En guise de conclusion

Résumé

Résumer les grandes lignes de ce cours dans une vue synthétique : méthodes et stratégies dans l'objectif d'une comparaison globale des méthodes sur les différents jeux de données (cancer, pollution, carte visa). Il évoque enfin les pièges fréquents de telles démarches et revient sur la place du statisticien ou data scientist.

Retour au plan du cours

1 Stratégies du data mining

Les chapitres précédents décrivent les outils de base du prospecteur ou scientifique des données tandis que les logiciels en proposent une intégration plus ou moins complète, plus ou moins conviviale de mise en œuvre. En pratique, l'enchaînement de ces techniques permet la mise en place de *stratégies de fouille* bien définies. Celles-ci dépendent essentiellement des *types* de variables considérés et des *objectifs* poursuivis.

Types de variables

Explicatives L'ensemble des p variables explicatives ou prédictives est noté X, il est constitué de variables

- $X_{\mathbb{R}}$ toutes quantitatives [1],
- X_E toutes qualitatives,
- $X_{\mathbb{R}\cup E}$ un mélange de qualitatives et quantitatives.

À expliquer La variable à expliquer ou à prédire ou *cible* (target) peut être

- Y quantitative,
- Z qualitative à 2 modalités,
- T qualitative.

1. Une variables explicative qualitative à 2 modalités (0,1) peut être considérée comme quantitative ; c'est l'indicatrice des modalités.

Objectifs

Trois objectifs principaux sont poursuivis dans les applications classiques de fouille / science des données :

1. **Exploration multidimensionnelle** ou réduction de dimension : production de graphes, d'un sous-ensemble de variables représentatives X_r, d'un ensemble de composantes C_q préalables à une autre technique.
2. **Classification** (clustering) ou segmentation : production d'une variable qualitative T_r.
3. **Modélisation (Y ou Z)/Discrimination (Z ou T)** production d'un modèle de prévision de Y (resp. Z, T).

D'autres méthodes plus spécifiques à certains objectifs peuvent apparaître : détection d'atypiques ou d'anomalies, imputation, préalables ou non à la modélisation.

Outils

Les méthodes utilisables se classent en fonction de leur objectif et des types de variables prédictives et cibles.

Exploration

`ACP` $X_{\mathbb{R}}$ et $\emptyset$

`AFCM` X_E et $\emptyset$

`AFD` $X_{\mathbb{R}}$ et T

Classification

`CAH` $X_{\mathbb{R}}$ et $\emptyset$

`NuéeDyn` $X_{\mathbb{R}}$ et $\emptyset$

...

Modélisation

1. Modèle linéaire généralisé

   `RLM` $X_{\mathbb{R}}$ et Y

   `ANOVA` X_E et Y

   `ACOVA` $X_{\mathbb{R}\cup E}$ et Y

`Rlogi` $X_{\mathbb{R}\cup E}$ et Z

`Lglin` X_T et T

2. Analyse discriminante

   `ADpar/nopar` $X_{\mathbb{R}}$ et T

3. Classification and regression Tree

   `ArbReg` $X_{\mathbb{R}\cup E}$ et Y

   `ArbCla` $X_{\mathbb{R}\cup E}$ et T

4. Réseaux neuronaux

   `percep` $X_{\mathbb{R}\cup E}$ et Y ou T

5. Agrégation de modèles

   Bagging $X_{\mathbb{R}\cup E}$ et Y ou T

   RandFor $X_{\mathbb{R}\cup E}$ et Y ou T

   Boosting $X_{\mathbb{R}\cup E}$ et Y ou T

6. Support Vector Machine

   SVM-R $X_{\mathbb{R}\cup E}$ et Y

   SVM-C $X_{\mathbb{R}\cup E}$ et T

Stratégies

Les stratégies classiques de la fouille de données consistent à enchaîner les étapes suivantes :

