Introduction

"Life is just a long random walk. Things are created because the circumstances happen to be right. More often than not, creations [...] are accidental". Ces phrases sont les toutes premières du livre de Devroye, Györfi et Lugosi, "A Probabilistic Theory of Pattern Recognition" [Devroye et al., 1996]. Le contenu de ce mémoire d'Habilitation tout entier en est l'illustration. J'éviterai d'ailleurs de m'étendre sur son titre : faire plus court eut été une gageure, je crois. Il ne manquerait sans doute que quelques ratons laveurs pour en faire un inventaire à la Prévert, mais ma thèse fût, à sa manière, déjà un inventaire, même si le manuscrit n'y fait aucunement référence : une partie de mes travaux de 1998, n'y figure en effet pas. A la liste des travaux que j'ai entrepris depuis l'obtention de ma thèse, s'ajoute donc dans cette Habilitation une partie de ceux qui ont occupé mon temps pendant la fin de ma thèse elle-même. Comment chercher à caractériser les problèmes auxquels je me suis intéressé ?

En fait, il s'agit systématiquement de problèmes de classification. Qu'il s'agisse en effet de classer des exemples, d'interpréter des formules, de réduire des jeux de données, ou bien de trouver des objets dans une image, il y a toujours, sous-jacent, le problème du partitionnement d'un ensemble d'objets en un nombre de "classes" plus ou moins réduit. Cette passion pour la classification au sens très large du terme est survenue alors que je me préparais à faire ma spécialisation de troisième année d'Ingénieur Agronome. Je devais opter pour un DEA de Biostatistiques, mais ce fut un conseil très avisé de Gilles Caraux qui m'orienta vers un DEA d'Informatique, celui du Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM). J'avais opté, avant mon DEA, pour des modules optionnels approfondis en statistiques, aussi un tel choix d'orientation pourrait sembler étrange, mais le DEA d'Informatique fut l'occasion de découvrir de nouveaux aspects de la classification, et en particulier, un champ d'action s'étendant bien au delà des problèmes que je connaissais : analyse d'images, apprentissage automatique, Data Mining, etc. . J'eus par la suite la chance, durant mon service militaire, de travailler dans un centre d'expertise de la DGA (alors sous la direction du médecin Général Jean-Paul Papin) qui se focalisait entr'autre sur des problèmes d'évaluation des individus, un autre domaine d'investigation idéal pour la classification. Depuis, mon intérêt pour cette discipline, au sens le plus large, n'a cessé de croître, avec la reconnaissance d'un de ses volets industriels les plus importants, issu en partie de l'apprentissage automatique : le Data Mining. La métaphore la plus répandue pour présenter ce domaine est celle de la mine d'or, le chercheur étant celui qui construit les bulldozers, les grues permettant de fouiller cette mine de données digitales pour en trouver les "pépites", l'information qu'elle contient. A la différence des mines d'or traditionnelles cependant, le Data Mining fait face à un phénomène tout à fait original : la montagne à fouiller est susceptible de doubler de taille tous les neuf mois, et ceci depuis dix ans maintenant. Quand on regarde l'évolution de cette discipline depuis les premiers projets réels de Data Mining (qui datent environ de cette époque), on s'aperçoit qu'il y a un gouffre s'accroissant entre les quantités de données qui peuvent être générées, et nos compétences pour les analyser. Les besoins en outils de plus en plus puissants pour analyser et comprendre vont donc certainement s'accroître encore d'avantage. Le Data Mining n'est, bien évidemment, pas le seul domaine qui soit mis à rude épreuve par l'augmentation de capacité (de stockage ou de traitement) des ordinateurs. Un autre domaine se trouve être celui du traitement d'images, et surtout de la segmentation d'images, domaine qui a occupé une partie significative de mon temps de recherche. Au travers des premiers travaux que j'ai effectué avec Christophe Fiorio, ce qui m'a attiré dans ce domaine, et m'a poussé à en faire un de mes sujets de recherche, c'est aussi, d'une certaine manière, la disponibilité croissante d'une quantité d'information industrielle et facile à saisir, le besoin de pouvoir analyser cette montagne d'information de plus en plus rapidement pour en déterminer les objets constituants, tout en essayant de conserver une qualité maximale au résultat. Sur bien des points, la segmentation d'images partage ses problématiques avec le Data Mining.

