Mini-symposium Big Data
Mégadonnées : quelques enjeux.

Mini-symposium industriel

Résumé

Le développement des réseaux sociaux, et plus généralement de l'utilisation du web, a généré un nombre très important de données que les avancées technologiques permettent de traiter de plus en plus efficacement : l'essor du "Big Data" est au coeur de l'actualité. Dans ce mini-symposium industriel, consacré à ces données en grande dimension, différents intervenants présenteront les problématiques qu'ils étudient tant du point de vue académique que professionnel. Puis, une table ronde leur permettra de débattre de certaines questions structurantes liées au Big Data. Cette table ronde sera aussi l'occasion d'échanger avec la salle.

Organisateur(s)

1. **Laurence Carassus**, Research Center, Léonard de Vinci Pôle universitaire et LMR, Université Reims Champagne-Ardenne.

Liste des orateurs

1. **Karine Tribouley**, Soft Computing
   Titre : Le Big Data révolutionne-t-il le CRM ?.
2. **Pierre Alquier**, ENSAE
   Titre : Systèmes de recommandation et algorithmes de complétion de matrices.
3. **Georges Hebrail**, EDF
   Titre : Clustering préservant la confidentialité de données individuelles de consommation électrique.
4. **Maguelonne Chandesris**, SNCF
   Titre : SNCF Big Data, Big Challanges.
5. **Joseph Salmon**, Telecom Paris Tech.
   Titre : Règles d'écrémage sûres et méthodes actives pour accélérer les solveurs du Lasso.
6. **Nicolas Gaussel**, Metori Capital Management
   Titre : Statistical learning et gestion de portefeuille.

Laurence Carassus, Research Center, Léonard de Vinci Pôle universitaire et LMR, Université Reims Champagne-Ardenne, `laurence.carassus@devinci.fr`
Karine Tribouley, Soft Computing et Université Paris Diderot, `ktr@softcomputing.com`
Pierre Alquier, CREST, ENSAE, Université Paris Saclay, `Pierre.Alquier@ensae.fr`
Georges Hebrail, Chercheur Senior à EDF Lab Saclay, `georges.hebrail@edf.fr`
Maguelonne Chandesris, SNCF Innovation & Recherche - Data, Mobilité et Territoires, `maguelonne.chandesris@sncf.fr`
Joseph Salmon, LTCI, Télécom ParisTech, Université Paris-Saclay, `joseph.salmon@telecom-paritech.fr`
Nicolas Gaussel, CEO, Metori Capital Management, `Nicolas.gaussel@metoricapital.com`

Dans ce mini-symposium industriel, les intervenants tant académiques que professionnels présenteront différentes problématiques liées aux données en grande dimension. Dans l'introduction, nous présentons les objectifs détaillés du mini-symposium. Puis, dans les différentes sections, un résumé de chaque intervention sera proposé. Enfin, dans la conclusion, nous présentons les objectifs de la table ronde.

Introduction

Karine Tribouley de Soft Computing nous parlera des apports du Big Data dans la connaissance client. Elle nous présentera deux cas d'usage, le premier en BtB concerne l'aspiration de données et le second en BtC un moteur de recommandation. Le deuxième intervenant, Pierre Alquier de l'ENSAE poursuivra sur le thème des moteurs de recommandation en nous parlant d'algorithmes statistiques de complétion de matrice pour ces problèmes en grande dimension. Une question importante est la confidentialité des données. On se souvient du concours Netflix et des recours légaux qui en avaient découlés, aboutissant au retrait de la base de données. Georges Hebrail d'EDF, après avoir évoqué différents problèmes en lien avec le Big Data étudiés par EDF, nous présentera une méthode de clustering de séries temporelles permettant de préserver la confidentialité des données. Un aspect important du traitement des données en grande dimension est leur visualisation. La quatrième intervenante, Maguelonne Chandesris de la SNCF nous présentera les nouvelles avancées dans ce domaine et comment elle arrive à faire parler les données en les visualisant de façon plus intelligente. Elle nous présentera également la R&D dans le domaine du Big Data à la SNCF. Joseph Salmon de Telecom Paritech nous parlera ensuite de la recherche en Machine Learning du point de vue de l'optimisation et des enjeux computationnels. Il nous présentera également la chaire ML4BG et le Master of Science Big Data pour illustrer les liens qui se développent entre les entreprises et la recherche et les enjeux pour la formation des étudiants à Telecom Paritech. Enfin, Nicolas Gaussel de Metori Capital Management nous parlera des possibilités d'application des techniques du Big Data en finance, où les données sont déjà très travaillées, et sur le traitement statistique qu'il faudrait leur appliquer en amont. Il nous montrera aussi comment une application "brutale" des techniques d'apprentissage peut être inefficace sur des données financières.

