Chapitre 4.2

« DATA MINING » ou FOUILLE DE DONNÉES

Gilbert Saporta

1. Définitions et historique

Le “data mining” que l’on peut traduire par “fouille de données” apparaît au milieu des années 1990 aux États-Unis comme une nouvelle discipline à l’interface de la statistique et des technologies de l’information : bases de données, intelligence artificielle, apprentissage automatique (« machine learning »).

David Hand (1998) en donne la définition suivante: *« Data Mining consists in the discovery of interesting, unexpected, or valuable structures in large data sets».* Voir également Fayyad & al (1996) et Friedman (1997).
La métaphore qui consiste à considérer les grandes bases de données comme des gisements d’où l’on peut extraire des pépites à l’aide d’outils spécifiques n’est certes pas nouvelle. Dès les années 1970 Jean-Paul Benzécri n’assignait-il pas le même objectif à l’analyse des données ? : *« L’analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature ».*
On a pu donc considérer que bien des praticiens faisaient du data mining sans le savoir.
On confondra ici le « data mining », au sens étroit qui désigne la phase d’extraction des connaissances, avec la découverte de connaissances dans les bases de données (KDD ou Knowledge Discovery in Databases) (cf Hébrail & Lechevallier, 2003).
Comme l’écrivent ces derniers auteurs :
« La naissance du data mining est essentiellement due à la conjonction des deux facteurs suivants :

- *l’accroissement exponentiel dans les entreprises, de données liés à leur activité (données sur la clientèle , les stocks, la fabrication, la comptabilité …) qu’il serait dommage de jeter car elles contiennent des informations-clé sur leur fonctionnement (...) stratégiques pour la prise de décision.*
- *Les progrès très rapides des matériels et des logiciels (…)*

L’objectif poursuivi par le data mining est donc celui de la valorisation des données contenues dans les systèmes d’information des entreprises. »

Les premières applications se sont faites dans le domaine de la gestion de la relation client qui consiste à analyser le comportement de la clientèle pour mieux la fidéliser et lui proposer des produits adaptés. Ce qui caractérise la fouille de données (et choque souvent certains statisticiens) est qu’il s’agit d’une analyse dite *secondaire* de données recueillies à d’autres fins (souvent de gestion) sans qu’un protocole expérimental ou une méthode de sondage ait été mis en œuvre.

Quand elle est bien menée, la fouille de données a apporté des succès certains, à tel point que l'engouement qu'elle suscite a pu entraîner la transformation (au moins nominale) de services statistiques de grandes entreprises en services de *data mining*.

La recherche d'information dans les grandes bases de données médicales ou de santé (enquêtes, données hospitalières etc.) par des techniques de data mining est encore relativement peu développée, mais devrait se développer très vite à partir du moment où les outils existent. Quels sont les outils du data mining et que peut-on trouver et prouver ?

2. Les outils

On y retrouve des méthodes statistiques bien établies, mais aussi des développements récents issus directement de l'informatique. Sans prétendre à l'exhaustivité, on distinguera les méthodes exploratoires où il s'agit de découvrir des structures ou des comportement inattendus, de la recherche de modèles prédictifs où une « réponse » est à prédire, mais on verra plus loin que l'acception du terme « modèle » diffère fondamentalement de son sens habituel.

2.1 Exploration « non supervisée »

2.1.1 Analyse des données : visualisation, classification

Les techniques de projection orthogonale sur des sous-espaces : analyse en composantes principales, analyse des correspondances, permettent de réduire efficacement la dimension du point de vue du nombre de variables. Les méthodes de classification visent à former des groupes homogènes d'unités en maximisant des critères liés à la dispersion (*k-means*). Des extensions non-linéaires (splines, noyaux, etc.) étendent le champ de ces méthodes classiques.

