

Université de Paris-Sud

Département de formation doctorale en informatique École doctorale d'Informatique de Paris-Sud

Apprentissage et Circulation d'Information

MÉMOIRE

présenté et soutenu publiquement le 8 décembre 2005

pour l'obtention de l'

Habilitation à Diriger des Recherches de l'Université Paris-Sud
(Spécialité Informatique)

par

Antoine Cornuéjols

Composition du jury

Rapporteurs : François Denis
Stan Matwin
Lorenza Saitta

Examinateurs : Christine Froidevaux
Jean-Gabriel Ganascia
Mirta Gordon
Andrée Tiberghien

Mis en page avec la classe thloria.

Remerciements

Lorsque la vocation de chercheur s'affirme vers quinze ans, comme ce fût le cas pour moi, elle s'accompagne souvent, je crois, du mythe du génie essentiellement solitaire visité par des intuitions fulgurantes et révolutionnaires. L'anecdote à propos d'Henri Poincaré concevant en un éclair de clairvoyance les fonctions Fuchsianes en montant la marche d'un bus, l'inconcevable créativité d'Albert Einstein écrivant cinq papiers fondamentaux en quelques semaines en 1905, le jeu du jeune Richard Feynman allongé sur son lit en 1947 et voyant au plafond toutes les histoires possibles d'une particule entre deux états, y compris celles remontant vers le passé, tous ces épisodes merveilleux confortent cette image d'une activité scientifique de nature héroïque. Au fil des jours et des nuits de travail et de vie, on prend graduellement conscience du caractère social de la science. Aujourd'hui, si je continue à penser qu'une grande partie du vrai travail s'effectue dans le silence et la méditation solitaire, je ne peux plus envisager la science sans l'échange et la collaboration avec ses pairs et ses élèves, sans la stimulation réciproque et la confrontation des idées. C'est là l'un des côtés les plus satisfaisants de notre métier.

Il est finalement peu d'occasions où, s'affranchissant d'une pudeur naturelle, on puisse reconnaître ses dettes. Des dettes qui sont des enrichissements. C'est avec plaisir que je profite de ces pages trop courtes pour remercier tous ceux qui ont contribué à ce que je fasse mieux mon métier de chercheur et d'éducateur.

Je veux d'abord remercier les membres du jury qui ont bien voulu prendre de leur temps pour évaluer mon travail et donner leur avis. Leur jugement m'est très important.

Merci à François Denis, à Stan Matwin et à Lorenza Saitta d'avoir accepté d'être rapporteurs de ce mémoire. Merci aussi aux examinateurs : Christine Froidevaux, Jean-Gabriel Ganascia, Mirta Gordon et Andrée Tiberghien. Tous sont des sources d'inspiration pour moi. Couvrant un large registre de préoccupations, des plus fondamentales et formelles aux plus directement utiles pour les applications, leur démarche exigeante, rigoureuse et créative est un exemple. J'en ai souvent senti la présence à la fois réconfortante, mais aussi parfois intimidante, lorsque j'essayai de décrire clairement mes travaux, ou de rédiger le livre de synthèse co-écrit avec Laurent Miclet. Chaque fois que, par facilité, la tentation d'une expression approximative ou d'une idée mal étayée me vient, et cela arrive, il me suffit de penser à eux, et à tous mes pairs dont la probité assure la solidité de l'édifice scientifique, pour reprendre ma pensée et essayer de la redresser.

À chacun des membres de mon jury, je dois une influence profonde sur mon travail et sur la manière de le faire. Je ne peux en rendre compte ici, dans ces quelques lignes. Je souhaite simplement que les occasions de cette influence se présentent encore nombreuses à l'avenir.

Je voudrais aussi exprimer la satisfaction profonde ressentie à travailler avec Laurent Miclet dans cette entreprise déraisonnable qui était d'écrire un livre de synthèse sur l'apprentissage artificiel. Nous avons du ajuster nos styles, articuler nos contributions, combiner nos forces et nos faiblesses et combattre les différences irréductibles de nos systèmes *LATEX*. Mais nous n'avons jamais eu de problèmes à nous entendre, et, *a posteriori*, l'existence de notre livre semble absolument naturelle. Je ne doute pas que notre collaboration portera d'autres fruits.

Une Habilitation à Diriger des Recherches marque une étape, une occasion de se retourner sur un itinéraire scientifique. Mes premiers pas en Intelligence Artificielle, alors que j'étais encore élève-ingénieur, ont été marqués par les articles et les ouvrages de Jacques Pitrat, qui, dans cette période pré-Internet, devaient être cherchés dans les bonnes bibliothèques. Je voudrais aussi témoigner de l'exemple qu'est pour moi Judea Pearl, un de mes professeurs, lorsque j'étais étudiant en Ph.D à U.C.L.A. Il est sans conteste l'un des plus grands dans le domaine de l'intelligence artificielle, et sa stature morale et humaine resteront à jamais un modèle pour tous ses étudiants et ses collègues.

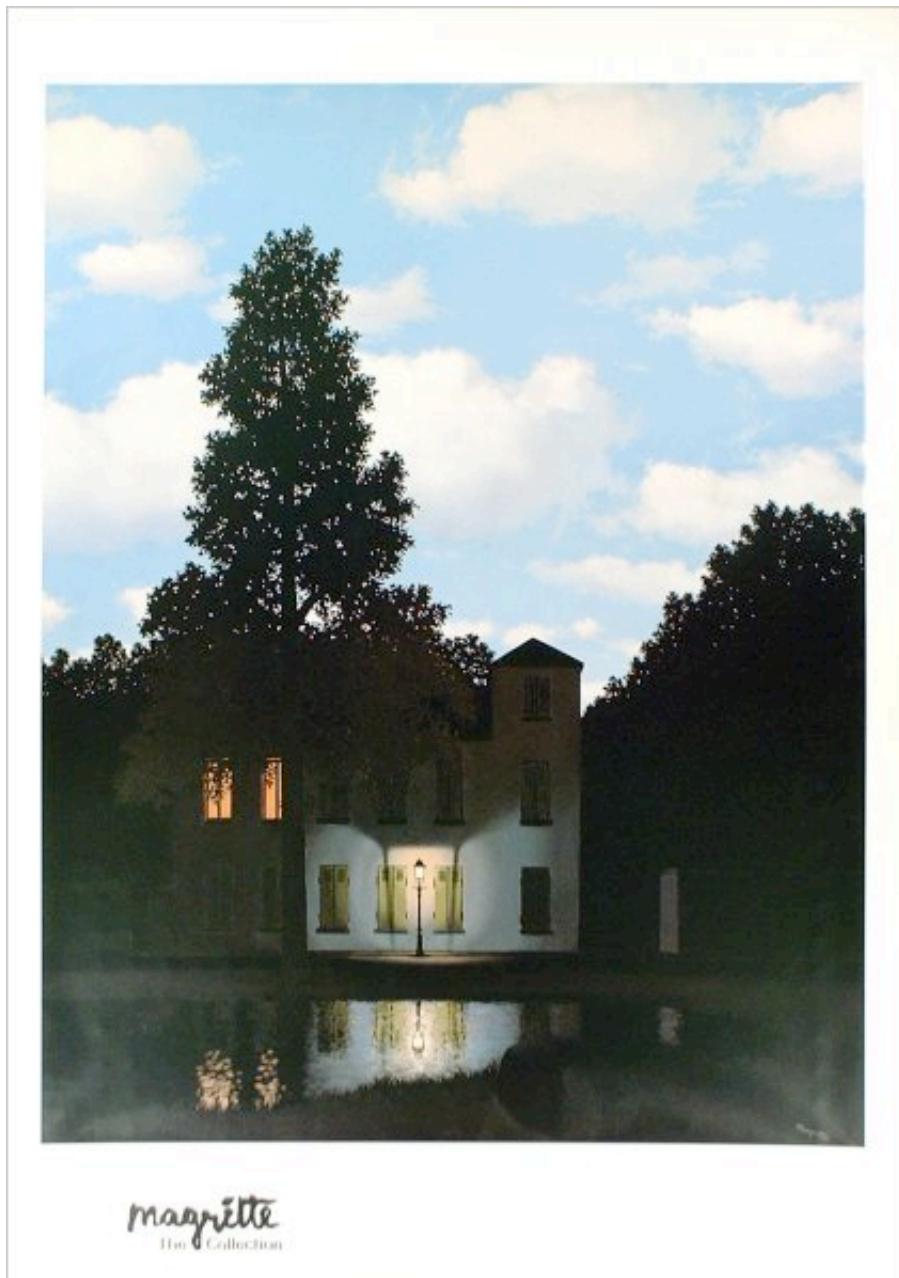
Je suis très reconnaissant à Yves Kodratoff de m'avoir accueilli dans son équipe à mon retour des États-Unis. Entre mille autres choses, je lui sais gré de son ouverture d'esprit, de son dynamisme et de la confiance qu'il sait accorder aux jeunes.

Je voudrais saluer tous ceux qui ont contribué à l'esprit particulier de l'équipe « Inférence et Apprentissage » au L.R.I. Un esprit de club, au meilleur sens du terme.

Merci aussi à mes collègues de l'I.I.E. (Institut d'Informatique d'Entreprise) qui en font un endroit agréable à vivre et à exercer ses activités de chercheur, d'enseignement et d'administration.

Je pense enfin à mes amis et à ma famille. Leur soutien, leur confiance, sont essentiels.

*Merci aux veilleurs, guetteurs, éclaireurs
aux artistes et aux ingénieurs.*



magritte
The Collection

FIGURE 1 – (René Magritte - L'empire des lumières, 1954) *En bas, quiétude immobile. Un réverbère. La clé est peut-être dessous. En haut, le ciel. Espace, transparence et mystère. Nuages et turbulences.*

Table des matières

Chapitre 1	
Introduction	1
1.1 Se pourrait-il	1
1.2 Des interrogations fondatrices	3
1.3 Un état des lieux en apprentissage	5
1.4 Une trajectoire scientifique	8
1.5 Plan du document	11
Chapitre 2	
Apprentissage et gain d'information	13
2.1 Apprentissage, exemples et information	15
2.2 Un phénomène de « transition de phase » trop ignoré ?	17
2.2.1 Test de couverture et induction	17
2.2.2 Phénomènes de transition de phase en informatique	18
2.2.3 Transition de phase en Programmation Logique Inductive	23
2.3 Transition de phase en Inférence grammaticale	29
2.3.1 Inférence grammaticale : quelques notions de base	30
2.3.2 Premiers résultats sur l'inférence grammaticale	33
2.3.3 Notions de base sur l'espace de recherche en inférence grammaticale	34
2.3.4 Transition de phase dans l'espace de recherche	39
2.3.5 Comportement des algorithmes standards d'inférence grammaticale	41
2.3.6 Bilan et perspectives	44
2.4 Conséquences et conjectures pour une stratégie d'enseignement	47
2.5 Un nouveau codage pour redécrire les données : FISICA	49
2.5.1 La reconnaissance de scènes naturelles	50
2.5.2 Les bases de la méthode FISICA	53
2.5.3 Application à la reconnaissance d'images	55
2.5.4 Bilan et perspectives	59

Table des matières

2.5.5 Analyse théorique	60
2.6 Publications, projets et stages liés à ces directions de recherche	62

Chapitre 3

Corrélations, repères et échange d'information	65
---	-----------

3.1 Mesures de corrélation et comparaison de repères	66
3.2 Corrélation et combinaison de détecteurs de régularités	69
3.2.1 Sélection d'attributs et apprentissage non supervisé	69
3.2.2 Approche par méthode d'ensemble	72
3.2.3 Mesure de corrélation entre détecteurs de régularité	73
3.2.4 Utilisation dans une méthode de combinaison de méthodes	77
3.3 L'analogie : construction dynamique d'une corrélation orientée	78
3.3.1 Phénoménologie de l'analogie	80
3.3.2 Analogie et induction	81
3.3.3 Une formalisation de l'analogie	82
3.3.4 Analogie et induction : les deux extrémités d'un spectre	84
3.4 Publications, projets et stages liés à ces directions de recherche	86

Chapitre 4

Trajectoires d'apprentissage et circulation d'information	89
--	-----------

4.1 Vers une étude de la dynamique de l'apprentissage	89
4.1.1 Apprentissage et dynamique des systèmes	90
4.1.2 Symétries et propagation d'information	91
4.1.3 Le cas des systèmes d'apprentissage indépendants de l'ordre des entrées .	92
4.1.4 Théorie de l'apprentissage en-ligne	95
4.2 L'apprentissage actif	96
4.2.1 L'existant en apprentissage actif	96
4.2.2 Le cas d'une stratégie active de sélection d'attributs	97
4.3 L'effet tunnel cognitif : comment construire un nouveau domaine conceptuel . .	100
4.4 Publications, projets et stages liés à ces directions de recherche	104

Chapitre 5

Perspectives	107
---------------------	------------

Annexes	111
----------------	------------

Annexe A (Publications sélectionnées)	111
--	------------

Annexe B (Curriculum Vitae)	113
Annexe C (Liste de publications)	115
Bibliographie	117

Table des matières

1

Introduction

1.1 Se pourrait-il ...

... que l'apprentissage ne soit pas seulement ce dont traite l'ouvrage *Apprentissage Artificiel. Concepts et algorithmes* [CM02] ?

Se pourrait-il que l'apprentissage ne soit pas circonscrit à la science des échantillons constitués d'exemples tirés aléatoirement par un oracle indifférent suivant une distribution inconnue (ce que l'on appelle un tirage i.i.d. pour indépendamment et identiquement distribué) ? Se pourrait-il que l'apprentissage ne se borne pas non plus à être la science du contrôle de la suradaptation (*overfitting*) ? Se pourrait-il enfin que l'apprentissage soit un processus dynamique et historique¹ ? Et qu'il reste à la science d'en rendre compte ?

Se pourrait-il ... ?

Mon penchant naturel est d'aborder les questions fondamentales, les principes essentiels. Mon entraînement d'ingénieur formé dans les Grandes Ecoles me pousse à chercher la belle théorie, celle qui permet d'obtenir tout un ensemble de théorèmes ou au moins de conséquences heuristiques qui prédisent et expliquent la réalité. Rien n'est plus extraordinaire que d'avoir l'impression soudain de comprendre un pan du monde. Plusieurs chemins existent pour cela. Il y a le chemin rigoureux, ascétique et prudent de Bourbaki et de Kolmogorov. Celui dans lequel avant de considérer une distribution continue de probabilité, il est nécessaire d'introduire la théorie des ensembles, les champs de Borel, la théorie de la mesure, l'intégrale de Lebesgue-Stieltjes et le théorème de Radon-Nikodym. Il y a aussi un chemin, tout aussi exigeant, qui cherche à deviner les sommets, les points de vue intéressants et trace des raccourcis, à consolider, pour y arriver. Poincaré et Einstein. Et puis il a la démarche de l'ingénieur qui trouve une solution et construit un outil pour répondre à un problème concret. Le problème est résolu ou il ne l'est pas. Autre type de compréhension et ingéniosité admirable de Gustave Eiffel, et de cette équipe de la Nasa qui en quelques heures a trouvé comment ramener l'équipage d'Apollo 13 sur Terre.

En informatique et en apprentissage artificiel, nous avons de la chance, il n'y a pas de dilemme, pas de choix déchirant à faire. Il faut les trois approches. Chacun peut donc doser comme il le souhaite, selon son tempérament. Prise dans son ensemble cependant, la discipline a modifié son équilibre et privilégié au cours du temps telle ou telle tendance.

Jusque vers la fin des années quatre-vingt, ce fût l'époque des ingénieurs, celle de l'invention de tout un ensemble d'outils, d'algorithmes² et de concepts : méthodes d'apprentissage de

1. Au sens où l'histoire du système joue un rôle dans l'état courant du système par delà les seules contingences actuelles.

2. Écrits en Lisp ! Merveilleuse époque.

règles, réseaux de neurones, algorithmes génétiques, Explanation-Based-Learning, etc. Chacun attaquait un type de problème, et démontrait la performance de son système sur une population de micro-mondes ou de micro-applications bien choisies. Évidemment, cela ne constituait pas une science, même si les liens avec la psychologie cognitive assurait un ancrage sur des concepts plus généraux : mémoires à long terme vs. mémoire à court-terme, connaissances déclaratives vs. connaissances procédurales, etc. Sont alors arrivées deux populations de théoriciens qui ont entrepris de réinterpréter ces travaux et de donner une armature conceptuelle et théorique à cette jungle, luxuriante et sympathique, mais jungle quand même. D'un côté, derrière Leslie Valiant [Val84], s'engouffrèrent des informaticiens spécialistes de la théorie de la complexité et de la cryptographie. En même temps qu'une idée majeure, celle d'approximation en probabilité, le PAC-learning³, ils mirent l'accent sur les problèmes de langages d'expression des hypothèses et des exemples (e.g. langages CNF, DNF) et sur le calcul de la taille, suffisante pour PAC-apprendre, des échantillons d'exemples supposés tirés aléatoirement. Le problème dans nombre de ces recherches était de régler juste comme il le fallait le protocole définissant l'apprentissage pour capturer des situations d'apprentissage « raisonnables ». Il faut bien dire que pendant un temps, ce furent surtout les théorèmes d'impossibilité qui s'empilèrent, sans que cela émeuve beaucoup les praticiens du domaine. De l'autre côté, derrière Vladimir Vapnik [Vap95], arrivèrent les mathématiciens spécialistes de la convergence des processus empiriques. Privilégiant l'analyse de l'apprentissage comme un processus par lequel on cherche, à partir d'estimations empiriques, à trouver des hypothèses ayant une espérance de performance optimale, ils focalisèrent l'attention sur les conditions de convergence des estimations empiriques sur leur espérance.

Si il est incontestable que ce mouvement vers une analyse théorique a amené à la fois une meilleure compréhension des propriétés des méthodes existantes, spécialement des réseaux de neurones, et de nouvelles idées, comme le boosting ou les méthodes à noyau, dont les célèbres SVM (Séparateurs à Vastes marges), il a aussi tendu à réduire et à figer le champ d'attention des chercheurs. En particulier, tous les outils nécessaires aux démonstrations impliquent l'hypothèse de tirage aléatoire des exemples suivant une distribution constante, et une analyse dans le pire cas (convergence uniforme), tout en ignorant toute structure de la connaissance ou même de l'espace des hypothèses. De ce fait, tout un champ de l'apprentissage concerné par l'apprentissage en présence de théorie pré-existante, ou l'apprentissage à partir d'échantillon choisi ou l'apprentissage sur des distributions changeantes, et bien d'autres problématiques encore, s'est trouvé comme oblitieré, tandis que se dissolvaient les liens avec les sciences de la cognition naturelle. La discipline est encore essentiellement dans cette période.

J'avoue mon étonnement et ma gêne face au grand nombre d'écoles d'être en apprentissage portant sur la théorie de l'échantillonnage et sur la concentration de la mesure, alors qu'il n'en existe pas une seule sur l'apprentissage actif, l'apprentissage de connaissances complexes, l'apprentissage à long terme, et plus généralement sur les différents types d'apprentissage qui semblent être en jeu dans les agents naturels ou les sociétés. Se peut-il vraiment que l'apprentissage soit épousé par le point de vue selon lequel il s'agit de trouver la meilleure manière de conditionner un problème d'estimation et d'optimisation ?

Une réaction face à cette domination paradigmatische est celle consistant à mettre en avant une autre légitimité, celle des applications en « vraie grandeur » et qui finit par opposer la fouille de données à l'apprentissage. On voit ainsi fleurir les « défis »⁴. Reste à voir si ils parviendront à stimuler des innovations réelles en apprentissage, ou s'ils ne resteront que des bancs d'essai et

3. Probablement Approximativement Correct.

4. *Challenges* pour ceux qui préfèrent l'académisme anglophone. Voir par exemple le très intéressant site <http://www.pascal-network.org/Challenges/>

des terrains de course, utiles mais non essentiels.

À la fois parce qu'il me semblait que des questions fondamentales étaient laissées de côté et parce que, je m'en rends compte maintenant, je suis davantage intéressé par l'étude des processus, des changements, des circulations, que par les analyses privilégiant les aspects statiques, j'ai suivi une autre voie.

Tout en gardant le contact avec les développements théoriques ainsi qu'avec les sources de problèmes réels⁵ : la bioinformatique (voir section 3.2), l'analyse de bases de données médicales, la vision (voir section 2.5), l'optimisation en continu de ressources pour la téléphonie mobile (contrat en préparation avec France Telecom R&D), etc., j'ai consacré une partie de mes efforts à des questions différentes, peu usitées, mais qui me paraissent fondamentales pour comprendre l'apprentissage.

1.2 Des interrogations fondatrices

Globalement, si tout le monde est d'accord pour dire que les notions d'information, de gain d'information, de transformation, de transmission, sont étroitement liées à ce qui définit le processus d'apprentissage, elles n'apparaissent quasiment jamais explicitement dans la littérature scientifique sur le sujet. Peut-être est-ce justement le signe d'une sagesse, d'une maturité, comme lorsque la notion de phlogistique a été abandonnée à la suite de Lavoisier. Ne parler que d'exemples, d'espace d'hypothèses et de mesure d'erreur permet effectivement de construire une science de l'apprentissage, au moins jusqu'à un certain point. Mais ne peut-on pas cependant à tout le moins faire semblant de prendre au sérieux la notion d'information et construire un programme de recherche qui prenne cette notion comme pivot ?

Un ensemble de questions peuvent se poser dans cette optique. En voici quelques unes.

1. Une situation d'apprentissage élémentaire est celle où le système modifie sa réponse lorsqu'il est confronté à la même situation à deux instants différents, avant et après avoir observé une séquence de données. Si l'on peut paramétriser cette séquence, on peut même obtenir un effet d'hystérésis (voir figure 1.1 pour un exemple).

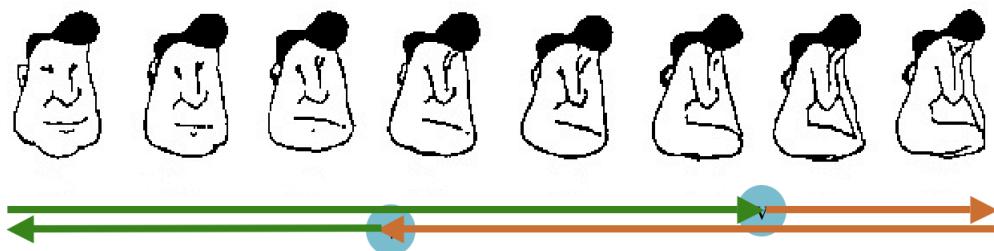


FIGURE 1.1 – *Cycle d'hystérésis mettant en jeu l'interprétation de formes. Il s'agit de parcourir les formes de gauche à droite et retour. Les deux points en bas indiquent les points extrêmes du cycle, pour moi au moins.*

Peut-on alors quantifier l'effet de mémoire, l'inertie de notre système cognitif? Peut-on même, comme en physique, associer à l'aire du cycle (qu'il faudrait définir) une quantité d'énergie⁶ associée à perte d'information ?

5. C'est-à-dire qui importent aussi à des personnes qui ne font pas partie de la communauté sur l'apprentissage

6. En physique, le parcours d'un cycle d'hystérésis s'accompagne d'une dépense d'énergie égale à l'aire du cycle lorsque les bons axes sont choisis.

2. Que signifie *oublier*? Ne plus avoir accès à une donnée ne signifie pas qu'elle est oubliée. Une trace, un ensemble de conséquences calculées, peuvent rester dans le système et influencer son comportement. Comment pourrait-on effacer cette trace seulement en fournissant d'autres données ? (une sorte de lavage de cerveau)
 3. Concernant l'apprentissage incrémental ou en-ligne⁷ :
 - Comment et pourquoi certaines séquences de données sont plus favorables que d'autres à un apprentissage ? La théorie actuelle, de nature statistique, de l'apprentissage ne donne un rôle qu'à l'ensemble des données fournies (vues comme un ensemble de contraintes), mais l'ordre ou la vitesse de présentation peuvent également influer. Comment en rendre compte ? Existe-t-il un lien, et si oui, lequel, avec le « priming effect » observé en psychologie⁸ ?
 - Comment peut-on caractériser la trajectoire d'apprentissage qui résulte de la présentation d'une séquence d'apprentissage ?
 - Qu'est-ce qui permet de rendre un apprentissage incrémental indépendant de l'ordre des entrées ? Cela est-il lié à une circulation de l'information entre les états ? (e.g. Il est évident que si le système se contente de mémoriser les entrées, l'ordre de la séquence ne change pas le résultat final. Mais existe-t-il d'autres types d'apprentissage ayant cette même propriété?).
 - Des données peuvent nuire à l'apprentissage d'une hypothèse ou d'un modèle du monde. La théorie actuelle ne considère cette possibilité que sous l'angle du bruit de description des données. Mais ne peut-on pas en rendre compte par une notion de corrélation ou d'anti-corrélation entre données, ou entre une donnée et un état du système ? Les définitions de corrélation existantes, qui se fondent sur les notions de redondance ou d'économie de description apportée par une donnée ne peuvent mesurer des corrélations négatives. Peut-on trouver d'autres notions de corrélation qui permettent de traduire la nuisance potentielle d'un exemple ?
 4. Existe-t-il des principes d'économie cognitive en jeu dans l'apprentissage, comme il existe des principes de moindre action en physique ? Des principes qui permettraient de rendre compte de certains types d'inférence et transferts d'information, comme en analogie, ou de caractériser les trajectoires d'apprentissage ?
 5. Quelles sont les conditions de circulation ou de construction de l'information ? Plus précisément :
 - Quel est le rôle respectif des langages d'expression des exemples et des hypothèses ?
 - Comment circule l'information dans l'analogie ?
 - Comment circule l'information entre domaines conceptuels ?
 - Comment circule l'information entre méthodes de détection de régularité, comme en co-apprentissage et dans des méthodes apparentées ?
 - Comment circule l'information entre états d'un système lors d'un apprentissage en ligne ?
-
7. J'établirai la distinction suivante :
- *Dérive de concept* : on suppose que le concept à apprendre, et la distribution des exemples, varie avec le temps.
 - *Apprentissage incrémental* : on suppose que le concept à apprendre est fixe, mais que le jeu de donnée est fourni en séquence. On impose au système d'apprendre (fournir des hypothèses ou être prêt à prendre des décisions) en continu, sans avoir reçu l'ensemble des données.
 - *Apprentissage en-ligne* : apprentissage à partir de données fournie en séquence avec possibilité de dérive de concept. On impose au système d'apprendre sans avoir reçu l'ensemble des données.
8. Lorsque la présentation de données ou de problèmes prépare et facilite la réponse à des données ultérieures (voir par exemple [Bad90]).

Toutes ces questions sont très liées à l'étude de l'apprentissage à partir de données qui ne sont pas i.i.d. (indépendantes et identiquement distribuées). Bien que la communauté en apprentissage soit consciente que l'hypothèse i.i.d. n'est généralement pas réaliste, peu de travaux ont encore abordé ce problème de front. Quelques techniques heuristiques visant à détecter les à-coup dans la dérive de concept ont été publiées, et des travaux sur la dépendance faible entre éléments d'une séquence sont présents dans la littérature orientée vers la théorie (voir [Com01] pour un panorama). Les études sur l'apprentissage actif et sur l'analyse de séquences temporelles fournissent également autant de passerelles permettant d'aborder ce problème.

Cependant, c'est une véritable science de la dynamique de l'apprentissage que j'appelle de mes vœux. Pour le moment, tout ou presque reste à faire. Je pense que la théorie du contrôle et la théorie des systèmes dynamiques pourraient utilement être convoquées. Mes propres travaux sont certainement inspirés par ces approches. Ils sont très parcellaires, jetant des lueurs sur certains aspects, mais bien loin encore d'une synthèse générale et de cette belle théorie que j'évoquais plus haut.

Avant de décrire plus précisément mes travaux, il est utile de rappeler brièvement les lignes directrices de la théorie actuelle de l'apprentissage.

1.3 Un état des lieux en apprentissage

Dans le paradigme actuel, l'apprentissage est considéré comme étant la science de l'induction. Le but d'un système apprenant est de trouver des régularités à partir d'une collection d'observations particulières de telle manière à pouvoir faire des prédictions sur des événements à venir. Dans le cas de l'apprentissage supervisé, qui représente une majorité des études, les observations sont des couples (*exemple, étiquette*), et le but est de pouvoir calculer l'étiquette d'exemples non encore vus mais supposés obéir aux mêmes lois de dépendance que celles dont on suppose l'existence entre les exemples et les étiquettes observées. Il est alors naturel d'adopter comme mesure de performance le risque encouru à choisir une régularité (que l'on appellera désormais hypothèse ou fonction hypothèse) parmi les régularités envisageables. Ce risque, appelé *risque réel*, correspond à l'espérance du coût associé au choix d'une hypothèse dans le monde à venir. Par exemple, sur la base d'un échantillon d'observations sur des malades, un médecin peut élaborer des règles afin d'établir un diagnostic sur ses futurs patients. La valeur de ces règles pourra être mesurée par le coût pour la société des diagnostics faits à partir de ces règles. Il s'agit là d'un critère statistique qui dépend de la distribution des maladies et des coûts associés à chaque décision (par exemple, il est moins grave de diagnostiquer une appendicite alors qu'il s'agit d'une indigestion que de passer à côté d'une vraie appendicite. Par ailleurs, ne pas savoir diagnostiquer le paludisme dans une région où l'anophèle femelle est absente n'a pas d'importance). En notant $l(h(\mathbf{x}), u)$ la fonction de coût associée au diagnostic $h(\mathbf{x})$ pour une entrée \mathbf{x} , alors que l'étiquette vraie est u , le *risque réel* s'écrit :

$$R(h) = \mathbb{E}_{D_{\mathcal{X} \times \mathcal{Y}}}[h] = \int_{\mathcal{X} \times \mathcal{Y}} l(h(\mathbf{x}), u) dx dy \quad (1.1)$$

Il exprime l'espérance du coût de la fonction de décision h lorsque la distribution des événements suit une loi $D_{\mathcal{X} \times \mathcal{Y}}$. Dans un espace d'hypothèses candidates \mathcal{H} , l'hypothèse optimale h^* est celle qui minimise le risque réel :

$$h^* = \operatorname{ArgMin}_{h \in \mathcal{H}} [R(h)] \quad (1.2)$$

La loi $D_{\mathcal{X} \times \mathcal{Y}}$ étant inconnue *a priori*, $R(h)$ ne peut être calculé, et il faut avoir recours à un critère de choix de l'hypothèse h qui s'appuie sur d'autres informations. On appelle *critère inductif* le critère que l'on choisit d'optimiser à la place du risque réel.

L'échantillon d'observations, noté S doit naturellement être pris en compte dans ce critère, en ce qu'il est une manifestation de la distribution inconnue. Une bonne hypothèse h devra être un minimum en accord avec les données. Ainsi, un critère inductif évident est celui qui consiste à choisir une hypothèse qui minimise le *risque empirique* mesuré sur l'échantillon d'apprentissage :

$$R_{\text{Emp}}(h) = \sum_{(\mathbf{x}_i, u_i) \in S} l(h(\mathbf{x}_i), u_i) \quad (1.3)$$

Ce critère s'appelle *minimisation du risque empirique* ou MRE. On notera \hat{h} l'hypothèse de \mathcal{H} minimisant le risque empirique.

Deux questions se posent alors :

1. Sous quelles conditions le choix d'une hypothèse optimisant le risque empirique conduit à une hypothèse proche de h^* selon le risque réel ? En d'autres termes, sous quelles conditions y a-t-il convergence du risque réel associé à l'hypothèse \hat{h} vers le risque réel associé à l'hypothèse optimale h^* ? C'est le problème de la consistance.
2. En supposant que cette convergence puisse se réaliser, quelle est sa vitesse ? C'est-à-dire, comment se caractérise cette convergence en fonction de la taille m de l'échantillon S ?

Ce sont les questions fondamentales étudiées dans le cadre de l'étude statistique de l'apprentissage qui est le cadre largement, pour ne pas dire entièrement, dominant actuellement, depuis les travaux initiés par Vapnik et Chervonenkis d'une part, et par Valiant et surtout par Blumer, Ehrenfeucht, Haussler et Warmuth (« the four germans gang »), d'autre part (voir par exemple [BEHW89, Val84, VC71, Vap82, Vap95]). On trouvera de belles synthèses sur ces études en particulier dans [AB99, AB92, BBL05, DGL96, KV94, Vap95, Vid03].

Si l'on suppose que les données sont tirées aléatoirement suivant une distribution de probabilité $D_{\mathcal{X} \times \mathcal{Y}}$ fixe, alors le risque empirique mesuré sur un échantillon est une variable aléatoire constituée de la moyenne d'une somme de coûts sur des variables aléatoires (les données) indépendamment et identiquement distribuées. Par ailleurs, on a :

$$\begin{aligned} R(\hat{h}) - R(h^*) &= R(\hat{h}) - R_{\text{Emp}}(\hat{h}) + R_{\text{Emp}}(\hat{h}) - R(h^*) \\ &\leq [R(\hat{h}) - R_{\text{Emp}}(\hat{h})] + [R_{\text{Emp}}(h^*) - R(h^*)] \end{aligned} \quad (1.4)$$

car par hypothèse, $R_{\text{Emp}}(\hat{h}) \leq R_{\text{Emp}}(h^*)$ et $R(h^*) \leq R(\hat{h})$.

Afin de borner le risque réel que l'on encourt à choisir une hypothèse \hat{h} en lieu et place de la meilleure hypothèse possible h^* , il suffit donc de borner la différence entre l'erreur empirique et l'erreur réelle pour toutes les hypothèses de \mathcal{H} . Il s'agit là d'un problème de convergence uniforme de moyennes vers leur espérance. Comme, par ailleurs, la distribution $D_{\mathcal{X} \times \mathcal{Y}}$ est inconnue, il faut assurer cette convergence pour toute distribution. C'est donc un cadre *MinMax*, dans lequel on cherche à garantir la convergence pour la pire des hypothèses et contre toute distribution possible.

La convergence uniforme requiert des propriétés de limitation de la diversité de l'espace des hypothèses candidates. Cette notion de diversité est cruciale en théorie de l'apprentissage statistique. Plusieurs mesures ont été proposées, comme la dimension de Vapnik-Chervonenkis, les nombres de couverture, des métriques spécifiques sur les espaces fonctionnels, ou les moyennes de Rademacher. L'application de méthodes et de théorèmes sur la déviation d'une fonction aléatoire par rapport à son espérance conduit essentiellement à des bornes de la forme :

$$R(h) \leq R_{\text{Emp}}(h) + \text{diversité}[\mathcal{H}] + \text{fct}(\delta, m) \quad (1.5)$$

où $\text{diversité}(\mathcal{H})$ dénote une mesure de la diversité de la classe d'hypothèses \mathcal{H} , par exemple une moyenne empirique de Rademacher ; δ est un seuil de confiance qui signifie que l'inégalité est vraie avec une probabilité au moins égale à $1 - \delta$; m est la taille de l'échantillon d'apprentissage, et où $\text{fct}(\delta, m)$ est une fonction décroissante de m (sinon le critère MRE n'est pas consistant). Le taux de convergence est une mesure de l'efficacité du processus inductif, généralement en $\mathcal{O}(1/\sqrt{m})$, mais en $\mathcal{O}(1/m)$ lorsque le processus est parfaitement efficace, ce qui peut se produire sous certaines conditions particulières (par exemple que l'hypothèse cible, inconnue, appartienne à la classe \mathcal{H} des hypothèses candidates. En général, le taux de convergence dépend de la dimension de Vapnik-Chervonenkis de \mathcal{H} et de la distribution $D_{\mathcal{X} \times \mathcal{Y}}$).

L'important est que le lien entre le risque empirique mesuré et le risque réel encouru dépend de la diversité de la classe \mathcal{H} dans son ensemble. Élaborée récemment, une analyse plus fine, dite locale, permet de prendre en compte le fait que seules les hypothèses de faible risque empirique sont généralement considérées dans l'apprentissage (voir par exemple [Bou03, BE02]). Si les bornes obtenues sont ainsi plus serrées, et surtout introduisent une mesure de diversité empirique, évaluable à partir des données d'apprentissage, le message reste cependant le même. Il est essentiel de savoir contrôler cette diversité. Les approches par régularisation, par sélection de modèles, par maximisation de la marge, sont des moyens d'y parvenir. On ne cherche plus à minimiser seulement le risque empirique, mais une expression composite, le risque régularisé, incorporant aussi la diversité de l'espace des hypothèses.

L'étonnant dans cette analyse de l'apprentissage, ce qui semble d'une intrépidité inouïe et qui a pu, dans un premier temps, désarçonner les praticiens de l'apprentissage, est qu'il n'est nulle part ici question d'algorithme d'apprentissage !!

On suppose que l'algorithme, quel qu'il soit, trouve sans difficulté une hypothèse optimale \hat{h} , au sens du risque empirique, dans \mathcal{H} . C'est évidemment traiter comme un détail une source majeure de problèmes et de travaux portant sur l'exploration effective de l'espace des hypothèses.

Grossièrement, il existe deux types de méthodes d'exploration : les *méthodes par gradient* et leurs variations stochastiques (recuit simulé, évolution simulée, ...), d'une part, et les *méthodes guidées par les relations de généralité* dans le cas d'espaces d'hypothèses dits symboliques, d'autre part (cas de l'inférence grammaticale ou de l'induction de programmes logiques par exemple). Les méthodes par gradient, particulièrement, sont sujettes à une convergence vers des optima locaux non optimaux.

L'un des grands apports, à la fois conceptuel, théorique et pratique, de ces dernières années consiste à avoir fait le lien entre le problème de l'induction par minimisation du risque empirique régularisé et les techniques d'optimisation de problèmes complexes. Il est en effet possible, en relâchant de manière appropriée les contraintes associées au risque empirique, par modification de la fonction de coût l , de parvenir à un problème d'optimisation convexe. On y gagne à la fois l'existence d'un optimum unique et des vitesses de convergence optimisées. Les méthodes à base de fonctions noyau (voir [STC04]), dont les SVM, ainsi que le boosting, peuvent être considérées comme des instances de cette nouvelle approche de l'induction (voir par exemple [Bou05, BBL05, BLV03]).

