantes :

- une réaction « à chaud » sur l'article de Besse *et al.*,
- la description rapide d'une application opérationnelle de data mining menée à EDF, mettant l'accent sur le rôle du statisticien dans le processus de mise au point de l'application puis dans le fonctionnement opérationnel,
- une discussion sur l'apport du data mining dans les entreprises, en terme d'aspects positifs et d'aspects négatifs,
- les regrets que nous avons par rapport à ce qui était annoncé au départ par le data mining,
- un court rappel de la nécessité d'une charte de déontologie du data mining.

La réaction à chaud sur l'article

Nous ne revenons pas sur la définition du data mining qui est bien présentée dans l'article et reprend deux références qui nous paraissent importantes : celle de Fayyad & Piatetski-Shapiro, et celle de Hand.

L'article met en évidence quelques différences entre le data mining et la statistique. De notre point de vue, le data mining se traduit dans les faits par une 'industrialisation' de l'analyse de données plus que par une discipline nouvelle ou un domaine nouveau. C'est pourquoi nous ne sommes pas étonnés de trouver dans l'article de Besse *et al.* des exemples de pièges à éviter lorsqu'une analyse statistique des données est effectuée. D'ailleurs, les pièges présentés ne sont pas spécifiques à une activité de data mining qui serait fondamentalement différente d'une analyse des données.

Si l'automatisation est présentée dans l'article comme une caractéristique majeure du data mining, nous ne l'observons pas dans la pratique courante

* EDF Recherche et Développement 1, Av. du Général de Gaulle, 92141 CLAMART CEDEX ; christian.derquenne@edf.fr, georges.hebrail@edf.fr

car les logiciels ne sont - à notre avis - utilisables que par des personnes compétentes en analyse statistique des données. Preuve en est la pénurie actuelle de ces profils sur le marché du travail. Les analyses ne sont pas automatiques et l'intervention du statisticien reste indispensable. A contrario, comme les éditeurs de logiciels ont beaucoup investi dans les développements de nouveaux logiciels (le marché le permet), nous observons que la productivité de l'analyste a fortement augmenté. Ce point sera développé dans la partie soulignant les aspects positifs du data mining.

Nous partageons l'avis de Besse *et al.* qui indique que le concept de data mining présente des spécificités par rapport à l'analyse statistique des données. Nous pensons ici par exemple aux points suivants :

- les données n'ont pas été prévues pour l'analyse statistique,
- les données sont volumineuses,
- le data mining déclare aborder l'identification de modèles, ainsi que la combinaison de modèles,
- un des objectifs est la recherche de patterns, de niches.

Mais les logiciels actuels ne permettent pas encore véritablement d'explorer les deux derniers points. Lorsque ce sera le cas, nous serons preneurs d'exemples pédagogiques mettant en garde l'utilisateur contre de nouveaux pièges.

Enfin, nous souscrivons tout à fait à la dernière remarque de conclusion des auteurs, qui encouragent les enseignants-chercheurs en statistique à s'investir dans le data mining afin d'y apporter tout leur acquis, et d'y puiser des problèmes concrets pour illustrer le contenu des enseignements théoriques.

Une application opérationnelle du data mining dans un environnement industriel

Besse *et al.* bâtissent leur discussion sur les divergences entre Statistique et Data Mining sur cinq piliers : les données a priori, la taille des données, l'automatisation, la validation et, la Statistique et les Mathématiques. Notre objectif ici est d'illustrer les quatre premiers points du message des auteurs, en présentant un autre cas industriel.