1 Le Big Data révolutionne-t-il le CRM ?

Le CRM - gestion de la relation client - désigne l'ensemble des techniques permettant de collecter et d'analyser les informations concernant des (futurs) clients dans un but de (conversion) fidélisation. Classiquement, les données utilisées proviennent de manière explicite du client ou sont extraites de bases internes à l'entreprise indiquant le comportement d'achat et permettent de proposer des modélisations type score et classification aux services chargés des opérations marketing. L'irruption du numérique dans notre monde entraine une ré-organisation complète de cette chaine organisationnelle dont nous allons discuter trois aspects essentiels

- aspect technologique : changement du volume des données ainsi que de leur nature, capacité de stockage, compétence scientifique algorithmique, capacité technique de calcul,
- aspect organisation de l'entreprise : apparition de nouveaux métiers, organisation transverse plutôt qu'en silo,
- aspect offre de service : demande très forte des clients pour de nouveaux services en échange de leurs données personnelles.

Chacun de ces aspects sera approfondi avec une présentation de deux projets spécifiques : la construction d'un IOT équipé d'un moteur de substitution/recommandation pour un grand distributeur et la mise en place de cartes d'identité "Entreprise" pour la prospection B2B d'une société de ré-assurance. Enfin, nous proposerons un tour d'horizon prospectif sur les futurs services imaginés par les entreprises afin de mieux satisfaire - et fidéliser - ses clients.

2 Systèmes de recommandation et algorithmes de complétion de matrices

Pour les sites de vente en ligne (Amazon, CDiscount etc.) être capable de faire de bonnes suggestions d'achats à ses utilisateurs est un enjeu crucial. Ce problème a fait irruption de façon assez spectaculaire

dans la communauté de recherche en mathématiques, statistique et machine learning lors du challenge Netflix : la société proposait une récompense d'un million de dollars à l'équipe de recherche qui améliorerait de 10% la performance de leur moteur de recommandation (Bennett et Lanning, 2007 [1]). Les statisticiens ont proposé de modéliser la recommandation comme une instance du problème de complétion de matrices : une ligne d'une matrice représente un individu, une colonne un produit, et chaque entrée est une valeur numérique qui quantifie l'intérêt de l'individu pour le produit. Les achats passés permettent de connaître un très petit nombre d'entrées de cette matrice, et reconstruire les entrées manquantes, c'est en particulier prédire les produits qui susciteront de l'intérêt chez un utilisateur.
D'un point de vue mathématique, un certain nombre de résultats (Candès et Tao 2009 [2], Candès et Plan 2009 [2]) ont montré que si la matrice à reconstruire est de rang faible, on peut la reconstruire à partir d'un nombre faible d'entrées. Par ailleurs, il est possible de formuler la méthode de reconstruction comme un problème d'optimisation convexe (Candès et Tao 2008 [2]), pour lequel des algorithmes rapides existent.
Il est à noter que depuis, les applications de ce modèle vont bien au delà de la vente en ligne. Les systèmes de recommandation sont par exemple testés par des organismes comme pôle emploi pour recommander des offres à des demandeurs d'emploi. Mais les applications du modèle de complétion de matrice et de variantes vont même bien au delà de la recommandation : physique quantique (Gross et al 2010 [4]), séparation de source en vidéo, etc.
Après avoir exposé brièvement les résultats mathématiques et les algorithmes d'optimisation utilisés en pratique, j'expliquerai également comment utiliser SoftImpute, un des packages R les plus connus pour la complétion de matrice.

3 Clustering préservant la confidentialité de données individuelles de consommation électrique

Au-delà de la facturation, les données de consommation électrique issues des compteurs communicants peuvent être utilisées pour développer de nouveaux services à destination des clients. Dans le but de réaliser des conseils personnalisés sur l'optimisation de leur consommation d'énergie, une approche efficace consiste à comparer chaque client à des profils types de clients construits par un clustering des données de consommation de l'ensemble des clients individuels. Cependant la divulgation de telles données individuelles pose un problème de protection de la vie privée, ces données pouvant révéler des informations sensibles sur l'activité des clients. Une nouvelle approche de clustering de séries temporelles préservant la confidentialité sera présentée et illustrée sur le cas des données individuelles de consommation électrique. Un rapide tour d'horizon des autres cas d'utilisation de la science des données dans le secteur électrique terminera la présentation.

