JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

DJAMEL ABDELKADER ZIGHED

Discussion et commentaires. Data mining et statistique

Journal de la société française de statistique, tome 142, n° 1 (2001), p. 85-88

<http://www.numdam.org/item?id=JSFS_2001__142_1_85_0>

DISCUSSION ET COMMENTAIRES

Data Mining et Statistique

Djamel Abdelkader ZIGHED *

Je me suis longtemps demandé comment j'allais positionner ma contribution à ce débat sur le data mining et la statistique. Est-ce que je dois faire une critique de l'article ? ou dois-je prendre un point particulier et l'approfondir ? ou encore proposer un texte parallèle sur un aspect non traité dans cet article ? ...

Mon choix s'est finalement porté sur un exercice, peut-être périlleux, qui vise à chercher un équilibre entre les points que j'ai évoqués.

On comprend à travers le titre que les auteurs, qui viennent plutôt du monde de la statistique, se posent la question de savoir si le data mining et la statistique sont une même chose ou pas. L'impression qui se dégage est que l'on se demande si le data mining n'est pas seulement un habillage de la bonne vieille statistique par quelques outils de bases de données que l'on a assaisonnés d'un nouveau jargon. Si le data mining est quelque chose de nouveau, s'agit-il alors d'une nouvelle science ou bien d'une technologie ?

Le data mining, dans sa forme et compréhension actuelles, à la fois comme champ scientifique et industriel, est apparue au début des années 90. Cette émergence n'est pas le fruit du hasard mais le résultat de la combinaison de nombreux facteurs à la fois technologiques, économiques et même socio-politiques. Il pourrait être utile de faire un historique et de proposer quelques clés qui pourraient contribuer au moins au débat épistémologique. Le point de vue adopté par les auteurs (*cf.* section 2.1) considère le data mining comme une nécessité imposée par le besoin des entreprises de valoriser les bases de données qu'elles accumulent. En effet, le développement des capacités de stockage et les vitesses de transmission des réseaux ont conduit les utilisateurs à accumuler de plus en plus de données. Certains experts estiment que le volume des données double tous les ans. Que faire avec ces données coûteuses à collecter et à conserver ?

Le data mining est l'un des maillons de la chaîne de traitement pour la découverte des connaissances à partir des données (Knowledge Discovery in Data Bases : KDD) que nous avons désigné par Extraction des Connaissances à partir des Données (ECD). Sur ce point, nous sommes en phase avec les auteurs (*cf.* section 2.3). Avant de parler du data mining, il convient alors

* Laboratoire ERIC, Bat. L, Université Lumière Lyon 2, 5 avenue Pierre Mendès-France, C.P.11, F 69676 Bron Cedex
e-mail : zighed@univ Lyon2 fr

de décrire ce processus et, ensuite, de donner aux lecteurs les points d'entrée pour approfondir les sujets aux frontières du data mining.

Le data mining est l'art d'extraire des connaissances à partir des données. Les données peuvent être stockées dans des entrepôts (*data warehouse*), dans des bases de données distribuées ou sur Internet. Le data mining ne se limite pas au traitement des données structurées sous forme de tables numériques, il offre des moyens pour aborder les corpus en langage naturel (*text mining*), les images (*image mining*) ou le son (*sound mining*). C'est une ingénierie pour extraire des connaissances à partir des données qui se met en place. Cette vision un peu plus élargie du data mining n'est pas mentionnée par les auteurs. Je crois que l'un des enjeux majeurs du data mining va justement se situer à ce niveau pour assurer une fusion des données hétérogènes.

Avant de lancer les techniques de data mining, il convient de réaliser une série d'opérations dites de pré-traitement. Cela comporte l'accès aux données en vue de construire des datamarts, la mise en forme selon le type d'entrée : numérique, symbolique, images, textes, son,..., le nettoyage des données, le traitement des données manquantes, la sélection d'attributs, la sélection d'instances, etc. Ce point est également d'un grand intérêt, car le choix des descripteurs et la connaissance précise de la population sont des éléments essentiels, notamment dans la mise au point de modèles de prédiction. L'information nécessaire à la construction d'un bon modèle de prévision peut être disponible dans les données, mais un choix inapproprié de variables et/ou d'échantillon d'apprentissage peut faire échouer l'opération.