Un projet de Data Mining au service de la connaissance des clients d'EDF

Dans le contexte actuel de développement du Groupe EDF, une meilleure connaissance des attentes et des besoins de la clientèle est un enjeu fort. Le Data Mining s'applique particulièrement bien, comme l'indiquent les auteurs, sur de grandes bases de données comme celles de gestion des clientèles de masse d'EDF. Pour mieux connaître la clientèle résidentielle, un projet a été mis en œuvre en 1998 à EDF R&D. A l'origine, ce projet avait pour but d'évaluer la faisabilité et l'intérêt de l'application des techniques de Data Mining sur les différentes données de la clientèle d'EDF. Parmi les nombreuses analyses réalisées dans ce projet, nous avons choisi de discuter

ici de l'enrichissement de certains champs de la base de facturation d'EDF (énergie de chauffage du logement, énergie de chauffage de l'eau sanitaire, ...). Ces informations sont très utiles pour les actions de marketing menées sur le terrain. L'enrichissement de telles variables se fait par prédiction à l'aide de la régression logistique. Mais il doit se faire de la façon la plus rigoureuse possible, afin d'offrir une bonne qualité et une bonne validité statistique. L'approche développée est donc un bon exemple de traitements mêlant Statistique et Data Mining.

Les données a priori : la pierre angulaire pour une bonne réussite de l'application de méthodes

Ici, les données a priori concernent la facturation de la clientèle résidentielle d'EDF. Certains champs sont renseignés à 100%, car ils correspondent à des données indispensables à la facturation (type de tarif, puissance souscrite, relevé de consommation, date de création du contrat du client, ...), alors que d'autres sont relatifs à des informations supplémentaires (énergie de chauffage du logement, énergie de l'eau chaude sanitaire, type de logement, ...), et présentent généralement beaucoup de données manquantes, étant remplis au fil de l'eau. Ces informations supplémentaires n'ont pas été saisies selon le processus classique d'une enquête par sondage : il n'y a donc pas de plan de sondage établi a priori en vue de traitements statistiques. En effet, si l'on désire estimer la proportion des clients chauffant leur logement à l'électricité, l'oubli d'un tel plan, et donc d'une absence de redressement, peut provoquer un biais sérieux sur les résultats. Par conséquent, nous avons construit un plan d'échantillonnage avec des variables disponibles dans la base de données, les plus indépendantes, discriminantes et significatives possibles pour stratifier la population des clients (dans notre exemple : date de création du contrat, type de logement, nombre de clients dans la ville d'habitation , ...). Ici, la démarche traditionnelle du statisticien a été employée sur des données non prévues pour un traitement statistique.

Taille des données : des pièges à éviter

La base de données de facturation d'EDF contient de nombreuses tables de données relatives au local du client, à ses caractéristiques, à ses usages de l'électricité, à son type de contrat, etc... La tentation aurait pu être grande de fusionner toutes ces tables par numéro de référence client, et de récupérer l'ensemble des variables. On aurait alors obtenu une grande table de données de 28.000.000 de clients, avec plus de 200 variables, dont beaucoup auraient été inutiles pour l'objectif d'enrichissement de données que l'on s'était fixé. Une sélection préalable de ces variables a été réalisée, au cas par cas, de façon à ne garder que des variables candidates à l'explication jugées pertinentes par les analystes. De plus, nous avons choisi de découper la population des clients en sous-populations naturelles correspondant à chacun des 100 Centres EDF (200.000 à 400.000 clients par Centre). Ce découpage a l'avantage de fournir des sous-populations homogènes, et surtout utilisables localement. Pour chaque Centre EDF, un plan d'échantillonnage a été construit, ce qui

permet bien de traduire la spécificité de chacun d'eux. Il s'agit d'un plan post-stratifié accompagné d'un redressement.