1. **Extraction** de l'entrepôt des données éventuellement par sondage pour renforcer l'effort sur la qualité des données plutôt que sur la quantité.
2. **Exploration**
   - Tri à plat, et étude des distributions : transformation, recodage éventuel des variables quantitatives, regroupement de modalités des variables qualitatives, élimination de variables (trop de données manquantes, quasi constantes, redondantes...). Gestion des données manquantes et valeurs atypiques.
   - Étude bivariée, recherche d'éventuelles relations non linéaires, de variables redondantes, d'incohérences.
   - Étude multivariée, représentations en dimension réduite (ACP, AFCM) et classification non-supervisée par classification ascendante hiérarchique (CAH) ou *k*means ou stratégie mixte.
3. **Apprentissage :** régression ou discrimination (classification supervisée).
   - Itérer les étapes suivantes :
     (a) Extraction d'un échantillon *test*,
     (b) Estimation, optimisation (validation croisée) des modèles pour chacune des méthodes utilisables.
     (c) Prévision de l'échantillon test.
   - Comparer les distributions et moyennes des erreurs de prévision, éventuellement les courbes ROC.
   - Choisir une méthode et le modèle associé de complexité "optimale" et le ré-estimer sur l'ensemble de l'échantillon.
4. **Exploitation** du modèle sur l'ensemble des données et diffusion des résultats.

2 Comparaison des résultats

La procédure décrite ci-dessus a été systématiquement mise en œuvre en automatisant dans R ou Python l'extraction aléatoire d'un échantillon test et les estimations, optimisations des différents modèles. Les codes sont disponibles sous forme de scénarios sur le site wikiwtat. La librairie `caret` (Kuhn, 2008)[1] et celle `Scikit-learn` de Python s'avèrent très efficaces pour mettre en œuvre cette démarche. L'optimisation des paramètres est réalisée par validation croisée.

Chaque échantillon test fournit donc une estimation sans biais de l'erreur de prévision. La distribution de ces erreurs est alors représentée par des diagrammes en boîtes. En discrimination binaire, des courbes ROC complètent les résultats. Les figures suivantes synthétisent les résultats pour les données de cancer du sein, de chromatographie NIR (cookies), de prévision du pic d'ozone et enfin bancaires (appétence carte visa premier). d'autres exemples sont traitées sur le dépôt wikiwtat.

3 Pièges

Les principaux pièges qui peuvent être rencontrés au cours d'une prospection peuvent être le résultat d'un *acharnement* en quête de sens (*data snooping*). Cela signifie qu'à force de creuser, contrairement à un prospecteur minier à la recherche de diamants bien réels, le prospecteur en données disposant d'un grand nombre de variables finit bien, en mode exploratoire, par trouver des relations semblant hautement significatives. Par exemple, au seuil classique, 5% des tests sont, à tort, significatifs et conduisent à des "faux positifs" ou des fausses corrélations. Il suffit donc d'en faire beaucoup, de croiser beaucoup de variables, pour nécessairement trouver du "sens" dans des données. Encore une fois, il est préférable d'éviter le fonctionnement "Shadock" (cf. figure 10) : *je n'ai qu'une chance sur un milliard de réussir ; je me dépêche donc de rater le plus d'essais possibles*.

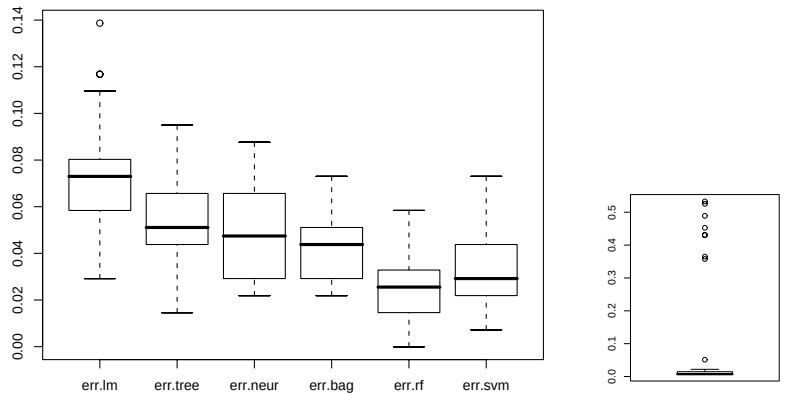

FIGURE 1 – Cancer : Diagrammes boîtes des taux d'erreurs. Le boosting es
mis de côté pour des problèmes d'échelle et de comportement erratique prove
nant d'une focalisation extrême sur une observation imprévisible.