Je me suis focalisé autour de deux grands types d'outils pour étudier tous ces problèmes de classification. Le premier, ce fut l'analyse de la complexité de résolution dans le pire des cas, autrement dit, en étant quelque peu rude, la NP-Complétude et ce qui en découle (l'étude de la difficulté d'approximation, et des liens entre NP et des classes de complexité généralisant la classe P, soit par la complexité en temps, soit par l'aspect de randomisation). Lorsque j'ai commencé cette Habilitation, beaucoup de problèmes d'apprentissage pour lesquels on conjecturait des résultats négatifs de complexité n'étaient pas encore résolus (beaucoup ne le sont d'ailleurs toujours pas). La théorie de la complexité structurelle, c'était sans doute un moyen original de s'attaquer à l'apprentissage automatique, *i.e.* un domaine dont la vocation est, après tout, de prouver l'existence d'algorithmes d'apprentissage efficaces, donc des bornes supérieures, et moins d'exhiber des bornes inférieures les plus grandes possibles ! C'était en tout cas un moyen plus facile que d'autres pour entrer dans ce domaine, surtout vues mes performances pitoyables en programmation au début de ma thèse (Christian Capelle s'en souvient sans doute encore). C'était même d'autant plus intéressant que les résultats qui allaient être utilisables, en théorie des graphes par exemple, étaient à peine sortis à l'époque, et l'on ne disposait pas encore des grands résultats d'inapproximabilité "probablement approximativement optimaux" de Håstad *et al.* autour de ces problèmes. Je ne fus d'ailleurs pas déçu de mon approche (on s'en rendra compte dans la première partie) : les résultats se sont rapidement accumulés, tous plus négatifs les uns que les autres. Aux articles ou se trouvent des résultats sur l'approximabilité de la formule minimale consistante, se succédèrent des résultats sur l'approximabilité de l'erreur minimale de formules de taille contrainte, puis des résultats sur la difficulté de remplacement des critères traditionnelles utilisés en classification, etc, etc. . Il ne se passait de mois sans qu'un résultat négatif ne tombe, sans qu'un théorème ne vienne "améliorer" un dernier résultat négatif obtenu. Durant les périodes ou je programmais, ce genre d'activité était d'ailleurs devenu une sorte de passe-temps : dès

que je m'escrimais trop à trouver des bugs dans mes programmes, je m'arrêtais parfois, le temps d'améliorer un de mes résultats négatifs ! Pourrait-il y avoir en effet de vengeance plus douce, devant un code refusant obstinément de marcher pour des raisons défiant la plus ordinaire des logiques, que de lui montrer que de toute façon, même quand il marchera, il existera toujours des cas pathologiques devant lesquels il ne pourra rendre qu'un résultat médiocre, voire à peine meilleur que des solutions triviales consistant à ne pas même l'exécuter ! C'était effectivement amusant, et sans doute la raison pour laquelle ces résultats ont été nombreux (!), mais souvent très localisés dans le temps. En effet, je ne me suis jamais acharné durant de longue périodes à démontrer des résultats négatifs d'approximabilité. Jamais, sauf pour l'un d'entre eux.
Un seul problème m'occupe en effet depuis le début (et encore aujourd'hui lorsque j'ai des bugs dans mes programmes...), à intervalles réguliers, et m'a aidé à trouver des outils pour résoudre une partie des autres : l'approximabilité de DNF, l'un des grands problèmes autour du modèle d'apprentissage PAC de [Valiant, 1985].
Les méthodes que j'ai développé autour de ce problème m'ont servi dans d'autres problèmes qui ne sont pas forcément du ressort direct du modèle PAC, mais elles ont quasiment toutes un point commun, qui est un parallèle à mon sens amusant avec une des techniques les plus puissantes en classification : le Boosting [Freund and Schapire, 1997]. Le Boosting est une méthode permettant de combiner de manière extrêmement puissante des *solutions* d'algorithmes d'apprentissage (au sens d'un modèle plus faible que PAC) pour en tirer une formule de beaucoup plus grande taille, mais qui est beaucoup plus fiable que n'importe laquelle de ses formules composantes. Si je présente ainsi cette technique de combinaison de résultats positifs dans un paragraphe consacré à des résultats négatifs, c'est tout simplement parce que la technique qui m'a servie à obtenir mes résultats négatifs les plus forts procède de la même manière, mais elle combine des *instances* de problèmes difficiles reliés à l'apprentissage, d'une manière telle que l'instance résultante (du *même* problème), de taille raisonnablement plus grande, soit cette fois d'une difficulté à résoudre sans commune mesure avec n'importe laquelle de ses composantes. Ces résultats font l'objet de la première partie de ce mémoire.

