École des Ponts
ParisTech

DOSSIER
RECHERCHE
de l'École des Ponts ParisTech
20

CERTIS

Comprendre des systèmes complexes *via* l'apprentissage statistique

Jean-Yves Audibert et Renaud Keriven

Centre d'Enseignement
et de Recherche
en Technologies
de l'Information
et Systèmes

Université Paris-Est /
École des Ponts ParisTech

L'apprentissage statistique est une discipline à l'intersection entre statistique, informatique et mathématiques appliquées. Elle vise à dégager des théories et techniques générales permettant de traiter les systèmes complexes pour lesquels il n'existe pas de modèle simple, précis et facile à mettre en œuvre.
Le développement de capteurs d'acquisition de données, l'accroissement des capacités de stockage et la puissance de calcul des ordinateurs offrent de nouvelles perspectives pour comprendre et modéliser des systèmes de plus en plus complexes à partir d'observations.
Cet apprentissage automatique est utilisé pour remplir des tâches en vision par ordinateur insolubles par les méthodes classiques (détection d'objets, reconnaissance de l'écriture manuscrite, segmentation, indexation, recherche d'images).
Ce domaine d'expertise du CERTIS a un large spectre d'applications : analyse et indexation de textes (par exemple, détection de spams), analyse des marchés financiers, moteur de recherche, bio informatique, reconnaissance de la parole, robotique, génie industriel, etc.

► **[Fig. 1]**
Quelques images numériques d'un même objet. L'apprentissage statistique est la solution en plein essor pour reconnaître automatiquement des objets malgré leur apparence variable

CERTIS
Ecole Nationale des Ponts et Chaussées
Centre d'Enseignement et de Recherche en Technologies de l'Information et Systèmes

UNIVERSITÉ PARIS-EST

http://certis.enpc.fr

Problématique

Le problème de l'apprentissage peut se formaliser ainsi : N couples entrée-sortie sont observés. Une nouvelle entrée X arrive. Quelle est la sortie Y lui étant associée ? L'entrée est le plus souvent un objet complexe alors que la sortie est simple (un nombre réel ou un élément d'un ensemble fini [Fig. 2 et 3]. La modélisation probabiliste la plus usuelle est de considérer que les données observées ainsi que le couple entrée-sortie (X,Y) sont des réalisations indépendantes et identiquement distribuées d'une loi (probabiliste) inconnue qui caractérise le problème.

« Fonction cible »

La meilleure fonction de prédiction ou « fonction cible » est définie comme une fonction de l'espace des entrées dans l'espace des sorties faisant l'erreur la plus faible en moyenne sur un jeu de données infini généré suivant cette loi.
Cette « fonction cible » n'a pas nécessairement une erreur nulle car une même entrée ne correspond pas nécessairement toujours à la même sortie. Par exemple, deux assurés ayant les mêmes caractéristiques (salaire, âge, profession, etc.) peuvent mener à des gains ou des pertes très variables.

Difficulté de l'apprentissage

La difficulté de l'apprentissage provient du caractère aléatoire du problème et de l'impossibilité de modéliser simplement la « fonction cible » qui rappelons-le, est définie sur un espace très complexe. À cela, se rajoutent les contraintes de mise en œuvre sur un ordinateur : rapidité en temps de calcul, faiblesse de la mémoire requise, simplicité de l'implémentation, robustesse numérique, etc. On se contente par conséquent, d'inventer des algorithmes fournissant une fonction de prédiction approchant la « fonction cible ».

Consistance universelle

Un algorithme est dit universellement consistant quand la prédiction qu'il fournit devient similaire à celle de la fonction cible lorsque le nombre N de couples entrée-sortie observés devient suffisamment grand.
De tels algorithmes existent. Malheureusement, il est impossible de savoir à l'avance la taille de la base d'exemples nécessaire pour atteindre un niveau de précision donné et un algorithme universellement consistant peut fournir de mauvaises prédictions par manque de données-exemples.

Deux grandes classes d'algorithmes universellement consistants

Une première classe d'algorithmes est basée sur le moyennage local des sorties observées. L'idée de ces algorithmes est que pour prédire la sortie Y associée à X, il suffit de « moyenner » les sorties correspondant aux entrées de la base d'exemples proches de X. Au début des années 1970, l'étude de ces algorithmes a permis de montrer leur consistance et de ce fait, l'existence d'algorithmes universellement consistants.