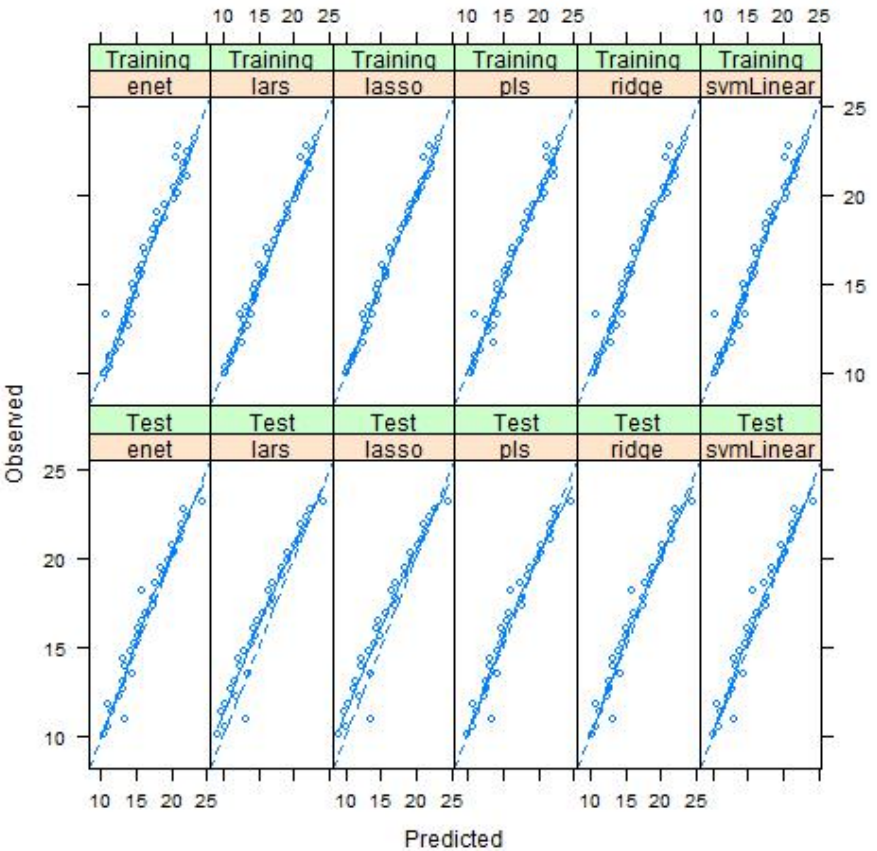

FIGURE 2 – Cookies : Résidus (apprentissage et test) des différents modèles mettant en évidence la forte linéarité des données ainsi que les aspects volontairement atypiques de l'échantillon test original.