4 SNCF Big Data, Big Challanges

SNCF manipule et exploite quotidiennement de très grands volumes de données, données de plus en plus nombreuses et variées. Les challenges induits seront présentés en mettant notamment l'accent sur les enjeux d'Innovation & Recherche. Un de ces enjeux est la visualisation de données massives qui sera illustrée par des exemples concrets d'applications.

5 Règles d'écrémage sûres et méthodes actives pour accélérer les solveurs du Lasso

Les méthodes de régression linéaire parcimonieuses ont connus un important succès au cours des deux dernières décennies années. D'abord principalement développés dans le contexte de la génomique, ces méthodes ont été popularisée dans des domaines varies allant du traitement des images à la publicité en ligne. Dans cet exposé, on s'intéressera aux enjeux computationnels que représentent les méthodes les plus simples, celles qui reposent sur l'optimisation convexe. En particulier, au delà de l'utilité concernant l'interprétation des modèles, on illustrera l'intérêt computationnel que peut avoir la parcimonie. On verra notamment comment des méthodes de types "screening" et d'ensembles actifs permettent d'améliorer simplement les solveurs actuels, en particulier pour le cas du Lasso ou de ces variantes multi-tâches.

6 Statistical learning et gestion de portefeuille

L'apprentissage statistique est un ensemble d'outils statistiques utilisés pour comprendre des ensembles de données plus ou moins larges et plus ou moins complets. Certaines approches récentes se sont révélées spectaculairement efficace dans des situations de complétion de base de données ou d'extrapolation.
Dans le domaine financier, les résultats semblent moins convaincants voire trompeur. Nul ne semble en mesure, à ce jour, d'organiser un concours "à la Netflix" sur données financières. Il ne paraitrait pas raisonnable non plus à se fier à une stratégie de portefeuille apprise de façon non supervisée.
L'objectif de cet exposé est de montrer comment certaines méthodes d'apprentissage appliquées sur des bruits blancs peuvent se révéler inefficaces voire dangereuses. Il est aussi un appel pour combiner certaines de ses méthodes avec des approches de type "test statistiques" dont l'hypothèse nulle serait que l'échantillon ne contient aucune information.

Conclusions

Lors de la table ronde, les intervenants débâteront de différents enjeux liés au Big Data. Le premier est organisationnel, les problématiques liées à la donnée vont pousser l'entreprise à passer d'une organisation en silo à une organisation transversale et à créer des lieux d'échange, les Data Inovation Lab. On note aussi le développement des hackathons sur des challenges collaboratifs liés à la donnée en grande dimension. Nous parlerons également du formidable moteur qu'est le Big Data pour la recherche en statistique et plus particulièrement sur le développement de méthodes numériques. Nous enchainerons sur les problématiques d'infrastructure de stockage et du traitement informatique de la donnée. Nous évoquerons ensuite les risques liés au monopole de certains algorithmes propriétaires de deep learning, permettant par exemple de prévoir de façon très efficace le comportement des consommateurs. Ces algorithmes, tenant plus du savoir faire, sont des boites noires inaccessibles aux chercheurs, qui ne peuvent reproduire et étudier rigoureusement leurs comportements. Enfin, nous conclurons sur une réflexion sur l'avenir à court et long terme du domaine tant du point de vue universitaire que de l'entreprise et sur une question plus philosophique ou épistémologique : "les données peuvent-elles tout expliquer ?"

Références

[1] Bennett, J. and Lanning, S., *The netflix prize*, Proceedings of KDD cup and workshop, 35, 2007.

[2] Candès, E. J. and Plan, Y., *Matrix completion with noise*, Proceedings of the IEEE, 98, 6, 925–936, 2010.

[3] Candès, E. J. and Tao, T., *The power of convex relaxation : near-optimal matrix completion*, IEEE Trans. Inform. Theory, 56, 5, 2053–2080, 2010.

[4] Gross, D. and Liu, Y.-K. and Flammia, S. T and Becker, S. and Eisert, J., *Quantum state tomography via compressed sensing*, Physical review letters, 105, 15, 150401, 2010.