Selon le point de vue de l'analyse statistique, l'apprentissage est donc un problème d'optimisation d'un critère inductif sous la contrainte de l'échantillon d'apprentissage à l'intérieur d'un espace d'hypothèses, éventuellement paramétré pour permettre le contrôle de sa diversité.

Toute l'étude repose sur l'hypothèse que la distribution des données est fixe aussi bien en apprentissage qu'en test et que l'échantillon d'apprentissage est issu d'un tirage aléatoire selon cette distribution.

L'analyse de la consistance du principe inductif permet parfois de calculer des vitesses de convergence vers le risque réel en fonction de la taille de l'échantillon d'apprentissage. C'est là toute l'information qu'elle peut nous apporter sur la dynamique de l'apprentissage.

Or tout indique que celle-ci peut être beaucoup plus intéressante. Peut-on se résigner à ignorer cette richesse ? Et moi, le pouvais-je ?

1.4 Une trajectoire scientifique

Ma thèse cherchait à mettre en évidence les possibles effets de l'ordre et de la vitesse de présentation des données dans des apprentissages incrémentaux. J'avais pour ce faire utilisé une analogie avec des systèmes dynamiques auto-organisés. Ces systèmes, des réseaux sémantiques en construction sous l'effet des données disponibles, offraient un joli champ d'expérience, mais ne permettaient pas de dégager des lois fondamentales. Il fallait trouver un modèle beaucoup plus simple. Or, au même moment je lisais d'excellents livres d'introduction à la mécanique (y compris quantique) et à la dynamique des systèmes. Il était clair qu'un concept essentiel était celui d'*action associée à l'évolution d'un système*. Cette action qui dans le cas de la mécanique du point matériel se résume à la circulation de la quantité de mouvement le long de la trajectoire suivie, permet de caractériser la trajectoire dans son ensemble⁹ et non seulement par l'équation d'évolution locale (i.e. équation des forces de Newton). Mais de manière plus importante encore, elle permet, grâce à un merveilleux théorème du à Emmy Noether (1882-1935), de relier des symétries de l'action à des quantités physiques. Dans le cas d'invariances de l'action par des transformations globales, on met en évidence des constantes du mouvement (énergie, quantité de mouvement, etc.); dans le cas d'invariance par des transformations locales (symétries de jauge locales), ces symétries impliquent des échanges de grandeurs (scalaires ou plus complexes : tenseurs, etc.) pour compenser ces transformations, échanges qui correspondent aux particules observées ou encore à observer en physique.

Il me fût immédiatement évident qu'avant d'étudier la complexité des dynamiques d'apprentissage possibles, il fallait *d'abord s'attacher à identifier les conditions dans lesquelles ces dynamiques présentent des symétries* par modification des séquences d'apprentissage, et en particulier par permutation de l'ordre des entrées. L'idée étant que ces permutations devaient être compensées par des échanges d'information entre états. J'ai alors beaucoup joué avec des modèles minimaux d'apprentissage, ainsi qu'avec le perceptron, pour essayer de caractériser les conditions d'indépendance sur l'ordre des entrées. Pour voir aussi ce qui facilitait le plus les apprentissages. Mais je n'arrivais qu'à des recettes peu sûres et indémontrables. C'est un article de Haussler, Kearns et Shapire, dans COLT-91¹⁰, qui curieusement est entré en résonance avec mes préoccupations et qui m'a mis sur une voie intéressante. Toute leur étude était fondée sur le « volume » de l'espace des versions. En caractérisant l'état d'un système par l'espace de version courant, je pouvais montrer que tout choix du système, en cours d'apprentissage (tout élagage de son espace de recherche), impliquait une dépendance potentielle sur l'ordre des entrées. En plus de cela, la démonstration passait par l'établissement d'une équivalence entre l'oubli (d'une par-

9. La trajectoire effectivement suivie étant celle qui minimise, ou plus exactement rend stationnaire, l'action qui lui est associée.

10. *Bounds on the sample complexity of bayesian learning using information theory and the VC dimension* (voir [HKS94] pour une version plus développée de ce papier).

tie de l'espace de recherche) et la prise en compte d'exemples supplémentaires. Cette symétrie, même si elle n'était pas surprenante dans ce contexte, était très satisfaisante.

J'ai écrit quelques papiers dont [Cor93a] et organisé une petite conférence sur l'apprentissage incrémental à Stanford dans le cadre des Spring Symposium de AAAI [Cor93b]. Le thème commençait à éveiller l'attention. J'obtins d'excellents contacts et d'excellents conférenciers et les participants manifestèrent un vrai désir de chercher à se comprendre et à travailler ensemble, mais le résultat évoquait davantage une tour de Babel qu'une parade organisée. Entre les théoriciens du PAC-apprentissage et ceux du paradigme de l'indentification à la limite, ceux de la logique non-monotone, moi-même avec mon approche bizarre, et tous les praticiens en particulier de l'apprentissage non supervisé, le cocktail était étonnant mais peu miscible. Le domaine n'était pas prêt pour une démarche unificatrice. Plus tard, j'ai retrouvé cette impression dans le cadre du projet « Learning in Humans and Machines » financé par l'European Science Foundation entre 1994 et 1997 et qui comprenait une importante proportion de spécialiste des sciences cognitives. Notre sous-groupe, focalisé sur les effets de séquence dans l'apprentissage, a eu du mal à prendre, et ce n'est que maintenant que va sortir un livre présentant notre contribution [LR06].

Je n'ai cependant jamais abandonné cette voie de recherche. Elle est difficile, mais fondamentale pour comprendre l'apprentissage. Les nouveaux (et moins nouveaux) travaux sur les marches aléatoires et les mouvements browniens qui prospèrent en particulier en finance, ainsi que les études sur les dépendances faibles dans les séquences, sont sans doute des voies à explorer.

Parallèlement je commençais à étudier la programmation génétique. Outre la possibilité d'utiliser cette technique pour découvrir l'équation d'une fonction de décision en apprentissage supervisé [BTC96], ce qui m'intéressait plus fondamentalement était la *caractérisation de l'information qui est transmise d'une génération à la suivante*. Comment régler au mieux la taille de la population, les opérateurs de reproduction et autres paramètres pour optimiser la capacité de ce canal de transmission ? En fait, il y avait là aussi un no-free-lunch theorem à l'œuvre. Aucune technique n'est uniformément supérieure aux autres. Je commençais à rédiger un papier lorsque je découvris que Thomas English venait de publier sur ce sujet [Eng96]. Une autre idée intéressante était de mesurer l'information transmise quand la population restait stable ou au contraire changeait sans cesse de territoire. Dans les deux cas aucune information ne passe, alors qu'un maximum est atteint pour des comportements intermédiaires. Il y avait là des choses qui rappelaient la complexité de Kolmogorov, or justement ...

Je m'intéressais également au *raisonnement par analogie* en ce qu'il pouvait fournir un autre éclairage sur la dynamique de l'échange d'information. En effet, il s'agit d'un cas très intéressant dans lequel la question (appelée cible en raisonnement par analogie) oriente la recherche de régularité dans la source. De plus, ce type d'inférence n'est pas symétrique, et on peut même s'amuser à construire des chaînes d'analogies infinies, ou périodiques. Par ailleurs, il me semblait important d'essayer de trouver un lien avec la théorie de l'induction dans laquelle on a un grand nombre de sources et un grand nombre de cibles potentielles. Un principe de minimisation d'un coût cognitif formalisé grâce à la notion de complexité de Kolmogorov répondait bien aux spécifications [Cor94c, Cor94b, Cor94a, Cor96a, Cor96b, CAB98]. Nous cherchâmes alors, avec mon étudiant, Jacques Ales-Bianchetti, à éliminer au maximum les paramètres à fournir, en systématisant la construction d'une sorte de réseau sémantique toujours en s'appuyant sur la théorie de la complexité algorithmique et des machines de Turing. Ce travail n'a jamais abouti de manière satisfaisante, et le raisonnement par analogie est trop éloigné des préoccupations de la discipline pour qu'il soit judicieux de s'y investir davantage pour le moment.

C'est vers cette période que j'ai eu la grande chance de rencontrer Andrée Tiberghien, didacticienne de la physique, alors à l'ENS de Lyon. Elle montait un projet pluridisciplinaire sur

l'apprentissage de connaissances complexes et sur les mécanismes cognitifs en jeu. Elle venait juste de faire une magnifique expérience avec des élèves de lycée apprenant le concept d'énergie et de chaînes énergétiques à l'occasion de petites expériences de physique. Analyser avec le plus de rigueur mais aussi le plus d'ouverture d'esprit possible les transcriptions des dialogues au sein des dyades d'élèves était quelque chose de complètement nouveau pour moi, ainsi que tout l'appareil conceptuel sous-jacent mis en jeu dans la construction de l'expérience et dans la recherche de son interprétation. Avec Evelyne Cauzinille-Marmèche et Gérard Collet, nous formâmes vite un groupe fonctionnant à merveille, sortant de chacune de nos réunions mensuelles stimulés et plein d'idées nouvelles. Nous découvrîmes un type de raisonnement permettant aux élèves de *fabriquer des interprétations du monde physique dans un domaine conceptuel en construction* pour eux. Ce nouveau type de raisonnement qui emporte en contrebande des matériaux et des inférences d'autres domaines dans le domaine cible, nous l'avons appelé *effet tunnel* par analogie à l'effet tunnel en physique [CMCCT97, Col00, CTC98, CTC99b, CTC99a]. Jamais décrit auparavant, ce joli mécanisme est, je crois, important tant pour les sciences cognitives que pour les études sur la découverte scientifique. Il mériterait que nous le poussions davantage dans le monde. Malheureusement, les travaux interdisciplinaires sont difficiles à faire passer dans chacune des disciplines concernées ...

Mais la recherche n'est pas la seule activité d'un enseignant-chercheur. À force d'enrichir mon cours d'apprentissage, et face au manque d'*un ouvrage de synthèse* de facture moderne, j'avais envie d'en écrire un. Laurent Miclet, que j'avais rencontré en 1997 en arpantant à pied avec lui Prague dans tous les sens, se résignait de son côté à ne pas pouvoir traduire le livre de Tom Mitchell [Mit97] (qu'au risque de me répéter, je ne trouve pas si moderne). En 1999, nous décidions d'unir nos efforts. Il nous fallut trois ans pour sortir un ouvrage de 630 pages [CM02]. Nous avons du beaucoup apprendre, sur l'apprentissage, sur *LAT_EX*, et sur l'harmonisation de nos points de vue et de nos styles, pour y parvenir. Ce fut une grande aventure, qui j'espère aura une suite.

Au printemps 2000, j'eus l'équivalent d'un coup de foudre ou d'une révélation lorsque je pris connaissance pour la première fois des travaux de Lorenza Saitta et de son groupe à Turin, ainsi que de Michèle Sebag, sur le *phénomène de transition de phase* en induction de programmes logiques. Pour des raisons qui étaient différentes des leurs, je perçus immédiatement que ce phénomène, lié aux variations du taux de couverture des hypothèses, était capital pour l'étude des conditions de l'induction, au même titre que la théorie statistique de Vapnik. Alors que celle-ci ne prend pas en compte le processus de recherche de l'espace des hypothèses, l'analyse des variations du taux de couverture permet de prédire les difficultés éventuelles des méthodes par gradient, donc de l'ensemble des méthodes d'induction. Encore une fois, on peut y lire un message sur les conditions de gain d'information en fonction du langage des exemples et de celui des hypothèses. Un travail récent, avec Michèle Sebag et un stagiaire de DEA, Nicolas Pernot, portant sur l'inférence grammaticale montre l'intérêt de ce type d'analyse [PCS05a, PCS05b, CS05]. À approfondir, cette investigation ouvre aussi des perspectives pour des stratégies d'enseignement ou d'apprentissage actif. Nous entreprenons à présent une étude systématique de ce phénomène et de voies possibles pour l'exploiter, avec mon étudiant, Raymond Ros, qui débute sa thèse.

Encore au gré des hasards ou des opportunités préparées et favorisées, je profitais à partir de 2001 d'une collaboration avec Philippe Tarroux et Jean-Sylvain Liénard du Limsi sur le thème de la vision de scènes naturelles. Ces scènes, traduites en entrées vectorielles de très grandes dimensions, présentent une distribution de leurs moments statistiques spécifique à laquelle le système visuel des mammifères est particulièrement adapté, capable de reconnaître des formes avant que la conscience n'en soit avertie. Le codage mis en jeu dans les aires visuelles semble être de type clairsemé, comme dans l'analyse en composantes indépendantes (ICA). Pouvait-on

reproduire ce type d'analyse sur des systèmes artificiels ? Michèle Sebag, Sébastien Jouteau, stagiaire de DEA, et moi avons proposé une méthode complètement originale, forçant une *analyse en composantes indépendantes sur un codage à partir d'items fréquents* (*Frequent Item Sets* ou FIS). La technique FISICA était née. Rendue opérationnelle grâce aux astuces de programmation trouvées par Sébastien, elle donnait des résultats dignes d'intérêt, et je la présentais avec succès dans plusieurs séminaires (et voir [Jou02, JCS⁺03, CSM04]). Essayant de comprendre les clés de son succès, je crus pendant 24h à Noël 2003, avoir montré que nous avions trouvé la recette miraculeuse : un risque empirique nécessairement nul et une dimension de Vapnik-Chervonenkis extraordinairement limitée dans des espaces pourtant de grandes dimensions. Las, le raisonnement était erroné et la d_{VC} pouvait, dans le pire des cas, croître linéairement avec le nombre d'exemples. Il fallait avoir recours à des analyses dépendant de la distribution des exemples pour borner la d_{VC} . Une entreprise fort délicate. Entre temps, nous découvrîmes que FISICA n'était en fait pas réellement supérieure à d'autres méthodes, comme les plus proches voisins ! À ce jour, je reste intrigué par les propriétés théoriques de ce codage qui dépend des données. D'autres, séduits par son originalité, en ont repris le principe [RLJS05].

Parmi les dernières entreprises récentes, la plus significative est celle conduite avec Christine Froidevaux sur l'*analyse du transcriptome*. Un jour, Christine m'a expliqué ce qu'elle faisait, avec son enthousiasme et sa chaleur communicative, et trois mois après, à l'occasion du stage de DEA de Jérémie Mary, nous commençons ce qui, je crois, est autant une complicité qu'une collaboration. Nous nous entendions aussi parfaitement avec Marie Dutreix, fougueuse biologiste de l'institut Curie à Orsay. À dire vrai, il fallait une solide foi, et un brin d'inconscience, pour oser croire qu'il était possible de tirer des informations de données dans lesquelles l'activité de milliers de gènes est mesurée à travers un dispositif expérimental et conceptuel qui introduit énormément de bruit, tandis que le nombre de mesures est de moins de deux dizaines !! Cette recherche a débuté comme une aventure d'ingénieur : tout était bon à essayer pour trouver quelque chose [DCCF04, JMC⁺04, MMC⁺03, MMC⁺04, MBM⁺04] Elle abouti maintenant à de nouvelles idées sur la comparaison et la coopération entre méthodes de détection de régularité qui non seulement améliorent incroyablement l'information que l'on peut tirer de ces données (précision sur le nombre de gènes pertinents et leur identité) par rapport aux approches précédentes, mais ouvrent la voie à de *nouvelles techniques d'ensemble* en apprentissage [CFM05]. Nous allons continuer.

Présentée de cette manière, presque chronologique, ma trajectoire scientifique peut ressembler à un « vol de Lévy », cette marche aléatoire décrite par le mathématicien Paul Lévy (1886-1971), dans laquelle des séries de zig-zag locaux sont entrecoupées occasionnellement de bonds plus importants vers d'autres territoires¹¹. Cependant, un fil directeur unit mes différents travaux de recherche : celui de comprendre comment s'opèrent les transferts d'information dans l'apprentissage. Ma démarche aussi suit un déroulement idéal, pas toujours atteint : d'abord commencer par choisir des questions intrigantes, des phénomènes fondamentaux pour l'apprentissage, puis les traduire sous une forme épurée, minimale, à partir de là, essayer d'en dégager les lois profondes, puis repartir vers des situations plus complexes. J'ajouterais un ingrédient essentiel, qui vient parfois tout bousculer pour le meilleur : travailler avec des personnes avec lesquelles s'écouter, s'entendre, n'est jamais un problème.

11. Techniquement, un vol de Lévy est une marche aléatoire dans laquelle la taille des pas suit une loi de puissance. Contrairement à la marche aléatoire classique, les pas peuvent être de n'importe quelle taille, mais avec une probabilité décroissant avec cette taille suivant une loi de puissance.

1.5 Plan du document

Le plan de ce rapport place mes contributions dans un ordre thématique et logique organisé autour du concept d'information dans l'apprentissage.

La **première partie** traite du problème initial : **comment peut-on gagner de l'information à partir des données**. Deux problèmes y sont abordés. D'abord celui de l'information qui peut être prise dans les exemples, c'est-à-dire la manière dont les exemples peuvent modifier l'espace des hypothèses candidates. Des recherches récentes avaient montré le rôle crucial des langages d'expression des exemples et des hypothèses, nos travaux étendent le champ de ces études et montre l'importance aussi de considérer les opérateurs de recherche utilisés pour explorer l'espace des hypothèses. Un phénomène de transition de phase peut ainsi menacer la possibilité même d'un apprentissage inductif dans certains cas. Le deuxième problème est celui du codage des données avant leur traitement par un algorithme d'apprentissage. Les recherches sur les méthodes à noyau ont mis à la mode le codage dans un espace de redescription intermédiaire de (très) grande dimension. Notre travail présente un codage utilisant des primitives dépendant des données et ayant les propriétés d'une analyse en composantes indépendantes, mais rendue possible ici dans des espaces de très grandes dimensions.

À partir de là s'amorce la recherche sur l'étude de la dynamique de l'apprentissage. Cette dynamique se traduit par des trajectoires dans un espace d'états ou de phase. Afin de les caractériser, il faut au moins savoir **mesurer des corrélations** qui pourront conduire à des produits scalaires et des métriques. C'est l'objet de la **deuxième partie**. Dans un premier temps, à l'occasion de nos travaux sur la sélection d'attributs en bioinformatique, je présente une nouvelle méthode de mesure de corrélation entre méthodes. Puis, dans un deuxième temps, je montre, avec une certaine formalisation du raisonnement par analogie, comment on peut effectuer des changements de repères entre situations.

La **troisième partie** aborde l'étude de l'**apprentissage en-ligne et des trajectoires d'apprentissage**. Je traite d'abord, dans un cas encore limité, des conditions à imposer pour obtenir une invariance par rapport aux permutations de l'ordre des entrées. Un moyen de sortir du cadre des données i.i.d. est d'étudier l'apprentissage actif dans lequel le système a l'initiative du choix des données ou, au moins de l'ordre de présentation. Abordant ce point de vue, je décris rapidement une technique de sélection active d'attributs. Finalement, je quitte le cadre des mondes de connaissance très pauvres étudiés jusque là pour montrer une étude du transfert d'information à l'œuvre dans l'apprentissage de nouveaux domaines conceptuels et de connaissances complexes.

La conclusion dresse un bilan de ces recherches et indique des directions de recherche à moyen terme.

2

Apprentissage et gain d'information

Il est évident que l'apprentissage est fondamentalement un processus par lequel un système élabore de l'*information* à partir de son interaction avec l'environnement. Il semble donc que toute étude de l'apprentissage devrait faire une large place à la notion d'information, à sa construction, son gain ou sa perte, sa transformation, sa transmission d'un système à l'autre, ou plus simplement d'un état du système au suivant. Pourtant, le mot information est quasiment entièrement absent de la littérature scientifique sur l'apprentissage. De même, il est capital d'étudier les caractéristiques du *canal de transmission* associé à l'interaction de l'apprenant avec son environnement. Ce sont ces deux points qui sont au centre de ce chapitre.

Pour aller vite, la notion d'information est liée à l'ensemble des hypothèses, et plus généralement au système des connaissances, entretenue par l'apprenant. L'information s'accroît lorsque la mesure de l'espace des hypothèses candidates décroît, c'est-à-dire quand diminue l'incertitude sur les hypothèses à considérer. L'apprentissage est lié à la dynamique de cette information. Dans le cas le plus simple de l'apprentissage supervisé, celle-ci dépend fondamentalement de deux éléments : d'une part, l'information apportée par l'échantillon d'apprentissage via le critère inductif, qui suppose qu'à chaque instant l'hypothèse choisie \hat{h} optimise ce critère (le canal de transmission), et, d'autre part, de l'efficacité de l'algorithme d'exploration de l'espace des hypothèses \mathcal{H} à trouver cette hypothèse optimale (l'élaboration de l'information).

Comme il a été rappelé dans l'introduction (section 1.3), dans le cas du critère inductif de la minimisation du risque empirique, les courbes d'apprentissage, qui décrivent l'évolution du risque réel (estimé sur un ensemble test) en fonction de la taille de l'échantillon d'apprentissage, devraient obéir à des lois en $\mathcal{O}(d/m)$ ou en $\mathcal{O}(\sqrt{d/m})$ d'après la théorie statistique de l'apprentissage. Il est connu que ce n'est souvent pas le cas : soit que la courbe d'apprentissage devienne bonne (beaucoup) plus rapidement que prévu par ces lois, parfois même pour des tailles d'échantillon d'apprentissage inférieures à la dimension de Vapnik-Chervonenkis de l'espace des hypothèses, ce qui devrait être impossible ; soit que la courbe présente des transitions soudaines vers de meilleures performances ; soit qu'au contraire, l'apprentissage soit extrêmement long, et ne parvienne même jamais à identifier une bonne hypothèse dans \mathcal{H} (voir [CM02] et [EdB01, RGG01, Gor02] pour le point de vue de la physique statistique sur l'apprentissage en-ligne.).

Plusieurs raisons ont été invoquées pour expliquer ces comportements imprévisibles.

La première serait le manque de finesse des mesures de la diversité de l'espace des hypothèses \mathcal{H} , telle que la dimension de Vapnik-Chervonenkis. Elles ne rendraient en effet pas compte du fait que l'algorithme d'apprentissage n'explore qu'un petit sous-espace de \mathcal{H} dont la diversité effective pourrait être bien inférieure à celle de \mathcal{H} . De plus, les mesures telles que la dimension de

Vapnik-Chervonenkis, valables contre toute distribution des exemples, seraient trop générales, et il faudrait pouvoir prendre en compte la distribution effective. C’est d’ailleurs ce que les mesures empiriques d’entropie ou les moyennes de Rademacher, réalisent maintenant en partie. Cependant, ces arguments sont insuffisants pour expliquer les transitions soudaines, ou au contraire la difficulté de certains apprentissages.

La deuxième raison est relative aux propriétés des algorithmes d’exploration de \mathcal{H} . Idéalement, selon la théorie statistique de l’apprentissage, ils devraient identifier l’hypothèse \hat{h} optimale selon le critère inductif appliqué à l’échantillon d’apprentissage. Cependant, l’exploration d’un espace d’hypothèses peut être très difficile, particulièrement si le critère inductif induit un paysage présentant de nombreux minima locaux et irrégularités. Inversement, l’algorithme peut aussi parfois retourner une hypothèse meilleure que \hat{h} au sens du risque réel. Ainsi, par exemple, la question des performances et du comportement des algorithmes opérant par gradient stochastique a fait l’objet de très belles études par Léon Bottou (voir par exemple [Bot03]). Cette analyse, qui est loin d’être encore complète, donne des vitesses de convergence, en particulier dans le cas de fonctions à optimiser convexes, et montre la bonne efficacité, en termes d’opérations sur les exemples, des algorithmes de gradient stochastique. Des études analogues portent sur les algorithmes par évolution simulée (voir [FRPT99]).

Soit qu’elles supposent l’optimalité de l’algorithme de recherche, soit qu’elles soient limitées au cas de problèmes d’optimisation convexes, ces études sont insuffisantes pour rendre compte de la richesse des phénomènes observés en apprentissage. Des travaux récents montrent en particulier qu’il faut aussi examiner l’interaction entre les exemples et les hypothèses et l’information que peuvent apporter les unes sur les autres.

On peut ainsi évaluer l’apport potentiel d’information d’un exemple aux changements qu’il doit provoquer, selon le critère inductif, sur la valeur de chaque hypothèse. Si ces changements sont minimes, l’exemple est peu informatif. On savait depuis au moins les travaux de Tom Mitchell sur l’algorithme d’élimination des candidats [Mit82] que les meilleurs exemples sont ceux qui permettent de réduire au maximum le volume de l’espace des versions, c'est-à-dire de l'espace des hypothèses cohérentes avec les données observées¹². Mais jusqu'en 1999 et le travail expérimental réalisé par Botta, Giordana, Saitta et Sebag [BGSS99], personne n'avait songé à examiner systématiquement la distribution de cet apport potentiel d’information qui lie les exemples, les hypothèses et le critère inductif. Il se trouve que, avérées dans le cas de l’induction de programmes logiques (PLI), certaines caractéristiques de cette distribution peuvent considérablement contrarier le processus de recherche d’une bonne hypothèse.

Dans ce chapitre, la notion d’information apportée par un exemple est d’abord brièvement discutée, puis l’étude de la distribution de gain d’information est présentée, avec la mise en évidence d’un phénomène de transition de phase pour la PLI. Nos travaux ont porté sur les conditions d’apparition d’un tel phénomène en inférence grammaticale et sur des parades possibles. La section 2.3 décrit cette investigation. La dernière section 2.5 présente une nouvelle méthode de codage des données qui modifie l’information que peuvent apporter les exemples. Les expériences réalisées pour l’analyse de scènes visuelles, par nous-mêmes et par d’autres chercheurs séduits par cette approche, offrent de nouvelles perspectives pour la fouille de bases de données (voir [JS04, SPBJ04, RLJS05]).

12. On dit d’une hypothèse qu’elle est cohérente avec des données étiquetées positivement ou négativement si, d’une part, elle couvre tous les exemples positifs (complétude) et si, d’autre part, elle ne couvre aucun exemple négatif (correction).

2.1 Apprentissage, exemples et information

Il est entendu que les observations sur le monde, les exemples, sont nécessaires à l'apprentissage. Mais que sait-on de l'information qu'elles peuvent apporter à l'apprenant ? Comment peut-on la caractériser ? Comment peut-on la mesurer ?

En apprentissage, l'information apportée par un exemple est relative à l'impact qu'il peut avoir sur le modèle du monde entretenu par l'apprenant. Un exemple qui ne change pas ce modèle ne véhicule pas d'information pour le système¹³. Plusieurs critères ont été proposés pour évaluer cette information dans le cadre de l'apprentissage.

1. Un critère naturel est lié à la précision avec laquelle les paramètres décrivant la ou les hypothèses candidates peuvent être estimés. Dans le contexte d'un apprentissage de concepts¹⁴, une mesure standard est le volume de l'espace des versions, c'est-à-dire l'espace des concepts cohérents avec les données d'apprentissage : ceux dont la définition est vérifiée par tous les exemples étiquetés positivement et par aucun des exemples étiquetés négativement. Dans le contexte de l'estimation bayésienne, qui prend en compte une distribution de probabilité sur l'espace des hypothèses, une mesure de précision correspond à l'entropie de cette distribution. Dans les deux cas, un exemple sera d'autant plus informatif qu'il conduira à une diminution de l'entropie de la distribution sur les hypothèses, soit du volume de l'espace des versions (qui de fait correspond au cas extrême d'une distribution de probabilité en $\{0, 1\}$). Dans le cas d'un espace des hypothèses doté d'une distribution de probabilité *a priori* $\mathcal{D}_{\mathcal{H}}$, et d'un apprentissage par élimination des hypothèses non cohérentes avec les exemples, le *gain d'information* du à l'exemple i dans une séquence d'apprentissage peut alors s'exprimer comme :

$$\mathcal{I}[(\mathbf{x}_i, u_i)] = -\log \frac{\mathcal{D}_{\mathcal{H}}(V_i)}{\mathcal{D}_{\mathcal{H}}(V_{i-1})} \quad (2.1)$$

où V_i dénote l'espace des hypothèses après la prise en compte des i premiers exemples d'une séquence d'apprentissage $\mathcal{S} = \langle (\mathbf{x}_1, u_1), (\mathbf{x}_2, u_2), \dots \rangle$ ([FSST97]).

Dans le cas d'un apprentissage par mise à jour de la distribution de probabilité sur \mathcal{H} , le gain d'information peut s'écrire :

$$\mathcal{I}[(\mathbf{x}_i, u_i)] = \text{Entropie}(\mathcal{H}_{i-1}) - \text{Entropie}(\mathcal{H}_i) \quad (2.2)$$

2. Un autre critère ne prend pas en considération l'espace des hypothèses, mais mesure, en aval, la précision avec laquelle l'apprenant est capable de répondre ou de faire des prédictions face aux situations à venir. Cela est évidemment très similaire au critère de risque réel utilisé en théorie de l'apprentissage.
3. Un troisième critère, que je n'ai pas vu évoqué directement dans la littérature sur l'apprentissage, pourrait être évalué par la variation de complexité de description (i.e. la complexité de Kolmogorov [Del94, LV97]) du modèle courant entretenu par l'apprenant avant et après prise en compte de l'exemple. Un exemple qui n'apporte rien ne modifiera

13. Il est bien entendu que la notion d'information n'est plus absolue ici comme en théorie de la communication, mais dépend du système qui interroge la nature et de sa capacité à interpréter les signes perçus.

14. Un *concept* est, selon la définition habituelle en apprentissage, considéré comme une fonction indicatrice (à valeur dans $\{0, 1\}$) définie sur un espace d'exemples.

pas cette complexité de description, tandis qu'un exemple surprenant devrait la modifier de manière importante¹⁵.

La mesure du gain d'information associée à un exemple et à un état de l'apprenant est à la base des **stratégies d'apprentissage actif** présentées dans la littérature. Rappelons que l'apprentissage étudié dans le cadre de la théorie statistique est un apprentissage que l'on peut qualifier de « passif ». L'apprenant reçoit des données i.i.d. suivant une distribution de probabilité sur laquelle il n'a aucun pouvoir. L'*apprentissage actif*, au contraire, décrit les protocoles dans lesquels l'apprenant a une initiative, plus ou moins grande suivant les protocoles, sur le choix des données. Dans le protocole standard, l'algorithme d'apprentissage peut choisir les exemples d'apprentissage, soit parmi un ensemble pré-défini, soit « à la volée », sur la séquence des données issues de l'environnement, soit encore en construisant lui-même des exemples [BK92]. On suppose évidemment que l'examen d'un exemple a un coût, et l'on cherche à minimiser ce coût pour l'obtention d'une performance visée.

Le critère relatif à la réduction de l'incertitude sur les paramètres des hypothèses candidates est utilisé par exemple dans la technique de *uncertainty sampling* de Lewis et Gale [LG94] qui choisit les exemples pour lesquels l'apprenant est le plus incertain, ou dans l'approche *query by committee* [SOS92, FSST97, AM98] qui cherche les exemples permettant de réduire au maximum le volume de l'espace des versions (en tenant compte de la distribution *a priori* dans l'espace des hypothèses \mathcal{H}), ou encore dans la méthode de Tong et Koller [Ton01] à base de calculs de marges à l'aide de SVM.

Le critère s'attachant à l'optimisation directe de la performance en généralisation exige une estimation du risque réel qui se fait généralement par échantillonnage d'exemples test [RM01], ou, quand c'est possible, par calcul analytique direct [CGJ96].

Toutes ces approches reposent sur une supposition qui semble tellement bénigne qu'elle n'est jamais discutée : celle selon laquelle il est possible, facile même, de trouver des hypothèses cohérentes avec les données d'apprentissage.

On retrouve le même type de supposition dans l'image désormais classique de l'apprentissage de concept par généralisation ou par spécialisation (voir la partie II et en particulier le chapitre 4 de [CM02]). Dans cette approche, on suppose qu'il existe une relation d'ordre partiel défini sur l'espace des hypothèses \mathcal{H} correspondant à une relation de généralité entre les hypothèses. Si elle existe, cette relation d'ordre permet de doter \mathcal{H} d'une structure de treillis et de guider puissamment l'apprentissage. En effet, étant donnée une hypothèse candidate à un instant t , si celle-ci est trop spécifique, c'est-à-dire ne couvrant pas tous les exemples positifs, on peut suivre les arcs du treillis pour chercher une hypothèse plus générale afin de couvrir tous les exemples positifs sans couvrir d'exemples négatifs ; inversement, si elle est trop générale, c'est-à-dire qu'elle couvre indûment des exemples négatifs, il suffit de chercher à la spécialiser afin de ne plus couvrir ces exemples tout en couvrant tous les exemples positifs. Tom Mitchell [Mit82] a même proposé de conserver en mémoire les deux bornes extrêmes de l'espace des versions selon la relation de généralité : le *S-set* (hypothèses les plus spécifiques) et le *G-set* (hypothèses les plus générales) cohérentes avec les données. Si les données ne sont pas bruitées et que l'espace des hypothèses contient le concept cible, on est garanti que ce concept est dans le sous-treillis borné par le *S-set* et par le *G-set*.

Dans ce cadre de l'apprentissage par exploitation d'une relation de généralité¹⁶ dans \mathcal{H} ,

15. Un exemple décorrélé du modèle entretenu par l'apprenant aura un coût d'incorporation dans ce modèle égal à sa propre complexité de description.

16. Il peut exister plusieurs relations de généralité syntaxique, définie sur l'espace des hypothèses décrit par un langage, correspondant à la relation de généralité sémantique qui est celle de l'inclusion dans l'espace des exemples \mathcal{X} (voir par exemple le chapitre 5 de [CM02]).

il est devenu si habituel de représenter l'espace des exemples par un losange (voir figure 2.1) que personne ne remet en question ce que cette image peut véhiculer comme présuppositions fallacieuses. Nous y reviendrons.

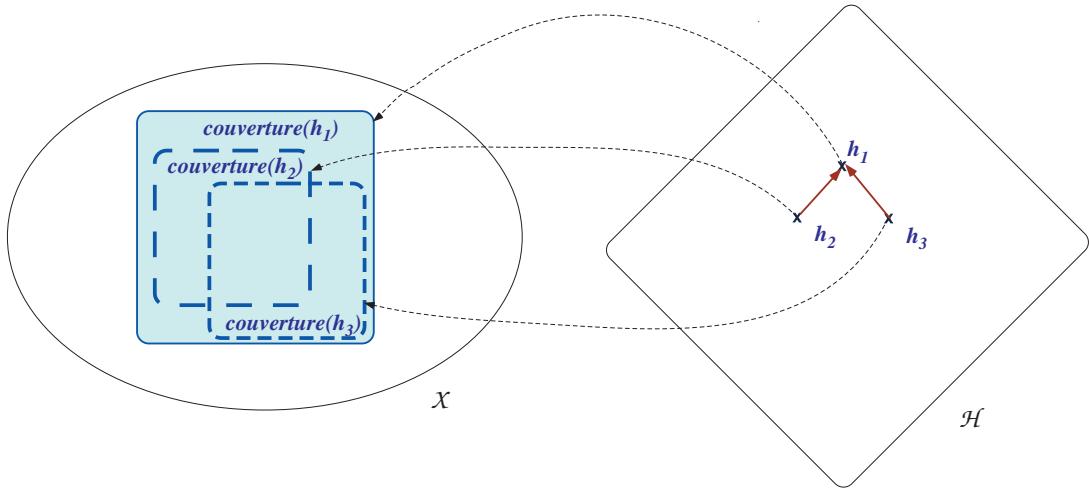


FIGURE 2.1 – La relation d'inclusion dans \mathcal{X} induit la relation de généralisation dans \mathcal{H} . Il s'agit d'une relation d'ordre partiel : ici, les hypothèses h_2 et h_3 sont incomparables entre elles, mais elles sont toutes les deux plus spécifiques que h_1 relativement à la relation d'inclusion dans \mathcal{X} .

La stratégie d'apprentissage consiste alors à suivre les directions de généralisation ou de spécialisation associées aux arcs du treillis dans l'espace des hypothèses \mathcal{H} , suivant que l'hypothèse candidate est trop spécifique ou trop générale. Il est évident que l'espace des hypothèses et le treillis de généralisation sont utiles dans la mesure où ils autorisent une exploration des parties de l'espace des exemples \mathcal{X} qui permette de trouver une bonne approximation, au sens du risque réel, du concept cible. Que les pas élémentaires de généralisation ou de spécialisation dans \mathcal{H} correspondent à de petites variations des parties correspondantes dans \mathcal{X} semble ainsi une propriété souhaitable.

On notera également que, dans tous les cas, on suppose qu'il est possible d'identifier les hypothèses cohérentes avec les données d'apprentissage. Mais ne peut-il pas y avoir des obstacles fondamentaux à la recherche de bonnes hypothèses dans \mathcal{H} ? C'est ce que nous allons voir dans la section suivante, et qui, à mon sens, correspond à une analyse aussi fondamentale que l'analyse statistique de l'apprentissage.

2.2 Un phénomène de « transition de phase » trop ignoré ?

Cette section concerne l'apprentissage supervisé de concepts, c'est-à-dire de fonctions : $\mathcal{X} \rightarrow \{0, 1\}$ à partir d'exemples positifs et négatifs du concept cible. Nous verrons dans quelle mesure il est possible d'en tirer des leçons valables pour d'autres contextes.

2.2.1 Test de couverture et induction

L'évaluation des hypothèses implique un contrôle vis-à-vis des exemples d'apprentissage, c'est-à-dire un test qui, pour chaque exemple, vérifie s'il appartient à la partie de \mathcal{X} définie par

l'hypothèses candidate.

Définition 2.1 (Couverture)

Nous appelons **couverture** d'une hypothèse $h \in \mathcal{H}$, la partie correspondante dans \mathcal{X} (voir figure 2.1).

On dira qu'*une hypothèse couvre un exemple* si cet exemple fait partie de sa couverture.