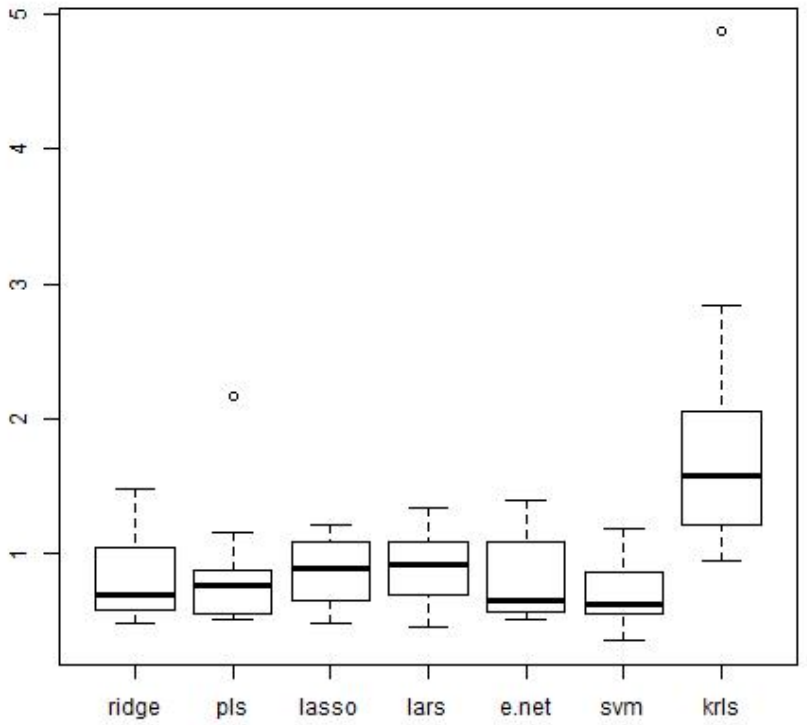

FIGURE 3 – Cookies : Diagrammes boîtes très proches des méthodes linéaires alors que les méthodes non-linéaires ne sont pas retenues car inefficaces. Les SVM (noyau linéaire) conduisent à la meilleure moyenne (0.70) devant la régression ridge (0.84), elastic net (0.85), lasso, PLS (0.86)

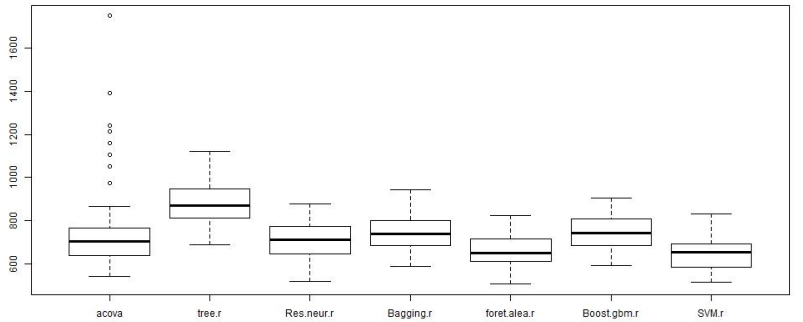

FIGURE 4 – Ozone : Diagrammes boîtes des taux d'erreurs en régression. Meilleur comportement des SVM avec noyau linéaire (649) devant random forest (666). L'analyse de covariance quadratique conduit à une moyenne élevée (774) mais reste utile pour l'interprétation.

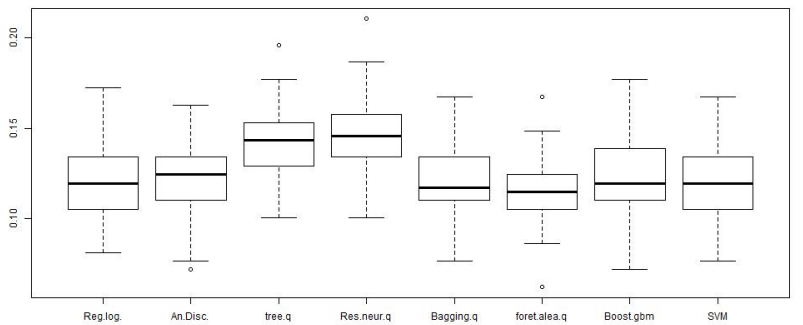

FIGURE 5 – Ozone : Diagrammes boîtes des taux d'erreurs pour la prévision des dépassements de seuil. En moyenne, les deux stratégies (prévision en régression ou directement du dépassement) sont finalement équivalentes pour les meilleures méthodes. Les moyennes se répartissent entre 11 % (random forest) et 14%.