2.1.2 Recherche de règles d'association

Cette méthode est une des innovations du data mining : introduite en 1993 par des chercheurs en base de données d'IBM, elle a pour but de rechercher des conjonctions significatives d'évènements. Typiquement une règle de décision s'exprime sous la forme : si (A et B) alors C mais il s'agit d'une règle probabiliste et non déterministe. On définit le support de la règle comme la probabilité d'observer à la fois la prémisse X et la conclusion Y : $P(X \cap Y)$ et la confiance comme $P(Y/X)$. Parmi les règles ayant un support et une confiance minimale on s'intéressera à celles où $P(Y/X)$ est très supérieur à $P(Y)$. Les premières applications ont concerné les achats dans les grandes surfaces : parmi les milliers de références disponibles et les millions de croisements, identifier les achats concomitants qui correspondent à des fréquences importantes. Cette méthode s'étend bien au delà de ce type d'application. L'originalité tient essentiellement à la complexité algorithmique du problème.

2.2 Prédiction ou apprentissage « supervisé »

Inutile d'évoquer ici les techniques de régression bien connues. La méthode la plus typique du data mining est certainement celle des arbres de décision : pour prédire une réponse Y, qu'elle soit numérique ou qualitative, on cherche tout d'abord la meilleure partition de l'ensemble des données (en général en deux sous-ensembles) issue d'une partition effectuées sur les prédicteurs et on itère dans chacun des sous-ensembles : la croissance exponentielle de l'arbre est contrôlée par des critères d'arrêt de type coût-complexité ainsi que par l'usage de données de validation qui permettent d'éliminer les branches non pertinentes.

Cette technique conduit à des règles de décision très lisibles, d'où son succès, et hiérarchise les facteurs explicatifs. A l'opposé en termes de lisibilité, les logiciels de data mining proposent souvent des méthodes hautement non-linéaires comme les réseaux de neurones, les machines à vecteurs de support (SVM). Même si les règles de décision ont une forme mathématique explicite, celle-ci est en général très complexe et ces méthodes sont utilisées comme des boîtes noires.

Une autre approche consiste à complexifier des méthodes simples : les arbres de décision étant souvent instables, on va en utiliser plusieurs obtenus sur des données rééchantillonnées par bootstrap : la décision finale s'obtient par une procédure de vote s'il s'agit d'un problème de classification, ou de moyenne pour un problème de régression : c'est le *bagging*. Citons également le *boosting*, qui consiste à améliorer des procédures de décision en surpondérant les unités mal classées, et en itérant le processus.

3. Quelques applications en épidémiologie et santé publique

L'utilisation des méthodes de « data mining » en épidémiologie et santé publique est en forte croissance. Comme dans d'autres domaines, c'est la disponibilité de vastes bases de données historiques (on parle maintenant d'entrepôts de données) qui incite à les valoriser, alors qu'au dire de beaucoup de spécialistes elles sont actuellement sous-utilisées.

Pour n'en citer que deux, la revue *Artificial Intelligence in Medicine* et le *Journal of the American Medical Informatics* Association y consacrent de plus en plus d'articles.

La plupart des publications portent sur les arbres de décision et les règles d'association (Lavrac 1999). Parmi les domaines traités, mentionnons la recherche de facteurs de risque pour les accidents domestiques, le diabète, les suicides, les infections nosocomiales (Brossette & al.1998), la détection de la fraude (Medicare, Australie). Ces publications mentionnent souvent la découverte de règles inattendues et efficaces.

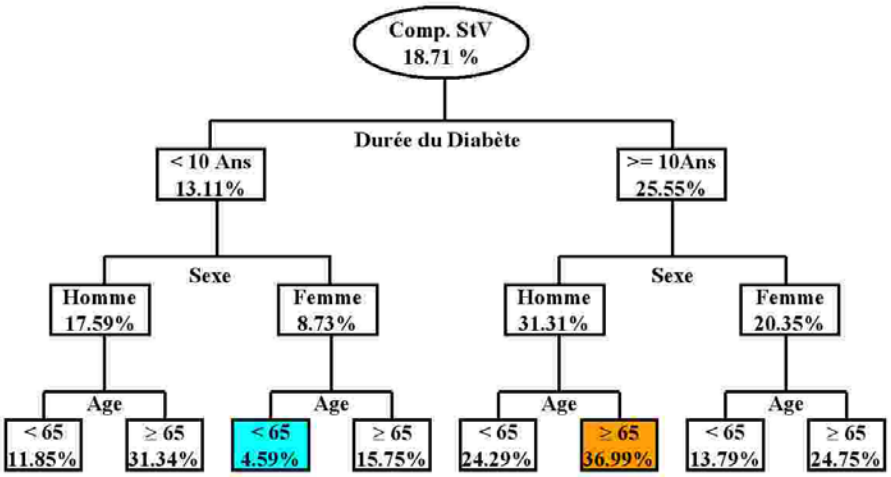

Figure 1 prévision de la complication de Saint Vincent
(Projet Data Diab, A.Duhamel, Lille 2)