Afin d'examiner la variation de couverture entre deux hypothèses, par exemple entre une hypothèse et ses voisines dans le treillis de généralisation, on fera appel au concept de taux de couverture :

Définition 2.2 (Taux de couverture)

Nous appelons **taux de couverture** d'une hypothèse $h \in \mathcal{H}$ relativement à une distribution de probabilité $\mathcal{D}_{\mathcal{X}}$, la mesure selon la distribution $\mathcal{D}_{\mathcal{X}}$ de la partie correspondante à h dans \mathcal{X} .

On comprend bien que pour que la recherche dans \mathcal{H} puisse être informée, il est nécessaire que la relation de voisinage dans \mathcal{H} (définie par exemple par les arcs d'un treillis de généralisation, mais aussi par le pas ε utilisé dans l'algorithme de rétro-propagation dans les réseaux de neurones) correspondent à une autre relation de voisinage dans \mathcal{X} , de telle manière qu'à de petites transformations dans \mathcal{H} correspondent des petites transformations dans \mathcal{X} . Ces dernières se définissant naturellement par la variation de la couverture. Idéalement, un petit changement de l'hypothèse candidate devrait entraîner une petite variation de la couverture, c'est-à-dire que seul un petit ensemble d'exemples change de statut, passant de couvert à non couvert ou vice-versa. Si, au contraire, le passage d'une hypothèse à sa voisine entraîne un changement très important de la couverture correspondante dans \mathcal{X} , l'exploration de \mathcal{H} ne pourra plus s'appuyer sur l'information de variation de couverture pour se guider, ce qui mettra en péril toutes les approches à base de technique de gradient, c'est-à-dire toutes les approches d'apprentissage existantes.

Il est donc clair qu'il est capital de s'assurer de cette *propriété de continuité* particulière pour rendre possible l'apprentissage inductif. Pourtant, ce n'est que très récemment, à partir de 1999, que cette propriété a été étudiée, et encore, indirectement et pour des motifs différents.

2.2.2 Phénomènes de transition de phase en informatique

Au cœur de la physique statistique, des mathématiques discrètes ainsi que de l'informatique, résident des problèmes de comptage et d'optimisation qui partagent essentiellement les mêmes propriétés. Ces problèmes, que l'on retrouve en planification, ordonnancement, apprentissage artificiel, conception de systèmes, bioinformatique, etc. sont souvent extraordinairement difficiles à résoudre. Ils appartiennent à la classe des problèmes NP (NP pour *Non deterministic Polynomial time*). Informellement, il s'agit de problèmes pour lesquels trouver une solution semble requérir un temps exponentiel (en certains paramètres définissant la « taille » du problème), tandis que la vérification d'une solution potentielle est facile. La propriété de complétude de la classe des problèmes NP-complet signifie que si un algorithme efficace est trouvé pour résoudre un seul de ces problèmes, alors il s'en déduira immédiatement des algorithmes efficaces pour tous les autres. Une conjecture fondamentale de l'informatique théorique est qu'il n'existe pas de tels algorithmes efficaces.

Cependant, alors que l'on pense que les problèmes NP-complets requièrent un temps exponentiel pour leur résolution dans le pire cas, il est plus difficile à caractériser les cas typiques.

Or ce sont ces cas qui sont les plus importants dans la pratique.

Fu et Anderson furent les premiers à émettre l'hypothèse d'une connexion profonde entre les problèmes NP-complets et les modèles étudiés en physique statistique [FA85]. Depuis, il a été découvert que les problèmes NP-complets aussi peuvent être sujet à un phénomène de transition de phase, comme des systèmes de physique statistique, avec les problèmes les plus difficiles se trouvant à proximité de la région de transition de phase (voir par exemple [EHW96, Hay97] pour un numéro spécial et pour un article d'introduction)¹⁷.

Les problèmes K-SAT : un banc d'essai

Parmi la classe des problèmes NP-complets se trouvent les problèmes K -SAT. Ceux-ci sont définis par un ensemble de N variables booléennes, et un ensemble de M clauses. Chaque clause est une disjonction logique de K variables, où chaque variable peut être niée. Le but est de trouver une manière cohérente d'assigner les valeurs des K variables de telle manière que toutes les clauses puissent être satisfaites. Par exemple, la formule 2-SAT : $(x \vee y) \wedge (\neg x \vee \neg y)$, qui contient deux 2-clauses, peut être satisfaite en posant $x=\text{Vrai}$ et $y=\text{Faux}$.

Dans le pire cas, pour $K \geq 3$, une recherche exponentielle dans l'espace des 2^N valeurs booléennes possibles pour les N variables est nécessaire afin, soit de trouver une solution au problème, soit de prouver qu'il n'en existe pas, auquel cas on dit que le problème est « non satisfiable ».

La complexité du cas typique pour le problème K -SAT s'étudie en considérant un ensemble de problèmes générés aléatoirement. Pour chaque formule, M clauses sont générées en choisissant aléatoirement K variables dans l'ensemble des N variables possibles, chaque variable choisie étant niée avec une probabilité 0.5. Les formules ainsi construites aléatoirement, avec un rapport $\alpha = M/N$ constant quand $M, N \rightarrow \infty$, constituent un ensemble de problèmes test.

La valeur de K joue un rôle important. Pour $K = 1$ ou 2 , les problèmes peuvent être résolus efficacement c'est-à-dire en temps polynomial (et même linéaire) dans la taille des formules. Les problèmes 1-SAT et 2-SAT sont dits appartenir à la classe des problèmes P (*Polynomial time solvable problems*). Le problème 2-SAT présente une transition de phase pour un rapport du nombre de clauses sur les variables égal à un. En-dessous de cette valeur du rapport, presque toutes les formules sont satisfiables (région SAT), tandis qu'au-dessus de cette valeur du rapport, presque toutes sont non satisfiables (région UNSAT). L'explication intuitive est que pour les faibles valeurs du rapport, le relativement faible nombre de clauses correspond à un faible nombre de contraintes et il est facile de trouver une valuation des variables permettant de satisfaire toutes les clauses. Pour les fortes valeurs du rapport α , au contraire, le problème devient sur-contraint et il n'est plus possible, en général, de trouver de solution (problèmes sur-contraints). De plus, l'espace de recherche est limité car l'impossibilité de trouver une solution apparaît rapidement. Les problèmes les plus intéressants se présentent pour des valeurs de α proches de 1.

Pour $K \geq 3$, les problèmes K -SAT sont NP-complets. Les expériences ont montré des transitions de phase pour $K = 3, 4, 5$ et 6 avec des valeurs de seuil de transition différentes. Les résultats mesurés pour la valeur $K=3$ sont illustrés sur la figure 2.2. On parle de transition *facile-difficile-facile* car on distingue trois régimes : (1) pour des faibles valeurs de α ($\alpha < \alpha_C$), il est relativement facile de trouver une solution ; (2) pour des valeurs $\alpha \approx \alpha_C$, il est très difficile de trouver une solution ou de prouver qu'il n'y en a pas ; (3) pour des valeurs $\alpha > \alpha_C$, il devient à nouveau relativement aisé de résoudre le problème c'est-à-dire, dans ce cas, de montrer que la formule est non satisfiable.

17. Il faut cependant noter que des phénomènes de transition de phase ont également été observés pour des problèmes « simples », plus précisément non NP-complets.

Étant donnée le caractère générique des problèmes K -SAT, il est intéressant de les utiliser comme banc d'essai pour la recherche de méthodes heuristiques de résolution de problèmes. Maintenant que la région des problèmes difficiles est précisée, c'est naturellement pour résoudre les problèmes dans cette région que l'on cherche de nouvelles idées et de nouvelles méthodes de résolution. Ces dernières années ont vu une floraison d'études qui ont mis en évidence par exemple la notion de squelette (*backbone* et qui ont conduit à une augmentation spectaculaire des performances des systèmes de satisfaction de contraintes. (Voir par exemple [Sel95, GS05, KS]).

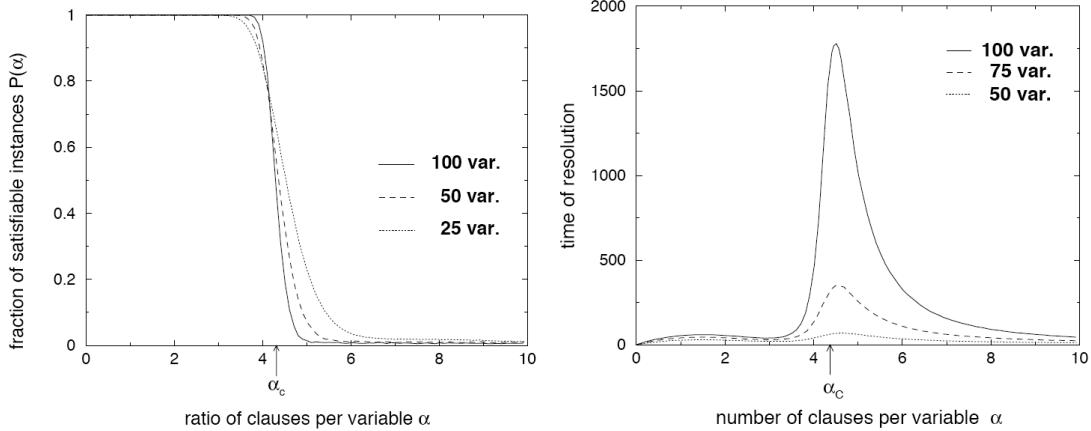


FIGURE 2.2 – Pour le problème 3-SAT, un phénomène de transition de phase se manifeste à la fois dans la fraction des problèmes satisfiables qui passe brutalement de presque 1 à presque 0 pour une valeur de seuil $\alpha_C \approx 4.31$ (à gauche), et dans le temps de résolution qui passe par un maximum pour cette même valeur de seuil (à droite). On note que la transition est d'autant plus marquée que le nombre N de variables est élevé. (Illustrations tirées de [Mon02])

Avant de montrer en quoi les problèmes de satisfaction de contrainte et de transition de phase ont une pertinence pour l'apprentissage artificiel, nous faisons un petit détour par la description de travaux qui visaient à mieux comprendre l'apparition d'une transition de phase en faisant varier, à l'intérieur d'une formule, le rapport de 2-clauses et de 3-clauses.

L'étude des problèmes (2+p)-SAT

Les problèmes (2+p)-SAT sont des problèmes dans lesquels les formules à satisfaire contiennent M clauses dont $(1-p)M$ contiennent deux variables et pM contiennent trois variables, avec $0 \leq p \leq 1$. Ils permettent d'examiner l'interpolation entre les problèmes 2-SAT et 3-SAT. Cela a conduit en particulier à deux idées intéressantes.

1. Verres de spin et « squelettes »

La première idée provient d'une analogie qui peut être faite entre les problèmes (2+p)-SAT et des verres de spin dilués, étudiés en physique statistique, avec N variables de spin S_i ($S_i = 1$ si la variable booléenne correspondante vaut Vrai et $S_i = -1$ si elle vaut Faux). À chaque configuration, c'est-à-dire à chaque valuation des variables booléennes, il peut être associée une énergie E , ou fonction de coût, égale au nombre de clauses violées dans cette configuration. La « méthode des répliques » [MPV87] peut être utilisée pour analyser le comportement de ces problèmes. Un paramètre d'ordre¹⁸ peut ainsi être mis

18. Cette appellation vient de l'analyse de transitions de phase magnétiques par Landau en 1937. Dans son

en évidence, qui caractérise la statistique des valuations optimales, c'est-à-dire des états fondamentaux (*ground states*) qui minimisent le nombre de clauses non satisfaites.

Soit par exemple, une instance d'un problème $(2 + p) - SAT$. En utilisant le nombre de configurations g optimales : N_{GS} , on définit le paramètre $m_i = 1/N_{GS} \sum_{g=1}^{N_{GS}} S_i^g$, c'est-à-dire, la valeur moyenne de variable-spin S_i sur toutes les configurations optimales. On a donc $m_i \in [-1, +1]$ avec $m_i = -1$ si la variable-spin x_i est toujours à Faux dans toutes les configurations optimales, et $m_i = 1$ si la variable-spin x_i est toujours à Vrai dans ces configurations. La distribution $P(m)$ de tous les m_i caractérise la structure microscopique des états fondamentaux. En particulier, les variables-spin associées à des valeurs $m_i = \pm 1$ correspondent à des variables-spin extrêmement contraintes, puisqu'elles prennent la même valeur dans toutes les configurations optimales. On appelle l'ensemble de ces variables-spin, le « squelette » (*backbone*). Ce squelette de variables joue le rôle de paramètre d'ordre. Il est nul en dessous du seuil et saute discontinûment à α_C , le seuil critique, à une valeur de l'ordre de 15%. La transition de 3-SAT peut être interprétée comme étant du 1^{er} ordre¹⁹.

Il est évident que la détermination du squelette, dans une première phase, permet d'accélérer la recherche d'une configuration pour les variables-spin restantes. Cette technique est à la base de nombreux progrès réalisés récemment dans la résolution de problèmes de satisfaction de contraintes (ref [CZ02, DD01, SW01, Zha01]).

Il est notable que la transition pour les problèmes 2-SAT aléatoires, à $\alpha_C = 1$ est du deuxième ordre et qu'aucun squelette n'émerge brutalement. La transition entre une transition de phase de 2^e ordre et une transition de phase de 1^{er} ordre s'effectue pour une valeur de $p \approx 0.4$.

2. La dynamique des algorithmes de résolution

Afin de comprendre en quoi la transition est liée à la complexité de résolution de 3-SAT, il est intéressant d'examiner brièvement la méthode de résolution dite de Davis et Putnam (DP). La procédure de DP correspond à une stratégie de base en optimisation, qui consiste en une recherche par essais et erreurs dont on peut rendre compte par l'exploration d'un arbre de possibilités. Rapidement, un noeud de l'arbre correspond au choix d'une variable et de sa valeur (1). Selon que cette dernière est vraie ou fausse, on emprunte l'une des deux branches possibles (2). Le long d'une branche, on analyse toutes les implications logiques du dernier choix effectué. Si une contradiction (une clause non satisfaites) surgit (3), on modifie le dernier choix (remontée dans l'arbre) et on poursuit sur une nouvelle branche en allant à l'étape 2 ; si toutes les clauses sont satisfaites, une solution est trouvée et on arrête la recherche, sinon, on retourne à l'étape 1.

La complexité computationnelle est fonction directe de la taille de l'arbre de recherche, plus précisément du nombre de noeuds explorés. Les performances peuvent être améliorées en jouant sur l'heuristique de choix des variables (étape 1).

L'algorithme de DP définit un processus dynamique complexe, non markovien et original par rapport aux évolutions classiques des systèmes physiques. Afin d'en étudier les

étude, Landau introduit un paramètre supplémentaire représentant l'ordre magnétique. Techniquement, un paramètre d'ordre est une fonction de l'énergie libre et de la température et change discontinuement de valeur à la transition de phase. (Voir [LL03, GH04] pour plus de détails). En apprentissage, il serait plus exact de parler de paramètre de contrôle.

19. C'est-à-dire impliquant une variation brutale d'une quantité. On parle de transition de 1^{er} ordre car la dérivée première d'un potentiel thermodynamique change brutalement de valeur. On peut aussi noter que seules les transition du 1^{er} ordre peuvent donner lieu à des cycles d'hystéresis.

caractéristiques, Rémi Monasson a proposé le modèle $(2 + p)$ -SAT en remarquant que la procédure récursive de DP transforme le problème 3-SAT de départ en un problème mixte $(2 + p)$ -SAT contenant des clauses de longueurs deux et trois, les clauses contenant une seul variable étant éliminées à l'étape 2 de l'algorithme ci-dessus ([AKKK01, BCM01, MMR01, Mon02, MRK⁺99]). Si l'on symbolise le problème 3-SAT initial comme un point de coordonnées $(p = 1, \alpha)$ dans le diagramme de phase de la figure 2.3, ce point va évoluer sous l'action dynamique de l'algorithme et décrire une trajectoire dans un espace de problèmes de caractéristiques différentes.

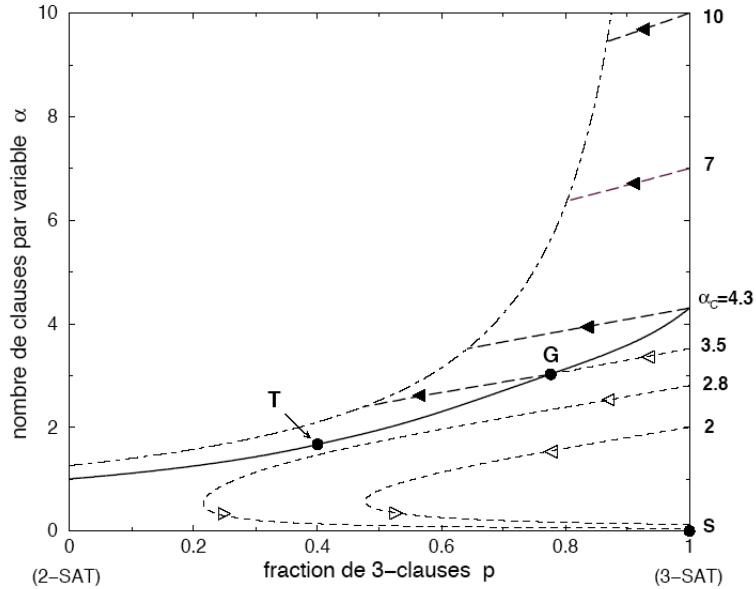


FIGURE 2.3 – Diagramme des phases du modèle $(2 + p)$ -SAT et flot dynamique de l'algorithme de Davis et Putnam. la ligne critique $\alpha_C(p)$ (ligne pleine foncée) sépare la région satisfaisable (en-dessous de la ligne) de la phase sans solution (au-dessus). Ses extrémités ont pour coordonnées 1 ($p = 0$, 2-SAT) et 4.3 ($p = 1$, 3-SAT). T est le point tricritique séparant les transitions continues (à gauche de T) de celles discontinues (à droite de T). les lignes en tirets représentent les trajectoires de branche unique (lignes minces, flèches évidées) et d'arbre touffu (lignes épaisses, flèches foncées). Les flèches indiquent la direction du flot. les points de départ pour les problèmes 3-SAT sont situés sur l'axe vertical $p = 1$ avec les rapports α correspondant. Pour $\alpha < \alpha_L \approx 3.003$, les trajectoires de branche unique restent confinées à la phase satisfaisable et se terminent en S de coordonnées $(1, 0)$, où une solution est trouvée. dans l'intervalle $\alpha_L < \alpha < \alpha_C$, la trajectoire de la branche unique traverse la ligne critique en G , dont la position dépend de la valeur initiale de α . Un arbre touffus croît alors dans la phase sans solution, comme lorsque le rapport initial α est au-dessus du seuil $\alpha_C \approx 4.3$. les trajectoires des branches dominantes des arbres touffus s'arrêtent sur la ligne de contradiction (trait mixte). Dès lors, DP atteint son point de remontée maximale dans l'arbre, c'est-à-dire le premier nœud si $\alpha > \alpha_C$ ou le nœud G si $\alpha_L < \alpha < \alpha_C$. Dans ce dernier cas, une solution peut être trouvée le long d'une nouvelle branche alors que, dans le premier cas, l'absence de solution est prouvée. (Illustration tirée de [BCM01])

Cette remarque me semble pouvoir ouvrir un champ d'études intéressant en apprentissage. Nous le verrons en section 2.4. Mais il faut d'abord voir en quoi les problèmes de satisfaction de contraintes sont similaires à des problèmes d'apprentissage et en quoi l'étude

de l'apprentissage peut bénéficier des concepts développés pour les problèmes K -SAT.

2.2.3 Transition de phase en Programmation Logique Inductive

L'observation clé qui relie l'apprentissage inductif aux problèmes K -SAT est que le test de couverture en apprentissage (voir section 2.2.1) consistant à vérifier qu'un exemple est couvert par l'hypothèse considérée est analogue à la recherche d'une solution pour une formule dans un problème K -SAT.

Plus précisément, en Programmation Logique Inductive (PLI), un exemple est usuellement représenté comme une conjonction d'atomes clos²⁰, par exemple :

$$e(a) : -r_1(a, a_1), r_1(a, a_2), r_1(a_1, a_{17}), r_1(a_2, a_8), \dots, r_2(a_8, a_{16}), r_2(a_{23}, a_{34}), \dots$$

définis à l'aide de constantes prises dans un ensemble donné. Une hypothèse ou concept est représenté par une formule qui est une conjonction de littéraux, par exemple :

$$h(X) : -r_1(X, Y), r_1(Y, Z), r_2(Z, T)$$

contenant des variables prises dans un ensemble donné. Le test de couverture cherche si il existe au moins un modèle de la formule correspondant au concept h dans l'univers défini par l'exemple e . Pour des raisons de décidabilité, la θ -subsumption est généralement utilisée pour le test de couverture, plutôt que l'implication.

Le groupe de recherche de Turin dirigé par Attilio Giordana et Lorenza Saitta²¹ a montré (voir [BGSS03]) que la θ -subsumption peut être traduite en un problème de satisfaction de contraintes. Dès lors, la question naturellement ouverte était de savoir si un phénomène de transition de phase affectait également le test de couverture. Dans cette perspective, le groupe de Turin a entrepris de mettre sur pied un protocole expérimental.

Étude de la variation du taux de couverture

L'idée, comme pour l'étude des problèmes K -SAT, est de générer des tests de couverture aléatoirement suivant une distribution uniforme dans un espace de paramètres « naturels ». Le mieux est de définir des paramètres de contrôle ayant un sens pour le problème, c'est-à-dire exprimant de manière condensée les paramètres influençant le comportement des solutions. C'est, par exemple, ce qu'avait fait Prosser ([Pro96]) en définissant la *densité de contrainte* et la *force de contrainte* comme paramètres pour les problèmes de satisfaction de contraintes booléennes. Son analyse reposait en particulier sur l'équivalence de ces problèmes avec des graphes dont les arcs et les nœuds représentent respectivement les contraintes et les variables. À défaut de paramètres synthétiques évidents, le test de couverture est gouverné par les paramètres :

- n : nombre de *variables* dans l'hypothèse h testée,
- m : nombre de *symboles de prédictats* dans h ,
- L : nombre total de *constantes* dans l'exemple e ,
- N : nombre de *littéraux* construits sur chaque symbole de prédictat dans l'exemple e .

Les expériences réalisées par le groupe de Turin ont consisté en une exploration systématique de l'espace décrit par ces quatre paramètres pour des valeurs pertinentes pour l'apprentissage. Ainsi, pour chaque 4-tuplet (n, N, m, L) (avec $m \geq n - 1$ et pour $n = 4, 6, 10, 14$; $N = 50, 80, 100, 130$; $m \in \{15, \dots, 50\}$; $L \in \{10, \dots, 50\}$), un millier de problèmes de subsumption ont été générés aléatoirement suivant un protocole particulier. Celui-ci assure que chaque hypothèse h contient exactement n variables et m littéraux, tous construits sur des prédictats différents, et telles que les variables sont toutes reliées entre elles par des chaînes de littéraux

20. Voir pour une introduction à la programmation logique inductive le chapitre 5 de [CM02].

21. Maintenant à l'Université du Piémont Oriental à Alessandria.

(pour éviter la possibilité du fractionnement du problème en sous-problèmes indépendants). De même, tous les littéraux construits sur un même symbole de prédicat sont générés par un tirage sans remise de N éléments parmi toutes les paires possibles (a_I, a_j) . Tous les littéraux sont de ce fait nécessairement distincts.

Conformément aux études sur les problèmes K -SAT qui étudient la probabilité de satisfaire une formule et la complexité de recherche d'une solution, en induction, deux variables sont intéressantes à mesurer. D'abord, la probabilité P_{sol} de couverture d'un exemple par une hypothèse, ensuite la complexité de recherche d'une substitution pour la θ -subsomption. Pour des raisons de lisibilité, les résultats sont rapportés pour une variation de m et de l tandis que les grandeurs n et N sont supposées constantes. La figure 2.4 illustre le type de comportement observé (ici dans des expériences rapportées dans [MS04]).

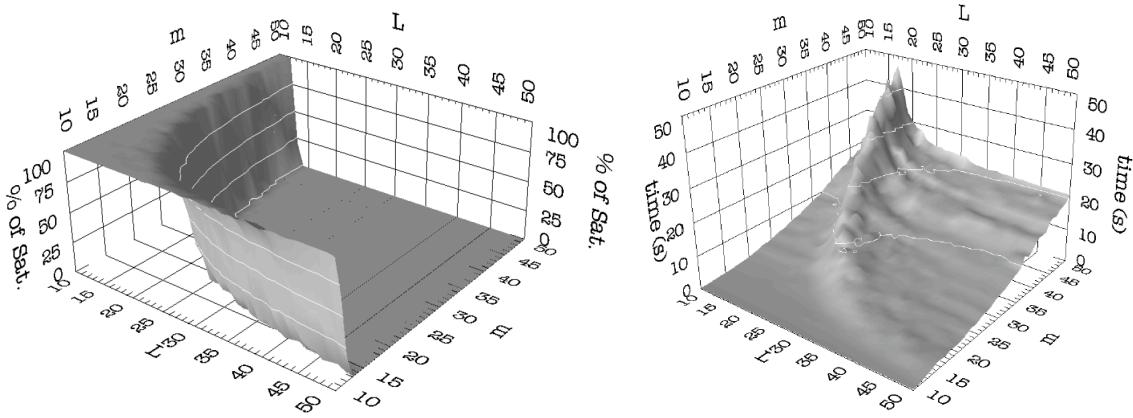


FIGURE 2.4 – (À gauche) *Pourcentage de tests de subsomption satisfaits sur 100 couples (h, e) où h comprend 10 variables et m littéraux ($m \in \{10, \dots, 50\}$), et e est généré uniformément avec $N = 100$ littéraux construits sur chacun des m symboles de prédicat et L constantes ($L \in \{10, \dots, 50\}$)*. (À droite) *Coût du test de θ -subsomption*. (Illustrations tirées de [MS04])

A posteriori, avertis par les observations réalisées à propos des problèmes K -SAT, les résultats obtenus étaient à attendre. Il y a effectivement une transition brutale de la probabilité de couverture d'un exemple par une hypothèse entre une région « YES » dans laquelle cette probabilité est presque égale à 1 et une région « NO » dans laquelle elle est presque égale à 0, avec, entre les deux, une région de transition, une « falaise », très étroite. De plus, le coût de recherche d'une solution s'accroît notablement au voisinage de la région de transition.

La découverte de ce phénomène dans le cadre de l'apprentissage inductif a été largement diffusée dans la communauté scientifique comme en témoigne une liste, sans doute incomplète, de publications sur ce sujet ([GBS99, BGSS99, GS00, GSSB00b, GSSB00a, SGS01, BGSS03, SZ00, BRS02, MS04]). Pourtant, on peut aussi dire que l'impact en a été limité. Ces travaux sont peu cités et ils ne semblent pas avoir modifié la perspective des chercheurs sur l'apprentissage inductif²². Ce relatif désintérêt est sans doute du au fait que ces travaux ont été perçus comme n'affectant que la programmation logique inductive (PLI), qui est une petite province, pour ne pas dire un valeureux petit village, dans l'univers de l'apprentissage artificiel. De plus, même à l'intérieur de la PLI, les nombreux problèmes conceptuels encore non résolus tendent à faire

22. En comparaison, par exemple, de la théorie statistique de l'apprentissage de Vapnik, pourtant relativement difficile et dont les leçons (nécessité du contrôle de l'espace des hypothèses) n'étaient pas nouvelles pour les praticiens de l'apprentissage.

passer au second plan des difficultés jugées d'ordre uniquement computationnel.

Informé pour la première fois de ces travaux par un exposé, en séminaire d'équipe, de Michèle Sebag au printemps 2000, j'ai pour ma part été immédiatement saisi par l'excitation que l'on ressent à pressentir une vérité profonde, une question essentielle. Pris par la fièvre de vouloir comprendre, j'ai alors passé des journées à réfléchir et à noircir des pages de mes cahiers de recherche.

Il faut dire qu'alors que Michèle et le groupe de Turin insistaient sur l'importance de l'explosion du coût computationnel de la subsomption dans la région dite de transition de phase, je considérais cet aspect de la question comme tout à fait secondaire par rapport à la signification du comportement de la probabilité de couverture d'un exemple par une hypothèse.

En effet, si les hypothèses couvrent soit presque tous les exemples possibles, soit presqu'aucun, alors c'est que l'on est en présence d'une propriété tout à fait remarquable du langage dans lequel on peut exprimer les hypothèses, qui interdit la description d'hypothèses de couverture intermédiaire. Par delà la notion de capacité de l'espace des hypothèses mis en évidence par la théorie statistique de l'apprentissage, il faudrait donc désormais également prendre en compte le pouvoir expressif de cet espace en terme de variation du taux de couverture des hypothèses de \mathcal{H} . Et cela pour deux raisons au moins.

D'abord, si un langage d'hypothèses n'autorise que la description d'« hypothèses ballons » ou d'« hypothèses confettis », alors il est sans espoir de chercher des « hypothèses balles » de couverture intermédiaire, celles qui correspondent pourtant au plus grand nombre de parties de l'espace des exemples \mathcal{X} ²³. Cette **limitation de l'expressivité** de \mathcal{H} peut très bien ne pas se refléter dans les mesures de capacité utilisées en théorie statistique de l'apprentissage. En fait, des petits calculs, non publiés, essayant de relier dimension de Vapnik-Chervonenkis d_{VC} et phénomène de transition de phase du taux de couverture montrent qu'une transition de phase, même radicale (de région intermédiaire de mesure nulle), peut avoir lieu sans être signalée par la $d_{VC}(\mathcal{H})$. Il serait donc souhaitable d'introduire une mesure plus précise de la capacité effective, fonction de l'échelle de description, de l'espace des hypothèses.

La deuxième raison pour laquelle le phénomène de transition de phase est capital, c'est que tous les algorithmes d'induction utilisent, d'une manière ou d'une autre, une technique d'exploration de \mathcal{H} par gradient s'appuyant sur la mesure du taux de couverture sur l'échantillon d'apprentissage. Dans la mesure où le phénomène de transition de phase signifie l'existence de deux régions dans lesquelles le gradient est quasi nul, avec une région intermédiaire presque négligeable, il implique que **les techniques de recherche par gradient sont impuissantes** dans ces conditions. La seule région de l'espace des hypothèses où l'on puisse effectivement tester les hypothèses est la région de transition de phase qui est extrêmement limitée et difficile à atteindre. C'est aussi le seul endroit où l'algorithme peut gagner de l'information.

Par conséquent, on voit ici apparaître des conditions de non apprenabilité qui sont très différentes de celles mises en avant dans la théorie statistique de l'apprentissage à savoir notion de capacité incontrôlée de l'espace des hypothèses, impossibilité d'exprimer de manière concise les exemples d'apprentissage (rasoir d'Occam), etc.

À l'image traditionnelle de l'espace des hypothèses comme un losange, il faut donc substituer l'image d'un yoyo (figure 2.5). Cela se traduit par le fait que les algorithmes « ascendants » : opérant par généralisation à partir d'une hypothèse très spécifique (par exemple à partir d'une hypothèse couvrant juste un exemple ou juste les exemples de l'échantillon d'apprentissage), ou, au contraire, « descendants » : opérant par spécialisation d'une hypothèse très générale, sont

23. Qui varie comme $\binom{|\mathcal{X}|}{p}$ où $|\mathcal{X}|$ est le cardinal de l'ensemble \mathcal{X} , et p le nombre d'exemples de la partie de \mathcal{X} considérée.

d'emblée confrontés à un énorme choix de directions possibles, respectivement de généralisation ou de spécialisation, au moment même où aucune information n'est disponible pour guider ces choix puisque toutes les hypothèses candidates semblent se valoir ayant toutes respectivement un taux de couverture presque égal à 0 ou presque égal à 1. Quand enfin l'exploration conduit la recherche dans la région dans laquelle le taux de couverture est intermédiaire, c'est-à-dire là où les hypothèses peuvent discriminer entre les exemples positifs et négatifs, tellement de choix ont déjà été faits (puisque une exploration exhaustive est inenvisageable) qu'il est très peu probable que les hypothèses considérées soient proches du concept cible.

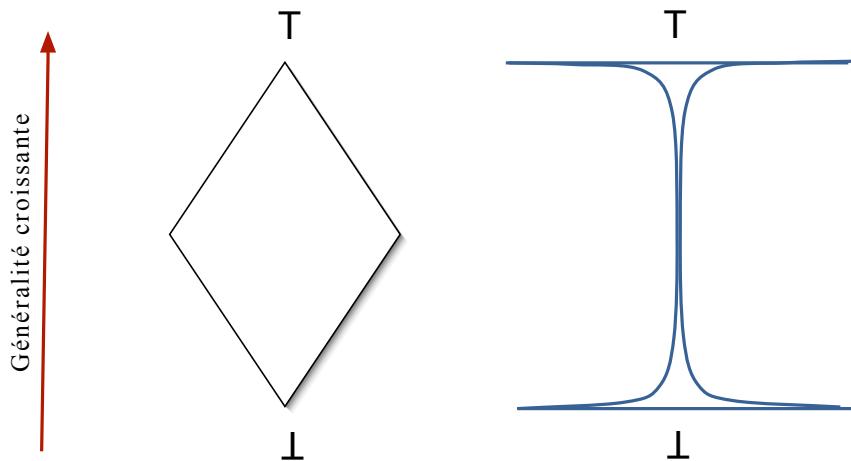


FIGURE 2.5 – (À gauche) On représente souvent l'espace des hypothèses par un losange dans lequel l'axe vertical correspond au degré de généralité (associé au taux de couverture) des hypothèses. Contrairement à l'idée répandue, lorsqu'il y a transition de phase du taux de couverture des hypothèses, il n'y a pas une majorité des hypothèses avec un taux de couverture intermédiaire (image du losange), mais la quasi totalité des hypothèses qui sont soit très générales, soit très spécifiques (image du yoyo).

En fait, les travaux publiés ont montré que le comportement des algorithmes, ascendants ou descendants, dans ces conditions est encore plus pervers.

Transition de phase et approximation de concept cible

Jusqu'ici, les expériences rapportées ne tiennent pas compte de la classe des exemples dans le test de subsomption. Une hypothèse de taux de couverture proche de 1 (resp. 0) couvre presque tous les exemples possibles (resp. presque aucun) et donc ne peut pas, sauf cas très particulier, discriminer entre exemples positifs et exemples négatifs. Cela semble donc exclure la possibilité d'un apprentissage comme nous l'avons vu. Pourtant, de nombreuses expériences d'apprentissage de concept avec les algorithmes existants en programmation logique inductive montrent que ces algorithmes retournent des hypothèses de taux de couverture intermédiaire et d'erreur empirique faible. Comment cela est-il possible ?

Il faut cette fois-ci introduire dans le protocole expérimental la notion de concept cible permettant d'étiqueter les exemples d'apprentissage (voir par exemple [BGSS03]).

Dans ce cadre, des problèmes d'apprentissage sont générés aléatoirement. Des concepts cibles sont générés de la même manière que les hypothèses dans le protocole décrit précédemment : à

2.2. Un phénomène de « transition de phase » trop ignoré ?

l'aide de n variables et de m symboles de prédicats²⁴. Par ailleurs, les exemples sont également générés aléatoirement comme précédemment : à l'aide de N littéraux et de L constantes.

Cependant, si aucune précaution n'était prise à ce stade, les problèmes impliquant des concepts cibles dans la région « YES » (resp. « NO ») seraient associés à des échantillons dont (presque) tous les exemples seraient naturellement étiquetés positivement (resp. négativement), et par conséquent les algorithmes d'apprentissage seraient dans l'incapacité de fonctionner, indépendamment même de l'existence d'une transition de phase. Pour ne pas surpénaliser les algorithmes, les échantillons d'apprentissage fournis sont donc ré-équilibrés artificiellement en forçant à retenir un nombre égal d'exemples positifs et négatifs. Cela revient à modifier localement la distribution de probabilité définie *a priori* sur l'espace des exemples \mathcal{X} .

La performance en généralisation est mesurée par le taux d'erreur obtenu sur un échantillon de test non bruité dans lequel les exemples positifs et négatifs sont équi-répartis comme pour l'échantillon d'apprentissage.

On observe trois phénomènes intéressants, quelques soient les algorithmes testés²⁵ :

1. La recherche d'hypothèses aboutit toujours dans la région de transition de phase.
2. La carte de performance des algorithmes (voir figure 2.6 pour un exemple) montre que les problèmes d'apprentissage difficiles se situent « au pied » de la falaise (concepts cibles de couverture presque nulle et à distance faible de la région de transition de phase).
3. Même lorsque les hypothèses retournées par les systèmes d'apprentissage ont une bonne performance en généralisation, elles sont le plus souvent très différentes des concepts cibles.

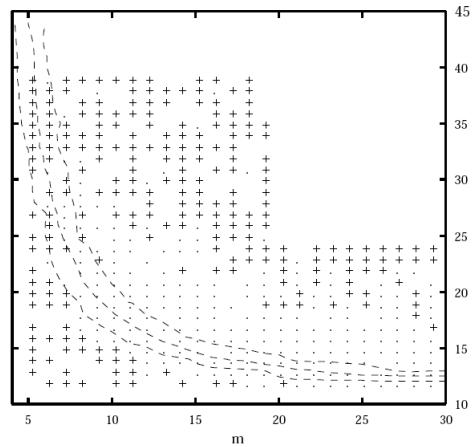


FIGURE 2.6 – *Carte de compétence obtenue pour FOIL pour les valeurs $n = 4$ et $N = 100$. La région de transition de phase est indiquée par les courbes en pointillés qui correspondent respectivement aux contours $P_{sol} = 90\%$, $P_{sol} = 50\%$ et $P_{sol} = 10\%$. Les '+' indiquent une erreur en généralisation < 20%, les '-' une erreur > 20%.*

Il est facile d'expliquer ces phénomènes. Comme il a été dit, seules les hypothèses qui se trouvent dans la région de la transition de phase sont capables de discriminer les exemples positifs des exemples négatifs. Le processus d'exploration de l'espace des hypothèses est donc contraint d'aller tester les hypothèses de cette région (phénomène (1) d'attraction).