Si l'apprentissage automatique fait la part "belle" à ce genre de résultats pour la partie algorithmique, il dispose d'une toute autre boite à outils pour étudier les capacités des algorithmes en généralisation : les phénomènes de concentration de variables aléatoires, dont les plus connus se trouvent dans les inégalités de Chernoff-Hoeffding. Si j'ai utilisé avec une parcimonie relative ce genre d'outils en apprentissage et Data Mining (j'étais sans doute trop occupé par ma quantité de résultats négatifs d'apprenabilité ou d'approximabilité), ce fut par contre une de mes préoccupations que de les utiliser dans un domaine de la classification où je les avait peu rencontrées : la segmentation d'images. Dès le fin de ma thèse, j'avais commencé à poursuivre quelques directions de recherche avec Christophe Fiorio, du LIRMM, sur de nouveaux algorithmes de segmentation possédant à la fois une complexité en temps et/ou espace presque optimale (voire optimale), et surtout permettant d'obtenir un minimum de propriétés (mêmes probabilistes) sur la segmentation finale. Christophe possédait à l'époque une structure de données perfectionnée, qui rendait tout algorithme de segmentation très rapide (*i.e.* beaucoup plus rapide que beaucoup de résultats que j'ai pu rencontrer dans la littérature consacrée), pourvu qu'on sache rapidement répondre au test de fusion de deux régions. Ayant déjà la structure de données, le code, et l'algorithme, il ne me restait plus qu'à trouver la pièce manquante : ce fameux test. Ce fut un terrain presque idéal pour expérimenter l'utilisation de quelques-uns des derniers résultats sur les phénomènes de concentration de variables aléatoires. En 1998, juste avant mon départ pour les Antilles, Michel Habib organisa une école autour des algorithmes randomisés et autres outils probabilistes de la conception et de l'analyse d'algorithmes. L'article de Colin McDiarmid, "Concentration" ([McDiarmid, 1998]) fut un déclic : il présentait clairement les tout derniers résultats en la matière. Avec Christophe, nous commençâmes à tester ces bornes, avec de bons résultats en général. Il ne me manquait plus, pour ainsi dire, qu'un lien avec l'apprentissage. Ce fut une courte discussion avec M. J. Kearns lors de ICML/COLT/ILP/UAI/AAAI'98 à Madison qui m'en donna l'idée. Dans un article qu'il présentait avec Y. Mansour, Il démontrait un résultat particulièrement élégant, permettant de savoir comment élaguer un arbre de décision avec une précision statistique redoutable. Chose amusante : ces résultats étaient en réalité difficilement utilisables en pratique[1] dans le cas des arbres de décision (!), mais il devait être possible de se servir des idées dans le domaine de la segmentation d'images, avec sans doute des raisons d'espérer que les idées s'adaptent mieux. Ce fut une de mes lignes directrices de recherche, exposée en détail dans la troisième partie.