Le data mining à proprement parler opère sur des tables bidimensionnelles appelés *datamarts* et fait appel à trois grandes familles de méthodes, issues de la statistique, de l'analyse des données, de la reconnaissance de formes, de l'apprentissage automatique,... On peut regrouper les méthodes couramment utilisées ou présentées comme faisant partie de l'arsenal du « *data mineur* » en trois catégories :

- Les méthodes de description uni, bi et multidimensionnelles. Cela comprend les méthodes numériques, pour l'essentiel issues de la statistique descriptive et de l'analyse des données, les techniques de visualisation graphiques dont certaines font même appel à la réalité virtuelle et à des métaphores de représentation assez élaborées.
- Les méthodes de structuration qui regroupent toutes les techniques d'apprentissage non supervisé et de classification automatique issues de la reconnaissance de formes, de la statistique, de l'apprentissage machine et du connexionisme.
- Les méthodes explicatives dont le but est de relier deux phénomènes : l'un dit à expliquer et l'autre dit explicatif. Généralement ces méthodes sont mises en œuvre en vue d'extraire des modèles de classement et/ou de prédiction. Une large variété de méthodes est disponible. Elles sont issues de la statistique, de la reconnaissance de formes, de l'apprentissage machine et du connexionisme.

En dehors des champs des statisticiens, on assiste à l'émergence d'outils plutôt que de méthodes exploratoires. On peut citer par exemple les algorithmes de recherche de règles d'associations dans les grandes bases de données. Les premiers algorithmes proposés dans ce domaine ont fait sourire quelques statisticiens et autres spécialistes de l'induction tant le matériel méthodologique utilisé est empreint d'une certaine naïveté. Les choses ont évolué car ces problèmes ont été ramenés dans un cadre méthodologique plus général : parcours de treillis de Gallois, recherche de décomposition optimale d'une relation binaire par des relations dites maximales,...

L'objectif de la mise en œuvre des techniques de data mining est d'aboutir à des connaissances opérationnelles. Ces connaissances sont exprimées sous forme de modèles plus ou moins complexes : une série de coefficients pour un modèle de prévision numérique, des règles logiques du type « Si Condition alors Conclusion » ou des instances. Pour que ces modèles acquièrent le statut de connaissances ils doivent être validés. Il s'agit alors de mettre en œuvre une série d'opérations dites de post-traitement qui visent à évaluer la validité des modèles, à les rendre intelligibles s'ils doivent être utilisés par l'homme ou à les exprimer dans un formalisme approprié pour une utilisation en machine. Au-delà de la validation statistique, l'intelligibilité des modèles est souvent un critère de survie de ceux-ci. En effet, un modèle compris par l'usager sera utilisé donc critiqué, perfectionné,... Les usager n'aiment généralement pas les boîtes noires. Se pose donc la question de savoir quel outil pour quel problème. Selon le type de problème, il existe de nombreuses méthodes de data mining concurrentes. Malgré l'acharnement normal des auteurs à vouloir prouver que leur méthode est supérieure à celles des autres, un consensus général semble se dégager pour reconnaître qu'aucune méthode ne surpasse les autres : elles ont toutes des forces et des faiblesses. On essaye davantage de faire coopérer des méthodes comme nous le ferions avec une équipe de spécialistes que de les mettre en concurrence (*cf.* section 2.4, comparaison de méthodes).

Les techniques de data mining ont été employées avec beaucoup de succès dans de nombreux domaines. Les grands secteurs d'application du Data mining sont la Gestion de la Relation Client (GRC) (*Customer Relationship Management* : CRM), ou la Gestion des Connaissances (*Knowledge Management*), l'indexation de documents,.... Aucun domaine d'application n'est a priori exclu. Dès que nous sommes en présence de données empiriques, le data mining peut se rendre utile.

Il existe une large panoplie de logiciels de data mining recensés sur Internet. L'un des meilleurs sites de référence est kd.nuggets.com. C'est un excellent portail pour entrer dans la jungle du data mining.

Le data mining est un domaine à la fois scientifique et technologique jeune où de nombreux défis sont encore à relever. Des problèmes de recherche mobilisent l'attention de la communauté des chercheurs dans ce domaine, comme la recherche de bons espaces de représentation, l'agrégation de prédicteurs,...

Grâce à Internet, l'accès est possible à une grande quantité de sites regroupant des logiciels, des données, des expertises, des cours, des communautés

d'échanges ou de la bibliographie. Certes cette bibliographie est encore relativement limitée en termes d'ouvrages mais très abondante en termes d'articles.

L'académisme n'aime pas beaucoup, du moins officieusement, le pluridisciplinaire (*cf.* section 2.4, Statistique et Mathématique). Pourtant, tous les chercheurs reconnaissent l'importance de la pluridisciplinarité dans le développement des sciences et de la technologie. La data mining réussira t-il à réunir statisticiens et informaticiens ? Les indices semblent montrer que cela va dans le bon sens.