24. À cause des limitations des algorithmes d'apprentissage existants, le nombre n de variables a du être limité à 4.

25. FOIL, SMART+ et G-NET.

Par ailleurs, il n'est pas surprenant que les problèmes pour lesquels les concepts cibles sont généraux et simples (impliquant peu de littéraux), soient faciles à apprendre, particulièrement pour les algorithmes descendants. Dans le cas de concepts cibles très spécifiques (dans la région « NO »), toutes les hypothèses situées dans leur « cône de généralisation » (c'est-à-dire plus générales qu'eux) sont nécessairement complètes (couvrent les exemples positifs) et sont très probablement correctes (ne couvrant pas les exemples négatifs) puisque leur taux de couverture est infime. Il existe donc toute une région d'hypothèses apparemment cohérentes (complètes et correctes) avec l'échantillon d'apprentissage. Lorsque le concept cible est très spécifique, le cône de généralisation est plus large que si le concept cible est un peu moins spécifique. Il est donc plus facile de trouver une hypothèse bonne en apprentissage et en généralisation (point 2), même si elle est, de fait, très différente du concept cible (point 3). (voir figure 2.7).

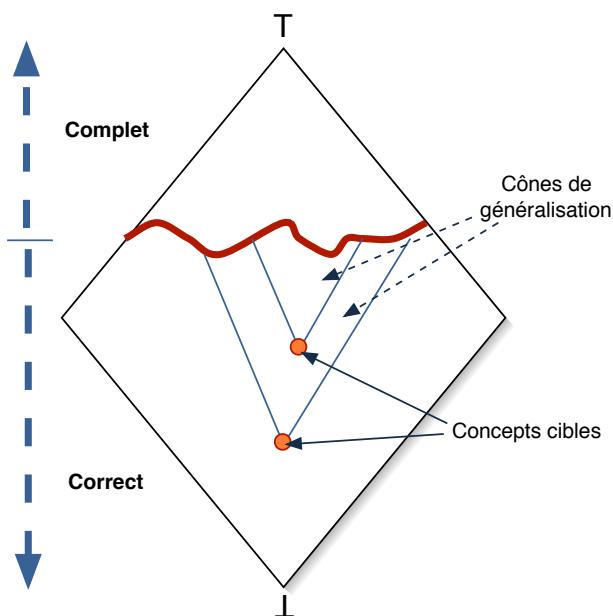


FIGURE 2.7 – Représentation de l'espace des hypothèses, l'axe vertical correspondant à une généralité croissante depuis l'hypothèse vide \perp jusqu'à l'hypothèse « Tout » T . La ligne curviligne représente la zone de transition de phase entre la région des hypothèses ne couvrant presque aucun exemple (en-dessous de la ligne) et donc probablement correctes, et la région des hypothèses couvrant presque tous les exemples (au-dessus de la ligne) et donc presque certainement complètes. On a représenté le cône de généralisation (jusqu'à la zone de transition de phase) pour deux concepts cibles.

Discussion

Ce travail, passionnant, laisse cependant plusieurs questions en suspens. Certaines sont techniques, comme celle, évoquée plus haut, du lien entre phénomène de transition de phase et capacité de l'espace des hypothèses. D'autres sont méthodologiques. Les variations du taux de couverture sont effectivement ici étudiées par rapport à une distribution uniforme des exemples dans \mathcal{X} . Un autre type de distribution pourrait-il changer qualitativement la nature du phénomène observé ?

D'autres encore touchent aux fondements. D'une part, les résultats obtenus sont de nature

empirique, expérimentale, il manque une formalisation théorique qui permette de modéliser ce phénomène et d'en comprendre les fondements. D'autre part, le phénomène même, qualifié de « transition de phase », est insuffisamment étayé tant que de vrais paramètres d'ordre n'auront pas été identifiés. Il est en effet facile d'objecter qu'un changement d'échelle des dimensions pourrait transformer une falaise en pente douce. Même si cette objection pourrait être considérée comme étant de mauvaise foi ici, il reste cependant à pouvoir lui faire pièce.

On peut donc comprendre que de nombreux chercheurs, sans remettre en doute les observations réalisées, puissent excuser leur molle indifférence par des interrogations sur la portée réelle de ces résultats.

Quelques exercices

Ces interrogations m'ont conduit à plusieurs travaux que je rapporte rapidement ici, avant de détailler davantage les travaux publiés.

Transition de phase et dimension de Vapnik-Chervonenkis. La première question, d'ordre théorique, que j'ai étudiée, est celle du lien entre l'expressivité de l'espace des hypothèses, telle qu'elle peut être mesurée par un paramètre combinatoire comme la dimension de Vapnik-Chervonenkis (d_{VC}), et le phénomène de transition de phase du taux de couverture des hypothèses. Plus précisément, en supposant que l'on ait une transition de phase radicale entre une région dans laquelle les hypothèses couvrent une proportion de plus de $(1 - \varepsilon)$ des exemples et une région où elles couvrent une proportion de moins de ε des exemples, cela implique-t-il une limite à la d_{VC} ? Inversement, une $d_{VC} = \infty$ implique-t-elle l'absence d'une transition de phase radicale?

La réponse aux deux questions semble négative.

Existence d'une transition de phase sous des distributions non uniformes. Afin de déterminer si l'apparition d'une transition de phase est un phénomène robuste, même sous des distributions non uniformes, un de mes stagiaires de DEA, Raymond Ros, a entrepris de reprendre les expériences de Saitta et de ses collègues dans le cadre d'un agent apprenant à se déplacer dans un labyrinthe, décrit en logique des prédictats, et en utilisant le système LEX décrit par Tom Mitchell [MU83]. Les expériences sont très longues et en cours.

En l'absence d'une théorie permettant de prédire pour quels espaces d'hypothèses et espaces d'exemples il peut y avoir transition de phase du taux de couverture, il est important de cerner, empiriquement dans un premier temps, les cas où l'on observe une telle transition. C'est dans cette intention que j'ai lancé, à partir de 2001, des recherches sur la possibilité d'une transition de phase en inférence grammaticale.

2.3 Transition de phase en Inférence grammaticale

Le stage de DEA de Sandra Pinto, au printemps 2001, a été l'occasion de commencer à explorer une question qui me semblait importante. Puisqu'il est établi qu'il existe un phénomène de transition de phase du taux de couverture des hypothèses en programmation logique inductive, et que ce n'est pas le cas en apprentissage dans des langages d'hypothèses propositionnels en général, peut-on supposer qu'il existe un seuil entre ces deux cas, un peu comme entre les problèmes 2-SAT et 3-SAT? Peut-on essayer de préciser ce seuil? En particulier, je voulais savoir

si ce phénomène de transition de phase était cantonné, si l'on peut dire, aux apprentissages en logique d'ordre un, ou si il pouvait également affecter des situations d'apprentissage pour lesquelles il n'était pas soupçonné.

En supposant que le spectre s'établissait en fonction du pouvoir expressif des langages d'hypothèses, il fallait donc trouver un langage intermédiaire entre la logique des propositions et la logique des prédictats²⁶. Il n'existe pas, à ma connaissance, de dimension paramétrée sur l'échelle de l'expressivité des langages de représentation. Pour des raisons plus intuitives que vraiment rigoureuses, il me semblait que les grammaires pouvaient être considérées comme d'un pouvoir expressif intermédiaire.

Je demandai donc à Sandra Pinto de répliquer le type d'expériences réalisées à propos de la P.L.I. dans le cas de l'inférence grammaticale.

Cela impliquait de définir dans ce cadre les notions de concepts ou d'hypothèses, d'exemples et de couverture. Il fallait ensuite déterminer des paramètres naturels permettant la description de l'espace des hypothèses et de l'espace des exemples, puis de faire un sondage systématique du taux de couverture dans ces espaces paramétrés, à la manière des expériences réalisées par le groupe de Turin en P.L.I. Si ce travail pouvait paraître assez trivial à première vue, et j'avoue l'avoir cru un moment, il s'est révélé à la fois plus exigeant et plus riche qu'attendu.

Mais il est nécessaire, au préalable, de prendre en compte quelques notions fondamentales en inférence grammaticale (voir pour plus de détails une abondante littérature, dans laquelle les Français se distinguent, [CM02] chap.7 ou [DM98] pour d'excellentes synthèses et listes de références).

2.3.1 Inférence grammaticale : quelques notions de base

Définition 2.3 (Alphabet, lettres, phrases)

Soit Σ un ensemble fini, appelé alphabet et les lettres a, b, c, \dots les éléments de Σ . On note u, v, w, x des éléments de Σ^* , c'est-à-dire des chaînes ou phrases²⁷ de longueur dénombrable sur Σ . On note ϵ la chaîne vide de longueur nulle et $|u|$ la longueur de la chaîne u .

Définition 2.4 (Préfixe)

u est un préfixe de v s'il existe w tel que $uw = v$.

Définition 2.5 (Langage)

Un langage L est un sous-ensemble quelconque de Σ^* . Les éléments de L sont des séquences de lettres de Σ , donc des chaînes.

Grammaires

Une grammaire est un objet mathématique auquel est associé un processus algorithmique permettant d'engendrer un langage.

Définition 2.6

Une grammaire est un quadruplet $G = (N, \Sigma, P, S)$ dans lequel :

- N est un alphabet composant l'ensemble des symboles non-terminaux de G .
- Σ est l'alphabet terminal de G , disjoint de N . On note : $V = N \cup \Sigma$.
- $R \subseteq (V^* N^+ V^* \times V^*)$ est un ensemble fini de règles de production.

26. Ou plutôt la forme limité de la logique des clauses de Horn sans symboles de fonctions utilisée en P.L.I.

27. On dit aussi mots ou séquences.

- $S \in N$ est l'axiome de G .

Une règle de production s'écrit $\alpha \rightarrow \beta$, avec $\beta \in V^*$ et $\alpha \in V^*N^+V^*$, ce qui signifie que α comporte au moins un symbole non-terminal.

Génération d'un langage par une grammaire

Soit $u \in V^+$ et²⁸ $v \in V^*$. On dit que u se réécrit en v selon la grammaire G , ou que la grammaire G dérive v de u en une étape si et seulement si on peut écrire u et v sous la forme :

- $u = xu'y$ (avec éventuellement $x = \epsilon$ ou $y = \epsilon$)
- $v = xv'y$
- avec : $(u' \rightarrow v') \in P$

La grammaire G dérive v en k étapes de u si et seulement s'il existe $k \geq 1$ et une suite $(v_0 \dots v_k)$ de mots de V^+ tels que :

- $u = v_0$
- $v = v_k$
- v_i se récrit en v_{i+1} pour $0 \leq i \leq k - 1$

La grammaire G dérive v de u s'il existe un entier k tel que u dérive v en k étapes. Pour $k \geq 1$, la séquence $u, \dots, v_i, \dots, v_k$ s'appelle une *dérivation* de v par u .

Définition 2.7 (Langage engendré par une grammaire)

On dit qu'un mot $v \in \Sigma^*$ est engendré par la grammaire G quand il peut se dériver à partir de l'axiome S de G .

Le langage engendré par la grammaire G est l'ensemble des mots de Σ^* engendrés par G . On le note $L(G)$.

Deux types de grammaires

Selon la forme de leur règles $\alpha \rightarrow \beta$, on distingue les grammaires suivantes :

Type 2 : grammaires algébriques²⁹ : $A \rightarrow \beta$, avec $A \in N$ et $\beta \in V^*$

Type 3 : grammaires régulières : $A \rightarrow wB$ ou $A \rightarrow w$, avec $w \in \Sigma^*$, $A \in N$ et $B \in N$

Un langage pouvant être engendré par une grammaire régulière (resp : algébrique) est appelé *langage régulier* (resp : *langage algébrique*). Un résultat classique de la théorie des langages est le suivant ([AU72]) :

Théorème 2.1

Tout langage régulier peut être engendré par un automate fini. Tout automate fini engendre un langage régulier.

Les automates finis

Comme l'assure le théorème précédent, les automates finis sont équivalents aux grammaires régulières. Ils sont d'un emploi beaucoup plus facile. Nous allons maintenant les définir rigoureusement.

Définition 2.8 (Automate fini)

Un automate fini est un quintuplet $(Q, \Sigma, \delta, q_0, F)$ où Q est un ensemble fini d'états, Σ est un alphabet fini, δ est une fonction de transition, c'est-à-dire une application de $Q \times \Sigma \rightarrow 2^Q$,

28. Étant donné un alphabet V , V^* désigne l'ensemble de tous les mots sur V , alors que V^+ exclut le mot vide.

29. Les grammaires algébriques sont aussi appelées *context-free* ou *hors-contexte*.

$Q_0 \in Q$ est le sous-ensemble des états initiaux et $F \in Q$ est le sous-ensemble des états finaux ou d'acceptation.

Définition 2.9 (Automate Fini Déterministe / Non Déterministe)

Si, pour tout q de Q et tout a de Σ , $\delta(q, a)$ contient au plus un élément (respectivement exactement un élément) et si Q_0 ne possède qu'un élément q_0 , l'automate A est dit déterministe (respectivement complet). Par la suite, nous utiliserons l'abréviation AFD pour « automate fini déterministe » et AFN pour un « automate fini non-déterministe ».

Un exemple d'automate fini est représenté à la figure 2.8. Il comporte cinq états, $Q = \{0, 1, 2, 3, 4\}$. Il est défini sur l'alphabet à deux lettres, $\Sigma = \{a, b\}$. Les états initiaux sont 0 et 5, $Q_0 = \{0, 5\}$ et les états 3 et 4 sont finaux, $F = \{3, 4\}$.

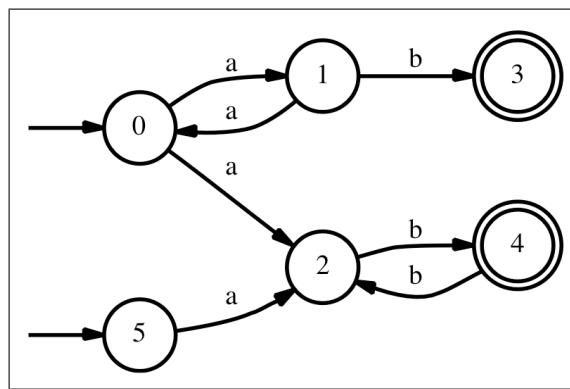


FIGURE 2.8 – Une représentation de l'automate A_1 .

Il s'agit ici d'un AFN (automate non déterministe) car il y a deux arcs étiquetés a à partir de l'état 0 et deux états initiaux. En d'autres termes, $\delta(0, a) = \{1, 2\}$.

Langage accepté par un automate fini

Définition 2.10 (Acceptation)

Une acceptation d'une chaîne $u = a_1 \dots a_l$ par un automate A , éventuellement non-déterministe, définit une séquence, éventuellement non unique, de $l + 1$ états (q^0, \dots, q^l) telle que $q^0 \in Q_0$, $q^l \in F$ et $q^{i+1} \in \delta(q^i, a_{i+1})$, pour $0 \leq i \leq l - 1$. Les $l + 1$ états sont dits atteints pour cette acceptation et l'état q^l est utilisé comme état d'acceptation. De façon similaire, les l transitions, c'est-à-dire des éléments de δ , sont dites exercées par cette acceptation.

Par exemple, la séquence des états $(0, 1, 3)$ correspond à une acceptation de la chaîne aab dans l'automate A_1 .

Cette définition est essentielle car elle fonde la notion de *couverture* d'une séquence ou chaîne par un automate.

Définition 2.11 (Langage accepté)

Le langage $L(A)$ accepté par un automate A est l'ensemble des chaînes acceptées par A .

Par extension, nous utiliserons la définition suivante :

Définition 2.12 (Couverture d'un automate)

Le langage $L(A)$ accepté par un automate A est l'ensemble des chaînes acceptées par A , et est appelé couverture de l'automate A .

2.3.2 Premiers résultats sur l'inférence grammaticale

Les grammaires régulières pouvant être représentées par des automates finis, il est tentant d'utiliser des caractéristiques de ces automates pour définir des paramètres de contrôle permettant de repérer l'espace des automates. Ainsi, les paramètres choisis sont :

- Le nombre Q d'états de l'automate
- Le nombre B d'arcs sortants pour chaque état (*branching factor*)
- Le nombre L de lettres sur chaque arc
- La fraction $a \in [0, 1]$ d'états acceptants de l'automate
- La taille $|\Sigma|$ de l'alphabet
- La longueur ℓ des exemples testés.

Le protocole expérimental consiste alors à générer aléatoirement des automates, soit déterministes, soit non déterministes, de caractéristiques $(Q, B, L, a, |\Sigma|)$:

1. Pour chaque état q , (i) B arcs sortants (q, q') sont créés, où q' est tiré suivant une distribution uniforme sans remise parmi les Q états ; (ii) $L \times B$ lettres distinctes sont tirés suivant une distribution uniforme dans l'alphabet Σ ; et (iii) ces lettres sont distribuées également entre les B arcs sortants de q .
2. Chaque état q est transformé en état acceptant avec une probabilité a .

Pour chaque quintuplet de valeurs des paramètres de contrôle, $N = 100$ automates ont été engendrés selon cette procédure. Dans les premières expériences, les valeurs des paramètres ont été prises dans les domaines suivants : $Q \in \{5, 10, \dots, 50\}$, $B \in [1..5]$, $L \in [1..4]$, $a \in \{0.1, 0.2, \dots, 1\}$.

La couverture de chaque automate a été estimée par le taux d'acceptation sur $M = 1000$ chaînes de longueur ℓ tirées aléatoirement (ici, dont chaque lettre est tirée suivant une distribution uniforme dans Σ)³⁰. Une chaîne est comptée comme acceptée lorsque, dans la séquence d'états rencontrés lors de sa lecture, un état acceptant est atteint. Le taux de couverture pour un point dans l'espace à 5 dimensions défini par les paramètres de contrôle est donc estimé par le rapport du nombre de chaînes acceptées sur le nombre de chaînes testées pour ce quintuplet de valeurs (ici, chaque estimation résulte donc de 100'000 tests d'acceptation).

Il est immédiatement apparent qu'un degré de liberté intervient dans ces expériences, qui n'existe pas en P.L.I. En effet, la probabilité d'acceptation d'une chaîne dépend de sa longueur. Un modèle mathématique de la probabilité d'acceptation montre clairement cette dépendance :

$$P(\text{accept}) = \begin{cases} a \cdot \left(\frac{B \cdot L}{|\Sigma|}\right)^\ell & \text{pour un DFA} \\ a \cdot [1 - (1 - \frac{L}{|\Sigma|})^B]^\ell & \text{pour un NFA} \end{cases} \quad (2.3)$$

Plusieurs expériences ont donc été menées en faisant varier la longueur ℓ des chaînes testées. Elles ont toutes résulté dans des comportements qualitativement similaires à ce qui est montré sur la figure 2.9 en projetant sur les axes : probabilité d'état acceptant a , nombre d'arcs sortants B .

³⁰. La notion de distribution uniforme sur des chaînes, et encore plus sur des arbres est problématique (voir par exemple [DFLS04]). Nous nous sommes contentés ici d'une définition élémentaire, suffisante pour notre propos.

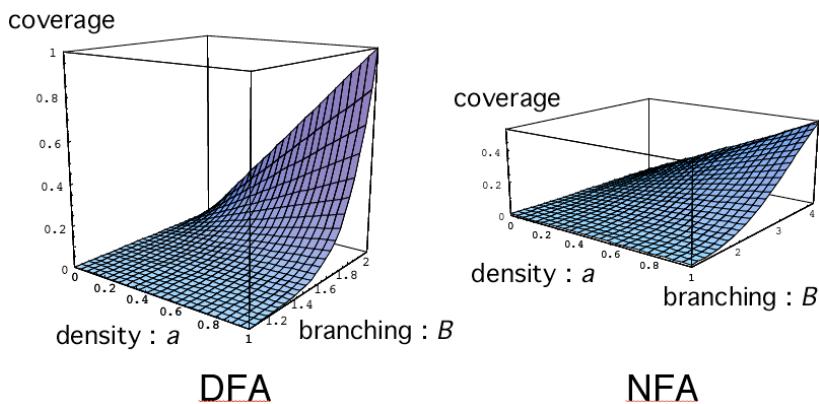


FIGURE 2.9 – Taux de couverture pour des automates déterministe (à gauche) et non déterministes (à droite). Ici, $|\Sigma|=2$, $L=1$ et $\ell=10$. La densité d'états acceptants a et le nombre d'arcs sortants B varient respectivement dans $[0, 1]$ et $\{1, 2\}$.

Le taux de couverture des automates décroît lorsque a et B décroissent, et la pente est plus abrupte dans le cas des DFA (automates finis déterministes) que pour les NFA (automates finis non déterministes). Mais, manifestement, il n'y a pas de phénomène de transition de phase.

On aurait pu en rester là ... et passer à côté de quelque chose de très intéressant.

En effet, si la variation du taux de couverture dans l'espace des paramètres défini ne présente pas de transition de phase, mais au contraire un gradient qui semble fort sympathique, ce n'est pas là l'espace exploré par les algorithmes d'inférence grammaticale existants.

Mais il faut, pour apprécier ce point, présenter d'abord quelques ...

2.3.3 Notions de base sur l'espace de recherche en inférence grammaticale

Nous définissons d'abord la notion de représentation minimale des langages.

Définition 2.13 (Automate canonique)

Pour tout langage régulier L , il existe un AFD noté $A(L)$ qui engendre L et possède un nombre minimal d'états. $A(L)$ est généralement appelé automate déterministe minimal ou automate canonique de L . On démontre que $A(L)$ est unique [AU72]. Par la suite, nous parlerons indifféremment de l'automate canonique ou de l'automate minimal déterministe de L .

Par exemple, l'automate canonique représenté à la figure 2.10 accepte le langage composé des phrases commençant par un nombre impair de a , suivi d'un nombre impair de b . Il s'agit du langage accepté également par l'automate A_1 de la figure 2.8. Il n'existe pas d'automate déterministe comportant moins d'états et acceptant ce langage.

Automates dérivés

Nous définissons maintenant une relation d'ordre partiel sur l'ensemble des automates, qui permettra un apprentissage par généralisation dans l'esprit de la méthode de l'espace des versions.

Définition 2.14 (Partition π)

Pour tout ensemble S , une partition π est un ensemble de sous-ensembles de S , non vides et disjoints deux à deux, dont l'union est S . Si s désigne un élément de S , $B(s, \pi)$ désigne l'unique

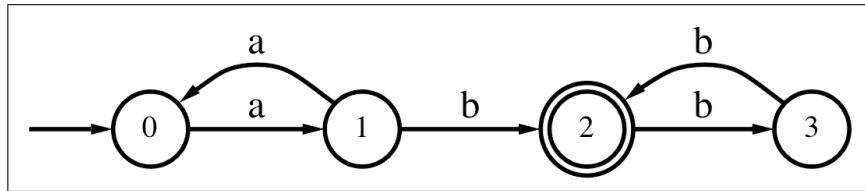


FIGURE 2.10 – Automate canonique du langage défini par l’expression régulière $L = a(aa)^*b(bb)^*$.

élément, ou bloc, de π comprenant s . Une partition π_i raffine, ou est plus fine que, une partition π_j ssi tout bloc de π_j est un bloc de π_i ou est l’union de plusieurs blocs de π_i .

Définition 2.15 (Automate quotient)

Si $A = (Q, \Sigma, \delta, q_0, F)$ est un automate, l’automate $A/\pi = (Q', \Sigma, \delta', B(q_0, \pi), F')$ dérivé de A relativement à la partition π de Q , aussi appelé l’automate quotient A/π , est défini comme suit :

$$\begin{aligned} Q' &= Q/\pi = \{B(q, \pi) | q \in Q\}, \\ F' &= \{B \in Q' | B \cap F \neq \emptyset\}, \\ \delta' : Q' \times \Sigma &\rightarrow 2^{Q'} : \forall B, B' \in Q', \forall a \in \Sigma, B' \in \delta'(B, a) \text{ssi } \exists q, q' \in Q, q \in B, q' \in B' \text{ et } q' \in \delta(q, a) \end{aligned}$$

Nous dirons que les états de Q appartenant au même bloc B de la partition π sont fusionnés.

Par exemple, reprenons l’automate A_1 , représenté à la figure 2.8 et définissons la partition de son ensemble d’états, $\pi_2 = \{\{0, 1\}, \{2\}, \{3, 4\}\}$. L’automate quotient A_1/π_2 , obtenu en fusionnant tous les états appartenant à un même bloc de π_2 , est représenté à la figure 2.11.

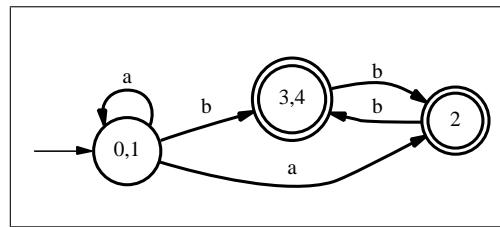


FIGURE 2.11 – L’automate quotient A_1/π_2 .

Une propriété fondamentale de l’opération de fusion est la suivante :

Proposition 2.1 (Dérivation d’automate et inclusion de langage)

Si un automate A/π_j dérive d’un automate A/π_i , alors le langage accepté par A/π_i est inclus dans celui accepté par A/π_j .

Par conséquent, en partant d’un automate A , on peut construire tous les automates dérivés de A en énumérant les partitions des états de A . Il existe sur cet ensemble une relation d’ordre partiel qui est cohérente avec l’inclusion des langages que reconnaissent ces automates. On peut ainsi construire un **treillis de généralité sur l’espace des automates**.

Proposition 2.2 (Automates dérivés et treillis)

L’ensemble des automates dérivés d’un automate A , qui est partiellement ordonné par la re-

lation de dérivation, est un treillis³¹. Les automates A et UA , l'automate universel, en sont respectivement les éléments minimal et maximal. On note ce treillis $\text{Lat}(A)$.

Illustration. En reprenant l'exemple A_1 de la figure 2.8, on a vu que le choix de la partition $\pi_2 = \{\{0, 1\}, \{2\}, \{3, 4\}\}$ permet de dériver l'automate quotient $A_2 = A_1/\pi_2$, représenté à la figure 2.11. On sait donc que le langage reconnu par A_2 inclut celui reconnu par A_1 . Prenons maintenant la partition $\pi_3 = \{\{0, 1, 2\}, \{3, 4\}\}$. Elle permet de dériver un automate A_3 , représenté à la figure 2.12.

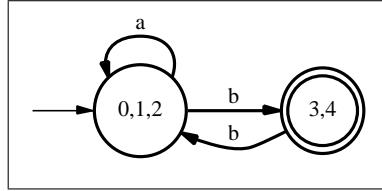


FIGURE 2.12 – L'automate quotient A_1/π_3 .

La partition π_3 est *moins fine* que π_2 , puisque ses blocs sont construits comme une union de blocs de π_2 ; on peut donc assurer que l'automate A_3 reconnaît un langage qui inclut celui reconnu par A_2 . En revanche, si on construit l'automate A_4 (figure 2.13) par dérivation de A_1 selon la partition $\pi_4 = \{\{0\}, \{1, 3\}, \{2, 4\}\}$, qui n'est ni moins fine ni plus fine que π_2 , on ne peut rien dire sur l'inclusion éventuelle des langages reconnus par A_4 et A_2 . Par exemple, la phrase abb est reconnue par A_4 et pas par A_2 , alors que la phrase b est reconnue par A_2 et pas par A_4 .

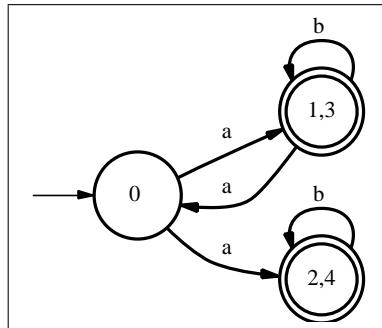


FIGURE 2.13 – L'automate quotient A_1/π_4 .

Conformément à l'astuce de représentation unique (*single representation trick*), l'élément minimal du treillis de généralité qui est exploré lors de l'apprentissage est un automate qui reconnaît exactement et uniquement les chaînes de l'échantillon positif.

Définition 2.16 (Échantillons positif et négatif)

Nous désignons par S^+ un sous-ensemble fini, appelé échantillon positif, d'un langage L quelconque. Nous désignons par S^- un sous-ensemble fini, appelé échantillon négatif, du langage complémentaire $\Sigma^* - L$. S^+ et S^- sont des sous-ensembles finis et disjoints de Σ^* .

Par ailleurs, nous introduisons ici une notion sur laquelle nous aurons à revenir plus tard

31. Cet ensemble n'est pas un treillis au sens algébrique, mais l'usage a jusqu'ici fait prévaloir ce terme.

lorsque nous parlerons de conditions d'apprentissage d'un automate cible (voir section 2.3.5). Tous les automates du treillis vont en effet être d'une certaine manière des « expressions économiques » de l'échantillon d'apprentissage.

Définition 2.17 (Échantillon structurellement complet)

Un échantillon S^+ est structurellement complet relativement à un automate A acceptant L , s'il existe une acceptation $\mathcal{AC}(S^+, A)$ de S^+ telle que :

- Toute transition de A soit exercée.
- Tout élément de F (l'ensemble des états finaux de A) soit utilisé comme état d'acceptation.

Par exemple, l'échantillon $S^+ = \{aab, ab, abbbb\}$ est structurellement complet relativement à l'automate A_1 .

L'élément minimal du treillis de généralité est l'automate maximal canonique.

Définition 2.18 (MCA)

On désigne par $MCA(S^+) = (Q, \Sigma, \delta, q_0, F)$ l'automate maximal canonique³² relatif à S^+ [Mic80].

Par construction, $L(MCA(S^+)) = S^+$ et $MCA(S^+)$ est le plus grand automate (l'automate ayant le plus grand nombre d'états) pour lequel S^+ est structurellement complet. $MCA(S^+)$ est généralement non-déterministe.

Sa construction est facile à observer sur un exemple. L'automate représenté à la figure 2.14 est l'automate maximal canonique relatif à l'échantillon $S^+ = \{a, ab, bab\}$.

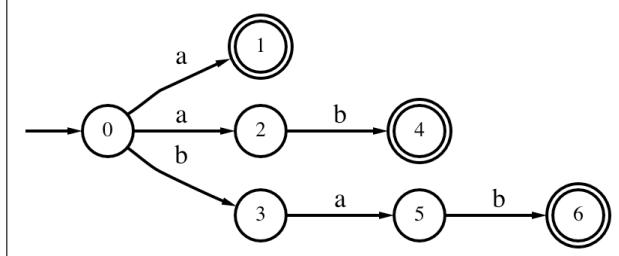


FIGURE 2.14 – L'automate maximal canonique de l'échantillon $\{a, ab, bab\}$.

Un automate dérivé du MCA, l'arbre accepteur des préfixes, est généralement utilisé en lieu et place du MCA comme élément minimal du treillis de généralité exploré lors de l'apprentissage.

Définition 2.19 (PTA)

Nous désignons par $PTA(S^+)$ l'arbre accepteur des préfixes³³ de S^+ [Ang82]. Il s'agit de l'automate quotient $MCA(S^+)/\pi_{S^+}$ où la partition π_{S^+} est définie comme suit :

$$B(q, \pi_{S^+}) = B(q', \pi_{S^+}) \text{ssi } Pr(q) = Pr(q').$$

En d'autres termes, $PTA(S^+)$ peut être obtenu à partir du $MCA(S^+)$ en fusionnant les états partageant les mêmes préfixes. $PTA(S^+)$ est déterministe.

32. L'abréviation **MCA** pour *Maximal Canonical Automaton* provient de la terminologie anglaise. Le qualificatif *canonique* se rapporte ici à un échantillon. Le MCA ne doit pas être confondu avec l'automate canonique d'un langage (voir déf. 2.13).

33. L'abréviation PTA pour *Prefix Tree Acceptor* provient de la terminologie anglaise.

À titre d'exemple, l'automate représenté à la figure 2.15 est l'arbre accepteur des préfixes relatif à l'échantillon $S^+ = \{a, ab, bab\}$.

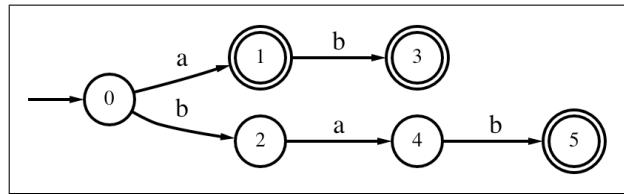


FIGURE 2.15 – L’arbre accepteur des préfixes PTA(S^+), avec $S^+ = \{a, ab, bab\}$.

Finalement, l’élément maximal du treillis est l’automate universel.

Définition 2.20 (Automate universel : UA)

Nous désignons par UA l’automate universel³⁴. Il accepte toutes les chaînes définies sur l’alphabet Σ , c’est-à-dire $L(UA) = \Sigma^*$. Il s’agit donc du plus petit automate pour lequel tout échantillon de Σ^* est structurellement complet.

L’automate universel défini sur l’alphabet $\Sigma = \{a, b\}$ est représenté à la figure 2.16.

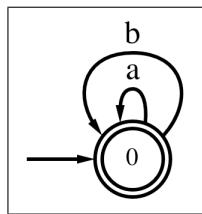


FIGURE 2.16 – L’automate universel sur l’alphabet $\Sigma = \{a, b\}$.

Les algorithmes d’inférence grammaticale s’appuient sur quelques propriétés et hypothèses fondamentales, que nous présentons rapidement.

Un premier théorème assure que, sous le biais de la complétude structurelle, l’ensemble des hypothèses compatibles avec l’échantillon est exactement le treillis construit sur $MCA(S^+)$.

Théorème 2.2

Soit S^+ un échantillon positif d’un langage quelconque régulier L et soit A n’importe quel automate acceptant exactement L . Si S^+ est structurellement complet relativement à A alors A appartient à $\text{Lat}(MCA(S^+))$. Réciproquement, si un automate A appartient à $\text{Lat}(MCA(S^+))$ alors S^+ est structurellement complet relativement à A .

Le second théorème assure que l’on peut réduire l’espace de recherche si on cherche l’automate canonique d’un langage.

Théorème 2.3

Soit S^+ un échantillon positif d’un quelconque langage régulier L et soit $A(L)$ l’automate canonique acceptant L . Si S^+ est structurellement complet relativement à $A(L)$ alors $A(L)$ appartient à $\text{Lat}(PTA(S^+))$.

34. L’abréviation UA pour *Universal Automaton* provient de la terminologie anglaise.

De plus, il existe une propriété d'inclusion entre ces deux treillis :

Proposition 2.3

$$\text{Lat}(\text{PTA}(S^+)) \subseteq \text{Lat}(\text{MCA}(S^+))$$

Cette propriété découle directement de la définition 2.19 du $\text{PTA}(S^+)$ qui est un automate quotient du $\text{MCA}(S^+)$. De plus, comme le treillis $\text{Lat}(\text{PTA}(S^+))$ est généralement strictement inclus dans le treillis $\text{Lat}(\text{MCA}(S^+))$, rechercher une solution dans $\text{Lat}(\text{PTA}(S^+))$ au lieu de $\text{Lat}(\text{MCA}(S^+))$ permet de considérer un espace de recherche plus restreint.

2.3.4 Transition de phase dans l'espace de recherche

Après les premières expériences effectuées en 2001 (voir [Pin01]), il était donc apparent que le paysage des variations du taux de couverture dans l'espace des automates ne présentait pas de transition de phase. Mais ce n'était pas l'espace réellement exploré par les algorithmes d'inférence grammaticale qui est celui du treillis de généralité défini par l'élément minimal qui est le PTA de l'échantillon positif S^+ et l'élément maximal : l'automate universel (UA).

J'ai donc décidé, avec Nicolas Pernot, lors de son stage de DEA au printemps 2004, d'examiner cette fois-ci plus précisément ce sous-espace des automates que constitue le treillis de généralité (Voir [PCS05a, PCS05b, CS05]).

Le protocole expérimental est par conséquent différent de celui présenté dans la section 2.3.2. Cette fois-ci, il s'agit en effet d'effectuer un sondage aléatoire dans des treillis de généralité, qui dépendent chacun d'un échantillon positif. On va donc engendrer des échantillons positifs, afin d'obtenir des PTA et ainsi des treillis de généralité qui seront ensuite sondés de manière aléatoire. Pour réaliser ce sondage, le principe est d'appliquer aléatoirement des opérations de fusion d'états, puisqu'elles sont l'opération de généralisation fondamentale en inférence grammaticale, pour obtenir des trajectoires de généralisation. On mesure alors le taux de couverture de chacun des automates ainsi obtenus par test de couverture sur un échantillon de chaînes aléatoirement tirées. Plus précisément :

1. N échantillons positifs de taille $|S^+|$ constitués de chaînes de longueur $= \ell$ ou de longueur $\leq \ell$ (dans ce cas, le nombre n_d de chaînes de longueur $d \leq \ell$ est égal à $|\Sigma|n_{d-1}$ pour respecter une distribution universelle sur les chaînes) sont tirés aléatoirement (par tirage uniforme sur les lettres de l'alphabet Σ), et les PTA correspondants sont construits.
2. Pour chaque PTA, K trajectoires de généralisation sont aléatoirement engendrées partant du PTA jusqu'à l'automate universel UA. Dans chaque trajectoire ($A_0 = \text{PTA}, A_1, \dots, A_t = \text{UA}$), l'automate A_i est construit à partir de A_{i-1} par fusion de deux états choisis aléatoirement suivant un tirage uniforme, et en appliquant ensuite des opérations de déterminisation si la tâche est de trouver un automate déterministe (DFA).
3. On mesure alors le taux de couverture de tous les automates ainsi obtenus, soit en testant T (e.g. 1000) chaînes de longueur ℓ tirées aléatoirement comme pour l'étape (1), soit en testant T chaînes de longueur $\leq \ell$ suivant la distribution universelle (comme en (1)).

Les résultats obtenus furent alors complètement différents de ceux observés pour l'espace des automates en général. Afin d'en rendre compte, nous avons tracé les graphes du taux de couverture en fonction du nombre d'états des automates testés. Une trajectoire de généralisation trace ainsi un chapelet de points partant du PTA et de son taux de couverture, jusqu'à l'automate universel (de droite à gauche sur les figures).

