Big Data : enjeux technologiques et impact scientifique

Par Stephan CLÉMENÇON
Professeur de mathématiques appliquées à Télécom-ParisTech,
Institut Mines-Télécom

L'évocation du terme Big Data provoque généralement une réaction ambivalente. Une crainte, fondée le plus souvent sur des dangers bien réels : une automatisation des processus de décision pouvant s'accompagner d'une perte de contrôle, un impact négatif sur l'emploi, la dépendance de certaines activités à l'égard des systèmes d'information et la disparition de la vie privée. Mais également un engouement certain pour ce que les masses de données aujourd'hui disponibles, combinées à des sciences et technologies de l'information en plein essor, le *machine learning* en particulier, pourraient permettre d'accomplir dans de nombreux secteurs (science, médecine, commerce, transports, communication, sécurité), à l'instar des progrès réalisés ces vingt dernières années dans des domaines tels que la vision par ordinateur ou la reconnaissance de la parole. S'il est encore aujourd'hui difficile de percevoir précisément comment organiser une régulation efficace sans pour autant brider les avancées promises, la maîtrise des risques passe en partie par l'éducation et la formation, par une plus grande diffusion d'une « culture des données et des algorithmes ». Les peurs suscitées par l'automatisation ne sont pas nouvelles. Dans le cas du traitement des masses d'informations numérisées, cette automatisation est pourtant inévitable et souhaitable. Perçue à tort comme une discipline visant à remplacer l'expertise d'un opérateur humain par des machines réalisant des tâches automatisées définies par des données, le *machine learning* a au contraire pour objectif de nous aider à exploiter les données brutes collectées par les capteurs modernes (télescope spatial, spectromètre de masse, téléphones mobiles), portant une information complexe qu'il nous est absolument impossible d'embrasser sans un traitement mathématique adéquat, mis en œuvre au moyen de programmes informatiques dédiés. Il est aujourd'hui à l'œuvre dans de nombreux domaines et s'incarne avec succès dans des applications telles que la vidéosurveillance, la maintenance prédictive des grands systèmes et infrastructures ou les moteurs de recommandation sur le web.

On peut prévoir que ce corpus de connaissances et techniques à l'interface des mathématiques et de l'informatique, en progrès constant depuis quelques décennies, sera encore à l'origine de nombreuses innovations à fort impact sociétal, économique ou scientifique, pour peu que son potentiel soit compris par un public de plus en plus large, qu'il soit maîtrisé par un nombre croissant d'ingénieurs et de cadres techniques, et qu'il se confronte aux enjeux de la société moderne. Le véritable danger de l'automatisation du traitement des données massives résiderait au contraire dans une pénurie d'expertise et des compétences qui permettent de vérifier les conditions dans lesquelles les données sont collectées, d'assurer leur véracité et le bien-fondé des modèles statistiques sur lesquels reposent les applications modernes, et d'interpréter les résultats.

Le paradigme de l'apprentissage statistique

L'un des exemples les plus éloquents de l'impact du Big Data est sans aucun doute celui de la reconnaissance de formes. Celle-ci s'incarne dans les applications de l'intelligence artificielle les plus fréquemment mises en avant aujourd'hui pour illustrer l'efficacité des solutions qu'elle permet de produire, telles que la vision par ordinateur, la reconnaissance automatique de la parole ou de l'écriture manuscrite.

Les concepts mathématiques et algorithmiques à l'œuvre pour mettre au point ces systèmes intelligents sont pourtant loin d'être nouveaux. Même s'ils ont fait l'objet d'une amélioration significative ces dernières décennies, leur élaboration remonte pour l'essentiel à plus d'un demi-siècle. Dans tous ces problèmes, la tâche que la machine doit accomplir consiste, à partir d'une donnée d'entrée X, en la reconnaissance automatique avec une marge d'erreur minimale d'une catégorie Y d'un certain type, spécifié à l'avance, et dont relève la donnée X. Pour reprendre l'exemple de la biométrie, X peut être, par exemple, une image pixellisée ou un signal sonore, et Y est l'identité de l'individu figurant sur l'image ou dont la voix a été capturée par le signal enregistré. Les mêmes technologies sont déployées désormais dans le cadre de l'aide au diagnostic ou pronostic médical ou dans la gestion du risque de crédit, mais on comprendra aisément que les données d'entrée X déterminant alors la catégorie Y dans une bien moindre mesure, le niveau d'erreur attendu est largement supérieur pour ces applications à celui des moteurs de reconnaissance biométrique évoqué précédemment. La reconnaissance de forme est un problème prédictif dans la mesure où les règles élaborées ne doivent pas seulement pouvoir être mises en œuvre au moyen des bibliothèques logicielles disponibles et minimiser l'erreur commise sur une base de données historiques contenant un certain nombre d'exemples (X,Y) appelés « données d'apprentissage », mais ces règles doivent aussi permettre de prédire efficacement le label Y pour de nouvelles entrées X, non encore observées mais issues de la même population statistique que les exemples d'apprentissage (on conviendra qu'il est toujours aisé de « prédire le passé »). On parle alors de capacité de généralisation de la règle prédictive. La formulation du problème d'apprentissage d'une telle règle convoque donc naturellement le langage des probabilités et sa résolution pratique consiste à sélectionner une règle prédictive, au moyen d'un algorithme d'optimisation opérant sur une classe donnée de règles candidates, minimisant une version statistique de la probabilité d'erreur calculée à partir des exemples stockés dans la base d'apprentissage. La théorie mathématique élaborée par Vladimir Vapnik à la fin des années 1960 garantit la capacité de généralisation des règles ainsi construites, pour peu que les classes à partir desquelles l'apprentissage automatique est réalisé soient d'une complexité contrôlée. Le cadre de validité qu'elle a permis de donner à l'apprentissage statistique a fait naître un courant de recherche très actif, mobilisant des chercheurs à l'interface de plusieurs disciplines, les mathématiques et l'informatique bien sûr, mais aussi les sciences cognitives.