La recherche en génomique et la protéomique fait également de plus en plus appel à des techniques de data mining.

4. Data Mining *versus* modélisation statistique

4.1 Le rôle des modèles

La notion de modèle en fouille de données prend un sens particulier : un modèle est une relation entre des variables exprimable sous une forme analytique ou algorithmique qui ne provient pas d'une théorie mais réalise un bon ajustement aux données. Ainsi il est courant d'explorer différents modèles (linéaires, non-linéaires) en faisant varier les paramètres (nombre de couches dans un réseau de neurones, noyau pour des SVM etc.) jusqu'à obtenir les meilleures prédictions. On est très loin de la démarche usuelle de modélisation, mais plutôt dans une optique pragmatique où il ne s'agit pas forcément de comprendre mais simplement de prévoir. Rappelons quand même qu'un modèle au sens classique, n'est qu'une simplification de la réalité et comme le disait George Box : « Tous les modèles sont faux, certains sont utiles ».

Cette démarche n'est pas pour autant du pur empirisme et se fonde sur une théorie solide, celle de l'apprentissage statistique : un modèle réalise un compromis entre sa capacité à rendre compte des données d'apprentissage et sa capacité de généralisation à de nouvelles données. Plutôt que des indices statistiques de type vraisemblance pénalisée (critères d'Akaïké ou de Schwarz) reposant sur des hypothèses distributionnelles, le choix d'un modèle en data mining se fait en fonction de ses performances sur d'autres données que celles qui ont servi à le choisir et le caler, d'où l'emploi de méthodes de validation croisée (les données sont divisées en plusieurs parties, chacune étant prédite à l'aide du reste des données) ou de mesures de capacité de type dimension de Vapnik-Cervonenkis.

4.2 Problèmes spécifiques d'inférence et de validation dans les grandes bases de données.

L'inférence statistique classique a été développée pour traiter des « petits » échantillons. En présence de très grandes bases de données le paradoxe est que tout

devient significatif : par exemple, pour un million d'individus, l'hypothèse d'indépendance entre deux variables sera rejetée au risque 5% si le coefficient de corrélation linéaire est supérieur en valeur absolue à 0.002, ce qui est sans intérêt pratique. L'inférence classique ne fonctionne plus et la fouille des grandes bases de données amène à repenser la notion de test et conduit ainsi à des recherches nouvelles.

L'échantillonnage ne perd cependant pas ses droits car il est souvent préférable de travailler sur une partie de la base que sur la totalité. L'exhaustivité des traitements n'est souvent qu'un argument commercial des éditeurs de logiciel. Un problème demeure cependant, celui de la représentativité de la base : même très grande on ne peut garantir que les futures observations se comporteront comme les passées, d'autant plus que la base n'a pas été constituée à des fins de traitement statistique. Des recherches originales portent sur ce point.

L'alimentation continuelle des bases de données de façon quasi automatique pose également des problèmes nouveaux connus sous le nom de flots de données (data streams) qu'il faut traiter « à la volée » sans devoir reprendre à chaque fois l'ensemble des données disponibles (Domingos & Hulten, 2000).

Quand en fouille de données, on a exhibé une structure ou une association intéressante et inattendue, on n'est pas certain de sa validité. En effet avec une exploration combinatoire il est inévitable de trouver toujours quelque chose ! Ce problème est proche de celui bien connu des comparaisons multiples, mais à une toute autre échelle. On trouve ce genre de situations dans la recherche de règles d'associations ou dans l'analyse des puces à ADN où on réalise des milliers de tests simultanément. La théorie pertinente est celle du contrôle du taux de fausses découvertes de Benjamini et Hochberg (1995) qui fait l'objet de recherches en plein essor, voir l'article de Ge & al.. (2003).