De nombreuses expériences ont été réalisées en variant la valeur des paramètres $|S^+|$, $|\Sigma|$ et ℓ en particulier (voir [Per04]). Les résultats sont qualitativement similaires à ceux des figures 2.18 (pour la recherche d'automates déterministes : DFA) et 2.17 (pour la recherche d'automates finis déterministes ou non : FSA).

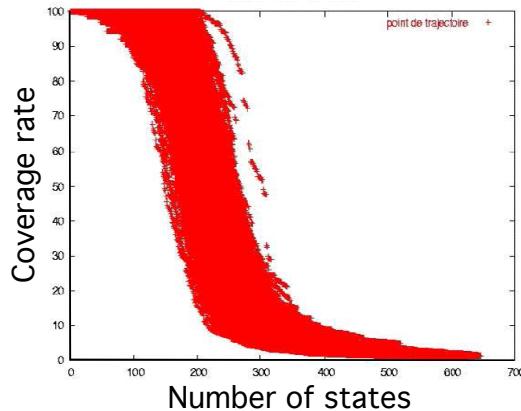


FIGURE 2.17 – Variations du taux de couverture dans le treillis de généralité des automates finis déterministes ou non (FSA). Les paramètres sont les mêmes que dans la figure 2.18.

La figure tracée pour les automates en général (figure 2.17, ne permet pas de distinguer les trajectoires individuelles. Elle montre cependant nettement que le taux de couverture varie très rapidement, de presque 0 à presque 1, sur moins de 20% de l'axe portant le nombre d'états. Des analyses plus fines sont à mener, afin d'examiner les trajectoires de généralisation individuelles et de voir en particulier si elles présentent des sauts importants dans le taux de couverture des automates explorés en cours de généralisation.

Le profil de l'évolution du taux de couverture est en revanche complètement différent dans le cas de la recherche d'automates déterministes (voir figure 2.18). La courbe exhibe, de manière spectaculaire, un « trou » dans les taux de couverture accessibles (ici dans [0.13, 0.54] environ) et dans la taille des automates (ici dans l'intervalle [180, 420] environ).

Quoique nous n'ayons pas, à ce stade, examiné précisément les automates engendrés le long d'une trajectoire (leur analyse, exigeant l'examen d'automates ayant des centaines d'états et des milliers d'arcs, est difficile), il semble aisément de comprendre la raison du saut observé. La fusion de deux états peut en effet produire un automate non déterministe. Il faut alors fusionner d'autres états jusqu'à obtenir un automate déterministe. Tant que les états de l'automate ont, en moyenne peu d'arcs sortants et/ou peu de lettres par arc, la déterminisation se réalise avec peu de fusions. Mais, au fur et à mesure des fusions, pour généralisation ou pour déterminisation, les nombres d'arcs sortants et de lettres par arc s'accroissent, et le nombre de fusions pour déterminisation aussi. Il se trouve qu'un **modèle probabiliste** que nous avons écrit et programmé, avec Michèle Sebag, prédit l'existence d'un seuil au-delà duquel une « réaction en chaîne » se produit, c'est-à-dire que chaque fusion pour déterminisation entraîne d'autres, et ainsi de suite pendant très longtemps. Le modèle, qui reste à publier, prédit le seuil de réaction en chaîne, donc de saut, avec une précision d'environ 15%, ce qui est remarquable compte tenu du fait qu'il ne tient pas compte de la structure des automates, mais seulement de valeurs moyennes d'arcs sortants et de lettres par arc.

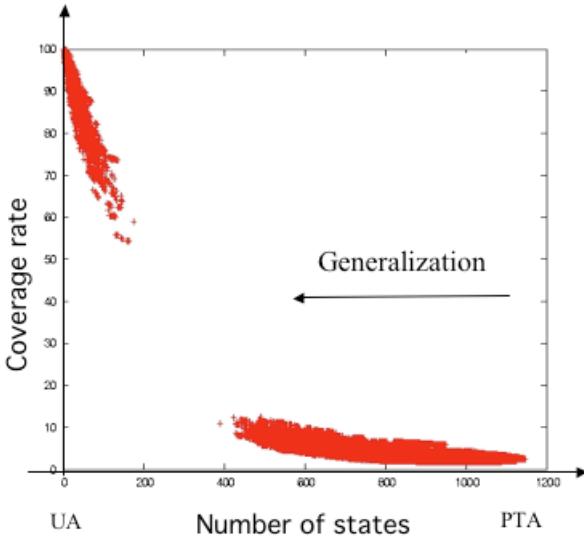


FIGURE 2.18 – Variations du taux de couverture dans le treillis de généralité restreint aux automates déterministes (DFA). Ici, $(|\Sigma| = 4, \ell = 8)$. À l’extrême droite se trouvent les départs des 50 trajectoires issues de 50 PTAs engendrés aléatoirement, chacun comprenant approximativement 1150 états. 1000 trajectoires aléatoires ont été engendrées, et le taux de couverture de 270'000 automates environ a été estimé sur un échantillon test de 1000 chaînes.

De cette courbe, et des résultats obtenus pour d’autres conditions expérimentales qui toutes confirment ce phénomène, il faut s’attendre à ce que les algorithmes d’inférence d’automates déterministes (DFA) opérant par généralisation dans le treillis de généralité aient de sérieux problèmes pour trouver des automates dans la région du « trou » si l’automate cible à apprendre se trouve dans cet intervalle de taux de couverture. Du moins, si ils opèrent par fusion aléatoire. C’est pourquoi, l’étape suivante de notre recherche a été d’examiner le comportement d’algorithmes canoniques en inférence grammaticale pour voir si ils étaient sujet à ce problème potentiel.

2.3.5 Comportement des algorithmes standards d’inférence grammaticale

Sachant que tout automate dérivé (par opération de fusion d’états) du $MCA(S^+)$ accepte un langage incluant S^+ , tout automate appartenant à $\text{Lat}(MCA(S^+))$ constitue une généralisation possible de l’échantillon positif. Mais rien ne permet, sans information supplémentaire, d’être à l’abri d’une surgénéralisation. Deux types d’approches sont utilisées pour fournir cette information. La première consiste à chercher la solution dans une sous-classe particulière des langages réguliers (e.g. recherche de langages k-réversibles, algorithme ECGI). la seconde utilise un échantillon négatif afin de borner la trajectoire de généralisation. C’est à ce dernier type d’approche, la plus utilisée, que nous nous sommes intéressés.

L’algorithme RPNI, publié par Oncina et Garcia en 1992 [OG92], est l’algorithme de base. Il effectue une recherche en profondeur d’abord dans $\text{Lat}(MCA(S^+))$ et trouve un optimum local au problème du plus petit automate déterministe cohérent (complet et correct) avec l’échantillon $S = S^+ \cup S^-$. Utilisant comme élément minimal du treillis le $PTA(S^+)$, il suppose donc que l’échantillon est structurellement complet par rapport à l’automate cible (condition suffisante, d’après le théorème 2.3, pour que celui-ci se trouve dans le treillis). Par ailleurs, RPNI retourne

un automate maximal sous la contrainte de S^- , c'est-à-dire de nombre d'états minimal (biais d'économie ou du rasoir d'Occam). Nous reportons à [CM02] chap.7, ou à [OG92] pour des détails sur l'algorithme.

L'algorithme RPNI est très efficace sous l'hypothèse de complétude structurelle de l'échantillon d'apprentissage³⁵. Mais si l'échantillon d'apprentissage ne contient pas d'échantillon caractéristique³⁶, par exemple un échantillon avec peu de données, alors l'automate inféré par RPNI peut être de mauvaise qualité³⁷. C'est pourquoi des heuristiques ont été développées pour guider l'apprentissage.

EDSM (Evidence-Driven State Merging) [LPP98] est une variante de RPNI, dans laquelle l'exploration se fait non plus en profondeur d'abord, mais sous le contrôle d'une heuristique qui privilégie les fusions d'états les plus prometteurs au sens du nombre de fusions (pour déterminisation) effectuées après fusion d'un couple d'états donné.

Expériences

Afin de disposer d'un échantillon négatif, il faut désormais définir un automate cible. Il sera, de plus, possible de contrôler ainsi son taux de couverture pour tester la capacité des algorithmes à approcher des automates cibles dans le « trou » de taux de couverture. Le protocole utilisé est celui défini par [LPP98] pour retenir des automates cibles d'un taux de couverture et d'un nombre d'états approximativement fixés. Pour chaque automate sélectionné, un nombre N d'échantillons d'apprentissage de taille $|S|$ sont tirés aléatoirement suivant le protocole défini plus haut (section 2.3.4). Le taux de couverture est calculé comme précédemment, sur un échantillon test tiré aléatoirement.

Les expériences réalisées ont porté sur des automates cibles de taux de couverture d'environ 56% (comme dans le défi *Abbadingo* [LPP98]) et d'environ 3%.

Les résultats correspondent à ceux rapportés dans la figure 2.19. Trois trajectoires de généralisation suivies par RPNI sont tracées, superposées aux nuages de points obtenus par exploration aléatoire, pour les mêmes échantillons d'apprentissage. Pour EDSM, les courbes montent légèrement plus haut, surtout pour les automates cibles de couverture d'environ 3%.

Ces résultats partiels indiquent que les heuristiques de contrôle utilisées dans RPNI et dans EDSM permettent, d'une part, d'explorer la zone de très faible densité, mais, d'autre part, tendent à produire des automates de couverture largement trop élevée lorsque l'automate cible a un faible taux de couverture (environ 15% avec RPNI, et 30% avec EDSM!). Ces résultats peuvent expliquer pourquoi le phénomène de trou dans le taux de couverture n'a pas été découvert précédemment.

Afin de mesurer l'impact de ces heuristiques sur la qualité d'apprentissage, nous avons également étudié l'erreur en généralisation des automates appris, en distinguant les erreurs de type I (faux positifs) et de type II (faux négatifs).

35. Il faut noter à ce sujet l'extraordinaire astuce derrière RPNI qui permet de tirer parti au maximum des données (sous l'hypothèse de complétude structurelle de l'échantillon). Même les fusions pour déterminisation sont valides en produisant des hypothèses dans l'espace des versions, ce qui n'était absolument pas évident *a priori*. (Je remercie Laurent Miclet pour m'avoir éclairé sur ce point et sur bien d'autres en inférence grammaticale).

36. Condition techniquement un peu plus restrictive que celle de complétude structurelle.

37. De la Higuera et al., [dlHOV96, dlH96, dlH97], ont montré que la taille de l'échantillon caractéristique peut être exponentielle en celle de l'automate cible.

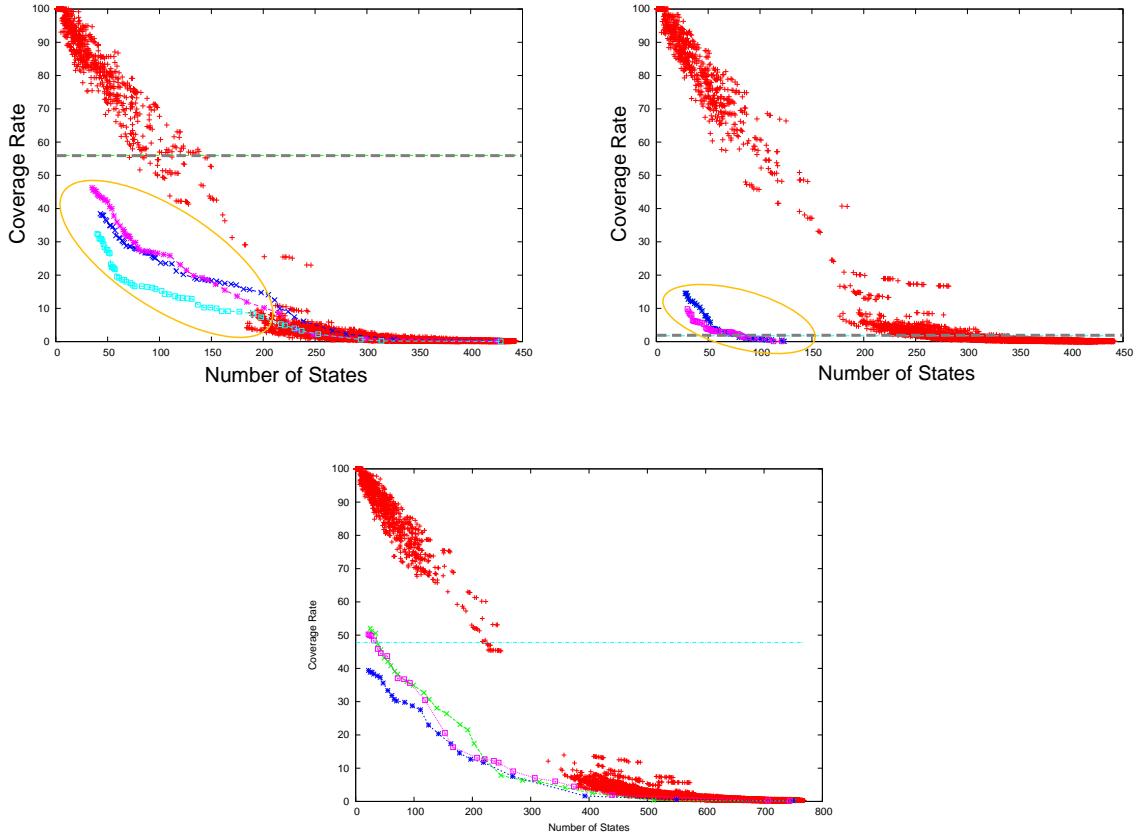


FIGURE 2.19 – Trajectoires de généralisation pour RPNI. (À gauche) : pour des automates cibles de taux de couverture $\approx 56\%$. (À droite) : pour des automates cibles de taux de couverture $\approx 3\%$. (En bas) : Trajectoires de généralisation pour EDSM. : pour des automates cibles de taux de couverture $\approx 56\%$.

Impact sur les performances en généralisation

Les expériences ont été réalisées avec des automates cibles de tailles variées (entre 25 et 100 états), et de propriétés structurales diverses (par exemple, en faisant varier le taux de récursivité du graphe). Par ailleurs, on a également fait varier la complétude structurelle des échantillons d'apprentissage utilisés, afin de voir si celle-ci avait une grande influence sur les performances obtenues. Les détails figurent dans [Per04].

La table 2.1 résume les résultats obtenus pour des échantillons d'apprentissage de complétude structurelle $> 40\%$, et pour des automates cibles de faible taux de couverture : entre 3 et 6%.

Ces résultats montrent qu'alors même que les heuristiques utilisées semblent permettre une exploration de la zone de trou de couverture, et induisent des surgénéralisations, en terme de taux de couverture, des automates appris, la couverture des exemples test positifs est faible : entre 50 et 66%, tandis que le taux de couverture des exemples test négatifs est élevé : entre 25 et 42% !

On a ici un phénomène comparable à ce qui est observé en programmation logique inductive :

Algo.	Q_c	$ucov_c$	Q_f	$ucov_f$	$pcov_f$	$ncov_f$
EDSM	15	5.97	10.38	33.81	60.93	34.69
EDSM	25	4.88	12.77	40.35	62.68	37.87
EDSM	50	4.2	14.23	45.38	66.14	42.23
EDSM	100	3.39	13.13	30.35	42.81	28.69
RPNI	15	5.95	5.14	22.9	57.51	26.99
RPNI	25	4.7	7.56	23.07	56.38	25.98
RPNI	50	3.87	14.08	23.45	51.89	24.42
RPNI	100	3.12	26.41	23.151	50.12	24.40

TABLE 2.1 – Performances de EDSM et RPNI pour des automates cibles déterministes de tailles $Q = 15, 25, 50$ et 100 états. Q_f , $ucov_f$, $pcov_f$ and $ncov_f$ dénotent respectivement la taille moyenne de l'automate appris, leur couverture moyenne, les taux de vrais positifs et de faux positifs.

des hypothèses surgénéralisées sont retenues par les algorithmes d'apprentissage et elles ont de mauvaises performances en généralisation, y compris sur les exemples positifs !

Dans la mesure où les automates cibles étaient par construction choisis dans la zone des taux de couverture de 50%, le défi *Abbadingo*, qui a servi de stimulant à la communauté en inférence grammaticale, a pu contribuer à ce que ce phénomène ne soit pas découvert plus tôt.

2.3.6 Bilan et perspectives

Les expériences rapportées ci-dessus s'inscrivent dans une investigation plus ample qui nous semble introduire un nouveau type d'analyse sur l'apprentissage supervisé, d'une grande importance potentielle.

Cette section donne un aperçu rapide des conclusions et des perspectives que nous voyons à ce stade.

Sur l'inférence grammaticale

Deux résultats peuvent être immédiatement notés à partir des expériences réalisées :

- Il existe une nette différence de comportement entre l'induction d'automates déterministes (DFA) et l'induction d'automates finis quelconques (FSA). L'induction de DFA demande des heuristiques spécifiques pour permettre l'exploration de la zone de trou dans le taux de couverture.
 - Les algorithmes RPNI et EDSM parviennent à visiter la zone de trou grâce à des heuristiques qui provoquent rapidement les opérations de fusion pour déterminisation ce qui permet d'éviter le phénomène de réaction en chaîne responsable du trou.
- Ces heuristiques sont étroitement liées à l'hypothèse de l'existence d'un échantillon d'apprentissage caractéristique impliquant la présence de l'automate cible dans le treillis³⁸. En supposant de plus que l'échantillon négatif est suffisamment informé, il devient intéressant de chercher à trouver le plus vite possible un automate maximalement général cohérent avec les données.

38. On est ici implicitement dans le cadre théorique de l'*identification à la limite* dans lequel l'apprentissage a pour but d'identifier exactement le concept cible. Le cadre PAC, qui s'attache aux conditions d'approximation du concept cible, n'est apparu que tardivement en inférence grammaticale, et n'a eu qu'un effet limité sur la définition des algorithmes d'apprentissage.

Dans le cas où l'échantillon d'apprentissage ne vérifie pas ces bonnes propriétés, il est intéressant de s'interroger sur la modification du critère d'arrêt, des opérations de généralisation et des heuristiques de contrôle.

Par exemple, une information sur le taux de couverture du concept cible permettrait de stopper la trajectoire de généralisation avant qu'elle ne dépasse ce taux. Par ailleurs, on pourrait également envisager des opérateurs plus prudents, éventuellement n'utilisant pas la fusion d'états³⁹. Cela demande cependant une analyse théorique sur les propriétés de tels opérateurs qui reste à faire, et qui n'est pas facile.

Il est également possible, sur la base de cette analyse, d'envisager des méthodes d'apprentissage différentes, n'utilisant pas le treillis de généralité fondé sur l'opération de fusion. Par exemple, dans le cas d'une exploration directe de l'espace des automates, éventuellement avec des algorithmes génétiques, on aurait un paysage du taux de couverture tel que présenté dans la figure 2.9, qui se prête bien à des techniques de gradient (voir [Dup94]). Bien entendu, l'immensité de cette espace de recherche devient un autre problème.

Une question intéressante est celle du rapport de ce travail avec les résultats rapportés par Kevin Lang [Lan92] sur le fait, qu'expérimentalement, des automates cibles tirés aléatoirement peuvent être appris avec un très petit taux d'erreur à partir d'un échantillon d'apprentissage restreint également tiré aléatoirement. Ce que Lang montre, c'est qu'un échantillon d'apprentissage très limité, et donc loin de pouvoir être caractéristique de l'automate cible, peut suffire à bien apprendre. Cependant, les automates tirés « aléatoirement » sont construits de telle manière que leur taux de couverture est d'environ 50% sur l'ensemble des chaînes de longueur bornée par une fonction de la taille de l'automate. Nos résultats montrent qu'effectivement les méthodes d'inférence grammaticale par fusion d'états fonctionnent bien pour des automates cibles ayant un tel taux de couverture. Il faudrait reproduire les expériences de Lang sur des automates de taux de couverture faible ($\leq 10\%$).

Finalement, ce qui limite nos investigations est le manque de bons paramètres (d'ordre ?) pour caractériser les automates. Il est clair que leur structure joue un rôle important, mais les paramètres utilisés pour le moment ne la reflètent pas suffisamment. Il y a là une priorité dans les recherches, et qui est d'un intérêt général pour l'inférence grammaticale.

Sur l'intérêt de l'étude de taux de couverture

La découverte du phénomène de transition de phase a permis d'expliquer ou d'éclairer les problèmes rencontrés en programmation logique inductive pour des problèmes complexes. Il est évident que la prise en compte des variations du taux de couverture dans l'espace des hypothèses apporte des informations que ne fournit pas l'analyse statistique de l'apprentissage. D'une part, elle peut montrer la relative rareté de certains types d'hypothèses (e.g. hypothèses de taux de couverture intermédiaire), d'autre part, et surtout, elle explique que les méthodes d'exploration par gradient peuvent être vouées à l'échec quand existe un phénomène de transition de phase.

D'un point de vue philosophique, l'*analyse statistique de l'apprentissage ne tient compte ni des caractéristiques des algorithmes d'apprentissage en dehors du fait qu'ils obéissent au principe de minimisation du risque empirique, ni de la structure de l'espace des hypothèses* (sa topologie) qui n'est caractérisée que par une mesure de capacité expressive.

L'étude des variations du taux de couverture exploite des informations supplémentaires sur ces deux points : d'une part, elle prend en compte le fait que les algorithmes

39. Par exemple, par ajout d'états acceptants, ou d'arcs. Nous avons fait des expériences sur huit opérateurs de généralisation afin de comparer leurs caractéristiques en terme d'évolution du taux de couverture lors du processus de généralisation. (Voir [Per04]).

s'appuient sur ce taux de couverture et utilisent tous, peu ou prou, une technique de gradient, d'autre part, elle prend en compte la structure de voisinage induite par les opérateurs de transformation d'hypothèses sur l'espace des hypothèses. Cette étude peut donc fournir des informations plus fines sur le comportement des algorithmes. C'est pourquoi je crois qu'il est très important de poursuivre cette direction de recherche.

La question est alors de savoir *quelles sont les classes de situations d'apprentissage susceptibles de présenter un phénomène de transition de phase*. Celui-ci est avéré pour la programmation logique inductive. Il l'est désormais pour l'inférence grammaticale, mais pas dans le sens que j'attendais en démarrant ce travail. J'étais tout à fait conscient que, contrairement à la P.L.I., la transition de phase du taux de couverture ne s'accompagnerait pas d'un pic de la complexité du test de couverture, puisque celui-ci, en inférence grammaticale, est une fonction linéaire de la taille de l'automate. Et, contrairement aux chercheurs qui avaient mis en évidence le phénomène de transition de phase en P.L.I., cela ne me semblait pas fondamental. En revanche, j'ai été surpris par la cause de cette transition, qui, dans le cas de l'induction d'automates déterministes, est due à une sorte d'artefact : la procédure de fusion pour déterminisation utilisée.

D'un certain côté, cela fait de ce phénomène une transition qui n'est certainement pas du même type qu'en P.L.I., et donc, peut-être, qu'en physique statistique. D'un autre côté, cette découverte ouvre de nouvelles perspectives. Elle montre en effet l'importance de l'interaction entre espace de recherche et algorithme d'exploration. Dans le cas de l'inférence grammaticale, il a fallu ne pas oublier de se concentrer sur l'étude de l'espace *effectivement exploré* par l'algorithme. Il a fallu de plus tenir compte de toutes les particularités du processus d'exploration : ici le processus de déterminisation des automates non déterministes. Si donc, intrinsèquement, le phénomène de saut observé dans le taux de couverture n'est pas du aux mêmes cause que pour la P.L.I., il n'en reste pas moins un déterminant essentiel de la performance en induction.

Deux types de recherche sont donc à mener. D'une part, il est nécessaire d'essayer de trouver un modèle théorique expliquant la transition de phase en P.L.I.. Cela devrait permettre de prédir les situations d'apprentissage potentiellement concernées par ce phénomène. Pour ce faire, il faut probablement s'inspirer des études réalisées en physique statistique sur les problèmes de satisfaction de contraintes, et en particulier sur leur traduction sous forme de graphe et de propagation de contraintes. D'autre part, il faut poursuivre des études, empiriques et théoriques, sur d'autres langages d'expression des hypothèses, comme par exemples les grammaires stochastiques ou les réseaux bayésiens. Les grammaires *context-free* s'annoncent plus difficiles car il a été montré que dans leur cas la relation *plus-spécifique-que* est indécidable [LB87].

Il faut enfin ne pas oublier que la transition de phase est le fruit d'un jeu entre les caractéristiques de l'espace des hypothèses et celles des exemples. La zone de transition n'a en effet pas lieu pour les mêmes valeurs de paramètres en fonction des caractéristiques des exemples, d'apprentissage et de test.

Sur les remèdes envisageables

Depuis que le phénomène de transition de phase a été découvert à propos de la P.L.I. et qu'il a été réalisé qu'il s'agissait sans doute de la source de l'incapacité des systèmes actuels à traiter des problèmes complexes, des chercheurs se sont attachés à découvrir des remèdes permettant de pallier ce problème.

Puisque ce phénomène est du aux propriétés du langage de représentation des hypothèses, une idée est de le modifier, soit pour le faire complètement disparaître (par exemple en apprenant des automates non déterministes plutôt que des automates déterministes), soit pour changer les

valeurs des paramètres pour lesquelles il se produit, afin de le rendre moins handicapant pour les algorithmes d'apprentissage (en diminuant par exemple la longueur du « plateau » à parcourir avant de trouver le gradient et l'information disponible sur la « falaise »). C'est cette deuxième démarche qui est défendue dans [SZ00], en faisant appel à une approche par redescription et abstraction. Malheureusement, il reste à montrer exactement comment la mettre en œuvre.

Une autre approche, imaginée par Michèle Sebag et moi-même, est de tirer parti du fait que les trajectoires d'apprentissage se terminent nécessairement dans la zone de transition de phase avec des hypothèses approchées pour utiliser une méthode d'ensemble pour combiner les prédictions de ces hypothèses. Nous n'avons malheureusement guère été plus loin. Il faudrait, d'une part, s'assurer que les hypothèses produites sont suffisamment diverses pour qu'une combinaison de leurs « expertises » soit bénéfique. L'emploi de méthodes stochastiques de choix d'opérateurs devraient y conduire. Il faut aussi, d'autre part, trouver une bonne méthode de combinaison des avis de ces hypothèses.

Récemment, j'ai commencé, à l'occasion du stage de Master de Raymond Ros [Ros05], à étudier encore une autre méthode. L'idée est de modifier, en cours d'apprentissage, la topologie de l'espace des hypothèses, c'est-à-dire les connexions entre hypothèses, pour privilégier une structure de réseau à invariance d'échelle (*small-world*) afin d'accélérer l'exploration de l'espace des hypothèses en particulier dans les zones *a priori* intéressantes. Il s'agit d'une technique d'apprentissage incrémental qui cherche à tirer parti des expériences passées pour améliorer l'apprentissage. On suppose donc que les divers problèmes d'apprentissage rencontrés sont issus d'une même distribution. Nous avons commencé une série d'expériences avec des agents simulés dans des labyrinthes et une représentation en logique des prédicats. Ces expériences ont également l'intérêt d'étendre le contexte expérimental utilisé jusqu'ici en PLI, par exemple en ne contraignant pas le nombre d'instances de chaque relation dans les exemples à être constant (variable L).

2.4 Conséquences et conjectures pour une stratégie d'enseignement

Les travaux sur les modèles d'apprentissage par *distributions bienveillantes* (voir [DG97, DGS97]) et par *exemples simples* (voir apprentissage PACS [DDG96, DGS97, PH97, CG98]) affinent l'idée initiale d'échantillon caractéristique et prouvent, par exemple, que les langages réguliers peuvent être appris à partir d'exemples positifs seuls tirés aléatoirement sous réserve que ces exemples vérifient des conditions raisonnables, à savoir, essentiellement, que les exemples simples soient les mieux représentés⁴⁰. Il est clair que la possibilité d'un apprentissage et sa facilité dépendent beaucoup des propriétés relatives des concepts cibles et des exemples d'apprentissage.

Jusqu'à présent, j'ai présenté les choses comme si le phénomène de transition de phase et, plus généralement, le profil de variation du taux de couverture, dépendait uniquement de l'espace des hypothèses. Il est temps de souligner sa dépendance aussi sur les caractéristiques des exemples.

En P.L.I., [BGSS03] montre que la position de la zone de transition de phase dépend du nombre de constantes dans les exemples (L), ainsi que du nombre de littéraux construits sur chaque symbole de prédicat dans les exemples (N) (voir figure 2.20).

40. Techniquement, [DDG96] suppose que les exemples du concept cible sont tirés selon une mesure de Solomonoff-Levin conditionnelle à l'une de ses représentations. Cela signifie en gros que la probabilité d'une chaîne x est (à peu près) égale à $1/K_U(x)$, où $K_U(x)$ est la longueur du plus petit programme capable de produire x sur une machine de Turing U .

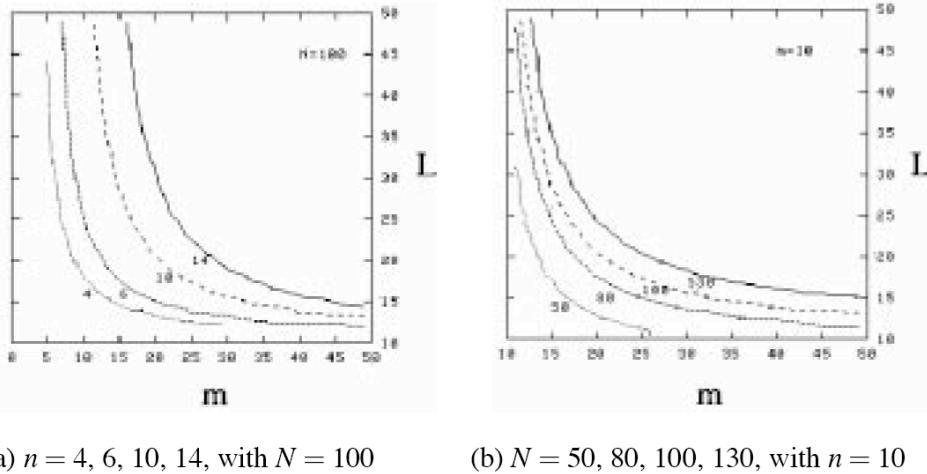


FIGURE 2.20 – Variation de la localisation de la zone de transition de phase (ici, les contours correspondent à $P_{sol} = 0.5$ dans le plan (m, L) pour diverses valeurs de n (à gauche), et de N (à droite). (Tiré de [BGSS03]).

Ces courbes suggèrent que la stratégie d'apprentissage actif ou d'enseignement devrait tenir compte de la stratégie d'apprentissage suivie. En effet, afin de pouvoir guider sa recherche de manière informée, le système a intérêt à rencontrer la zone de transition de phase le plus vite possible. Si le système opère par spécialisation (c'est-à-dire en partant en bas à gauche dans les figures), il vaut donc mieux que les exemples soient « simples » au sens où le nombre L de constantes est faible ainsi que le nombre N de littéraux construits sur chaque symbole de prédicat. Inversement, dans le cas d'un apprentissage par généralisation, les exemples à privilégier, au moins au début de l'apprentissage, sont les plus complexes !

En **inférence grammaticale**, nos expériences (dont l'essentiel est encore non publié) sont encore trop limitées pour être concluantes. Il semble cependant que le trou de couverture est d'autant plus important que la longueur des chaînes test utilisées pour évaluer la couverture est grande par rapport aux chaînes utilisées pour l'apprentissage. Cela suggère que c'est une mauvaise idée d'apprendre avec des chaînes courtes par rapport à celles qui seront rencontrées en généralisation (!). Je ne crois pas que l'on puisse pour le moment en conclure que cela va à l'encontre des études théoriques sur l'apprenabilité à partir d'exemples simples. Mais cela invite à poursuivre l'étude de cette question.

L'étude des variations du taux de couverture peut-elle conduire à de nouvelles idées sur des stratégies d'enseignement? En dehors des observations faites ci-dessus, je pense que le potentiel pour de nouveaux principes est important.

L'étude statistique de l'apprentissage ne tient compte que d'une mesure combinatoire de capacité expressive de l'espace des hypothèses et du critère inductif de minimisation du risque empirique utilisé par les systèmes d'apprentissage. Les résultats portant sur l'enseignabilité⁴¹ s'expriment uniquement en termes de taille d'échantillon d'apprentissage tirés suivant une distribution i.i.d. Dans les cas du modèle par distributions bienveillantes ou du modèle PACS, s'y ajoutent des prescriptions sur la distribution des exemples (tout exemple représentatif doit avoir une probabilité non nulle d'apparaître, les exemples doivent être d'autant mieux repré-

41. Une traduction pas très belle de l'anglais *teachability* (voir par exemple [SM90]).

sentés qu'ils sont « simples »), ceci dans le but d'obtenir un apprentissage efficace (en temps polynomial de certaines caractéristiques du problème).

Les travaux initiés en programmation logique inductive sur les variations du taux de couverture dans l'espace des hypothèses prennent en compte davantage d'information sur le problème d'apprentissage que l'étude statistique. Elle tient compte, d'une part, de la topologie de l'espace des hypothèses induite par les opérateurs d'exploration (e.g. opérateurs de spécialisation ou de généralisation, ou bien domaine de variation élémentaire des poids d'un réseau de neurones), et, d'autre part, de la technique de gradient suivie par la plupart des systèmes inductifs. On peut donc attendre une caractérisation plus fine des conditions d'apprentissage, de même que des prescriptions plus fortes sur le meilleur choix des exemples d'apprentissage. On a vu que les résultats sont encore très limités et préliminaires.

Notre étude, faite à propos de l'inférence grammaticale, correspond à une étape supplémentaire dans la prise en compte des caractéristiques de l'apprentissage. Au lieu d'étudier les variations du taux de couverture sur l'ensemble de l'espace des hypothèses, elle montre l'intérêt qu'il y a à se focaliser sur l'espace de recherche effectif.

Mais, on peut envisager d'aller encore plus loin, selon une idée similaire à celle développée par Rémi Monasson à propos de l'analyse des problèmes K -SAT lorsqu'il a proposé d'étudier la trajectoire des résolutions dans l'espace des problèmes $(2+p)$ -SAT. L'idée est alors d'examiner si le système ne traverse pas des sous-espaces de problèmes, en cours d'apprentissage, dont les caractéristiques varient (par exemple, le profil du taux de couverture pourrait varier pour les sous-espaces traversés). Il pourrait alors y avoir avantage à adapter dynamiquement soit les heuristiques d'apprentissage, soit les exemples d'apprentissage fournis.

Je trouve cette idée de prise en compte de plus en plus précise des conditions de l'apprentissage très séduisante, mais tout reste encore essentiellement à faire.

Une leçon de ce travail est que l'information que peut apporter un exemple dépend de l'algorithme d'apprentissage, et en particulier du gradient du taux de couverture autour de l'hypothèse courante. Si ce gradient est quasi nul, alors l'exemple n'apporte aucune information. Faciliter l'apprentissage signifie trouver une bonne interaction entre l'espace des hypothèses, la méthode d'exploration et les exemples d'apprentissage sélectionnés.

Dans ce cadre général, une idée peut aussi être de recoder les données dans une base de caractéristiques qui optimise l'information qui peut être exploitée à partir des exemples d'apprentissage. C'est ce que tente la méthode FISICA présentée ci-dessous.

2.5 Un nouveau codage pour redécrire les données : FISICA

La méthode FISICA est issue d'un projet sur la reconnaissance de scènes naturelles en vision. Nos yeux sont des capteurs qui nous abreuvant de données sur le monde extérieur. Grâce à notre système visuel, nous devons devenir conscient des objets et des êtres vivants qui nous entourent, et nous sommes capables de nous représenter leurs formes et leurs propriétés dans notre cerveau. La vision par ordinateur cherche à reproduire cette capacité dans les machines.

La vision est une tâche très difficile. Les images d'une tête humaine et d'un melon peuvent être très similaires sous certains éclairages, tandis que deux images de la même tête dans des conditions d'éclairage différentes peuvent être très dissimilaires. Pourtant, nous n'avons généralement aucun problème pour les identifier correctement. L'image d'un arbre est composée d'une structure complexe d'objets lumineux et sombres, verts et bruns, et cependant nous sommes capables de la percevoir comme un seul objet ou comme une collection de branches et de feuilles. Ces exemples montrent clairement qu'une métrique naïve dans l'espace de très haute dimension

des images ne peut pas être très informative pour extraire les informations pertinentes et les concepts des images. Différents objets peuvent produire la même image et, vice-versa, le même objet peut engendrer des images très différentes en fonction de la prise de vue et des conditions d'éclairage. Il faut donc introduire des connaissances sur le monde, sous la forme de régularités à attendre et de biais de recherche par exemple, pour rendre possible l'extraction de l'information qu'une image peut fournir. De nombreuses illusions d'optique bien documentées (voir par exemple [PlS03]) montrent les effets de ces attentes et de ces biais à l'œuvre dans notre système visuel.

Pour comprendre comment le système visuel est capable de telles performances, il faut non seulement étudier la question du codage de l'information sensorielle, mais intégrer cette question avec celle de son traitement et celle de la représentation de propriétés plus abstraites des stimuli.

Mon intérêt pour les problèmes de vision datait de longue date puisque j'avais déjà fait des recherches sur la détection d'objets en mouvement par le calcul de flot optique en 1987 [CC87], une méthode tout à fait nouvelle à l'époque. Avec Jean-Pierre Cassou, à l'E.N.S., nous avions eu d'excellents résultats, mais bien loin du temps réel exigé par les applications (nous utilisions une méthode de recuit simulé!). En 2002, à l'occasion d'un projet BQR mené par Philippe Tarroux du Limsi à Orsay⁴², je retrouvais ce domaine de recherche sous une autre perspective. À l'inverse de mes premières incursions sur ce sujet, il s'agissait cette fois, *a priori*, d'étudier une solution plausible d'un point de vue cognitif et même neurobiologique. Je ne sais pas si ce que nous avons finalement trouvé répond à ce critère, mais ce travail nous a conduit à quelques idées originales et à des résultats jugés prometteurs par d'autres chercheurs [JCS⁺⁰³, CSM04, JS04, RLJS05].