Automatisation : vers une industrialisation contrôlée

Dans ce projet, nous avons mis au point une démarche complète, comprenant un certain nombre d'étapes, afin de pouvoir l'appliquer de la façon la plus automatique possible sur l'ensemble des 100 Centres EDF. Cependant, si certaines étapes sont « automatisables », d'autres ne peuvent pas l'être, ou du moins doivent être contrôlées rigoureusement par un statisticien. Les étapes d'extraction des tables de données, de fusion de ces tables en une seule, ont été facilement automatisées. La phase de construction du plan de sondage post-stratifié est semi-automatisée, car le découpage des variables est adapté à chaque Centre EDF en fonction des caractéristiques de sa clientèle. Les étapes de redressement de l'échantillon renseigné, de constitution d'un échantillon d'apprentissage pour la régression logistique, ont été également complètement automatisées. Les deux phases suivantes concernent la sélection des variables « explicatives » (avec un modèle logit), et la modélisation sur l'ensemble réduit des variables sélectionnées. Cela permet d'obtenir un modèle parcimonieux et robuste. Ces deux phases sont essentiellement manuelles car primordiales pour assurer la qualité du modèle et donc des résultats de l'enrichissement. Puis, la phase de validation du modèle statistique sur l'échantillon test a été semi-automatisée, car elle peut remettre en question la qualité prédictive du modèle. Enfin, les prédictions sur la population totale à l'aide des règles issues du modèle, ainsi que l'exportation de ces prédictions dans la base de facturation d'EDF, ont été automatisées.

Validation : importante, même en data mining

La validation est une phase importante dans ce type d'étude même si, comme le précisent les auteurs, une très grande précision n'est pas un réel enjeu dans les applications de ciblage en marketing. Cependant, la qualité de l'étape de validation est essentielle pour un bon enrichissement des données clientèle, car les données prédites peuvent être utilisées à d'autres fins à l'insu de l'analyste. Ainsi, nous avons établi un niveau de confiance individuel que nous attribuons à chaque information prédite de chaque client. Ce niveau de confiance individuel varie entre 0 et 1. Dans notre cas, nous avons décidé qu'une valeur supérieure à 0,8 correspondait à un niveau de confiance acceptable. Seules les prédictions supérieures à ce seuil sont « reversées » dans la base marketing.

Les aspects positifs de l'arrivée du data mining dans les entreprises

Comme annoncé plus haut, la première retombée de l'introduction du concept de data mining a été un engouement sans précédent des entreprises pour exploiter les informations contenues dans leurs bases de données, à des fins

d'aide à la décision. Cet engouement a permis : (1) aux analystes de données de convaincre leurs décideurs de l'intérêt de cette approche, (2) et par conséquent aux éditeurs de logiciels d'investir fortement dans l'amélioration des outils d'analyse statistique des données. Le résultat est là aujourd'hui : les logiciels de data mining, même s'ils sont encore chers et s'ils ne reprennent dans les faits que des méthodes statistiques assez classiques, sont bien plus faciles à utiliser et permettent un fort gain de productivité de l'analyste statisticien de données. Plus précisément, ont été améliorées les fonctions suivantes :

- la collecte des données, avec la mise au point de passerelles très commodes - mais encore méconnues de beaucoup d'analystes - permettant de constituer des tableaux de données en adressant des requêtes à des bases de données relationnelles.
- L'amélioration de la phase de nettoyage et de sélection des données, par le développement d'outils 'd'audits' des variables, par la facilitation des transformations des variables, par des outils de constitution d'échantillons d'apprentissage, de validation, de test, ...
- La mise en place de démarches statistiques *reproductibles* depuis l'extraction des données jusqu'à la publication des résultats, par l'intermédiaire d'enchaînements de traitements divers. La reproductibilité est assurée par la possibilité de sauvegarde de ces enchaînements. C'est de notre avis un progrès important car la reproductibilité – même si elle était déjà possible en conservant les programmes d'analyses – s'avérait dans les faits quasiment impossible à assurer.
- L'amélioration de la phase d'évaluation des performances des méthodes de prédiction, ces performances étant automatiquement et facilement calculées sur les échantillons de validation et de test.
- L'amélioration de la phase de publication (pour l'analyse exploratoire) et de déploiement (pour l'analyse décisionnelle) des résultats. Aujourd'hui, il suffit d'appuyer sur un bouton pour que le graphique construit soit accessible sur un serveur intra ou internet. Il suffit d'appuyer sur un bouton pour générer un programme C ou SQL qui va de façon opérationnelle calculer le score de chaque client.