La « découverte » de règles intéressantes par la fouille de données doit donc être considérée comme une phase exploratoire, nécessitant une validation ultérieure, mais avec des outils différents. Même en cas de validation, le problème de la causalité reste posé.

4.3 Penser la complexité

Il est illusoire de croire que des modèles simples peuvent toujours convenir à des situations complexes. Les modèles de régression (linéaire ou logistique) ont l'avantage de l'interprétabilité, mais cela ne suffit plus en présence de phénomènes fortement non-linéaires. Il faut alors souvent plonger les données dans des espaces de grande dimension en utilisant des opérateurs de régularisation.

Le traitement d'images médicales ou de puces à ADN, en est une illustration frappante : il pose un défi dû à la fois à la complexité des données et au rapport inhabituel entre le nombre de variables et le nombre d'observations. Le nombre de variables est souvent considérablement plus grand que celui des observations : une image d'un megapixels en couleur correspond à trois millions de variables… La théorie de l'apprentissage déjà évoquée (Hastie & al., 2001) fournit le cadre théorique adapté tout en faisant le lien avec des aspects bien connus des statisticiens: estimation fonctionnelle (splines de lissage, estimateurs à noyaux) et régression non-paramètrique.

5. Conclusions et recommandations

La disponibilité accrue de bases de données médicales de plus en plus vastes (carte Vitale, données hospitalières, grandes enquêtes, etc.) sera un domaine de prédilection pour les méthodes de fouille de données, et des découvertes pourront certainement en être tirées. Ces découvertes doivent être validées par des techniques différentes des tests de la statistique classique.

Une nouvelle forme d'inférence pour les grands ensembles de données est en train d'émerger et le « data mining » est aussi une source de recherches théoriques et pas seulement un ensemble de techniques empiriques. Le data mining n'est certainement pas une mode éphémère, mais une démarche et des outils appropriés à l'analyse des très grandes bases de données. Il serait dommageable de laisser ce champ aux seuls informaticiens, car de par leur formation à l'aléatoire et à la compréhension de la variabilité, statisticiens et épidémiologistes sont les plus à même d'en tirer profit et d'en déjouer les pièges.

Quelques recommandations :

- Associer dans les formations universitaires épidémiologie et bioinformatique.
- Enseigner aux épidémiologistes les techniques et les outils de la fouille de données, ainsi que les bases de données
- Organiser des groupes de travail interdisciplinaires à l'instar du « DIMACS Working Group on Data Mining and Epidemiology » de Rutgers.

Références :

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSSB* 57, 289-300.

Brossette, S.E., & al. (1998) Association rules and data mining in hospital infection control and public health surveillance, *J Am Med Inform Assoc*. **5**(4):373-81.

Domingos P., Hulten G. (2000) Mining high-speed data streams. ACM SIGKDD; Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (eds.) (1996) *Advances in Knowledge Discovery and Data Mining.* Menlo Park, California: AAAI Press

Friedman, J.H. (1997) Data mining and statistics : what's the connection ? *http://www-stat.stanford.edu/~jhf/ftp/dm-stat.ps*

Ge,Y., Dudoit,S., Speed, T.P. (2003). Resampling-based multiple testing for DNA microarray data analysis. *Test*, 12, 1-77.

Hand, D.J. (1998) Data mining: statistics and more ?, *The American Statistician*, **52**, 112-118

Hastie,T. Tibshirani, R., Friedman J. (2001) *The Elements of Statistical Learning*, Springer.

Hébrail, G., Lechevallier, Y. (2003) Data Mining et Analyse des données in *Analyse des données*, G.Govaert éditeur, Hermes, 323-355

Lavrac N. (1999) Selected techniques for data mining in medicine, *Artificial Intelligence in Medicine*, **16,** 3 - 33

Saporta, G. (2000) *Data Mining and Official Statistics*, Quinta Conferenza Nationale di Statistica, ISTAT, Rome

Journal de la SFdS 142,1, (2001) : n°spécial sur le data mining