2.5.1 La reconnaissance de scènes naturelles

Les scènes naturelles constituent une minuscule fraction de l'ensemble des images possibles. Elles ont été dédaignées dans les premiers temps de la robotique car les robots, engins fragiles, étaient consignés à l'intérieur de bâtiments, donc à des environnements à la fois pauvres et remplis d'arêtes et de surfaces planes. Avec une nouvelle génération de robots plus robustes et plus autonomes, de nouvelles applications en milieu naturel deviennent envisageables, or il apparaît vite que les systèmes de vision développés ne sont pas performants dans ces environnements. En revanche, le cortex visuel des animaux a évolué pour fonctionner précisément dans ces milieux. Pourquoi alors ne pas essaer de s'inspirer, pour une fois, des solutions développées par la nature et qui semblent incroyablement performantes ?

L'un des objectifs des recherches est ainsi de caractériser la structure des images naturelles, et de chercher des stratégies de codage efficaces pour ce type de structure et qui semblent étayées par nos connaissances sur le cortex visuel, en particulier l'aire « V1 » (voir [OF96]).

L'hypothèse de base est que les régularités présentes dans le monde sont traduites par des dépendances statistiques complexes, et des redondances. Un objectif des systèmes de vision est d'extraire ces dépendances statistiques de telle manière que les images puissent être expliquées en termes d'une collection d'événements indépendants (**1**). Cette hypothèse de décomposabilité est évidemment très forte, quoiqu'en disent par exemple des chercheurs comme Bruno Olshausen ou David Field, mais elle permet des traitements simples et elle semble compatible avec nos connaissances sur le cortex visuel.

Par ailleurs, (**2**) les régularités présentes dans les images naturelles se trahissent par des corrélations statistiques particulières, avec, spécialement, des corrélations non linéaires (e.g. des corrélations de trois régions ou plus). L'idée est donc de chercher un codage des images en terme

42. Auquel participaient Jean-Sylvain Liénard et Nathalie Denquive (doctorante) du Limsi et Michèle Sebag et moi-même du LRI.

de combinaisons linéaires (selon (1)) de fonctions de base représentant ces corrélations de haut degré (selon (2)).

De plus, les études de l'aire V1 du cortex visuel, qui font apparaître qu'une scène visuelle est codée par un petit nombre de neurones à l'intérieur d'une collection qui en comprend plus d'un milliard, suggèrent que le codage utilisé est économique ou « clairsemé » (*sparse-coding*) dans un système de fonctions de base sur-complet (*over-complete*) c'est-à-dire dont la dimension est supérieure à celle des formes codées.

Le problème est alors de trouver une base de telles fonctions qui soient adaptées au codage des images de scènes naturelles et telles qu'elles soient indépendantes et permettent un codage clairsemé des images attendues.

Le problème de recherche d'un codage clairsemé

On peut décrire le problème de recherche du codage linéaire d'une image $I(x, y)$ comme celui d'une base de fonctions $\phi_i(x, y)$ satisfaisant :

$$I(x, y) = \sum_i a_i \phi_i(x, y) \quad (2.4)$$

Le but est de trouver une base qui soit à la fois une base *complète* (permettant la description de l'espace d'entrée) et *clairsemée* (permettant la représentation des images avec peu de coefficients $a_i \neq 0$). En d'autres termes, la distribution de probabilité sur chacun des coefficients devrait être très « piquée » autour de 0, avec des queues épaisses. Une telle distribution a une faible entropie et permet également de réduire les dépendances statistiques entre les fonctions de base.

Ce problème peut être formulé comme un problème d'optimisation avec la fonction de coût :

$$E(a, \phi) = \sum_{x,y} [I(x, y) - \sum_i a_i \phi_i(x, y)]^2 + \beta \sum_i S\left(\frac{a_i}{\sigma_i}\right) \quad (2.5)$$

où $\sigma_i^2 = \langle a_i^2 \rangle$. Le premier terme mesure l'adéquation du code avec l'image, suivant une mesure d'écart quadratique, tandis que le second terme tend à favoriser les codes clairsemés, en fonction de la fonction de coût S .

Deux démarches existent. Soit utiliser un code linéaire choisi *a priori*, comme les transformées de Fourier ou les représentations en ondelettes⁴³. Soit utiliser des représentations dépendantes des données, qui sont ajustées automatiquement aux statistiques des entrées. De telles représentations sont apprises directement à partir des données en optimisant des mesures qui quantifient leurs propriétés désirables. Cette classe de méthodes inclue l'*analyse en composantes principales* (ACP), l'*analyse en composantes indépendantes* (ICA) [Hyv99, HO00, HKO01] et la *factorisation non-négative en matrices* (NMF) [LS99]. Si l'analyse ACP est trop limitée pour s'appliquer au domaine de la vision, en revanche des résultats intéressants ont été obtenus par ICA et par NMF.

Analyse en composantes principales L'idée de base de l'analyse en composantes principales est de trouver les composantes s_1, s_2, \dots, s_n qui rendent compte au mieux de la variance du signal par n composantes transformées linéairement. La méthode consiste à chercher les n vecteurs propres correspondant aux n plus grandes valeurs propres de la matrice de covariance.

43. Elles ont à la fois d'intéressantes propriétés mathématiques et une certaine plausibilité biologique, qui en font une représentation de choix dans les travaux sur le système visuel.

Lorsque $n < m$ (m : dimension de l'espace d'entrée), on peut prouver que la représentation en composantes principales effectue une réduction de dimension linéaire optimale au sens des moindres carrés. Mais cette méthode suppose que le signal est fortement (totalement gaussien) car seule cette partie du signal est prise en compte.

Analyse en composantes indépendantes Les méthodes de plus haut degré (que 2) cherchent l'information qui n'est pas présente dans la matrice de covariance. Cela correspond à des familles de distribution de densité plus générales. Parmi celles ci (qui incluent les méthodes de *projection pursuit* et *blind deconvolution*) l'analyse en composantes indépendantes cherche une décomposition de la fonction jointe de densité de distribution :

$$f(y_1, \dots, y_m) = f_1(y_1)f_2(y_2) \dots f_m(y_m) \quad (2.6)$$

où $f_i(y_i)$ représente la densité marginale de y_i . Afin de distinguer cette forme d'indépendance d'autres concepts d'indépendance, par exemple la dépendance linéaire, cette propriété est parfois appelée indépendance statistique.

L'indépendance doit être distinguée de la *non corrélation*, qui, elle, signifie :

$$E(y_i y_j) - E(y_i)E(y_j) = 0 \quad \forall i \neq j \quad (2.7)$$

L'*indépendance* correspond en général à une condition beaucoup plus forte : si les y_i sont indépendants alors :

$$E\{g_1(y_i)g_2(y_j)\} - E\{g_1(y_i)\}E\{g_2(y_j)\} = 0 \quad \forall i \neq j \quad (2.8)$$

et pour toute paire de fonctions (g_1, g_2) ⁴⁴.

Définition 2.21 (Analyse en Composantes Indépendantes (ICA))

L'analyse en composantes indépendantes d'un signal (vecteur) \mathbf{x} consiste à trouver une transformation linéaire $\mathbf{s} = \mathbf{W}\mathbf{x}$ telle que les composantes s_i soient aussi indépendantes que possible, au sens de la maximisation d'une fonction $F(s_1, \dots, s_m)$ qui mesure l'indépendance.

Une méthode d'analyse en composantes indépendantes peut donc être vue comme la somme :

$$\text{Méthode d'ICA} = \text{Fonction objectif} + \text{Algorithme d'optimisation} \quad (2.9)$$

Il faut évidemment traduire le problème et le domaine sous forme d'un modèle adéquat et d'une fonction objectif ayant les bonnes propriétés statistiques. Il faut ensuite que l'algorithme d'optimisation soit adapté au problème d'optimisation posé. Nous ne rentrerons pas davantage dans les détails ici (voir [Hyv99, HO00, HKO01]).

L'analyse en composantes indépendantes a été appliquée aux problèmes de séparation de sources (identification de m sources produisant par mélange linéaire un signal mesuré en m points) et d'extraction de composantes d'un signal. L'analyse d'images entre dans ce dernier cadre. L'idée est alors d'extraire un code, c'est-à-dire un catalogue, constitué d'« images de base », dans lequel toute image du domaine considéré (correspondant à une certaine famille de distributions de probabilité) peut être représentée fidèlement.

Étant donnée la complexité des calculs et la taille limitée souhaitée pour le catalogue d'images de base, les expériences réalisées ne travaillent pas directement sur les images, mais sur des « imagettes » (par exemple 8×8 ou 12×12). Les figures 2.21 et 2.22 montrent respectivement la décomposition linéaire d'une imagette 12×12 et une base d'imagettes obtenues par ICA à partir d'images en couleur (Transparents fournis par Patrick Hoyer et Aapo Hyvärinen).

44. Les fonctions doivent être mesurables.

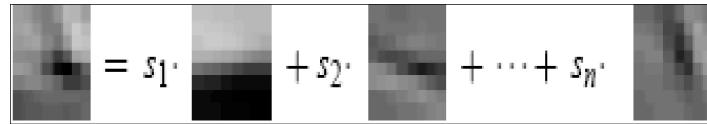
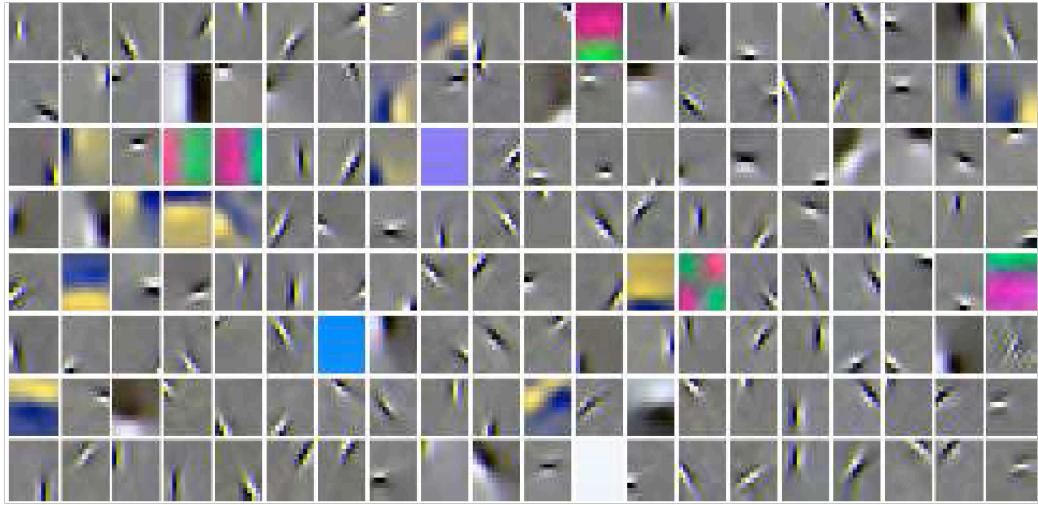


FIGURE 2.21 – Principe de décomposition linéaire d'une imagette.


 FIGURE 2.22 – Base d'imagettes 12×12 de base obtenue par Patrick Hoyer et Aapo Hyvriinen sur une base d'images en couleur (signal RGB).

Les résultats obtenus de cette manière sont intéressants, car ils sont en particulier confrontables avec les champs perceptifs des neurones tels qu'on peut essayer de les interpréter. D'un autre côté, la limitation au traitement d'imagettes oblitère peut-être des corrélations spatiales à plus grande échelle importantes pour le domaine. C'est pourquoi, nous avons décidé d'essayer une attaque directe sur ce problème, en s'inspirant de l'analyse ICA (analyse en composantes indépendantes) mais en employant la technique récemment développée en fouille de données des FIS (*Frequent Item Sets*). C'est la méthode FISICA [Jou02, JCS⁺03, CSM04].

2.5.2 Les bases de la méthode FISICA

Si l'approche par l'analyse en composantes indépendantes d'un codage clairsemé est impossible, une approche directe est-elle envisageable ?

En faisant l'hypothèse que les données résultent d'une somme de formes latentes, et que ces formes pour être intéressantes doivent figurer suffisamment souvent dans les données, le rapprochement avec la technique de recherche de motifs fréquents s'impose. Les données étant décrites par un ensemble de descripteurs (attributs-valeur), on cherche les conjonctions d'attributs-valeur présentes dans un certain pourcentage des données. Grâce à certaines contraintes, on peut guider la recherche de tels motifs de manière à favoriser la découverte d'un ensemble de primitives (les motifs) permettant un codage clairsemé des données.

À ces contraintes peuvent s'ajouter des critères supplémentaires liés au domaine d'application permettant de sélectionner les motifs fréquents les plus intéressants. Par exemple, dans le domaine de l'analyse d'images, on pourra favoriser la recherche de fonctions de base ou motifs

correspondant à des régions connexes, ou à des lignes (voir plus bas, les résultats expérimentaux).

Les *fonctions de base* cherchées sont des conjonctions d'attributs-valeur, on parlera aussi d'atomes. Dans le cas de l'analyse d'images, il s'agira de collections de pixels, chacun de ceux-ci étant associé à un niveau de gris donné. Une fonction de base est ainsi une fonction booléenne prenant la valeur **vrai** si la collection d'atomes (pixel = valeur) correspondante est satisfaite dans l'image étudiée. On dira que le *support* d'une fonction de base est de $\varepsilon\%$ si cette fonction prend la valeur **vrai** dans $\varepsilon\%$ des images testées. Réciproquement, on appellera *code* d'une forme d'entrée (ici une image), l'ensemble des fonctions de base vérifiant cette forme.

Dans ce nouveau cadre, les propriétés désirées du système de codage se traduisent comme suit :

1. **Représentativité.** Chaque fonction de base a un support supérieur à $\varepsilon\%$. Elle est donc suffisamment représentée dans la base d'exemples pour être utile.
2. **Parcimonie.** Peu de fonctions de base sont vérifiées par un exemple (e.g. image). On vérifie ainsi l'une des propriétés du codage clairsemé.
3. **Suffisance.** Tout exemple rend vrai un nombre minimal de fonctions de base.
4. **Orthogonalité.** Pour chaque paire de fonctions de base, l'intersection des exemples qui rendent **vrai** l'une et l'autre est réduite. Les exemples sont donc décrits par des codes différents.

Nous avons adapté la méthode de recherche de motifs fréquents dans une base de données pour chercher un codage tendant à vérifier les propriétés ci-dessus.

La recherche de motifs fréquents de taux de couverture faible dans des données décrites par de nombreux attributs ne peut s'effectuer sans précautions. C'est pourquoi nous présentons rapidement l'algorithme développé à cet effet.

Échec d'une approche naïve

Une approche consiste à utiliser un algorithme existant de recherche de motifs fréquents, tel qu'**APRIORI** [AS94], pour chercher tous les motifs ayant un certain taux de couverture, puis ensuite à sélectionner parmi eux ceux qui satisfont les contraintes soulignées dans la section 2.5.2. Mais cette approche se révèle impraticable pour la reconnaissance d'images car le nombre de motifs fréquents croît exponentiellement en fonction de la taille des motifs comme le montre la table ci-dessous pour des images de taille 32x32 en 64 niveaux de gris (où m signifie mille, M million et MM milliard) :

1	2	3	4	5	6
2m	110m	3,8M	80M	1,15MM	12,5MM

Il faut donc renoncer à une méthode de recherche exhaustive des motifs.

Approche randomisée

Nous avons développé une méthode de construction incrémentale de motifs fréquents par ajouts successifs d'atomes (ici, de pixels d'un certain niveau de gris) en les sélectionnant à chaque pas afin que le motif en construction satisfasse aux critères désirés. L'exploration des motifs fréquents est donc maintenant stochastique, guidée mais non exhaustive. Des essais successifs peuvent ainsi produire des bases de motifs différentes.

Algorithme 2.1 Recherche itérative et stochastique de motifs fréquents.

Paramètres : taux de couverture $\varepsilon\%$. Nombre de motifs cherchés = N .

Nombre de motifs trouvés = $n \leftarrow 0$.

while $n \leq N$ **do**

Choix dans un exemple x_i encore peu couvert, d'un premier atome a_0 présent dans au moins $\varepsilon\%$ des exemples.

$motif \leftarrow a_0$

while Taux de couverture de $motif > \varepsilon\%$ **do**

Tirer au hasard un atome a de x_i couvrant au moins $\varepsilon\%$ des exemples et peu utilisé dans les motifs existants et satisfaisant des contraintes additionnelles sémantiques (voir section 2.5.3).

if $motif + a$ couvre au moins $\varepsilon\%$ des exemples **then**

$motif \leftarrow motif + a$

end if

end while

end while

Exploitation pour l'apprentissage supervisé

Une fois N fonctions de base trouvées sur un ensemble d'apprentissage, chaque exemple est recodé, devenant un vecteur de N booléens prenant la valeur **vrai** ou **faux** selon que la fonction de base correspondante couvre l'exemple ou non.⁴⁵

Dans le nouvel espace d'exemples ainsi construit, il est possible d'utiliser n'importe quelle méthode d'apprentissage supervisé. Dans les expériences rapportées ici, nous avons utilisé une méthode de classification par plus proche voisin. Les exemples d'apprentissage utilisés pour la recherche de fonctions de base sont également employés comme exemples étiquetés servant à la classification des exemples testés.

2.5.3 Application à la reconnaissance d'images

La méthode développée a été testée sur des tâches de classification d'images de scènes naturelles et de chiffres manuscrits. Elle implique deux phases : d'abord une étape de détermination d'une base de fonctions de base permettant de redécrire les données, ensuite l'emploi du système de codage ainsi obtenu pour classer de nouvelles formes.

Dans le cas de la reconnaissance de scènes naturelles, le problème consiste à apprendre à reconnaître des images de scènes naturelles classées en 12 catégories (voir la figure 2.23). Ces images proviennent de la base COREL (http://www.corel.com/gallery_line/). Les images sont redécrivées par 128×128 pixels en 128 niveaux de gris. Pour ces expériences, la base utilisée comportait 1082 images réparties également entre les 12 classes. Nous considérerons dans la suite que chaque pixel est un attribut pouvant prendre une valeur parmi 128. La dimension de l'espace d'entrée est donc dans ce cas de $32768 = 128 \times 128$.

Pour l'application étudiée, nous avons fixé à 1000 le nombre de fonctions de base recherchées. Plusieurs bases ont été obtenues en faisant varier les paramètres suivants :

- *Taux de couverture : 1%, 2%, 5% et 10%*

45. Nous avons aussi utilisé une formule d'appariement plus souple utilisant une fonction sigmoïde à valeur dans $[0, 1]$ qui tient compte du nombre d'atomes de la fonction de base qui couvrent l'exemple, et donnant un recodage dans $[0, 1]^N$ au lieu de $\{0, 1\}^N$



FIGURE 2.23 – Échantillon d’images utilisées dans cette étude. Noms des classes : avions (1), plats (2), Utah (3), minéraux (4), chiens (5), poissons (6), verres (7), papillons (8), porcelaines (9), figurines (10), voitures (11), fleurs (12). (Cette figure est reprise de [DT03]).

— Critère sémantique additionnel. Nous avons introduit des contraintes supplémentaires sur la construction des fonctions de base afin de tester des équivalences possibles avec d’autres types de codages classiques en traitement d’images. Quatre conditions ont été testées :

1. Aucune contrainte.
2. Les fonctions de base doivent correspondre à des *régions connexes* sur l’image : un nouveau pixel n’est ajouté à la fonction de base courante que si il est contigu à un pixel déjà sélectionné.
3. Les fonctions de base doivent correspondre à des *lignes* de l’image (régions de dimension 1). L’idée ici est de voir si l’on peut forcer le système de codage à retenir des contours dans l’image.
4. Les fonctions de base doivent correspondre à des *lignes raisonnables* de l’image, c’est-à-dire plus contraintes dans les changements de directions possibles. Cette contrainte a été imposée lorsqu’il s’est avéré que la précédente produisait des « vermineaux » remplissant des régions et non pas des lignes.

Environ la moitié des images de la base initiale de 1082 images, soit 500, ont été utilisées pour le calcul des fonctions de base. (Note : notre algorithme calcule une base de 1000 motifs en quelques minutes sur un PC équipé d’un Pentium II à 266 Mhz et 384 Mo de RAM). Les figures 2.24 et 2.25 illustrent le type de fonctions de base obtenues pour certaines conditions expérimentales. Des résultats plus complets sont disponibles sur le site <http://www.lri.fr/~antoine/Research/FISICA/egc-03.html>.

L’histogramme présenté dans la figure 2.26 (à gauche) permet de contrôler l’orthogonalité des fonctions de base obtenues. Ces fonctions de base sont orthogonales lorsqu’elles sont rarement vérifiées par les mêmes images. La figure montre que les différentes bases obtenues pour des conditions différentes peuvent effectivement être considérées comme orthogonales. Inversement, l’histogramme de la figure 2.26 (à droite) indique le nombre de fonctions de base qui sont vérifiées par les images. On constate que ce nombre varie autour d’une dizaine, ce qui traduit bien que le codage obtenu est clairsemé.

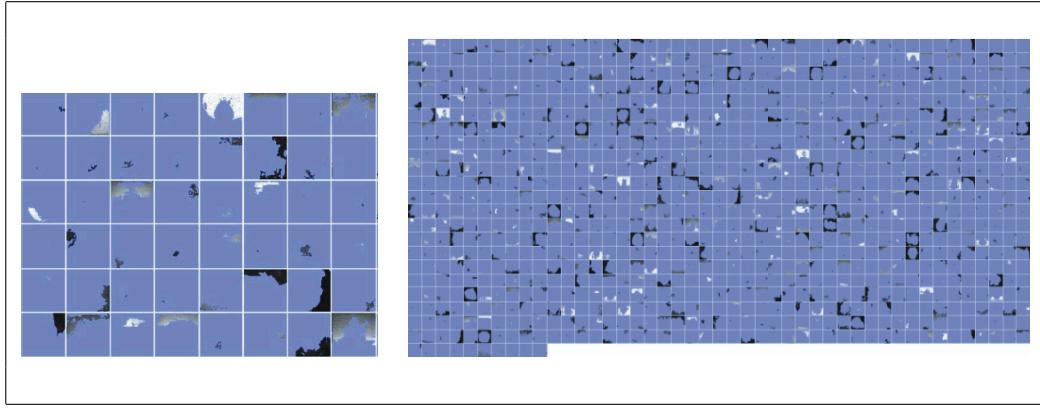


FIGURE 2.24 – (À gauche) Le détail de quelques unes des fonctions de base obtenues sur des images 64×64 en 16 niveaux de gris avec un taux de couverture de 1%, et en cherchant des régions connexes de l'image. Dans les images accessibles sur le site internet, le fond bleu correspond aux zones qui ne font pas partie des fonctions de base, tandis que les fonctions de base sont figurées par des pixels de niveaux de gris variés. Ici, le gris moyen correspond au fond. (À droite) Figure une base de 1000 fonctions de base.

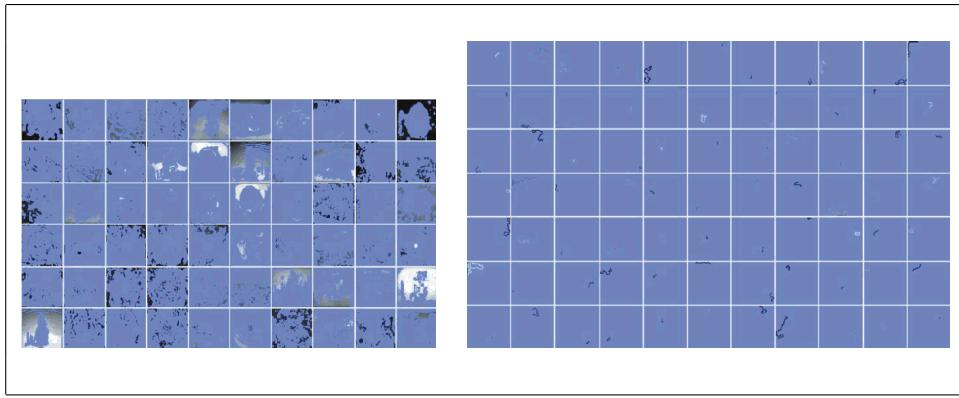


FIGURE 2.25 – (À gauche) Des exemples de fonctions de base obtenues avec un taux de couverture de 1%, et sans contrainte. (À droite) Des exemples de fonctions de base obtenues avec un taux de couverture de 1%, et sous contrainte de linéarité raisonnable. L'examen des motifs trouvés montre qu'ils ne correspondent pas à des contours des images de la base d'exemples.

Les résultats en classification

Les performances en classification ont été calculées sur les 582 images non utilisées pour déterminer la base des fonctions de base. Toutes les images sont recodées en utilisant les fonctions de base obtenues, et s'expriment donc sous la forme d'un vecteur de 1000 valeurs booléennes (en fait, dans certaines expériences, cette valeur booléenne était remplacée par une mesure plus continue d'appariement de l'image avec une fonction de base). Les 500 images employées pour la détermination des fonctions de base sont également utilisées comme base d'exemples étiquetées. Les images à classer sont alors étiquetées en utilisant une méthode de plus proche voisin. Dans les expériences rapportées ici, la distance utilisée est la distance L_1 .

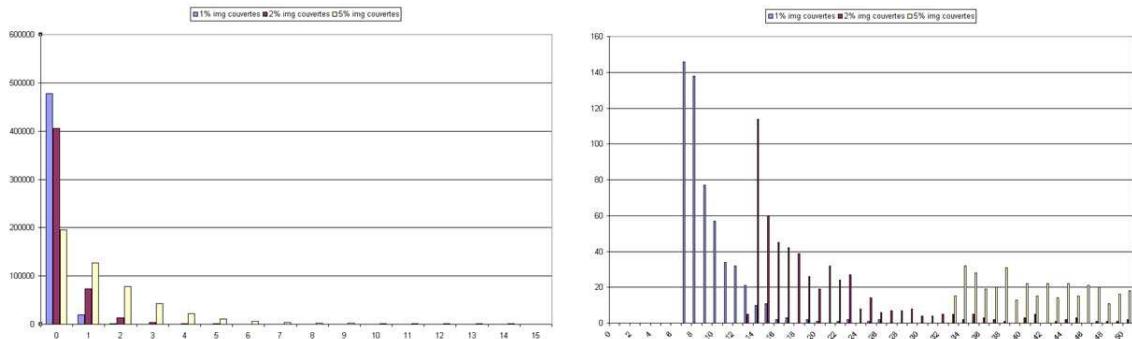


FIGURE 2.26 – À gauche : Histogramme représentant le nombre y de couples de fonctions (en ordonnée) ayant x images en commun (en abscisse). Ces fonctions ont été calculées à partir d’images de taille 64×64 en 16 niveaux de gris et sous la contrainte de connexité. Les résultats sont présentés pour des fonctions de base de taux de couverture 1%, 2% et 5%. Les fonctions de base obtenues pour $\varepsilon = 1\%$ sont les plus orthogonales entre elles. À droite : Histogramme représentant le nombre y d’images (en ordonnée) activant x motifs (fonctions de base) en abscisse. Ces fonctions ont été calculées à partir d’images de taille 64×64 en 16 niveaux de gris et sous la contrainte de connexité. Les résultats sont présentés pour des fonctions de base de taux de couverture 1%, 2% et 5%. Plus le taux de couverture est élevé, plus chaque exemple est couvert en moyenne par un nombre élevé de fonctions. On peut ainsi régler la parcimonie de la représentation et donc son caractère clairsemé.

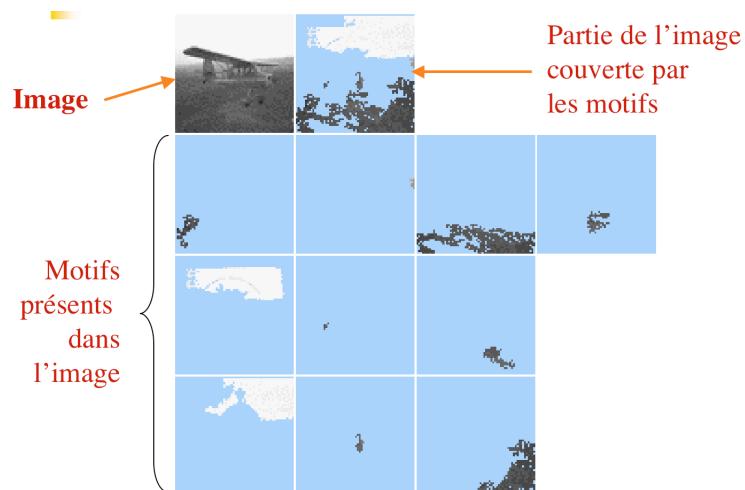


FIGURE 2.27 – Représentation d’une image (ici un avion) à l’aide des motifs fréquents sélectionnés. Il est clair que le codage obtenu ici ne permet pas de reconstruire l’image d’origine et ne vérifie donc pas la propriété d’approximation des méthodes classiques de codage.

2.5. Un nouveau codage pour redécrire les données : FISICA

	Av	Pl	Ut	Mi	Ch	Po	Ve	Pa	Por	Fi	Vo	Fl
Avi	67%	2%	-	-	2%	2%	10%	10%	4%	2%	-	-
Pla	-	21%	-	2%	7%	19%	10%	12%	5%	-	19%	5%
Uta	17%	-	33%	-	7%	-	-	3%	10%	10%	13%	7%
Min	-	-	-	100%	-	-	-	-	-	-	-	-
Chi	26%	5%	7%	-	14%	9%	12%	9%	5%	2%	12%	-
Poi	5%	13%	3%	8%	-	13%	18%	21%	-	3%	10%	8%
Ver	2%	2%	-	-	10%	7%	43%	-	21%	5%	7%	2%
Pap	6%	6%	-	-	2%	14%	14%	35%	6%	-	12%	4%
Por	2%	2%	-	-	-	2%	-	12%	70%	10%	-	2%
Fig	-	-	-	-	-	-	6%	-	24%	70%	-	-
Voi	21%	6%	-	-	4%	4%	8%	4%	4%	29%	19%	-
Fle	2%	9%	-	-	-	9%	21%	14%	-	-	16%	28%

TABLE 2.2 – Matrice de confusion obtenue avec des fonctions de base de taux de couverture $\epsilon = 5\%$ sous contrainte de connexité, en utilisant une formule d'appariement continu entre les images et les fonctions de base.

La table 2.5.3 fournit les résultats obtenus avec une base de 1000 fonctions de base de taux de couverture de 5% soumis à la contrainte de connexité. Quoique les différents nombres puissent varier sensiblement, on observe en général que les résultats obtenus dans une grande variété de conditions sont assez similaires en moyenne. Ils sont très sensiblement supérieurs à ceux rapportés dans [DT03] utilisant un réseau de neurones à bases radiales, ou à ceux que nous avons obtenus avec des Séparateurs à Vastes Marges (SVM) sur un codage à base d'ondelettes de Gabor⁴⁶. Des résultats plus complets sont accessibles à l'url :

<http://www.lri.fr/~antoine/Research/FISICA/core1.html>.

2.5.4 Bilan et perspectives

FISICA est donc une méthode de précodage des données par l'utilisation de fonctions de base correspondant à des motifs fréquents (de faible taux de couverture) trouvés dans les données d'apprentissage. Ce codage est inhabituel : il n'est pas défini *a priori*, mais au contraire dépend des données ; il ne permet pas de reconstruire les données codées et n'a donc pas de capacité d'approximation ; finalement, l'orthogonalité des fonctions de base est définie par rapport aux données. Par ailleurs, il est singulier dans la mesure où il n'effectue aucune « abstraction », mais au contraire semble lié aux coïncidences de bas-niveau dans les données⁴⁷.

Ce codage semblait se révéler très performant sur la tâche réputée difficile de reconnaissance de scènes naturelles. Et d'autres résultats, comparables à ceux de l'état de l'art, furent également obtenus⁴⁸ dans plusieurs domaines. En revanche, dans une tâche de reconnaissance des chiffres manuscrits⁴⁹, les résultats mesurés se révélèrent médiocres (de l'ordre de 80% de reconnaissance au lieu d'environ 90% pour l'état de l'art sur ces données, et de 96% avec une distance L_1 et un vote majoritaire sur les 10 plus proches voisins dans l'espace des pixels!).

46. Un travail non publié réalisé avec Olivier Bousquet au printemps 2001. Nous utilisions autant de classificateurs « un contre tous » que de classes. Les descripteurs consistaient en 20 filtres de Gabor (4 orientations \times 5 fréquences).

47. Paradoxalement, le codage effectué dans FISICA est proche de celui proposé dans les études sur l'abstraction par Nicolas Bredèche, Jean-Gabriel Ganascia et Jean-Daniel Zucker en particulier [Zuc96, ZG96, ZBS02, Bre02, BSZar], une différence étant que leurs travaux incorporent la possibilité de jouer sur la granularité de la représentation.

48. Avec la collaboration de Jérémie Mary en thèse avec moi, de 2002 à 2005.

49. Voir les données à : <http://ftp.ics.uci.edu/pub/machine-learning-databases/optdigits/>

Nous étions donc face à un certain nombre de questions. D'abord, est-ce que cette méthode de recodage permettait réellement une amélioration des performances en classification ? Et dans certains cas⁵⁰ ? Ensuite, quelles étaient les bonnes ou mauvaises propriétés théoriques de cette approche ?

Pour le moment, le « jury est dehors » (*the jury is out*) comme diraient les anglophones. Ces questions restent à élucider.

Du point de vue expérimental, de nouvelles expériences, réalisées en 2004, en reconnaissance de scènes visuelles utilisant une approche par plus proches voisins dans l'espace des images, sans précodage, donne des résultats analogues à ceux que nous avons obtenus avec codage par FISICA. Cette contre-expérience, que nous aurions du faire plus tôt, semblait marquer la fin de grands espoirs dans une méthode originale. Cependant, au même moment, nous étions contactés par Jean-Michel Jolion de l'INSA à Lyon et spécialiste reconnu du traitement d'images. Intéressé (et intrigué !) par notre approche, il l'avait essayée sur de l'indexation et de la reconnaissance automatique de séquences d'images de télévision. Ses résultats dépassaient de loin ceux de toutes les autres méthodes connues. Mais, au lieu d'utiliser comme nous la méthode directement sur les pixels, il effectuait d'abord un prétraitement hiérarchique des images en les codant par les « points d'intérêt »⁵¹, et ce sont ces points d'intérêt qui étaient ensuite utilisés en entrée de FISICA (voir [JCS⁺03, CSM04, JS04, RLJS05] pour des indications sur cette approche).

Ces résultats restent à confirmer et à expliquer. C'est pourquoi, il était crucial d'étudier les propriétés théoriques du recodage effectué par FISICA.

2.5.5 Analyse théorique

Pour résumer, la méthode FISICA consiste à calculer un codage Φ à partir des données d'apprentissage initialement décrites dans l'espace \mathcal{X} , puis à utiliser ce codage pour redécrire les données, d'apprentissage et de test, dans un nouvel espace que nous noterons $\Phi(\mathcal{X})$. Un algorithme de classification est alors utilisé sur les données décrites dans $\Phi(\mathcal{X})$ (voir figure 2.28). Dans nos expériences, nous avons essentiellement utilisé des méthodes de classification par plus proches voisins, mais toute autre méthode de classification supervisée pourrait *a priori* être utilisée.

Il est intéressant de noter d'emblée certaines caractéristiques de ce codage :

- Le codage dépend des données (il reflète la distribution de probabilité des images).
- Il ne permet pas de reconstruire les données. Il n'y a donc pas, comme dans d'autres méthodes (infomax⁵² par exemple [BS95]), la recherche de garder l'information permettant d'approcher les formes d'entrée.
- Les fonctions de base ou motifs sont orthogonaux entre eux, mais au sens d'une orthogonalité définie par rapport aux données !
- L'espace de redescription $\Phi(\mathcal{X})$ comporte $\binom{m}{p}$ points, avec m étant le nombre de fonctions de base (par exemple 1000 dans nos expériences) et p étant le nombre maximal de fonctions de base utilisées pour coder les données (de l'ordre de quelques dizaines).
- Tous les points correspondant à des exemples sont orthogonaux entre eux dans cet espace⁵³.

50. En application du *no-free-lunch-theorem*, il existe nécessairement des cas où la méthode marche mieux que d'autres méthodes.

51. Sortes de points saillants de l'image qui sont aussi ceux sur lesquels se portent naturellement les saccades oculaires de notre système visuel.

52. Qui consiste à maximiser l'entropie jointe entre l'entrée et la sortie.

53. En fait, on a ici un code correcteur idéal. Or Tom Dietterich a montré qu'utilisés pour le codage de sortie des Perceptrons Multi-couches, ils conduisent à des résultats améliorés [DB95].

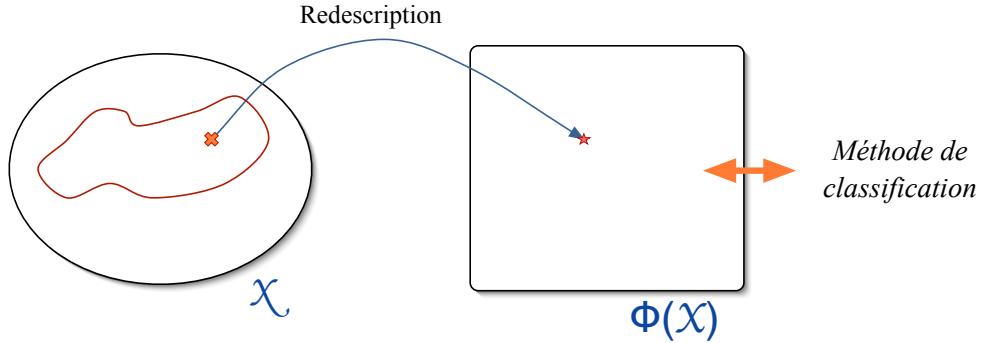


FIGURE 2.28 – La méthode FISICA consiste à redécrire les données (ici prises dans le sous-espace des scènes naturelles) de l'espace des images \mathcal{X} dans un espace de redescription $\Phi(\mathcal{X})$ sur lequel pourra être utilisé une méthode de classification.

