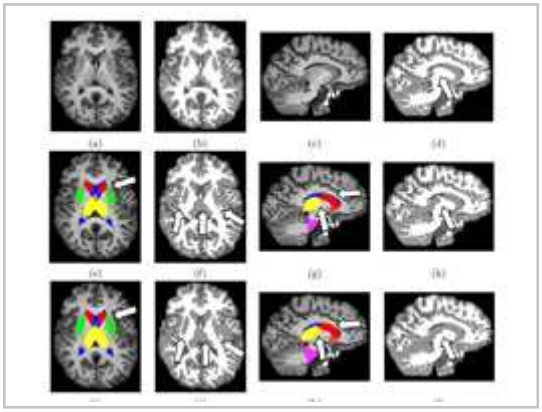
L'apprentissage statistique de plus en plus performant

Florence Forbes, Zaid Harchaoui, Françoise Breton

Florence Forbes, Zaid Harchaoui, Françoise Breton (Technoscope) - 25/10/2012

L’apprentissage statistique de plus en plus performant



Logiciel Locus Mistis - *©Mistis*

Les algorithmes d’apprentissage statistique sont aujourd’hui capables de fournir des outils très performants pour réaliser des tâches de reconnaissance sur de très grands ensembles de données complexes. De nombreux commerces ou services web, par exemple, utilisent aujourd’hui ces technologies pour ajuster leur offre de produits aux goûts de l’internaute ou fournir des traductions automatiques tout à fait honorables.

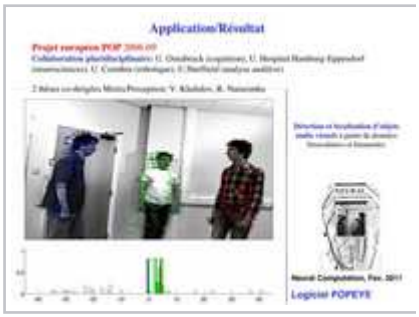
Témoignages de **Florence Forbes,** équipe-projet Mistis et de **Zaid Harchaoui**, équipe-projet Lear

Jusqu’aux années 1990, l’apprentissage automatique, utilisé notamment pour reconnaître des caractères ou des images, s’appuyait sur un modèle dit « génératif » de ce qu’il fallait reconnaître. Ce fut un réel changement de vision que de s’affranchir de ce type de modélisation, souvent difficile à définir mathématiquement, et de lui préférer des méthodes statistiques s’attachant plutôt à optimiser un critère de performance. Ces dernières permettent, à partir d’exemples, d’identifier des éléments communs qu’il s’agira, ensuite, de reconnaître sur de nouveaux objets.
L’approche s’est révélée très performante pour résoudre des tâches sur des données ou des phénomènes complexes, présentant une grande variabilité et susceptibles d’évoluer au cours du temps. Elle s’est particulièrement développée dans les années 1990, par exemple autour du problème de la reconnaissance automatique des codes postaux manuscrits grâce aux bases de données fournies par les services postaux américains. Par la suite, de nouvelles compétitions, académiques (comme les « Pascal challenges ») ou lancées par des entreprises pour résoudre un problème particulier, ont contribué à stimuler la recherche dans ce domaine.

> *« De l’identification de spam au e-commerce, en passant par la bio-informatique »*

La recherche dans ce domaine a longtemps été limitée par la très faible puissance des ordinateurs car les algorithmes d’apprentissage doivent s’entraîner sur un grand nombre d’exemples pour être performants.

Avec le développement fulgurant de l’électronique, les applications ont explosé, autant pour les services grands publics que pour les usages scientifiques. Les algorithmes d’apprentissage sont couramment employés aujourd’hui dans le domaine de la reconnaissance d’images ou de vidéo, tout comme en bio-informatique qui utilise cette approche pour estimer, à partir de bases de données des molécules connues, la fonction chimique probable d’une nouvelle molécule.



"Détection et localisation, sur des images vidéo, d'objets mobiles émettant des sons par apprentissage non supervisé. Ici il s'agit de personnes ; celle qui parle est marquée d'un point blanc" - *© Inria / Mistis - Perception*

Cette approche a également été utilisée avec succès pour détecter les *spams* d’après le contenu des messages, et le web, avec ses immenses quantités de données complexes et non organisées, est un domaine d’application de choix. Les techniques statistiques permettent par exemple à des services d’e-commerce de proposer des produits selon les goûts de l’internaute, inférés par ses choix antérieurs et ceux des autres clients. Certains services en ligne, comme Deezer, permettent même d’assister en direct au processus d’apprentissage : ce service radio propose des airs similaires à la chanson choisie au départ et affine progressivement ses propositions en fonction du jugement de l’utilisateur.

Et dans 20 ans ?

Zaid Harchaoui, chercheur de l'équipe projet Lear

« J’espère que l’on arrivera un jour à développer des algorithmes capables de détecter l’information pertinente dans des jeux d’exemples extrêmement variables et avec un fort bruit de fond. Cela permettrait, par exemple, de concevoir des moyens efficaces pour interpréter l’activité électrique cérébrale et offrir ainsi à des personnes handicapées, par exemple ayant perdu l’usage de la parole, un moyen de recouvrer leur autonomie.»

Florence Forbes, responsable scientifique équipe projet Mistis

« Jusqu'à présent, les chercheurs ont développé de l'apprentissage supervisé, c'est-à-dire que les exemples fournis à l'algorithme d'apprentissage sont accompagnés d'une description textuelle très précise de ce qu'ils représentent (annotations). Le défi des prochaines années est d'arriver à faire travailler un algorithme sur un flux de données non annotées ou partiellement annotées. C'est un domaine nouveau pour lequel les aspects théoriques sont très peu développés et qui est souvent décrit comme une limite indépassable. Nous commençons à explorer les marges de cette question en soumettant des exemples dont les annotations peuvent être manquantes ou comporter des erreurs, et avec parfois l'objectif moins ambitieux de regrouper les données selon leur similarité plutôt que de les identifier. »

Dates clés

- **1990 :** l'essor des compétitions sur la reconnaissance des codes postaux suscite de nombreux travaux dans le domaine.
- **1995** : sortie du livre « The Nature of Statistical Learning Theory » de Vladimir N. Vapnik, théoricien russe travaillant à ce moment-là chez AT&T. Ce *best-seller* a contribué à l'engouement pour le domaine de l'apprentissage statistique.
- **2003** : le réseau d'excellence européen Pascal contribue à constituer la communauté de l'apprentissage statistique. Son succès encore aujourd'hui témoigne de la vitalité du domaine.

Numérique & société

- **2009 :** Netflix, un loueur de DVD américain, lance un défi de 1 million de dollars à toute équipe parvenant à améliorer son système de 10% en performance. Très médiatisé, ce défi a suscité de nombreuses vocations chez les étudiants aux Etats-unis, et a contribué à illustrer la puissance des méthodes d'apprentissage statistique pour un problème aussi difficile que prédire les goûts cinématographiques d'utilisateurs en fonction de leurs choix passés. C'était aussi la première entreprise à soumettre à la communauté scientifique un problème sous forme de défi.
- **2009 :** EDF Osiris développe outils d'aide à la décision. Réalisation d'études sur les prévisions de consommation par le département EDF Osiris (Optimisation simulation risque et statistiques pour les marchés de l'énergie)
Source : innovation EDF

1992 - 2012

- Collection "20 ans d'avancées et de perspectives en sciences du numérique" par les chercheurs d'équipes Inria de Grenoble et Lyon.
- www.inria.fr/20ansgrenoble