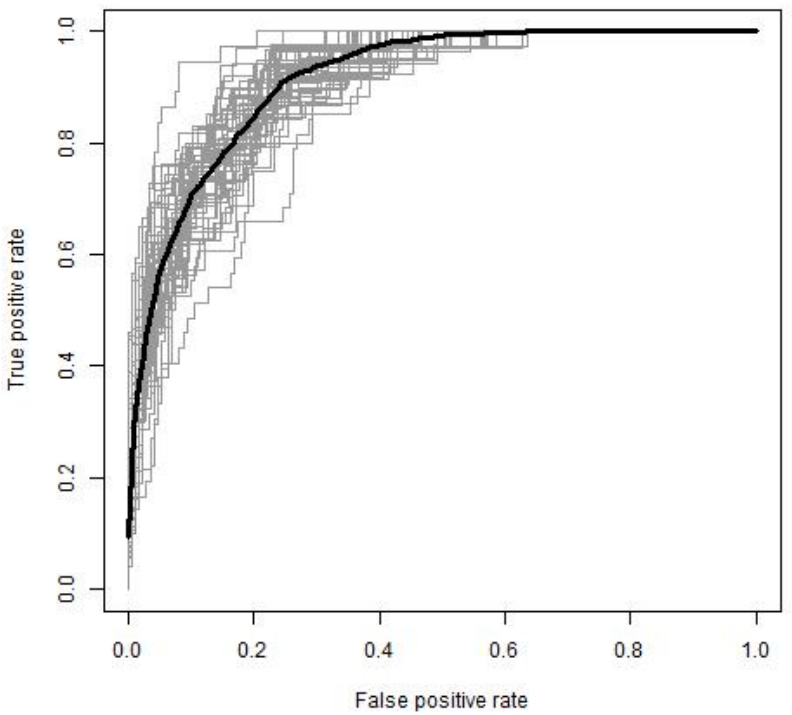

FIGURE 6 – Ozone : Attention,l'échantillon test est petit et les courbes ROC sont fortement dispersées. Il est important d'en calculer une moyenne sur les 50 échantillons tests.

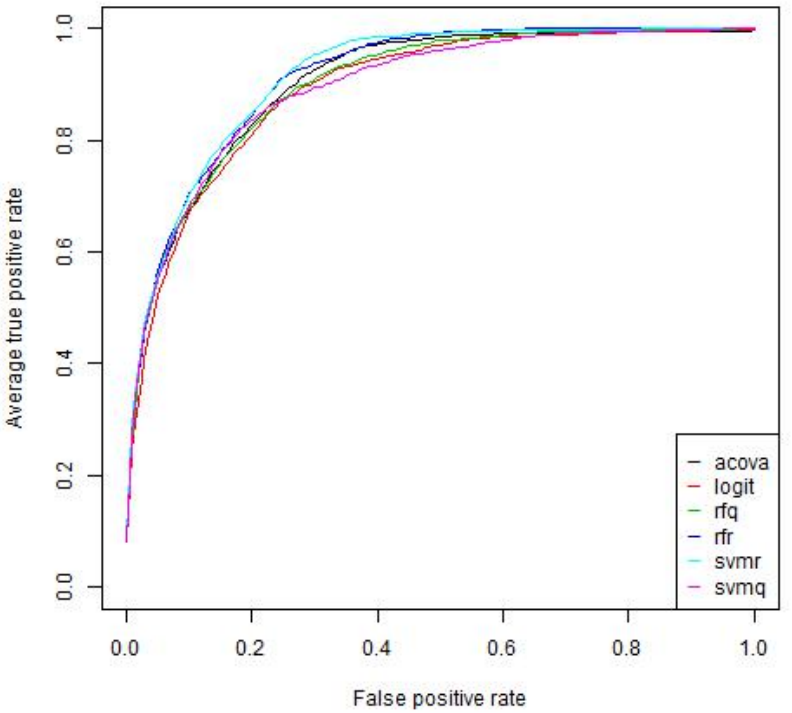

FIGURE 7 – Ozone : Les courbes ROC moyennes, qui permettraient de déterminer un seuil de déclenchement d'alerte, soulignent les meilleures comportements des SVM et de Random forest après régression.