La question est pourquoi la redescription des données dans un tel codage devrait conduire à de bons résultats en classification ?

Ce codage, dans un espace de redescription de très grande dimension, fait inévitablement penser au codage induit par les méthodes à base de fonctions noyau (*kernel machines*, voir [CM02, SBS99, SBSS00, SS02, STC04]). Celui-ci aussi est construit à partir des données, et traduit celles-ci, par l'intermédiaire des fonctions noyaux choisies, dans une base de corrélations statistiques. Le pouvoir de généralisation est expliqué par le fait que la capacité effective de l'espace de redescription est limitée, et ce d'autant plus que la marge entre les classes est grande. En automne 2003, lorsque nous ne doutions pas que notre méthode de redescription avait des vertus particulières, Michèle Sebag et moi étions donc confrontés à la question : est-ce qu'une raison du même type que celle des méthodes à noyaux peut s'appliquer dans notre cas, ou bien existe-t-il une autre explication aux performances obtenues ?

Les méthodes à base de fonctions noyau mettent en avant l'importance de la définition d'une bonne distance pour faire les comparaisons entre exemples⁵⁴. La question est alors de comprendre en quoi la distance dans le nouvel espace $\Phi(\mathcal{X})$ est plus adaptée à la classification des données que dans l'espace d'origine \mathcal{X} . Il était clair que la distribution des distances entre exemples d'apprentissage devait être modifiée, plus gaussienne dans l'espace d'origine \mathcal{X} et plus piquée vers la valeur $2 \times p$ dans l'espace $\Phi(\mathcal{X})$ (avec p : nombre moyen de motifs servant au codage de chaque exemple). Cette transformation est confirmée expérimentalement, mais nous n'avons pas pu pour le moment déterminer pour quels types de classes, ce recodage pourrait être favorable.

La deuxième piste que j'ai explorée est celle de la détermination de la capacité effective du nouvel espace $\Phi(\mathcal{X})$. Se pouvait-il qu'il y ait une sorte de régularisation automatique qui prédisse un bon pouvoir de généralisation ?

Dans l'espace de redescription $\Phi(\mathcal{X})$, en supposant l'utilisation d'une méthode par 1-plus-proche-voisin, la frontière de décision entre classes est une frontière de Voronoï, dans laquelle les cellules frontières sont similaires à des exemples support pour les Séparateurs à Vastes Marges. Ce qui semble intéressant immédiatement, c'est que, en conséquence de la propriété de codage

⁵⁴. Puisque ces méthodes remplacent la vision frontière de décision sélectionnée dans un modèle paramétré ou espace de fonctions, par la vision frontière de décision définie par une fonction de distance à des points privilégiés, par exemple les exemples de support dans les Séparateurs à Vastes Marges.

clairsemé (chaque exemple n'ayant qu'environ $p << m$ coordonnées $\neq 0$ (p : nombre de motifs utilisés pour coder chaque exemple dans la base de m motifs au total)), chaque paroi d'une cellule de Voronoi n'exige que $2 \times p$ paramètres pour être spécifiée (il y a $2p$ coordonnées distinctes au plus entre deux exemples). Si donc le nombre de parois était limité, on aurait immédiatement une belle propriété de capacité limitée de $\Phi(\mathcal{X})$. Malheureusement, l'hypothèse d'orthogonalité entre exemples d'apprentissage dans le nouvel espace implique que le nombre de cellules de Voronoi devrait croître à peu près comme le nombre d'exemples d'apprentissage, ce qui correspond à une dimension de Vapnik-Chervonenkis d_{VC} infinie ! Une adaptation au cas discret des résultats de Lin et Vitter [LV94] sur une théorie des mémoires à base de cas dans des espaces continus pourrait peut-être nous sortir de cette situation périlleuse, mais elle reste à faire. Encore un projet en état de sommeil momentané, qui attend un peu plus de temps et un prince charmant ou du moins un stagiaire intéressé.

2.6 Publications, projets et stages liés à ces directions de recherche

Un certain nombre de travaux présentés dans ce chapitre sont liés à des projets auxquels j'ai participé :

BQR-2002 : « Vision de scènes naturelles et codage clairsemé »

Responsable : Philippe Tarroux (CNRS-Limsi)

Participants : Antoine Cornuéjols (LRI), Nathalie Denquive (Limsi), Jean-Sylvain Liénard (Limsi) et Michèle Sebag (LRI)

BQR-2003 : « FISICA »

Responsable : Antoine Cornuéjols (LRI)

Participants : Christine Froidevaux (LRI), Jean-Sylvain Liénard (Limsi), Céline Rouveiro (LRI), Michèle Sebag (LRI) et Philippe Tarroux (CNRS-Limsi)

INDANA (2001-2004) : « Étude du risque cardio-vasculaire »

Responsable : Marie-Christine Jaulent (INSERM)

Participants : Florence d'Alche-Buc (LIP6), Isabelle Colombet (INSERM), Antoine Cornuéjols (LRI), Fabien Torre (LIFL), Rémi Gilleron, LIFL, Yves Grandvalet (UTC, Compiègne), François Gueyffier (SPC, Lyon), Christophe Marsala (LIP6), Mario Ota (INSERM), Michèle Sebag (LRI)

ACCAMBA (ACI-IMPBio) (2005-2007) : « Prédiction de bioactivité de molécules »

Responsable : Gilles Bisson (CNRS-IMAG)

Participants : Mirta Gordon (CNRS-IMAG), Laurence Lafanechère, Sylvaine Roy, Bernard Rousseau, Samuel Wieczorek (CEA-Grenoble), Antoine Cornuéjols (LRI), Laurent Meijer (CNRS-Roscoff), David Grierson , Nathalie Marie(Curie-Paris), Dragos Horvath (U. Lille)

Les **stagiaires et doctorants** encadrés ayant travaillé sous ma direction sur ces sujets sont :

- Jérémie Mary (doctorant : 2002-2005)
- Raymond Ros (doctorant : 2005 -)

- Sandra Pinto (stagiaire DEA : 2001)
- Sébastien Jouteau (stagiaire DEA : 2003)
- Nicolas Pernot (Stagiaire DEA : 2004)

Les **publications** concernées auxquelles j'ai participé sont :

1. (CC87) J.-P. Cassou and A. Cornuéjols, « Statistical filtering of motion field from image sequences », *GRETSI-87*, Nice, France, 1987
2. (Pin01) S. Pinto, (Encadrement A. Cornuéjols) *Etude du phénomène de transition de phase dans l'induction supervisée*, Rapport de DEA (LRI, Univ. Paris-Sud, Orsay), 2001
3. (Jou02) S. Jouteau (Encadrement A. Cornuéjols) *Reconnaissances de scènes naturelles*, Rapport de DEA (Dept. Optimisation, Paris-6), 2002
4. (CM02) A. Cornuéjols and L. Miclet, *Apprentissage Artificiel. Concepts et Méthodes*, Eyrolles, 2002
5. (JCSTL03) S. Jouteau, A. Cornuéjols, M. Sebag, P. Tarroux and J.-S. Liénard, « Nouveaux résultats en classification à l'aide d'un codage par motifs fréquents », *Revue d'Intelligence Artificielle (Proc. of the EGC-03 Conf.)*, vol. 17, No.1-3, 521-532, 2003
6. (ASM04) A. Cornuéjols, M. Sebag and J. Mary, « Classification d'images à l'aide d'un codage par motifs fréquents », *RFIA-04* (Workshop sur la fouille d'images), Toulouse, France, 2004
7. (PCS05a) N. Pernot, A. Cornuéjols and M. Sebag, « Phase transition in grammatical inference », *CAP-05 (Conférence Francophone d'Apprentissage)*, Nice, France, 2005, PUG (Ed. F. Denis), pp.49-60
8. (PCS05b) N. Pernot, A. Cornuéjols and M. Sebag, « Phase transition within grammatical inference », *Int. Joint Conf. on Artificial Intelligence (IJCAI-05)*, Edinburgh, UK, 2005, (Ed. L. P. Kaelbling), pp.811-816
9. (CS05) A. Cornuéjols and M. Sebag « A Note on Phase Transitions and Computational Pitfalls of Learning from Sequences », *Second Franco-Japanese Workshop on Information Search, Integration and Personalization (ISIP-05)*, Lyon, France, 2005.
10. (Ros05) R. Ros, (Encadrement A. Cornuéjols) *Transition de phase en apprentissage artificiel. Pistes pour sa mise en évidence en robotique* Rapport de Master (LRI, Univ. de Paris-Sud, Orsay), 2005

3

Corrélations, repères et échange d'information

Les mesures de corrélation sont nécessaires pour comparer les données entre elles, les fonctions de base d'un codage ou bien des méthodes de détection de régularité ou de classification. Elles servent à évaluer des redondances, à construire des repères pour des espaces de données, de fonctions ou de méthodes.

Le concept même de méthodes d'ensemble repose sur l'hypothèse de diversité suffisante des méthodes utilisées [Bre98]. L'étude des trajectoires d'apprentissage lors d'un apprentissage en-ligne requiert également la définition de repères dans l'espace, et bénéficie de la notion de corrélation entre états successifs du système (voir chapitre 5).

De nombreuses définitions de la corrélation reposent de manière ultime sur la notion d'information mutuelle ou d'information extractible. De même, la mise en œuvre des méthodes d'ensemble implique une notion d'échange d'information, tout comme l'apprentissage incrémental qui suppose une information circulant d'un état du système au suivant.

Il est donc clair que les notions de corrélation, d'une part, et d'information, d'autre part, sont liées et que leur élucidation est essentielle pour la compréhension d'un certain nombre de mécanismes d'apprentissage. Ce thème, de manière consciente ou non, a souvent, sinon toujours, été sous-jacent à mes préoccupations et à mon inspiration. Ce chapitre rend compte de mes investigations dans ce domaine et prépare ainsi les études du chapitre suivant sur la dynamique de l'apprentissage (chapitre 5).

Dans un premier temps, j'aborde la notion de corrélation telle que je l'ai définie et utilisée pour la sélection d'attributs dans une application d'analyse du transcriptome, en collaboration avec Christine Froidevaux du LRI et Marie Dutreix de l'institut Curie à Orsay. En utilisant cette nouvelle notion, j'ai pu développer une méthode originale de combinaison de méthodes d'estimation d'attributs qui permet d'obtenir des informations inaccessibles jusque là. D'un point de vue formel, cette nouvelle technique est très belle, et pourtant si simple. Il nous faut maintenant attendre le verdict des biologistes pour voir si ils confirment l'intérêt des résultats obtenus.

Dans un deuxième temps, je présente le raisonnement par analogie comme une construction de corrélation orientée (donc dissymétrique) basée sur un principe d'économie cognitive, c'est-à-dire de l'information qu'il faut apporter pour interpréter l'analogie et la résoudre.

3.1 Mesures de corrélation et comparaison de repères

Afin de pouvoir considérer la notion de trajectoire associée à un apprentissage, il faut pouvoir parler de « points » ou d'états, mais il faut aussi pouvoir les comparer entre eux. Un ensemble minimal de relations est donc exigé, qui définit une structure géométrique.

La panoplie minimale de propriétés constitue la topologie. On peut alors définir la dimension d'une variété et son caractère ouvert ou fermé. On ne sait pas encore y manipuler des vecteurs, y mesurer des angles ou des longueurs, y définir une courbure. Pour cela, il faut introduire un niveau de structure supplémentaire : une *métrique*.

À ma connaissance, aucune étude systématique n'a été faite des types de structures dont on pourrait doter les objets de l'informatique : données et programmes. Il y a là un programme de recherche sans doute très intéressant. En attendant qu'il soit mené à bien, il faut se contenter de consulter le catalogue des mesures et des distances parfois inventées de manière *ad hoc*. J'en esquisse ici un aperçu minimal.

Métriques pour espaces vectoriels

Définition 3.1 (Métrique)

On appelle *métrique* sur \mathcal{E} un espace vectoriel réel de dimension finie une forme bilinéaire, symétrique, définie et positive.

Soit $\mathcal{B} = (e_1, e_2, \dots, e_n)$ une base vectorielle de \mathcal{E} . En toute généralité, on définit une métrique par un tenseur : $\mathbf{g} = g_{\mu\nu} dx^\mu \otimes dx^\nu$ où dx^μ est la base duale de \mathcal{B} .

Dans le cas des espaces euclidiens (espace de courbure nulle), la matrice $g_{\mu\nu}$ associée au tenseur est diagonale, et on retrouve le produit scalaire classique.

Ce produit scalaire est souvent utilisé en apprentissage dès que les objets manipulés sont exprimés sous forme de vecteurs.

Dans le cas de vecteurs aléatoires \mathbf{x} et \mathbf{y} , on peut faire appel aux moments statistique d'ordre deux.

Définition 3.2 (Corrélation entre deux vecteurs aléatoires)

La corrélation entre deux vecteurs aléatoires \mathbf{x} et \mathbf{y} peut être définie par leur matrice de corrélation \mathbf{C}_{xy}

$$\mathbf{C}_{xy} = \mathbb{E}[(\mathbf{x} - \mathbf{m}_x)(\mathbf{y} - \mathbf{m}_y)^\top] \quad (3.1)$$

Les vecteurs sont indépendants si $\mathbf{C}_{xy} = 0$.

Une extension intéressante est utilisée dans le *Latent Semantic Analysis* qui tient compte des ressemblances entre objets du domaine. La métrique devient alors non linéaire et dépend des données.

Métriques pour variables aléatoires et pour distributions de probabilité

Lorsque les objets manipulés sont des distributions de probabilité, il est souvent nécessaire de comparer ces distributions, soit avec comme objectif de montrer qu'elles sont significativement différentes, soit au contraire pour mesurer la qualité de l'approximation d'une distribution cible par une distribution hypothèse.

Les mesures naturelles de distance ne sont plus alors les normes ℓ_p et leurs variantes, mais des mesures liées à la théorie de l'information comme la divergence de Kullback-Leibler ou la distance de Hellinger.

La théorie des tests statistiques [Dod03] est concernée par le premier cas. En faisant certaines hypothèses paramétriques sur les distributions en jeu, il est possible de comparer un écart à ce qui serait attendu dans une proportion jugée raisonnable des tirages (notion d'intervalle de confiance) si les populations comparées étaient issues de la même distribution (hypothèse nulle).

La notion de corrélation entre variables aléatoires donne lieu à plusieurs mesures [Mac04].

Définition 3.3 (Entropie jointe)

Soient deux variables aléatoires X et Y , leur entropie jointe est définie par :

$$H(X, Y) = \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x, y) \log \frac{1}{P(x, y)} \quad (3.2)$$

L'entropie est additive pour des variables aléatoires indépendantes.

Définition 3.4 (Entropie conditionnelle de X sachant $y = b_k$)

L'entropie conditionnelle d'une variable aléatoire X connaissant la valeur de la variable aléatoire $Y : y = b_k$, est égale à l'entropie de la distribution de probabilité $P(x|y = b_k)$:

$$H(X|Y) = \sum_{y \in \mathcal{A}_Y} P(y) \left[\sum_{x \in \mathcal{A}_X} \log \frac{1}{P(x, y)} \right] \quad (3.3)$$

Elle mesure l'incertitude moyenne qui reste sur x quand y est connue.

Définition 3.5 (Entropie conditionnelle de X sachant Y)

L'entropie conditionnelle d'une variable aléatoire X connaissant la variable aléatoire Y est la moyenne, sur y , de l'entropie conditionnelle de X sachant y :

$$H(X|Y) = \sum_{y \in \mathcal{A}_Y} P(y) \left[\sum_{x \in \mathcal{A}_X} \log \frac{1}{P(x, y)} \right] \quad (3.4)$$

Elle mesure l'incertitude moyenne qui reste sur x quand y est connue.

Définition 3.6 (Règle de chaînage pour l'entropie)

L'entropie jointe, l'entropie conditionnelle et l'entropie marginale sont reliées par :

$$I(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \quad (3.5)$$

En d'autres termes, cette formule exprime que l'incertitude de X et Y est égale à l'incertitude de X plus l'incertitude de Y sachant X .

Définition 3.7 (Information mutuelle entre X et Y)

L'information mutuelle entre X et Y est définie comme :

$$I(X; Y) \equiv H(X) - H(X|Y) \quad (3.6)$$

Cette formule satisfait $I(X; Y) = I(Y; X)$ et $I(X; Y) \geq 0$. Elle mesure la réduction moyenne d'incertitude sur x qui résulte de la connaissance de y ou, vice-versa, la quantité moyenne d'information que x transmet à propos de y .

Dans le cas de la mesure de l'approximation d'une distribution $p(\mathbf{x})$ par une distribution $p'(\mathbf{x})$, on utilise souvent la mesure d'entropie jointe de Kullback-Leibler :

$$D(p, p') = \int p(\mathbf{x}) \log\left(\frac{p(\mathbf{x})}{p'(\mathbf{x})}\right) d\mathbf{x} \quad (3.7)$$

ou, dans le cas discret :

$$D = \sum_{\mathbf{x}} p(\mathbf{x}) \log\left(\frac{p(\mathbf{x})}{p'(\mathbf{x})}\right) \quad (3.8)$$

où la somme est prise sur toutes les valeurs possibles de \mathbf{x} .

Métriques pour programmes

La théorie de la complexité algorithmique [LV97] définit des notions de « distance » entre programmes étroitement liées aux notions de distance fondées sur les mesures d'entropie définies pour les variables aléatoires.

Intuitivement, il est séduisant de considérer que la distance informationnelle minimale entre deux objets (souvent des chaînes de caractères) x et y par la longueur du programme le plus court pour une machine de Turing universelle pour transformer x en y : $K(y|x)$. À un facteur additif logarithmique près, cette distance de Kolmogorov est définie comme :

$$D(x, y) = \max\{K(y|x), K(x|y)\} \quad (3.9)$$

Limites et perspectives

Lorsque l'on essaie de caractériser une trajectoire d'apprentissage et de définir l'effet favorable ou non d'une donnée sur l'apprentissage d'un concept ou d'une performance cible, il est nécessaire de disposer d'une mesure de corrélation qui puisse prendre des valeurs négatives. Or les notions de distances entre distributions ou entre programmes classiquement proposées ne permettent pas de rendre compte de l'aspect néfaste d'une donnée ou d'une hypothèse (i.e. un programme) existant. Du point de vue de ces distances, au pire on ne pourra pas s'aider de la donnée ou de l'hypothèse existante, mais il n'y aura pas de coût associé à leur prise en compte.

Cela est lié à un principe fondamental de la théorie de l'information : le *principe de non-création de l'information*. Selon celui-ci, il est impossible de créer de l'information au moyen de manipulations quelconques de traitement de données. Ce principe à lui tout seul rend inutile la science de l'apprentissage. Il suffit de mémoriser les données. Tout traitement ultérieur est vain et ne sert à rien selon la théorie de l'information. Il y a là, je crois, le symptôme le plus net qu'il faut amender la théorie de l'information pour qu'elle puisse servir de fondement à une théorie de l'apprentissage. Et il faut faire cela en partant des principes utiles en apprentissage.

Par exemple, il est clair qu'il faut étendre les définitions de distance et de corrélation pour en faire des mesures capables de rendre compte des effets bénéfiques ou au contraire néfastes que peuvent jouer des données ou des connaissances *a priori* dans l'apprentissage. C'est un critère important pour toute définition de mesure de corrélation dans le cadre de l'apprentissage. La mesure que nous avons proposée pour la comparaison de méthodes (voir section 3.2) satisfait cette contrainte.

Par ailleurs, les espaces dans lesquels évoluent les systèmes d'apprentissage ne sont certainement pas bien représentés par des espaces euclidiens. Il faut pouvoir tenir compte de ce que le référentiel dans lequel une situation est considérée par un système peut changer. Il faut donc pouvoir également comparer les référentiels en des points différents. Peu de travaux prennent cela

en compte, une exception considérable étant constituée par les recherches de Shun-Ichi Amari [ABNK⁺87, MR93] sur l'estimation dans les espaces de distributions exponentielles.

Nos propres travaux sur une formalisation du raisonnement par analogie (voir section 3.3) essayent d'apporter une contribution à cette nécessaire étude.

3.2 Corrélation et combinaison de détecteurs de régularités

3.2.1 Sélection d'attributs et apprentissage non supervisé

Plus l'apprentissage artificiel s'est attaqué à des problèmes réels, et plus l'importance des prétraitements s'est trouvée soulignée. Fournir des données qui soient aussi peu bruitées que possible, dans lesquelles les données aberrantes (*outliers*) sont retirées, éventuellement avec une distribution de probabilité rectifiée, est un préalable qui facilite l'apprentissage, et peut dans certains cas le rendre quasi trivial. Redécrire les données fait parfois parti du prétraitemet, comme nous l'avons vu au chapitre 2. Cela peut aussi être constitutif de l'apprentissage. Parmi les codages possibles, l'un des plus simples consiste à sélectionner un sous-ensemble de descripteurs au sein des descripteurs d'origine. C'est ce que l'on appelle souvent la *sélection d'attributs*⁵⁵.

Lorsque les données disponibles sont en nombre suffisant par rapport au nombre de descripteurs, c'est-à-dire pour éviter le sur-apprentissage, la sélection d'attributs vise essentiellement à faciliter l'apprentissage en diminuant la complexité des calculs et/ou la taille des données à stocker. Mais, la sélection d'attributs peut également devenir l'objectif et non plus un outil ou une étape auxiliaire lorsque le problème consiste à découvrir les facteurs clés liés à un phénomène (voir par exemple le projet INDANA piloté par l'INSERM pour découvrir les facteurs déterminants pour les risques cardio-vasculaires [Col02]).

Un cas extrême est celui dans lequel les descripteurs sont bien plus nombreux que les exemples. Des applications aussi importantes que l'analyse du génome ou l'indexation de textes sur l'Internet ou des banques d'images ont brisé un tabou et stimulé de nouvelles recherches sur l'évaluation des attributs dans des contextes que les statisticiens s'interdisaient.

J'ai été confronté à ce problème lorsque, pour la première fois, j'ai entendu parler de l'analyse du transcriptome. Je lisais depuis longtemps avec intérêt les articles de vulgarisation et les livres écrits par de grands scientifiques sur la biologie et la révolution de la génétique, mais c'est une conversation avec Christine Froidevaux, au début 2002, qui m'a fait franchir le pas et chercher à participer également à cette quête que Christine présentait de manière si enthousiasmante. Parmi les problèmes dont elle me parla ce jour-là, il y avait donc ce problème d'analyse du transcriptome.

Brièvement, les cellules réagissent aux contingences de leur milieu en produisant en particulier des protéines. La production de ces protéines obéit à un processus d'expression des gènes par l'intermédiaire d'ARN messagers. En mesurant l'activité de ces ARNm à un instant donné, il est donc possible d'avoir des indications à la fois sur l'état de la cellule et sur celui du milieu. Les techniques de génomiques, et particulièrement la technologie des puces à ADN, ont révolutionné les recherches en rendant possible la mesure simultanée de l'activité liée à tous les gènes d'une cellule. On se retrouve donc face à un instantané de l'activité de la cellule exprimé par le degré d'activité de milliers de gènes. L'un des objectifs actuels des biologistes est d'isoler les gènes qui sont spécifiquement mis en jeu par les cellules face à certaines situations, par exemple lorsque la cellule est cancéreuse, ou lorsqu'elle est placée dans un milieu pollué ou faiblement radioactif. Là où les problèmes commencent, c'est que, d'une part, les données sur les degrés d'activité

55. Dans la suite de ce chapitre, j'utiliserai indifféremment les termes d'« attribut » ou de « descripteur ».

obtenues dans les puces à ADN sont très bruitées aussi bien par des artefacts expérimentaux et méthodologiques que par les limites de la technologie actuelle, et, d'autre part, que les données sont en nombre extrêmement réduit, typiquement quelques dizaines. Dans le cas des données dont me parlait Christine, et fournies par Marie Dutreix de l'Institut Curie à Orsay, nous disposions de 18 « lames » mesurant chacune l'activité d'environ 6135 gènes. Douze lames avaient été obtenues dans une condition de non irradiation, et six dans une condition de très faible irradiation⁵⁶. Le défi que nous proposait tranquillement Marie était d'identifier les gènes impliqués spécifiquement dans la réponse à l'irradiation à partir de ces données. Le premier réflexe, sain, était évidemment d'éclater de rire, et de retourner à des études sérieuses avec des échantillons normaux et bien constitués, c'est-à-dire ceux qui existent dans la littérature théorique.

Cependant Marie était persuasive et l'aventure trop tentante. Jérémie Mary, en stage de DEA, Christine et moi avons donc commencé par faire un état de l'art sur la sélection d'attributs.

Techniques de sélection d'attributs

Les méthodes de sélection d'attributs ont pour but d'identifier les attributs qui sont utiles en vue d'une tâche de classification. Chaque exemple, ou forme d'entrée, est décrit par d attributs (e.g. gènes) et appartient à une classe (e.g. tumeur ou non tumeur). L'échantillon d'apprentissage fournit des exemples avec leur classe (supposée correcte). Le problème est de découvrir les attributs les plus informatifs pour la détermination de la classe des exemples d'apprentissage, et aussi pour les exemples à venir, encore inconnus. De plus, on peut chercher à déterminer un ensemble d'attributs minimal permettant de classer les exemples, ou au contraire, vouloir connaître tous les attributs corrélés à la classe des entrées, même si ils sont redondants. Ce dernier cas est plus représentatif des objectifs de l'analyse du transcriptome.

Il faut noter que les attributs peuvent être informatifs à propos de la classe des exemples indépendamment les uns des autres (on parle de *corrélation linéaire*) ou en combinaison (il s'agit de *corrélations d'ordre supérieur*). Il est évident que les corrélations d'ordre supérieur sont plus difficiles à découvrir que les corrélations linéaires, et exigent généralement plus de données d'apprentissage. Pour cette raison, les méthodes de sélection d'attributs utilisées en bioinformatique sont le plus souvent orientées vers la découverte de corrélations linéaires entre les attributs (e.g. l'activité des gènes) et les classes.

Il existe trois grandes classes de méthodes de sélection d'attributs : les « méthodes intégrées » (*embedded*), les « méthodes symbiose » (*wrapper*) et les « méthodes de filtre » (*filter*) [BL97, GE03, KJ97]. Les premières consistent à utiliser directement le système d'apprentissage dans l'espoir que le système découvrira automatiquement les descripteurs utiles pour la classification. Ainsi par exemple, un système d'induction d'arbre de décision [CM02] effectue une sélection automatique des descripteurs en choisissant ceux qui sont suffisants pour la construction de l'arbre. Malheureusement, ce type d'approche est condamné à produire des résultats peu fiables lorsque les données sont très rares par rapport au nombre d'attributs.

Les méthodes de type *symbiose* (comme un parasite et son hôte) évaluent les sous-ensembles d'attributs en fonction des performances des méthodes de classification qui les utilisent. Ainsi, étant donné une méthode de classification (e.g. un perceptron multi-couche) et un ensemble d'attributs \mathcal{F} , la méthode symbiose explore l'espace des sous-ensembles de \mathcal{F} , utilisant la validation croisée pour comparer les performances des classificateurs entraînés sur chaque sous-ensemble. Intuitivement, les méthodes symbiose présentent l'avantage de sélectionner les sous-ensembles d'attri-

56. Si faible en fait, que certaines cellules de levure (*S. cerevisiae*) n'étaient irradiées que presque à la fin de la période d'exposition de 20h, et avaient donc moins de temps pour développer une réponse au stress radioactif. Par conséquent certaines lames classées dans la classe (I) : irradié, pouvaient l'être à tort.

but pertinents qui permettent les meilleures performances en généralisation, ce qui est souvent le but final. Cependant, tandis qu'il a été souligné récemment que cette approche pouvait être biaisée et trop optimiste sur le vrai contenu informatif des attributs sélectionnés [AM02b, XJK01], le principal inconvénient de ces méthodes est leur coût computationnel attaché à l'exploration de l'espace des sous-ensembles de \mathcal{F} .

C'est pourquoi les *méthodes de filtre* conservent leur attrait. Elles sont utilisées dans une phase de prétraitement, indépendamment du choix de la méthode de classification. La plupart d'entre elles évalue chaque attribut indépendamment en mesurant la corrélation (selon une métrique à définir) de leurs valeurs sur les exemples avec la classe de ces exemples. En d'autres termes, ces méthodes évaluent l'information apportée par la connaissance de chaque attribut sur la classe des exemples. Sous certaines hypothèses d'indépendance et d'orthogonalité, les attributs ainsi estimées comme informatifs peuvent être optimaux par rapport à certains systèmes de classification. Un avantage important de cette approche est son faible coût computationnel, puisqu'elle ne requiert qu'un nombre d'évaluations linéaire en le nombre d d'attributs, plus une opération de tri⁵⁷.

Au vu des avantages et des inconvénients mentionnés, nous avons choisi de nous concentrer sur les méthodes de filtre, et plus précisément d'utiliser trois méthodes particulières : ANOVA (*ANalysis Of VAriance*), SAM (*Significance Analysis of Microarray*) ([TTC01]) et RELIEF ([KR92, RSK03]) que nous avons adapté pour en faire un outil dédié à la bioinformatique *Bio-RELIEF*⁵⁸. Les méthodes ANOVA et SAM reposent essentiellement sur l'hypothèse que les classes sont associées à des distributions normales des valeurs de chaque attribut, et utilisent donc des tests statistiques adaptés, tels le F -test. En revanche, *Bio-RELIEF* ne fait pas de telles hypothèses sur la forme des distributions de probabilité, mais repose sur une hypothèse plus faible de compacité des nuages de points associés aux classes dans l'espace des d descripteurs. Les descripteurs mettant le plus en évidence cette compacité et la séparation des nuages de points sont jugés les plus pertinents.

Réflexions sur une étude théorique de la sélection d'attributs

En supposant d'une part que les exemples sont booléens et appartiennent à la classe **Vrai** ou à la classe **Faux**, et, d'autre part, que les attributs utilisés pour décrire les exemples puissent être classés comme « pertinent » ou « non pertinent », la tâche d'identifier les attributs pertinents à partir d'un échantillon peut sembler plus simple que celle d'identifier la règle de classification des exemples.

En effet, le nombre de règles de classification binaires d'exemples décrits à l'aide de d descripteurs booléens est de 2^{2^d} , tandis que le nombre de tris différents de d descripteurs est seulement égal à $d!$ (soit environ $d^d e^{-d} \sqrt{2\pi d}$). Ainsi, pour $d = 10$, le nombre de règles de classification est de : $1,8 \cdot 10^{308}$ (environ 4 lignes de chiffres) ; tandis que le nombre de tris est de : 3'628'800 seulement (!).

Il semble donc évident qu'il est plus facile de découvrir les descripteurs pertinents à partir d'un échantillon d'apprentissage, que de découvrir la règle de classification sous-jacente.

La théorie statistique de l'induction montre sous quelles conditions l'erreur observée sur l'échantillon d'apprentissage (le risque empirique) est représentatif de l'erreur réelle. Sous des conditions bien identifiées, il est donc possible d'extrapoler à partir d'un échantillon d'apprentissage limité une hypothèse sur la règle de classification sous-jacente avec certaines garanties sur

57. C'est ainsi que cette technique peut aussi être adaptée à l'ILP [AM02a].

58. Disponible à : <http://www.lri.fr/~chris/bioinfo/BioRelief>

la qualité de cette hypothèse. Curieusement, en revanche, il n'existe pas, à ma connaissance, une telle théorie pour lier la qualité d'un tri hypothétique des descripteurs avec le vrai tri sous-jacent.

Quelle peut en être la raison et quelles sont les perspectives envisageables ?

Une raison vient du fait que les exemples d'apprentissage fournissent des exemples de lien *description* → *classe*, tandis qu'ils ne fournissent qu'une information indirecte sur la valeur, « pertinent » ou « non pertinent », des attributs. On ne dispose donc pas directement d'une mesure équivalente au risque empirique.

Ensuite, on peut se poser la question de l'équivalent d'une mesure de capacité sur l'espace des tris possibles associés à une méthode de sélection des attributs. Comment mesurer une telle capacité ? Que signifie la corrélation de deux méthodes de tris ?

Ces questions me trottaient dans la tête au printemps 2004, accompagnées du sentiment diffus que la solution devait être simple. Mais chaque fois que je croyais enfin comprendre, la résolution se révélait erronée et le désespoir actif et impatient me poussait à de nouvelles soirées fiévreuses. Je n'ai toujours pas trouvé de solution, mais indirectement, sans que je m'en rende compte sur le moment, ces réflexions ont contribué à la découverte, en 2005, de la méthode exposée en section 3.2.4.

Mais, même si la chronologie n'est pas respectée, il est bon de faire un détour par un autre travail, lié à ces questions.

3.2.2 Approche par méthode d'ensemble

En fin 2003, motivée, d'une part, par le défi organisé pour NIPS-03 par Isabelle Guyon et aiguillonnée, d'autre part, par le problème d'identification des facteurs de risques cardio-vasculaires (projet INDANA, voir 3.4), Michèle Sebag proposait une nouvelle méthode de sélection d'attributs. Celle-ci, sans surprise venant de Michèle, reposait sur une méthode d'évolution simulée, et sur le concept novateur de fonction d'évaluation (*fitness function*) mesurant l'aire sous la courbe ROC et non un risque empirique, d'où son nom : ROGER⁵⁹. L'idée était de faire émerger un ensemble de classificateurs quasi-linéaires⁶⁰ permettant de classer les points d'apprentissage suivant une série de valeurs de précision et rappel correspondant à un chapelet de points définissant une courbe ROC. Chaque fonction de décision pouvant être interprétée comme un tri sur les descripteurs, en fonction du poids w_i associé à chacun d'eux.

L'algorithme produisant une population de tris fondés sur des critères différents, une question naturelle était de savoir si il était possible de combiner ces tris, par une méthode d'ensemble, pour obtenir un tri de meilleure qualité. Une sorte de boosting appliqué à des fonctions de préférence plutôt qu'à des classificateurs.

Nous avons alors montré ([JMC⁺04]) que le tri résultant d'un vote majoritaire pour décider de l'ordre sur chaque paire de descripteurs (le descripteur i est-il plus pertinent que le descripteur j : noté $i \prec j$?) avait de bonnes propriétés, à savoir : (i) cohérence du tri final (si $i \prec j$ et $j \prec k$ alors $i \prec k$), (ii) convergence du tri final vers le tri vrai, et (iii) convergence exponentiellement rapide avec le nombre de tris utilisés.

Les preuves mathématiques sont belles. Cependant, elles reposent sur un certain nombre d'hypothèses plus ou moins bénignes. La plus importante est la supposition selon laquelle les tris obtenus par l'algorithme génétique sont des perturbations aléatoires (chaque paire d'attributs ayant la même probabilité d'être intervertie dans le tri perturbé) du tri vrai. Or rien

59. ROc-based GEnetic learnerR.

60. Les fonctions de décision étant de la forme : $h_Z : \mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d \mapsto \mathbb{R}$, $h_Z(\mathbf{x}) = \sum_{i=1}^d w_i \times |x_i - c_i|$ où c_i est la i -ème coordonnée d'un point c .

n'indique que ce soit le cas. Le type de perturbation supposé est lui-même très hypothétique. Enfin l'indépendance supposée des différents tris est certainement non vérifiée.

Comme, de plus, le système d'évaluation des attributs repose sur l'utilisation de classifieurs particuliers, dont on ne peut prévoir à l'avance l'adéquation aux données, il est surprenant que les résultats empiriques sur des bases de données non artificielles [SAL04], aient été apparemment aussi bons. Les résultats n'étant pas meilleurs que ceux de techniques beaucoup plus simples sur d'autres données [SDDO03].

Ma conclusion, pour le moment, est que l'idée de combiner des fonctions de préférences imparfaites (c'est-à-dire des tris) est intéressante. Mais il faut s'appuyer sur des hypothèses plus réalistes et avoir un moyen d'estimer la diversité (ou la corrélation) entre les fonctions de préférence.

3.2.3 Mesure de corrélation entre détecteurs de régularité

Une méthode d'évaluation des descripteurs relatifs à une classification peut être considérée comme un détecteur sensible à un certain type de régularité. Ainsi, SAM et ANOVA sont efficaces si les classes ont une distribution normale lorsqu'elles sont projetées sur chaque descripteur et si la variance de ces projections est petite par rapport aux distances entre les moyennes pour les descripteurs pertinents. Par ailleurs, SAM et ANOVA ne détectent que des corrélations linéaires entre descripteurs et classe. Ces méthodes sont incapables de détecter la pertinence de deux attributs, liés par exemple par une fonction OR à la classe, si chacun d'eux indépendamment est décorrélé de la classe. *Bio-RELIEF*, quant à lui, ne fait pas d'hypothèse sur la distribution des classes dans l'espace des descripteurs, mais suppose cependant que la distance des points à l'intérieur d'une classe est petite par rapport à la distance des points entre classes, et que cela est apparent par projection sur les descripteurs pertinents. *Bio-RELIEF* est partiellement sensible aux corrélations d'ordre supérieur entre attributs et classes.

Les méthodes d'évaluation d'attributs calculent une valeur numérique pour chaque attribut qui représente à quel degré la régularité, à laquelle est sensible la méthode, est représentée dans la fonction qui lie le descripteur et la classe.

Deux questions se posent immédiatement :

1. *Le type de régularité mesuré par la méthode est-il approprié au domaine d'application ?*
Cette question est centrale par exemple pour le biologiste spécialiste du transcriptome et qui doit décider si il doit plutôt utiliser SAM, par exemple, que *Bio-RELIEF*, ou l'inverse, ou encore une autre méthode.
2. *En supposant qu'une méthode d'évaluation d'attributs soit jugée comme adaptée, comment doit-on fixer le seuil de valeur distinguant les attributs jugés pertinents de ceux qui ne le sont pas ?*

Une autre question essentielle est : *y a-t-il de l'information utile dans les données ?* Mais, sauf à essayer d'appliquer la théorie de Kolmogorov-Chaitin⁶¹, cette question est indissociable du choix préalable d'une méthode et de la question du seuil de pertinence.

