JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

YVES LECHEVALLIER

Discussion et commentaires. Data mining et statistique. Le data mining, une mise à niveau « informatique » des méthodes de l'analyse de données et de la statistique ?

Journal de la société française de statistique, tome 142, n° 1 (2001), p. 77-80

<http://www.numdam.org/item?id=JSFS_2001__142_1_77_0>

DISCUSSION ET COMMENTAIRES

Data Mining et Statistique

Le data mining, une mise à niveau «informatique» des méthodes de l'analyse de données et de la statistique ?

Yves LECHEVALLIER *

Ma contribution à cette discussion comprend les parties suivantes :

- Quelques remarques générales sur l'approche retenue par l'article de Besse *et al.*
- Le data mining se résume-t-il à la mise à niveau «informatique» des méthodes d'analyse de données ? La volonté du data mining de rendre plus compréhensibles, pour le client, les résultats d'analyse entraîne-t-elle une automatisation trop importante des processus d'analyse ?
- Le data mining peut-il générer de nouveaux axes de recherche en statistique ?
- Réponses ponctuelles à l'évolution des logiciels de statistique et à l'enseignement de la statistique.

Remarques générales

Le texte de Besse *et al.* présente le data mining comme une utilisation de logiciels commerciaux, ces logiciels ayant récupéré des techniques de l'analyse de données et de l'intelligence artificielle. Je pense que c'est une vision très réductrice du data mining. Il est regrettable que les auteurs n'aient pas présenté, ni établi une discussion autour des textes de Friedman et de Hand. Je pense que les exemples présentés par les auteurs sont plutôt des exemples «d'analyse des données» au sens qu'une analyse correcte peut être réalisée en utilisant uniquement des logiciels d'analyse des données. Comme ces exemples occupent une grande partie de ce texte, c'est dommage qu'ils n'illustrent pas plus précisément l'approche data mining.

* INRIA-Rocquencourt, Domaine de Voluceau BP 105, Rocquencourt, 78153 Le Chesnay Cedex
Email Yves.Lechevallier@inria.fr

Data mining et analyse des données

Dans les années 80-90 le développement des moyens de stockage et de calcul a permis de mettre en œuvre de nouvelles méthodes en analyse de données mais c'est l'approche data mining qui a, par la structuration de l'information (utilisation de la méthodologie des bases de données dans la définition de l'architecture et la construction des relations), introduit les méthodes d'analyse des données dans les entreprises en y ajoutant quelques approches originales. Et c'est cette modélisation de l'information qui a permis d'interfacer facilement, avec un contrôle renforcé de la qualité de l'information, les données aux outils de statistique ou d'analyse de données. Les bases de données relationnelles et le langage SQL permettent au statisticien de ne plus perdre de temps dans la phase de définition de la représentation des données, et par suite, j'y crois personnellement, une possibilité réelle de consacrer plus de temps à la construction et au choix de modèles.

Le data mining a remis en lumière l'approche exploratoire dans le processus de modélisation. Par ce fait, l'utilisation de cette approche exploratoire, proposée dans le data mining, inclut naturellement l'approche statistique en tant que principale démarche de traitement de l'information, notamment dans sa phase confirmatoire.

Automatisation des procédures

Il est à noter que le data mining n'a pas commis les même erreurs que les systèmes experts qui ont confondu automatisation et expertise. On ne peut qu'encourager l'automatisation des procédures, proposée par le data mining, car elle devrait permettre, à court terme, une recherche efficace de modèles statistiques. De même, il faut amplifier la formation en statistique afin d'améliorer l'expertise statistique qui est indispensable dans toute modélisation. L'inférence des modèles et de changement de méthodes au cours d'une analyse en data mining demande effectivement une bonne formation à la statistique. A ce sujet, on ne peut que regretter que l'enseignement de la statistique ne soit pas illustré plus fréquemment par des exemples issus du data mining.