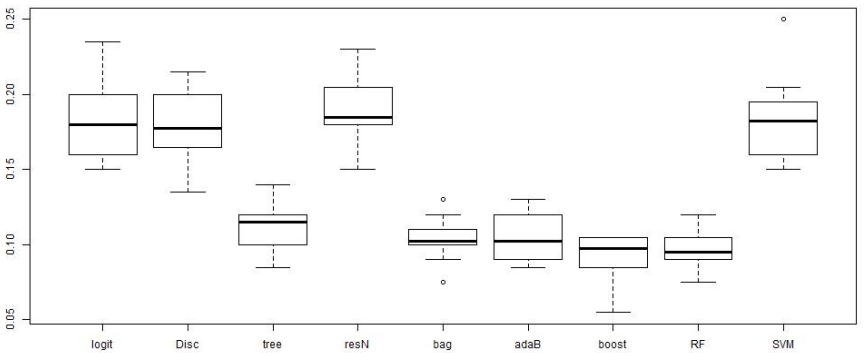

FIGURE 8 – Banque : Diagrammes boîtes des taux d'erreurs. En moyenne, les méthodes basées sur des arbres l'emportent nettement avec celle d'agrégation de modèles (boosting 9%, ranfom forest et bagging 10 %) devant un arbre seul (11 %) très utile pour l'interprétation.

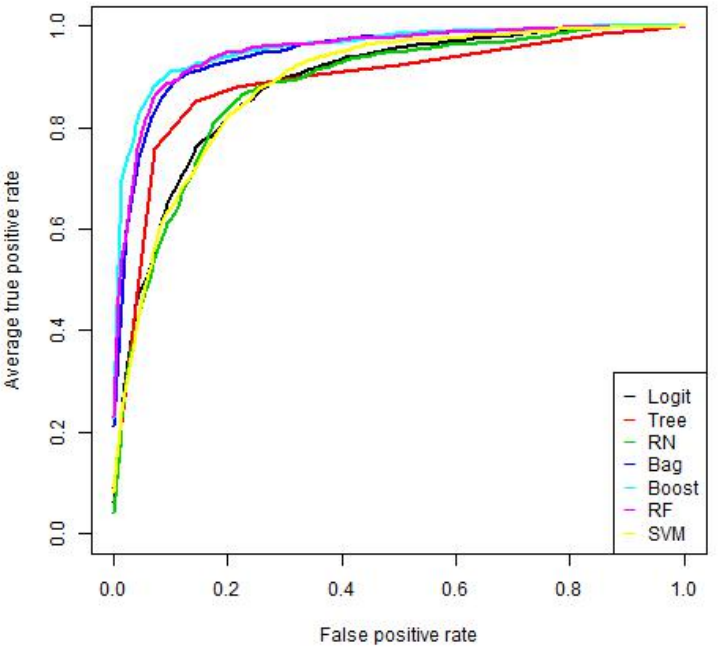

FIGURE 9 – banque : Les courbes ROC moyennes insistent sur le très bon comportement des agrégations de modèles (boosting, random forest, bagging) pour une très grande variété de choix de seuils contrairement à un arbre de discrimination dont les qualité se détériorent pour des seuils faibles.

En phase de modélisation, une sur-paramétrisation ou un sur-ajustement du modèle peut parfaitement expliquer des données sans pour autant que les résultats soient extrapolables ou généralisables à d'autres données que celles étudiées. Les résultats de prévision seront donc entachés d'une forte erreur relative liée à la variance des estimations des paramètres. C'est toujours le problème de trouver un bon compromis entre le biais d'un modèle plus ou moins faux et la variance des estimateurs. Nous insistons donc sur les indispensables phases de choix de modèles et comparaison des méthodes.