Une deuxième classe d'algorithmes a été développée ces vingt dernières années sur la base des travaux précurseurs de Vapnik et Cervonenkis, mathématiciens russes actuellement professeurs d'informatique à l'université de Londres.
Elle repose, d'une part, sur la définition d'un espace de fonctions suffisamment grand pour pouvoir approcher « fonction cible » et, d'autre part, sur le calcul de la fonction f de cet espace qui explique le mieux les données observées. La prédiction associée à X est alors f(X). Les réseaux de « neurones », les méthodes d'agrégation et les machines à « vecteurs supports » font partie de cette classe d'algorithmes et sont également universellement consistants.

Domaine	Tâche d'apprentissage	Entrée	Sortie
Vision par ordinateur	Reconnaissance de chiffres manuscrits	Une image	le chiffre présent dans l'image
Vision par ordinateur	Reconnaissance d'objets	Une image	l'objet présent dans l'image
Analyse de textes	Détection de spams	Un texte	"spam" ou "non spam"
Finance	Evaluation du taux d'emprunt	Un client (âge, profession, revenus,...)	le taux d'équilibre pour la banque
Assurance	Evaluation de la prime d'un contrat	Un assuré (âge, profession, historique,...)	la prime d'équilibre pour l'assureur

[Fig. 2]
Domaines d'application et exemples de problèmes traités par l'apprentissage statistique

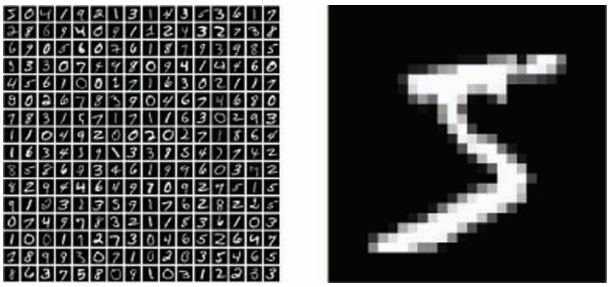
CERTIS

La connaissance *a priori*

L'universelle consistance est une propriété souhaitable. Néanmoins, elle ne garantit pas l'efficacité de l'algorithme sur un problème réel comprenant une base de peu d'exemples et des entrées complexes.
Face à de telles données, le statisticien se doit d'introduire ses connaissances *a priori* sur la « fonction cible ». Son travail de modélisation cherche à identifier les régularités et invariants de la « fonction cible ». Par exemple, en vision par ordinateur, une variation de luminosité ou de point de vue change l'aspect d'une image numérique mais n'en change pas le contenu sémantique [Fig. 1].
L'intégration de cette connaissance dans la procédure d'apprentissage se fera le plus souvent en redéfinissant la notion de similarité entre les entrées.

Sélection de modèles

Il est parfois difficile de savoir à l'avance quel est le « bon » *a priori* ou plus généralement quelle est la modélisation qui mènera à la meilleure prédiction. Il est alors raisonnable de proposer plusieurs modèles explicatifs des données. C'est pourquoi, tout un pan de la recherche en apprentissage est consacré à la création de procédures de sélection automatique de modèles. Ce problème de sélection se résume ainsi. Nous avons une collection de modèles plausibles. Pour chacun de ces modèles, nous avons un algorithme qui, lorsque le modèle est correct, fourni une prédiction presque parfaite. La question est comment reconnaître au vu des données observées, lequel des modèles est approprié pour notre problème.



[Fig. 3]
À gauche, exemple d'images de caractères manuscrits: chacune de ces vignettes est un tableau de 28 sur 28 pixels. À droite, une entrée X possible. La sortie Y associée est « 5 ». Le but d'un algorithme d'apprentissage est d'utiliser les images à gauche et la connaissance de ce qu'elles représentent pour trouver automatiquement ce que représente l'image de droite.

L'approche naïve consistant à choisir le modèle qui semble correspondre le mieux à la base d'exemples n'est pas satisfaisante car elle tend à favoriser les modèles très complexes qui, certes, explique bien les données observées, mais qui bien souvent ne fournira pas de bonnes prédictions. Les méthodes de sélection de modèles visent à quantifier la complexité d'un modèle et réaliser un bon compromis entre cette complexité et l'adéquation du modèle aux exemples observés.