Et puis, il y eut la plus grosse partie de toutes, celle qui s'est doucement élaborée depuis la fin de ma thèse, puis a littéralement explosé à la faveur d'une collaboration sans faille avec un collègue rencontré en Guadeloupe, Marc Sebban, autour d'un travail qu'on pourrait aisément qualifier "de dingues" lorsqu'on le regarde rétrospectivement. Cette partie concerne tous les algorithmes que j'ai étudié, autour de l'induction de concepts, depuis les algorithmes de réduction de données jusqu'aux algorithme d'induction de nouvelles classes de représentation de concepts. J'avais commencé à étudier les algorithmes d'induction durant ma thèse (mon premier article porte précisément sur ce sujet, [Nock and Gascuel, 1995]). Mon objectif était à l'époque de construire des formules faisant partie d'une nouvelles classe de représentation de concepts, qui avait été appelée les "comités de décision". Ce travail s'est ensuite étendu en fin de thèse aux listes de décision, et plus récemment à diverses classes, mais en utilisant un type d'algorithme "normalisé", car relativement indépendant du formalisme utilisé. L'idée était simple : utiliser et adapter les derniers résultats théoriques de Kearns *et al.* autour de l'induction et de l'élagage de certaines classes de représentation de concepts (surtout les arbres de décision). Ma collaboration avec Marc Sebban a vu mes activités autour de ce sujet "Boostées" tant nous avions d'idées autour de deux domaines de l'apprentissage automatique et du Data Mining qui ont reçu tout récemment un regain d'intérêt : la sélection de variables et de prototypes, c'est à dire des algorithmes permettant de simplifier des jeux de données sans perdre en capacité de généralisation (parfois même, en y gagnant). Notre collaboration n'en a été que plus fructueuses, car nous arrivions en plus avec des idées largement complémentaires et différentes, sur des sujets communs. Marc apportait sa connaissance du traitement et de la simplification des structures de voisinage à l'aide de critères informationnels, j'apportais ma connaissance des méthodes d'induction telles le Boosting, avec la

[1]En réalité, ça ne marchait pas du tout...

volonté d'en faire autre chose qu'un (excellent) algorithme d'induction, mais bel et bien une procédure puissante permettant aussi de réduire un jeu de données en en supprimant soit des variables, soit des exemples. Notre collaboration a montré l'intérêt de ces deux grands types d'approches. Cette partie étant sans aucun doute centrale, elle se trouve entre les deux précédentes. Il s'agit également de la thématique dans laquelle s'insère Patrice Lefaucheur, qui vient tout juste de s'inscrire en thèse sous la direction officielle d'Ahmed Ainouche, et sous mon encadrement. Ses travaux d'investigation portent sur des résultats récents autour des algorithmes de Boosting, afin de construire des procédures particulièrement rapides d'induction et de fouille de données sur de grandes bases de données. Ses travaux sont d'ailleurs co-supportés par des partenaires industriels Martiniquais des domaines de l'assurance et du *credit management*.

Enfin, tout le travail exposé dans cette Habilitation a donné lieu a plusieurs dizaines de publications (au moment de sa rédaction) internationales, référées, éditées, et publiées internationalement (par AAAI Press, Elsevier, IEEE Press, Kluwer, MIT Press, Morgan Kaufmann, Springer, etc.). Sachant qu'un article est, par essence, policé, "optimisé", et *self-contained* dans sa présentation (aux écarts éventuels des *sketches* de preuves), j'ai choisi de garder une taille raisonnable au mémoire (une quarantaine de pages), tout en incluant suffisamment de mes publications en annexes (une quinzaine), pour permettre au lecteur de trouver les idées originales, les preuves, les algorithmes, les expérimentations, sans surcharger inutilement le corps de l'exposé. Ceci implique aussi que, si toutes les publications sont référencées (pages 252-256), certaines ne sont que brièvement citées dans le texte, et leurs résultats ne sont parfois pas même exposés. J'ai fait le choix de privilégier un corps de texte uniforme, plutôt que d'en faire, en quarante pages environ, un catalogue d'abstracts. Je me suis juste laissé aller à ajouter dans la troisième partie de ce mémoire quelques images issues de mes derniers algorithmes de segmentation. Quelques-unes de ces magnifiques images ont été saisies par Pascal Jappy, et réutilisées avec sa permission, cela va de soi.