À l'inverse des systèmes experts l'approche data mining valorise le savoir faire du statisticien en lui demandant d'émettre un jugement (analyse exploratoire) ou de prendre une décision (analyse confirmatoire) sur la nature d'un phénomène à partir des données expérimentales.

Dans l'enseignement actuel de la statistique, ce problème inverse (la distribution de la population est inconnue, ou bien « learning from data » ou « what the data says » nouveaux slogans des statisticiens américains) est trop souvent jugé moins important que le problème direct incluant ainsi la statistique dans la théorie des probabilités. Avec l'accroissement de la précision des expériences, la multiplication des appareils de mesure et de calcul, le rôle du

problème inverse devient de plus en plus important et le data mining a parfaitement compris cette tendance et, surtout, propose des outils qui ont pour conséquence d'amplifier cette problématique.

Data mining et les nouveaux axes de recherche en statistique

L'approche data mining permet un renouvellement important des axes de recherche en statistique. Je propose de donner une liste, sûrement non exhaustive, de ces nouveaux axes :

- **Fusion de données et données manquantes :** Comme il faut savoir intégrer des sources d'informations hétérogènes dans un entrepôt de données, les méthodes de fusion de données deviennent maintenant indispensables.
- **Entrepôt de données :** la granularité (« facteur d'échelle ») du recueil de l'information est un problème important entre l'analyse statistique et les bases de données car elle entraîne souvent une incompatibilité logique.
- **Classification automatique :** Le data mining pose le problème de recherche de classes pour des structures complexes dans un cadre exploratoire (recherche automatique de « patterns ») comme un problème déterminant dans le processus du traitement de l'information. Dans son approche « combinaison de modèles » il intègre les procédures classificatoires comme une phase importante dans la recherche de modèles. Ceci devrait entraîner un renouveau des méthodes de classification.
- **Choix de modèle :** C'est un thème actuel et il est très important pour la statistique. L'introduction de l'échantillon de validation est une des solutions pour résoudre le problème du choix de modèle. En général, l'évaluation d'un modèle est réalisée par l'estimation de ses paramètres à partir de l'ensemble d'apprentissage, puis sa qualité ou son efficacité est mesurée à partir de l'ensemble test. Ayant le choix de plusieurs modèles, cette sélection nécessite l'utilisation d'un autre ensemble qui est l'ensemble de validation.
- **Données a priori :** Le data mining évite de faire la confusion entre protocole de recueil des données et construction d'un échantillon représentatif en affirmant qu'un échantillon n'est pas intrinsèquement représentatif mais qu'il l'est par rapport à l'objectif de l'étude. Ceci justifie la possibilité de constituer un échantillon représentatif à partir de données non nécessairement récoltées pour l'étude.
- **Méthodes de « scoring » :** Comme pour le problème « biais-variance » le statisticien doit choisir entre des méthodes de type « boîtes noires » vérifiant des propriétés statistiques bien connues et définies a priori, et des méthodes facilement compréhensibles et donc utiles pour les utilisateurs. En d'autres termes plutôt que de se limiter à des modèles

statistiques bien connus le data mining offre une possibilité d'utiliser des modèles statistiques alternatifs plus adaptés aux problèmes des utilisateurs.

Logiciels et data mining

L'approche data mining montre que le développement de logiciels de statistique spécifiques n'est pas une bonne solution et que l'avenir est plutôt sur une approche plurielle des logiciels de traitement de l'information. Il est indispensable d'intégrer dans ces logiciels les aspects bases de données et d'extraction de connaissances.

Les opportunités d'emplois des étudiants en statistique

Le développement du data mining est une bonne opportunité d'insertion dans le monde des entreprises pour les étudiants en statistique car il démontre la nécessité d'introduire un lien entre le gestionnaire de bases de données et le décideur. Ce lien manquant devrait être un « producteur de résultats d'analyses et/ou d'extraction de connaissances ». Pour mener à bien cette tâche une compétence en statistique est nécessaire. C'est sur ce créneau que nos étudiants en statistique auraient avantage à se positionner.