Vision par ordinateur et apprentissage statistique

Une image numérique est pour une machine un tableau de pixels (*picture elements*) où chaque pixel est représenté par un réel (pour les images en niveaux de gris [Fig. 3]) ou par trois réels (pour les images en couleur).
La vision par ordinateur s'intéresse à toutes les techniques de manipulation et de compréhension de ces images : recherche interactive d'images [Fig. 4], annotation automatique d'images, reconnaissance d'objets et de personnes et identification de leur contours, etc. L'informatique a réussi, au cours des dernières décennies, à traiter et automatiser un grand nombre de tâches. Cependant, elle ne parvient pas à comprendre le contenu sémantique à partir de sa représentation numérique faite de pixels. Le grand défi scientifique actuel est de développer des méthodes d'apprentissage qui, par la prise en compte des spécificités des images numériques et les avancées scientifiques effectuées en vision artificielle, puissent parvenir à une compréhension automatique des images et vidéos.

[Fig. 4]
Recherche interactive d'images. L'utilisateur recherche un type d'images dans une base de données. Plusieurs images lui sont présentées. L'utilisateur spécifie les images intéressantes (ronds verts) et les images qui ne lui correspondent pas (croix rouges). L'algorithme apprend de ses informations le type d'images recherchées et propose un nouvel affichage répondant mieux à l'attente de l'utilisateur. Ce dernier spécifie à nouveau ses préférences et ainsi de suite.

Perspectives

La recherche actuelle menée au CERTIS vise à développer des algorithmes robustes d'apprentissage, notamment dans les situations où les données observées sont peu nombreuses et complexes. Nous nous concentrons sur les thématiques suivantes :

• la sélection de modèles et de variables : dans de nombreux domaines d'application des procédures d'apprentissage. En particulier, dans les problèmes liés à la vision par ordinateur, le nombre de caractéristiques de l'objet d'intérêt est bien supérieur au nombre d'objets observés. Dans ces situations, il est indispensable de considérer des modèles qui ne fassent intervenir qu'un petit nombre de caractéristiques bien choisies et développer des méthodes identifiant les variables pertinentes ainsi que le meilleur modèle de prédiction ;

• l'optimisation stochastique : la vision par ordinateur travaille sur des données coûteuses en espace mémoire, dont la manipulation nécessite des processeurs rapides et des algorithmes performants. Le problème apparemment simple de mise en correspondance de deux images comprenant le même type d'objet vu dans des conditions différentes (variation d'éclairage, de point de vue, du décor, déformation de l'objet, etc.) est essentiel au développement de méthodes d'apprentissage de reconnaissance d'objets et de scènes.

CERTIS
École des Ponts ParisTech
19 rue Alfred Nobel
Cité Descartes – Champs-sur-Marne
77455 Marne-la-Vallée cedex 2

contact : Renaud Keriven / 01 64 15 21 72
renaud.keriven@certis.enpc.fr

Pour en savoir plus

http://certis.enpc.fr

László Györfi, Michael Kohler, Adam Krzyzak, Harro Walk . A Distribution-Free Theory of Nonparametric Regression. *Springer*, 2004, 656 p.

Arnak Dalalyan, Anatoly Juditsky et Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, Vol. 9, pp. 1647 - 1678, 2008

Jean-Yves Audibert, Rémi Munos et Csaba Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 2008

Hichem Sahbi, Jean-Yves Audibert, Jaonary Rabarisoa et Renaud Keriven. Context-Dependent Kernel Design for Object Matching and Recognition. In : *Conference on Computer Vision and Pattern Recognition,* Anchorage, Alaska, 2008

Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *Annals of Statistics,* 2008

Le CERTIS

Le CERTIS, créé en 2004, est le laboratoire d'informatique de l'École des Ponts ParisTech. Ses thématiques de recherche sont :

• la reconstruction 3D à partir d'images numériques avec un accent mis sur la qualité de la reconstruction 3D des bâtiments ;

• les procédures d'apprentissage statistiques - en particulier les aspects probabilistes et algorithmiques - dans l'optique de résoudre le défi que représente la compréhension automatique de notre environnement visuel.

- Chercheurs permanents : 6
- Doctorants : 12
- Post-doctorants : 3
- Personnel ITA : 1