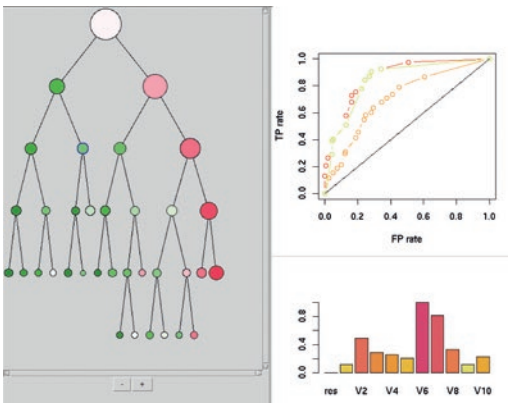
Apprendre automatiquement
à partir des données

Octobre 2014

Le *Big Data* est un phénomène sociétal autant que technologique. Convaincu que les données ont de la valeur, l'Institut Mines-Télécom a fait de ce sujet une thématique importante de recherche. Car, pour stocker ces données, les partager et les exploiter, pour en tirer le meilleur parti, il faut adopter une approche nouvelle. À l'instar du *Machine Learning*, l'objet des recherches de Stéphan Clémençon, professeur à Télécom ParisTech.

Chaque année, plusieurs zettaoctets de données sont produits. C'est-à-dire plusieurs milliards de milliards de milliers d'octets. Si cette manne d'informations permet l'émergence de nouveaux services, elle modifie également considérablement les besoins. Les outils d'hier sont dépassés, il faut imaginer d'autres façons d'extraire de la valeur de cette surabondance. C'est à ces fins qu'a été pensé le « *Machine Learning* », ou apprentissage automatique, une discipline qui mêle mathématiques et informatique pour concevoir des algorithmes de traitement des données massives et dont les applications industrielles sont très nombreuses.

C'est d'ailleurs pour en permettre l'essor que Télécom ParisTech a recruté Stéphan Clémençon, un mathématicien spécialiste de la modélisation et des statistiques. Car, comme le précise Stéphan, « *Dès lors qu'on manipule un très grand nombre de données, c'est le langage probabiliste qui s'impose.* » Un domaine longtemps négligé, en particulier par les écoles d'ingénieurs, où les mathématiques qui y étaient apprises étaient très déterministes, mais qui intéresse de plus en plus les étudiants. Et c'est heureux, tant « *les questions liées à la grande dimension soulèvent des problèmes ardus exigeant beaucoup de créativité !* » De nouvelles méthodes s'imposent, dont la différence principale avec les anciennes tient au fait que celles-ci, basées sur les statistiques traditionnelles, reposaient sur une modélisation *a priori* des données. Elles furent largement développées à une époque, les années 30, où les moyens et les problématiques étaient autres, où les ordinateurs avaient une capacité de calcul très limitée et où les données étaient coûteuses à produire.



Prototype R de la solution logicielle TreeRank

• Trouver le sens caché des données massives

Aujourd'hui, les capteurs sont partout et les données sont prélevées automatiquement. Sans usage prédéfini, mais avec l'idée qu'elles recèlent une information précieuse. L'idée est de considérer ces données avec intérêt et d'en tirer le meilleur parti. L'objectif du *Machine Learning* est de concevoir des algorithmes adaptés au traitement des données massives. Puisque ces données sont trop volumineuses pour qu'il soit envisageable de confier à un expert chaque stade de leur traitement, parce qu'on veut aussi voir émerger des services et des enseignements innovants, sans *a priori*, l'idée est venue de permettre aux machines d'apprendre automatiquement.

Le *Machine Learning*, c'est « *comment une machine peut-elle apprendre à décider toute seule ?* ». Comment compresser, représenter et prédire de l'information à partir de données choisies pour servir d'exemples ? Voilà tout l'enjeu du *Machine Learning*, nourri de modélisation probabiliste et d'optimisation et servi par une théorie de

Une chaire pour le *Machine Learning*

La chaire *Machine Learning for Big Data*, créée fin 2013, regroupe quinze enseignants, tous issus de Télécom ParisTech. Son objectif est d'informer sur le *Machine Learning*, d'illustrer l'ubiquité des mathématiques et de poursuivre un programme de recherche mené avec quatre partenaires privés qui apportent un financement de deux millions d'euros sur cinq ans et des problématiques réelles et concrètes :

- Criteo, leader mondial dans le ciblage publicitaire, tente de proposer à chaque internaute, en fonction de son historique de navigation, le lien sur lequel il a le plus de chances de cliquer. Comment explorer l'espace gigantesque du Web ?
- Le groupe Safran fabrique 70 % des moteurs d'avions civils ou militaires dans le monde. Comment détecter en temps réel les anomalies et proposer le remplacement d'une pièce avant la panne générale ?
- PSA Peugeot Citroën réfléchit à relier les données aux usages. Comment abaisser les coûts de construction, optimiser les offres commerciales et proposer des modèles qui correspondent à l'attente du marché ?
- Un groupe bancaire français lance une banque exclusivement numérique. Comment monitorer en temps réel les comptes des clients, proposer des offres financières adaptées ou simplifier les usages ?