4 Rôle du statisticien

4.1 Des compétences multiples

Une bonne pratique du *Data Mining* nécessite de savoir articuler toutes les méthodes entrevues dans ce document. Rude tâche, qui ne peut être entreprise qu'à la condition d'avoir très bien spécifié les objectifs et buts de l'étude. On peut noter que certaines méthodes poursuivent les mêmes objectifs prédictifs. Dans les bons cas, données bien structurées, elles fourniront des résultats très similaires, dans d'autres, une méthode peut se révéler plus efficace compte tenu de la taille de l'échantillon ou géométriquement mieux adaptée à la topologie des groupes à discriminer ou encore en meilleure interaction avec les types des variables. Ainsi, il peut être important et efficace de découper en classes des variables prédictives quantitatives afin d'approcher de façon sommaire une version *non-linéaire* du modèle par une combinaison de variables indicatrices. Cet aspect est par exemple important en régression logistique ou avec un perceptron mais inutile avec des arbres de décisions qui intègrent ce découpage en classes dans la construction du modèle (seuils optimaux). D'autre part, les méthodes ne présentent pas toutes les mêmes facilités d'interprétation. Il n'y a pas de meilleur choix *a priori*, seule l'expérience et un protocole de *test* soigné permettent de se déterminer. C'est la raison pour laquelle la librairie `caret` de R ne font pas de choix et offrent ces méthodes en parallèle pour mieux s'adapter aux données, aux habitudes de chaque utilisateur.

La librairie `caretEnsemble` comme les fonctionnalités *pipeline* de `Scikit-learn` permettent également de combiner des méthodes dans une architecture complexe telle qu'elles se retrouvent généralement dans les solutions gagnantes des concours *Kaggle*. Un ensemble de modèles construisent des prévisions ou vairables qui sont ajoutées comme nouvelles variables à l'entrée d'autres algorithmes. C'est également une des explications du succès de l'apprentissage profond..

FIGURE 10 – Shadoks : Tant qu'à pomper, autant que cela serve à quelque chose !

4.2 De l'utilité du statisticien

Le travail demandé déborde souvent du rôle d'un statisticien car la masse et la complexité des données peuvent nécessiter le développement d'interfaces et d'outils graphiques sophistiqués permettant un accès aisés aux données, comme à des résultats, par l'utilisateur finale à l'aide par exemple d'un simple navigateur sur l'intranet de l'entreprise. Néanmoins, au delà de ces aspects plus "informatiques", l'objectif principal reste une "quête de sens" en vue de faciliter les prises de décision tout en en préservant la fiabilité. Ainsi, la présence ou le contrôle d'une expertise statistique reste incontournable car la méconnaissance des limites et pièges des méthodes employées peut conduire à des aberrations discréditant la démarche et rendant caducs les investissements consentis.

4.3 Vers le Big Data

Le volume des données générées et stockées pas les entreprises industrielles et celles du e-commerce font franchir une nouvelle étape. Nous passons du TéraOctet au PétaOctet. Comme expliqué rapidement en introduction, cette nouvelle étape engendre de nouvelles approches tant pour les architectures des bases de données, la parallélisation des calculs, que pour les algorithmes et méthodes mises en œuvre.

D'un point de vue informatique, une connaissance du nouveau standard *Hadoop* [2] est vivement souhaitée. Il permet la création d'applications distribuées et "échelonnables" (*scalables*) sur des milliers de nœuds pour gérer des pétaoctets de données. Le principe est de découper et paralléliser (distribution) des tâches en lots de données afin de réduire linéairement le temps (scalable) de calcul en fonction du nombre de nœuds. *Hadoop* devient l'outil de référence du web mining et l'e-commerce.

D'un point de vue statistique / mathématique, le nouveau défi est la construction de bases de représentation fonctionnelle et de modèles pertinents pour aborder et prendre en compte des structures de données complexes : géolocalisation sur des graphes, signaux en temps réels, images 3D, séquences... Chaque problème, surtout industriel, nécessite une approche spécifique issue d'une recherche originale dans le cadre souvent d'une thèse, par exemple CIFRE, qu'un d'un développement d'ingénierie classique. Dans le cas de flots de données, l'aide à la décision devient adaptative ou séquentielle.

Références

[1] Max Kuhn, *Building Predictive Models in R Using the caret Package*, Journal of Statistical Software **28** (2008), n° 5.

2. Créé en 2009 et développé en Java par Doug Cutting au sein des projets de la fondation des logiciels libres Apache. Il est inspiré des principes de MapReduce de Google.