Mais soyons bien d'accord. Le progrès réside ici dans l'automatisation de tâches où l'analyste n'avait pas de plus-value, et perdait systématiquement du temps pour les réaliser. Sa productivité s'est améliorée et il peut mieux qu'avant se concentrer sur les tâches où sa plus-value d'analyste est certaine.

Les aspects négatifs de l'arrivée du data mining dans les entreprises

Le principal aspect négatif de l'arrivée du data mining réside bien dans les pièges classiques de l'analyse de données et qui sont rapportés dans l'article de Besse *et al.* Mais ceci n'est pas nouveau. De notre point de vue, le danger provient surtout de la pénurie de main d'œuvre statisticienne, ce qui peut

tenter quelques entreprises de faire réaliser les projets de data mining par des personnes non compétentes.

Nos regrets par rapport aux avancées annoncées du data mining

Lorsque le concept de data mining a été introduit, de nombreuses promesses ont été faites sur les objectifs du data mining, et n'ont malheureusement pas été tenues, du moins jusqu'ici. C'est là que nos regrets résident. Ces regrets portent sur les points suivants :

- Le data mining au carrefour des bases de données et de l'analyse des données

On pouvait espérer le développement de méthodes d'analyses portant sur des données à structure complexe, ou bien tenant compte des méta-données disponibles dans les bases de données. Les logiciels disponibles aujourd'hui n'ont que très peu progressé sur ce plan : la structure des données soumises au data mining est toujours celle d'un tableau standard, et les seules méta-données récupérées de la base de données sont les types de données. Il reste du chemin à parcourir, par exemple pour analyser des individus composites.

- Le data mining au carrefour de l'analyse de données et de l'apprentissage symbolique automatique.

On pouvait espérer voir se développer des méthodes dites symboliques, comme la classification conceptuelle, qui permettent de construire des classes plus interprétables et de prendre en compte des connaissances a priori sur les données. Ces méthodes, ainsi que d'autres fournissant des interprétations faciles des résultats, ne sont pas encore disponibles.

- Le data mining pour la recherche de niches, pour la détection de déviations.

D'autres idées ont été émises lors de l'émergence du data mining, et ne sont pas encore développées dans les logiciels. La recherche de niches, qui consiste à rechercher dans un jeu de données des phénomènes qui s'écartent des grandes tendances, est d'une grande utilité pour les entreprises. Une méthode de base pour résoudre ce problème est la mise en évidence des 'outliers'. Mais on peut faire mieux et il serait souhaitable de disposer dans les logiciels de data mining de méthodes adaptées à la recherche des niches. Dans le même ordre d'idées, avait été développé - par G.Piatetski-Shapiro lui-même - un système de 'détection de déviations' qui recherchait des explications aux écarts entre prévisions et réalisations de dépenses de santé. Les chercheurs en statistique ont certainement un rôle à jouer pour éviter à la communauté du data mining de réinventer tout ce que la statistique peut apporter comme solutions à ces problématiques.

Quid d'une charte déontologique de l'utilisation du data mining ?

La banalisation de l'activité de data mining dans les entreprises pose de graves questions déontologiques. L'enrichissement des fichiers de clients, le scoring des clients pour leur appétence à tel ou tel produit : a-t-on le droit de le faire ? sous quelles conditions ? doit-on en informer le client ? Voilà de nombreuses questions qu'il faut se poser et auxquelles il faut répondre. Si la réflexion doit avancer sur ces problèmes relevant de la déontologie, il faut également que les logiciels de data mining évoluent pour assurer toute la confidentialité nécessaire pour l'accès aux données. En effet, la mise à disposition des données aux analystes oblige en général à déverrouiller toutes les protections mises en œuvre sur les données détaillées, aussi bien au niveau de l'autorisation de constitution du fichier (la déclaration à la CNIL) qu'au niveau de l'accès informatique aux données. Des progrès sont donc attendus dans les logiciels de data mining pour mieux protéger les données nécessaires aux analyses ainsi que les résultats produits.