En savoir plus : http://machinelearningforbigdata.telecom-paristech.fr/fr/

l'apprentissage qui donne des garanties sur la robustesse des résultats. Le souci principal est de parvenir à concevoir des algorithmes qui aient une bonne chance de se généraliser. Proposer des règles trop complexes risque d'induire du sur-apprentissage, c'est-à-dire produire des modèles qui conviennent parfaitement pour les exemples fournis, mais qui ne sont pas généralisables. De l'autre côté de la balance, des règles trop frustes ont une capacité prédictive trop pauvre. Dans le cas du *Machine Learning*, ce bon niveau de complexité doit bien sûr être déduit automatiquement des données.

• Créer de nouveaux services efficaces à partir des données

« *Les nombreuses applications du Machine Learning sont un vrai moteur pour la recherche* », souligne Stéphan Clémençon, énumérant quelques exemples qui illustrent la diversité des domaines dans lesquels sont collectées et exploitées des données massives : « *La reconnaissance automatique des visages en biométrie, la gestion des risques en finance, l'analyse des réseaux sociaux en marketing viral, l'amélioration de la pertinence des résultats produits par les moteurs de recherche et de recommandation, l'offre de sécurité dans les bâtiments intelligents ou encore, dans les transports, la surveillance des infrastructures et la maintenance prédictive réalisées à l'aide de systèmes embarqués…* »

En *Machine Learning*, viennent ainsi dans un premier temps les applications, puis dans un second temps les mathématiques qui permettent de les comprendre et de les formuler proprement, et ainsi d'améliorer sensiblement les processus. Il faut donc développer une connaissance « métier » de ces applications. C'est dans cet esprit qu'a été créée la chaire *Machine Learning for Big Data* de Télécom ParisTech (voir encadré) autour des quatre partenaires que sont Criteo, PSA Peugeot Citroën, le groupe Safran et une grande banque française. L'idée est de partir de leurs problématiques pour parvenir ensemble, industriels et universitaires, à des solutions performantes. Avec, à la clé, pour les premiers, une meilleure connaissance de l'état de l'art, et pour les autres, une compréhension accrue des enjeux applicatifs.

Le *Big Data* dans son ensemble désigne à la fois toute une infrastructure dédiée et une liste de problèmes ouverts. Stéphan Clémençon regrette le fait qu'en France, « *nous avons raté le train de l'équipement* », mais heureusement, tempère le chercheur : « *Nous disposons de nombreuses PME innovantes et d'étudiants bien formés, notamment en mathématiques appliquées.* » L'ingénierie des données est nécessairement multidisciplinaire et une école comme Télécom ParisTech, où les enseignements sont variés, a une carte à jouer en proposant des cursus adaptés. D'autant que, comme le souligne Stéphan Clémençon, « *Le Machine Learning correspond à des problématiques d'entreprise très claires et il y a beaucoup de débouchés dans cette filière.* »

Stimuler la recherche et l'enseignement du *Machine Learning*

Stéphan Clémençon a rejoint Télécom ParisTech en 2007 pour y développer la recherche et l'enseignement du *Machine Learning*, l'apprentissage automatique à partir des données. Il est à la tête du groupe STA (Statistiques et applications), en charge du programme de master spécialisé « *Big Data* : Gestion et analyse des données massives ». Il a conçu le certificat d'études spécialisées « *Data Scientist* » délivré par l'école au titre de la formation continue et destiné aux ingénieurs en poste désireux de monter en compétence sur les techniques du *Machine Learning*. Stéphan enseigne également à l'ENS Cachan et à l'université Paris Diderot et il est professeur associé à l'École des Ponts ParisTech et à l'ENSAE ParisTech.

À TÉLÉCOM PARISTECH

Six axes stratégiques ont été identifiés comme présentant un fort potentiel de fédération des forces de recherche :

- *Big Data, Dynamique des données et des connaissances*
- *Confiance numérique, sécurité, sûreté et risques*
- *Très grands réseaux et systèmes*
- *Interactions réel-virtuel*
- *Modélisation*
- *Approche interdisciplinaire de l'innovation*

Le Machine learning est l'une des expertises de l'axe Big Data.

TELECOM ParisTech

Rédaction : Umaps, Frédéric Woirgard • Mise en pages : implica CR&A

Suivez l'actualité recherche & innovation de l'Institut Mines-Télécom

http://blogrecherche.wp.mines-telecom.fr et www.twitter.com/Mines_Telecom

INSTITUT Mines-Télécom

CONTACT INFORMATION RECHERCHE & INNOVATION
recherche@mines-telecom.fr
Institut Mines-Télécom
46 rue Barrault - 75634 Paris cedex 13
France
www.mines-telecom.fr

À PROPOS DE L'INSTITUT MINES-TÉLÉCOM

L'Institut Mines-Télécom est un établissement public dédié à l'enseignement supérieur, la recherche et l'innovation dans les domaines de l'ingénierie et du numérique. Il est composé des dix grandes écoles Mines et Télécom sous tutelle du ministre en charge de l'industrie et des communications électroniques, de deux écoles filiales, de deux partenaires stratégiques et d'un réseau de treize écoles associées. L'Institut Mines-Télécom est reconnu au niveau national et international pour l'excellence de stes formations d'ingénieurs, managers et docteurs, ses travaux de recherche et son activité en matière d'innovation.

L'Institut Mines-Télécom est membre des alliances nationales de programmation de la recherche Allistene, Aviesan et Athena. Il entretient des relations étroites avec le monde économique et dispose de deux Instituts Carnot. Chaque année une centaine de start-up sortent de ses incubateurs.