L'impact du Big Data

Mais si les concepts fondamentaux du *machine learning* et certains algorithmes tels que les réseaux de neurones sont présents sous des formes très abouties dès la fin des années 1970, ce n'est qu'au commencement de l'ère du Big Data, il y a une dizaine d'années, que le *machine learning* a pu commencer à rencontrer le succès qu'on lui connaît aujourd'hui. Les obstacles principaux résidaient d'une part en la rareté de l'information numérisée, la collection de données s'effectuant alors le plus souvent à travers des plans de sondage coûteux, et d'autre part en des capacités de mémoire et de calcul limitées, interdisant la mise en œuvre de programmes d'optimisation opérant sur de vastes classes de règles pour réaliser un apprentissage efficace. Dans bien des situations, les faibles capacités prédictives des règles produites par le *machine learning* pouvaient ainsi être imputées tout à la fois à une erreur statistique inhérente au faible nombre d'exemples à partir desquels l'apprentissage s'effectue, et au caractère fruste des modèles prédictifs constituant les classes sur lesquelles les programmes d'optimisation peuvent être appliqués. Les briques technologiques ayant permis le développement du web, comme les systèmes de fichiers distribués du *framework* Hadoop ou les langages de programmation tels que MapReduce, ont en effet engendré des progrès considérables dans le domaine de la collecte et du stockage de données et du traitement massivement distribué et parallélisé. Les mégadonnées du web, les immenses bibliothèques d'images, de sons ou de textes « étiquetés » auxquelles il permet d'accéder, entraînent ainsi les moteurs de reconnaissance de contenu avec d'innombrables exemples. Les avancées réalisées dans la gestion

de la mémoire ou dans le domaine du calcul parallélisé grâce en particulier aux processus graphiques permettent la mise en œuvre de programmes d'apprentissage opérant sur des classes très flexibles, telles que les réseaux de neurones profonds (*deep learning*), susceptibles, pour de nombreux problèmes, de rendre compte très efficacement de la façon dont l'information en entrée X permet de prédire la sortie Y. L'ubiquité des capteurs et le développement de l'*Internet of Things* (IoT) facilitent désormais l'accès à l'information numérique, et d'innombrables applications sont développées aujourd'hui sur le modèle de la reconnaissance de forme.

Les infrastructures de collecte, de gestion des masses de données et de calcul ne conditionnent cependant pas à elles seules les progrès réalisés dans le domaine du *machine learning*, et l'avenir ne se bornera pas à simplement décliner les applications du *deep learning*. Par exemple, la volonté d'embarquer des moteurs de reconnaissance biométrique performants dans des smartphones sans compromettre leur autonomie incite les chercheurs à comprendre comment « compresser » ces réseaux profonds de manière à limiter les échanges d'énergie sans pour autant dégrader la qualité de la reconnaissance.

Si le Big Data correspond pour l'apprentissage statistique à une sorte de nirvana, dont les méthodes sont d'autant plus fiables qu'elles sont fondées sur l'observation d'expériences « en grand nombre », le contrôle des conditions d'acquisition des données et des hypothèses de validité des algorithmes prédictifs est indispensable au succès des modèles calculés par les machines. La culture probabiliste et statistique devrait ainsi prendre une place de plus en plus importante dans la plupart des cursus universitaires, et pas seulement dans celui des *data scientists*, ces nouveaux spécialistes des statistiques algorithmiques. La diffusion accrue de cette culture ferait en particulier s'évanouir la crainte d'un monde où le Big Data permettrait de prédire sans erreur nos comportements ou la date de notre mort... Les « grands nombres » permettent en effet d'estimer la performance prédictive des modèles, d'évaluer les risques avec précision et d'optimiser les décisions en univers incertain, mais pas de réduire le caractère intrinsèquement aléatoire de certains phénomènes.