Pour distinguer la présence de l'absence d'information significative, les statisticiens utilisent le concept d'« hypothèse nulle » (voir par exemple [Dod03, Tas04]). L'idée est que si un dispositif de mesure produit, sur les données réelles, des résultats indistinguables (à une approximation près)

61. Qui distingue un signal contenant de l'information d'un autre n'en contenant pas par le fait que le premier est compressible contrairement au second. Mais cette théorie est difficile à appliquer. (Voir [LV97] pour une synthèse générale.)

de ceux obtenus sur un échantillon aléatoire, alors ces données ne fournissent pas d'information par le canal du dispositif de mesure.

Nous avons suivi cette démarche sur les données fournies par Marie Dutreix. Utilisant par exemple la méthode de *Bio-RELIEF* à la fois sur les vraies données et sur ces mêmes données dans lesquelles les étiquettes (6 '+' et 12 '-') ont été permutées au hasard (1000 fois), nous avons évalué les 6135 gènes et dressé la courbe du nombre de gènes dont la valeur dépasse un certain seuil. La courbe de la figure 3.1 à gauche montre la différence entre les deux situations. Il est apparent que le nombre de gènes pour lesquels la régularité détectée par *Bio-RELIEF* est présente à un certain degré est significativement plus important dans les vraies données que dans les données aléatoires. Le rapport entre les deux courbes qui est tracé sur la figure à droite peut être interprété comme un rapport signal/bruit. Il est ici maximal pour un seuil de 0.5. Seuls 30 gènes obtiennent un score supérieur pour les vraies données, tandis que l'on passe à moins de 1 pour les données aléatoires⁶²

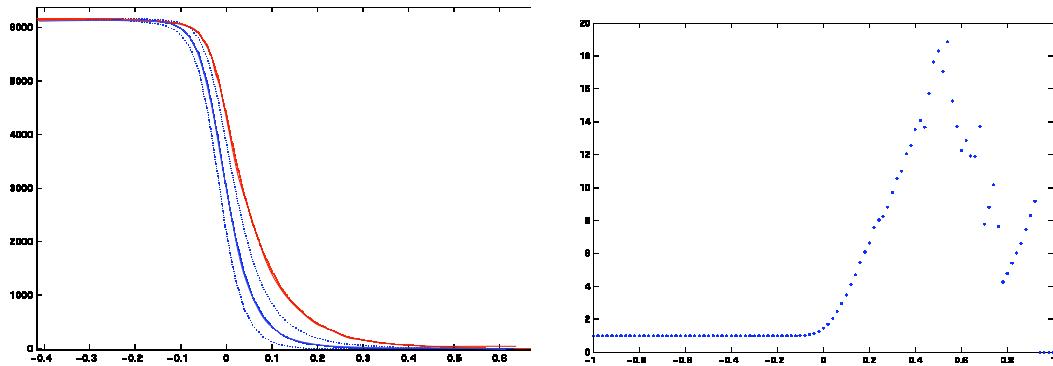


FIGURE 3.1 – (À gauche) Nombre de gènes (axe y) de score dépassant la valeur indiquée sur l'axe x (avec Bio-RELIEF). L'hypothèse nulle (courbe bleue) est significativement en-dessous de la courbe obtenue avec les vraies données (courbe rouge). (À droite) Courbe du rapport entre les courbes rouge et bleue. Ce rapport est maximal pour un score de 0.5, soit pour 30 gènes dont il est peu probable qu'ils comptent des faux positifs.

Ces résultats indiquent que *Bio-RELIEF* détecte la présence des régularités auxquelles il est sensible dans les données réelles et qui ne se trouvent pas dans les données aléatoires. Les données expérimentales de Marie, pour imparfaites et limitées qu'elles sont, n'en portent pas donc moins une information mesurable. Par ailleurs, ces premières études soulignaient cependant le caractère arbitraire de la détermination d'un seuil de pertinence. Ainsi, pour des raisons ayant trait tant aux attentes des biologistes sur le nombre de gènes mis en œuvre dans la réponse au stress radioactif que de caractère pratique sur le nombre de gènes analysable à la main pour espérer comprendre les mécanismes en jeu, Marie estimait à environ 150 à 200 le nombre de gènes pertinents, ce qui correspondait pour nous à un score d'environ 0,3 et un rapport signal/bruit de 9 environ.

Le caractère arbitraire de ces choix n'était pas satisfaisant pour un informaticien. Et la question se posait de savoir si il était possible de fournir des estimations plus précises sur le nombre probable de gènes pertinents en utilisant plusieurs méthodes d'évaluation d'attributs

62. La courbe pour les données aléatoires étant obtenue par moyenne de 1000 mesures, le nombre de gènes rapporté pour un score donné est une moyenne et peut donc être une valeur non entière.

plutôt qu'une seule.

Une nouvelle mesure de corrélation entre méthodes d'évaluation d'attributs

Supposons que je dispose de deux détecteurs de régularité et qu'ils donnent des résultats similaires sur un jeu de données, dois-je en conclure que je peux faire confiance aux résultats mesurés ? C'est ce que Marie voulait croire en constatant que sur les 500 gènes classés comme les meilleurs par Bio-RELIEF (ensemble noté top₅₀₀), 281 étaient communs aux 500 meilleurs selon ANOVA. Mais, par ailleurs, dans le cas de SAM et d'ANOVA, l'intersection était de 409 (figure 3.2). Une analyse plus fine était clairement nécessaire.

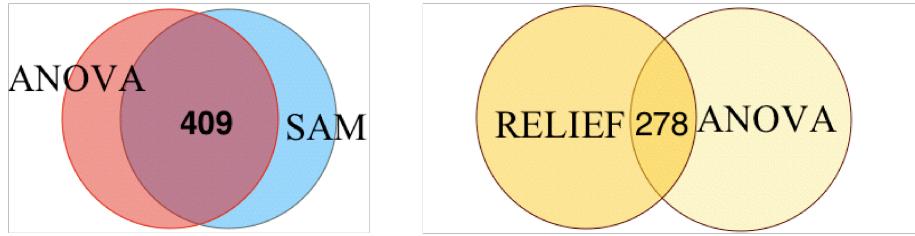


FIGURE 3.2 – Intersection des top₅₀₀ pour les couples de méthodes SAM et ANOVA (à gauche) et pour Bio-RELIEF et ANOVA (à droite).

L'intersection peut résulter de trois causes :

1. **Le hasard.** Le tirage aléatoire de deux sous-ensembles de n (e.g. 500) éléments parmi d (e.g. 6135) a une intersection dont la taille k suit la distribution :

$$H(d, n, k) = \frac{\binom{n}{k} \cdot \binom{d-n}{n-k}}{\binom{d}{n}} \quad (3.10)$$

2. **La corrélation *a priori* des méthodes.** (Deux méthodes identiques seront toujours d'accord sur leurs top _{n} . Plus les méthodes sont décorrélées, plus cette intersection est petite). On mesure cette corrélation sur les données étiquetées aléatoirement.
3. **Les régularités dans les données** sur lesquelles les méthodes sont d'accord, au-delà de leur corrélation *a priori*.

Si le calcul de l'intersection due au hasard était classique (voir notre papier [MBM⁺04]), la mesure de corrélation *a priori* est, à ma connaissance, une contribution originale que nous avons mise au point au printemps 2004 (et publiée dans [CFM05]).

L'idée est de définir à nouveau une sorte d'hypothèse nulle. La corrélation entre deux méthodes M_1 et M_2 peut s'évaluer par l'espérance de la taille de l'intersection de leur top _{n} si les données étaient aléatoires.

$$\langle M_1, M_2 \rangle = E_{\mathcal{D}}(|\text{top}_n(M_1) \cap \text{top}_n(M_2)|) \quad (3.11)$$

où $|.|$ est utilisé pour dénoter le cardinal d'un ensemble, et $\text{top}_n(M)$ dénote les n gènes les mieux classés par la méthode M . Cette espérance est définie sur une distribution \mathcal{D} de données.

Afin d'assurer que le biais introduit par la distribution des données est le même pour les données aléatoires et pour les données vraies, une solution simple est de calculer l'espérance sur les données dont on permute aléatoirement les classes. On peut également obtenir une estimation

empirique de cette espérance à partir de mesures des tailles d'intersection pour un certain nombre de jeux de données aléatoires.

La table 3.1 montre les résultats obtenus pour diverses valeurs de n pour les méthodes ANOVA et *Bio-RELIEF*, à la fois pour des jeux de données aléatoires (valeur moyenne et déviation standard) et pour les vraies données sur les très faibles irradiations.

n	100	200	300	400	500	600	700	800	900	1000
$\mu_{\mathcal{H}_0}$	21.2	54.2	93.2	135.4	180.3	226.9	276.3	326.2	378.9	432.5
$\sigma_{\mathcal{H}_0}$	8.0	16.9	24.5	32.3	41.8	50.3	57.7	64.1	71.3	78.0
k	37	93	149	210	281	339	406	470	535	605

TABLE 3.1 – Intersection of two top _{n} from ANOVA and *BioRELIEF* for various values of n , under the null hypothesis ($\mu_{\mathcal{H}_0}$ and $\sigma_{\mathcal{H}_0}$) and observed for the true data set (k).

Les courbes correspondant à l'intersection observée avec les vraies données, celle attendue à partir de la corrélation *a priori* et celle attendue par l'effet du hasard sont tracées dans la figure 3.3. L'abscisse correspond à la taille n des top _{n} et l'axe des ordonnées au rapport des tailles d'intersection sur n (e.g. pour $n = 500$, l'intersection observée est de 281, soit $0,562 \times 500$, d'où la valeur 0,562 reportée sur l'axe).

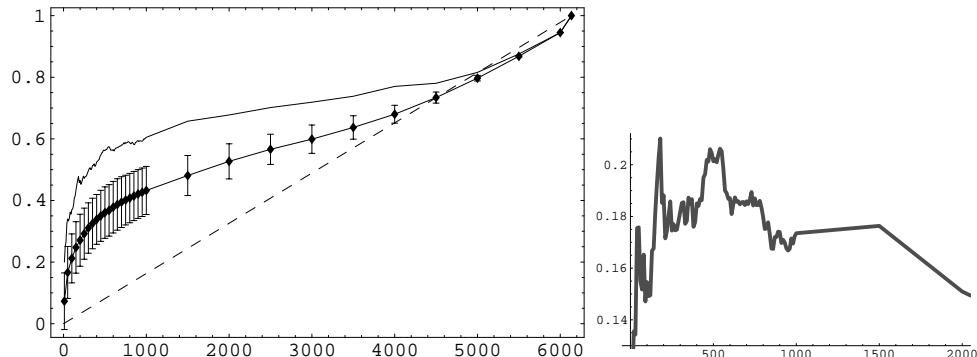


FIGURE 3.3 – L'abscisse représente le nombre n des attributs les mieux notés. L'ordonnée représente le rapport de la taille de l'intersection à n . (À gauche) : (courbe du haut) le rapport mesuré; (courbe du milieu) le rapport du à la corrélation *a priori* des méthodes (avec des barres d'écart-type); (courbe du bas) le rapport explicable par le hasard. (À droite) : Courbe de la différence relative, par rapport à n , de la taille de l'intersection observée k et de l'espérance de taille d'intersection due à la corrélation *a priori*. La courbe n'est représentée que pour $n < 2000$ car c'est la partie la plus intéressante.

De ces courbes obtenues sur les données particulières de Marie, il est possible de tirer deux conclusions. D'une part, il y a bien une information spécifique dans ces données, explicables ni par le hasard, ni par la corrélation des méthodes *a priori* : k est nettement supérieur à l'espérance $\mu_{\mathcal{H}_0}$ (de plus de deux écarts-type). D'autre part, il est possible de déterminer la valeur de n pour laquelle la plus grande proportion d'attributs sélectionnés est attribuable aux régularités détectées dans les données et non au hasard ou à la corrélation *a priori*. (Ici, la courbe de droite a deux pics, l'un pour $n \approx 180$ et l'autre pour $n \approx 540$, le premier pic étant sans doute un point aberrant sur la courbe).

Plus généralement, la mesure de corrélation *a priori* entre méthodes présentée a plusieurs propriétés remarquables :

- Contrairement aux mesures de corrélation fondées sur la notion d'information mutuelle, cette mesure **peut donner un résultat négatif** correspondant à deux méthodes *anti-correlées*. En effet, si l'espérance de taille d'intersection est inférieure à la taille d'intersection attribuable au hasard seul, c'est que chaque méthode produit un tri qui est comme informé sur le tri de l'autre méthode et s'arrange pour éviter de sélectionner les mêmes attributs. Cette capacité à mesurer une anti-corrélation est à mon avis très importante. Elle est cruciale pour l'étude de l'apprentissage incrémental et de l'utilité des exemples. (Voir le chapitre 5).
- Elle est d'usage général. J'ai montré ici son utilisation pour comparer des méthodes de sélection d'attributs, et, plus généralement, des fonctions de préférence ou de tri, mais elle est applicable également à la comparaison d'algorithmes d'apprentissage. En effet ...
- Elle peut-être considérée comme une application du *no-free-lunch theorem*⁶³. En effet, une méthode pour mesurer l'« alignement » de deux algorithmes est de mesurer l'espérance de l'accord de leurs prédictions sur l'ensemble des problèmes possibles. (Deux méthodes sont orthogonales quand cette espérance est égale à zéro).

Je crois donc que cette mesure de corrélation est intéressante et mérite d'être considérée dans d'autres contextes. Mais elle permet aussi d'obtenir des résultats bien plus forts dans la combinaison de deux méthodes.

3.2.4 Utilisation dans une méthode de combinaison de méthodes

D'un certain côté, utiliser deux méthodes de sélection d'attributs au lieu d'une semble n'apporter aucun progrès. Nous avions des méthodes dont nous ne savions pas si le biais, c'est-à-dire le type de régularité auxquelles elles sont sensibles, est adaptée à la tâche. Nous avons maintenant une troisième méthode, qui considère l'intersection des tris produits par les deux premières, et dont nous sommes tout aussi incapables d'estimer la valeur. Où est le progrès ?

En hiver 2005, j'ai réalisé que nous étions en fait en bien meilleure posture.

Il est en effet possible d'écrire un modèle génératif simple de la taille de l'intersection k observée quand on fait l'intersection de deux top_n produits par deux méthodes d'évaluation d'attributs.

Supposons que d soit le nombre d'attributs, p le nombre inconnu d'attributs effectivement pertinents (on suppose ici que les attributs sont pertinents ou non : par exemple des gènes qui sont impliqués dans un processus biologiques ou qui ne le sont pas), et n la taille des ensembles top_n des attributs les mieux notés par chacune des deux méthodes considérées. Faisons de plus l'hypothèse (sans doute discutable) que les deux méthodes sont également adaptées au problème (c'est-à-dire éventuellement également mal adaptées) et retournent chacune m attributs pertinents parmi les n les mieux évalués. (Voir figure 3.4). Alors, la distribution de probabilité de la taille k de l'intersection des deux top_n est donnée par la formule suivante :

$$p(\cap = k | d, p, n, m, \mu_{\mathcal{H}_0}) = \frac{\binom{p}{m} \binom{d-p}{n-m} \sum_{k+=2m-p}^m \binom{m}{k+} \binom{p-m}{m-k+} \binom{n-m}{k-k+} \binom{d-n-(p-m)}{n-m-(k-k+)}}{\binom{d}{n} \cdot \binom{d}{n}} / C(\mu_{\mathcal{H}_0}) \quad (3.12)$$

63. Le no-free-lunch-theorem [CM02, Wol92, Wol96] affirme qu'en moyenne sur tous les problèmes de classification possibles, aucune méthode de classification n'est supérieure à une autre, et par conséquent supérieure à une décision au hasard. L'interprétation positive de ce théorème est que chaque méthode est adaptée à certains types de problèmes et qu'il faut donc étudier les algorithmes et leurs mérites relativement à des tâches et des environnements particuliers.

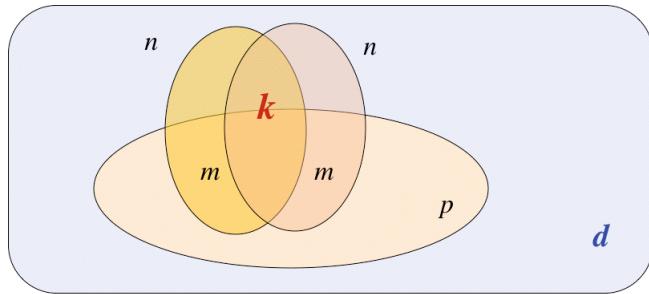


FIGURE 3.4 – Les ensembles impliqués dans le modèle génératif de la taille k de l'intersection.

Les détails de calcul figurent dans [CFM05]. Ce qui est intéressant, c'est qu'il est maintenant possible d'utiliser cette formule par le *principe de maximum de vraisemblance* pour calculer les valeurs les plus probables des inconnues p (nombre d'attributs effectivement pertinents) et m lorsque l'on connaît les valeurs de d (nombre total d'attributs), n , k taille observée de l'intersection des top n et $\mu_{\mathcal{H}_0}$ (taille de l'intersection des top n attribuable à la corrélation *a priori* des méthodes).

Appliquée aux données de Marie ($d = 6135$) et en prenant les résultats obtenus avec *Bio-RELIEF* et *ANOVA*, avec $n = 500$ et $\mu_{\mathcal{H}_0} = 181$, on obtient : $p = 420 \pm 20$ et $m = 340 \pm 20$ comme valeurs les plus probables du nombre total de gènes pertinents et du nombre de ces gènes présents dans les 500 meilleures de chaque méthode (*rappel* = $340/420 = 0,81$ et *précision* = $340/500 = 0,68$).

Un calcul simple, non encore publié, permet même de prédire que le nombre d'attributs pertinents dans l'intersection de 281 attributs obtenus pour les deux top 500 est le plus probablement d'environ 265 ! La précision atteint cette fois $265/281 = 0,94 !!$ Il est donc possible de soumettre au biologiste une liste de gènes presque tous certainement pertinents.

Évidemment, la précision de ces estimations dépend d'une part de la validité de l'hypothèse selon laquelle *Bio-RELIEF* et *ANOVA* sont également adaptés à la tâche, et, d'autre part, de la qualité des données, qui sont en très petit nombre. Il n'empêche que nous disposons maintenant d'une méthode permettant d'exploiter de manière beaucoup plus puissante qu'une simple comparaison directe, les résultats produits par deux méthodes.

Il reste, d'une part, à vérifier auprès des biologistes que les conclusions tirées sont sinon valides du moins plausibles, et, d'autre part, à faire de nombreuses expériences avec des données artificielles pour préciser les propriétés et la précision de cette approche⁶⁴.

Les méthodes d'ensemble pour la sélection d'attributs peuvent grandement bénéficier des hypothèses faibles faites ici et de l'exploitation du concept de corrélation *a priori* pour transformer des méthodes de sélection faible en méthode forte. C'est une direction de recherche pour des travaux de futur immédiat.

64. Il serait également intéressant d'essayer de couper les informations obtenues avec nos méthodes avec des informations obtenues par analyse des textes en bioinformatique [KMF04].

3.3 L'analogie : construction dynamique d'une corrélation orientée

Comment ne pas étudier l'analogie ? Elle est à l'induction ce qu'une randonnée est à la photo aérienne. J'argumenterai qu'elle est aussi à l'induction ce qu'une généralité est à un cas particulier.

L'analogie est omniprésente dans notre entendement du monde. D'un certain côté (notons la métaphore), toute interprétation d'événement du monde se fait par analogie, par rapprochement, par adaptation d'un autre cadre de référence, par torsion d'interprétations déjà là. (Voir la figure 3.5).



FIGURE 3.5 – *Qu'est-ce qui est spécial, amusant, dans cette scène ?* (Merci à Louis Gacogne, du LIP6 et de l'ENSIIE, pour cette photo et cette charade).

D'un autre côté, elle est certainement liée à l'induction. Il s'agit dans les deux cas d'une forme d'inférence non logiquement valide, ampliative, en ceci qu'elle calcule des conclusions qui vont au-delà de ce que la déduction logique autorise.

L'analogie est difficile à définir, à cerner et donc à modéliser (voir [Gen83, Hof85, Gen89, FFG89, HKT89, CG91, Mit93, FKGL94, KLD94, Bur95, Fre95, Hof95, HGK], pour une sélection de recherches). Elle met potentiellement en jeu un ensemble vaste et complexe de connais-

sances⁶⁵, et, de ce fait, il est ardu de reproduire les analogies élaborées par des sujets humains. Il est périlleux d'étudier l'analogie. Pour les psychologues, les modèles des informaticiens sont toujours trop restrictifs et souvent incapables de simuler complètement des analogies humaines. Pour les informaticiens, les théories et modèles élaborés sont incapables jusqu'ici d'être utiles dans des applications, sont quasi impossibles à évaluer objectivement et sont suspects d'incorporer trop de notions et principes issues des sciences humaines, par nature « mous », pour être susceptibles d'une vraie théorie.

Il était cependant impossible de résister à chercher à mieux comprendre cette forme si importante d'inférence dont la phénoménologie est si riche et intéressante. Ma ligne de vie dans cette étude a toujours été de chercher des principes fondamentaux qui puissent rendre compte de l'analogie, en étant persuadé que ces principes devaient être compatibles avec la théorie de l'induction, dans la mesure où l'analogie peut être considérée comme un cas limite de l'induction dans lequel on ne dispose que d'un exemple et d'un cas test.

Après avoir passé des centaines d'heures, sinon des milliers, à étudier l'analogie, je ne pense pas avoir atteint le modèle définitif qui rende compte de toutes ses propriétés, loin de là, mais je crois que mon travail [Cor94c, Cor94b, Cor94a, Cor96a, Cor96b, CAB98] peut stimuler la réflexion. J'en donne ici les lignes directrices.

3.3.1 Phénoménologie de l'analogie

Une décision importante à prendre lors de l'étude de l'analogie est le monde dans lequel on fera des expériences, dans lequel on illustrera les problèmes et les solutions produites par le modèle que l'on propose. En dehors de la cohérence de la théorie proposée, interne et externe par les liens avec d'autres théories, le critère d'évaluation d'un modèle de l'analogie réside en effet dans la comparaison avec les analogies que les sujets humains trouvent valides.

Le problème est compliqué. Il faut à la fois que le monde expérimental soit suffisamment riche et complexe pour qu'il permette de tester de manière satisfaisante le modèle, tout en restant cependant suffisamment limité afin qu'il soit possible de coder dans le système les connaissances relatives à ce monde. La tentation est grande de faire comme si le modèle s'appliquait à l'« Univers », par la convocation dans l'esprit des examinateurs de concepts naturels (comme le système solaire ou le modèle de l'« atome de Bohr » [Gen89]) alors que les connaissances codées dans la machine sont extrêmement pauvres et contraintes. Les psychologues choisissent souvent des champs expérimentaux beaucoup mieux délimités dans lesquels ils peuvent tester leurs hypothèses, mais ils ne proposent généralement pas de modèle informatique complet. Douglas Hofstadter s'est fait le champion de micro-mondes [Hof85, Hof95], tels le « monde des lettres », dans lesquels l'ensemble des connaissances à coder est limité et contrôlable tout en permettant des expériences complexes et représentatives des raisonnements d'analogie de notre expérience quotidienne.

L'inconvénient de l'utilisation des micro-mondes est qu'ils semblent éloignés des applications et de la vie réelle, et suspects de simplifications abusives. Je pense cependant, comme beaucoup d'autres, que les micro-mondes proposés par Hofstadter sont d'excellents champs d'expérience. C'est pourquoi je les ai choisis comme banc d'essai pour mes réflexions et mes modèles.

Le cas typique d'un problème d'analogie dans ce micro-monde est celui de la figure 3.6. Un *cas source* est fourni sous la forme $x \rightarrow y$ (ici $abc \rightarrow abd$) et un *cas cible* énigme ressemble à une question $x' \rightarrow ?$ (ici $ijjkkk \rightarrow ?$). Le problème est de compléter le cas cible en remplaçant $?$ par une chaîne de caractères en faisant appel à une analogie avec le cas source.

65. Pensez aux connaissances impliquées dans le processus interprétatif de la figure 3.5.

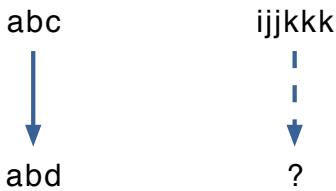


FIGURE 3.6 – Un exemple de problème analogique.

Ce problème est extraordinairement compliqué. Quel est le lien entre le cas source et le cas cible ? Quelle ressemblance entre `abc` et `ijjkkk` ? Quand bien même en trouverait-on, à quoi cela nous servirait-il puisque l'on ne sait pas quelle est la relation de dépendance utilisée pour produire la chaîne `abd` à partir de la chaîne `abc`. Une infinité de règles de réécriture peuvent expliquer cette transformation :

1. Remplacer `abc` par `abd`
2. Remplacer `c` par `d` dans les chaînes de caractères
3. Remplacer le *dernier élément* d'une chaîne par son *successeur*
4. ...

Pourquoi en préférer une plutôt qu'une autre ? Et supposons que l'on choisisse la troisième règle, comment pourrions-nous l'appliquer au cas cible ? Il faut une traduction et une interprétation : comment identifie-t-on un *élément* ? Le *dernier* ? Que signifie la *relation de succession* invoquée ?

Il est clair que, posé en ces termes, l'analogie est un problème extraordinairement sous-constraint. Avant de décrire mon approche pour mieux poser ce problème, je vais d'abord exhiber un ensemble de propriétés que semble posséder l'analogie. Pour ce faire, je fais appel à la capacité d'inférence analogique du lecteur. Car si il apparaît difficile de rendre compte de l'analogie et de résoudre de manière algorithmique des petits problèmes comme celui ci-dessus, ceux-ci ne posent le plus souvent que peu de difficulté aux sujets humains, et il existe généralement un très fort agrément sur ce qui est une bonne réponse et ce qui est « tiré par les cheveux » (si j'ose encore une métaphore).

Quelques propriétés de l'analogie

1. **L'analogie est non symétrique**, contrairement à ce qui est généralement considéré comme une propriété fondamentale et indiscutable.
Il est possible d'avoir : $(x \rightarrow y) \nleftrightarrow (x' \rightarrow y')$ (à lire comme : « par analogie au cas source $x \rightarrow y$, le cas cible $x' \rightarrow ?$ se complète par y' »)⁶⁶, et d'avoir réciproquement : $(x' \rightarrow y') \nleftrightarrow (x \rightarrow y'')$ où $y'' \neq y$.
2. **L'analogie est une transformation non linéaire** : une petite modification de la source peut entraîner une profonde réinterprétation de la cible.
3. **L'analogie est une transformation qui dépend du chemin suivi**.
À partir de : $(x \rightarrow y) \nleftrightarrow (x' \rightarrow y')$ et $(x' \rightarrow y') \nleftrightarrow (x'' \rightarrow y'')$, on peut avoir $(x \rightarrow y) \nleftrightarrow (x'' \rightarrow y''')$ avec $y'' \neq y'''$.

Des manifestations de ces propriétés sont illustrées dans la figure 3.7.

⁶⁶. J'utilise le symbole \nleftrightarrow pour souligner que le raisonnement analogique, bien qu'orienté de la source vers la cible, implique aussi un raisonnement bi-directionnel comme nous verrons.

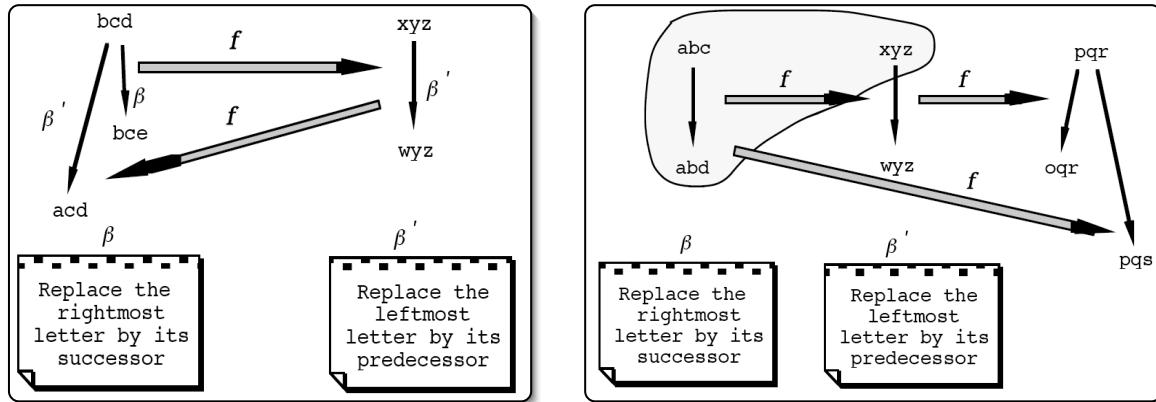


FIGURE 3.7 – Exemples illustrant la non symétrie de l'analogie (à gauche), et la dépendance sur le chemin suivi (à droite).

3.3.2 Analogie et induction

Il est également important de réaliser que l'analogie n'est pas un mécanisme de généralisation, ni même implique une généralisation dans son fonctionnement.

L'exemple suivant l'indique. Soient le cas source : $(abc \rightarrow abd)$ et le cas cible incomplet : $(xyz \rightarrow ?)$. En supposant que l'alphabet muni de la relation de succession classique ne soit pas un torre (i.e. a ne suit pas z), alors la réponse favorisée par les sujets humains est : $(xyz \rightarrow wyz)$. L'interprétation du cas source est : « remplacer la dernière lettre par son successeur », et celle du cas cible : « remplacer la première lettre par son prédécesseur ». Il n'y a pas d'inférence d'une règle générale, mais découverte d'une loi pour le cas source qui peut être traduite pour rendre compte du cas cible et le compléter.

Pourtant, il est tentant de voir l'analogie comme un cas extrême sur un spectre partant de l'induction. Dans l'induction, on dispose d'une collection de cas sources indépendamment et identiquement distribués (i.i.d.) et on suppose, par l'intermédiaire d'une espérance de risque, l'existence potentielle d'une collection de cas cibles également i.i.d. Dans l'analogie, il n'y a qu'un cas source et qu'un cas cible dont on connaît les « coordonnées » (x').

Dans l'induction, on cherche une règle générale qui explique au mieux les cas sources et qui permette de compléter les cas cibles. Le critère de performance implique une *fonction de coût* sur les réponses. L'induction devient un problème « bien posé » lorsque l'on choisit un *critère inductif* (e.g. la minimisation d'un risque empirique pénalisé) et un *espace d'hypothèses* \mathcal{H} (ou une hiérarchie d'espaces d'hypothèses).

Comme pour l'induction, il faut identifier les contraintes qui rendent le problème de l'analogie bien posé. Il est également souhaitable que ces contraintes soient compatibles avec celles qui définissent l'induction.

C'est là le défi que je me posais dans les années quatre-vingt-dix. Je n'étais d'ailleurs pas le premier. Stuart Russell avait déjà abordé l'analogie sous cet angle [Rus87, Rus89] dans sa thèse. Plus récemment, Laurent Miclet et Arnaud Delhay [MBD05], ainsi que François Yvon et Nicolas Stroppa [SY05] ont posé le problème de l'analogie dans des termes proches. Les principes qui les inspirent : ramener l'analogie à une déduction dans le cas de Russel⁶⁷, ou à une équation dans le cas de mes collègues français, conduisent cependant à des solutions qui limitent le champ du

67. Voir mon analyse de la proposition de Russell dans [Cor96a].

raisonnement analogique et ne leur permet pas, en particulier, de rendre compte des propriétés mentionnées plus haut.

3.3.3 Une formalisation de l'analogie

Il n'est pas facile d'adapter les principes de l'induction à l'analogie.

Dans l'induction, on fait l'hypothèse que l'échantillon d'apprentissage est représentatif de la distribution de probabilité des cas (i.e. des exemples), ce qui autorise, sous les conditions énoncées par la théorie statistique de l'apprentissage, l'extrapolation du risque empiriquement mesuré à une espérance de risque. Ne disposant que d'un cas source en analogie, cela devient impossible.

Dans l'induction, le risque empirique (éventuellement pénalisé) permet d'évaluer et de comparer les mérites des différentes hypothèses⁶⁸. Dans l'analogie, comme nous l'avons vu, une infinité d'hypothèses permettent de rendre compte du cas source. Pour toutes ces hypothèses donc, le risque empirique est nul.

En induction, la théorie statistique prescrit l'utilisation d'espaces d'hypothèses de capacité limitée, ce qui est une condition suffisante pour contrôler le lien entre risque empirique et espérance de risque. Pourquoi aurait-on une condition de cet ordre en analogie ? Quel critère peut mesurer la valeur de l'interprétation du cas source et de sa traduction dans le cas cible ?

Je ne sais pas comment répondre à ces questions en faisant la démarche d'aller de l'induction vers l'analogie. Mais j'ai fait le chemin inverse. J'ai proposé, dans les années quatre-vingt-dix, une formalisation raisonnable de l'analogie et contrôlé ensuite qu'elle était compatible avec la formalisation de l'induction.

Les principes sous-jacents à la formalisation proposée sont simples.

1. On suppose que la source est entièrement utile pour l'interprétation et la résolution de l'analogie. Il faut donc l'interpréter entièrement.
2. Comme en induction, on suppose que l'on dispose d'un code (dictionnaire de primitives ou de fonctions de base) pour décrire le cas source, de même qu'un code (pas nécessairement le même) pour décrire le cas cible.
3. On suppose que les éléments de chaque code sont associés à un coût (qui peut être la complexité algorithmique, le coût de mobilisation cognitive, ...).
4. On suppose que la traduction des éléments d'un code dans un autre code est associée à un coût.
5. On dit que le coût d'une analogie est la somme du coût de codage de la source, du coût de traduction des éléments de codage de la source utiles pour le codage de la cible, et du coût de complémentation de la cible étant donnés les éléments précédents (voir la formule 3.13 ci-dessous).
6. Les analogies candidates sont d'autant meilleures que leur coût est faible.

Proposition 3.1 (Coût associé à une analogie)

Le coût associé à une analogie est donné par la formule :

$$Coût((x \rightarrow y), x') = K(C_S) + K(x|C_S) + K(h_x|C_S) + K(C_T|C_S) + K(x'|C_T) \quad (3.13)$$

où $(x \rightarrow y)$ et x' dénotent respectivement le cas source et le cas cible incomplet, C_S et C_T les codes (ou fonctions de base) utilisés pour décrire la source et la cible, h_x et $h_{x'}$, les hypothèses

68. En supposant que l'on ne soit pas dans le cas de transition de phase (voir section 2.1).

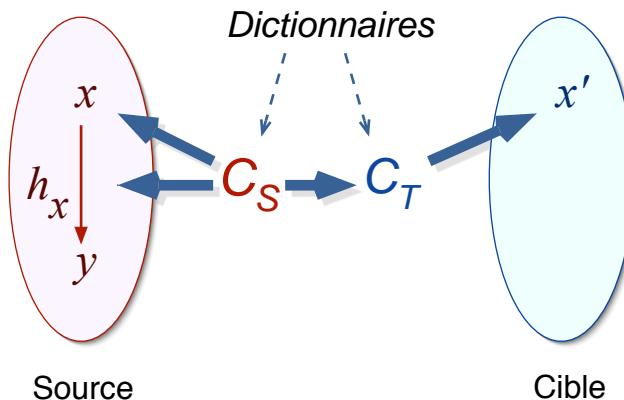


FIGURE 3.8 – Éléments en jeu dans le calcul du coût d'une analogie.

choisies rendant compte respectivement de la source ($x \rightarrow y$) et de la cible x' . $K(\cdot)$ dénote la fonction coût, et $K(x|y)$ le coût de description (ou de traduction) de x connaissant y .

La figure 3.8 souligne les éléments en jeu dans la formule 3.13.

D'un point de vue pratique, l'application de cette formule pour trouver les meilleures analogies implique que l'on se donne *a priori* des coûts associés aux éléments des codes et des coûts de traduction. De manière similaire à l'application du principe de minimisation de la longueur de description (MDLP : *Minimum Description Length Principle*) [LV97, GV03, GMP], cela revient essentiellement à se donner une distribution de probabilité *a priori* sur les éléments des codes de description. Cela peut paraître permettre l'entrée d'une grande part d'arbitraire dans l'évaluation des analogies que l'on peut ainsi biaiser de manière considérable. Mais, c'est aussi ce qui permet de rendre compte des préconceptions et des effets de l'histoire des sujets. Par ailleurs, nous avons essayé de réduire au maximum cet arbitraire en montrant comment on peut construire un dictionnaire et des codes de manière systématique en cherchant à se conformer le plus possible aux prescriptions de la théorie des codes et de la complexité de Kolmogorov (voir la thèse de Jacques Ales-Bianchetti [AB00]).

Des illustrations de cette approche sont détaillées en particulier dans [Cor96a, CAB98]. Elles montrent comment calculer effectivement le coût de diverses analogies candidates et prédisent les analogies qui sont préférées par les sujets humains sur divers problèmes.

Il est important de noter que la formulation proposée rend compte de l'asymétrie du raisonnement analogique, ainsi que de sa non linéarité et de la dépendance sur le chemin suivi, les trois propriétés décrites plus haut et qui échappent à la plupart des modélisations concurrentes de l'analogie.

Mais, par ailleurs, et de manière encore plus fondamentale sans doute, cette modélisation est compatible avec la théorie de l'induction.

3.3.4 Analogie et induction : les deux extrémités d'un spectre

Selon le modèle proposé, l'analogie implique le coût de l'interprétation de la source et conjointement celui de la cible, mais en ayant le droit de réaliser des économies en partageant et/ou en traduisant une partie des deux interprétations.

Par ailleurs, nous avons ici une nouvelle notion, inconnue en induction, d'une fonction hypo-

3.3. L'analogie : construction dynamique d'une corrélation orientée

thèse dont la définition dépend du point d'application : h_x . Pour une même source, l'interprétation et sa traduction par analogie en un cas cible dépend des coordonnées de ce cas.

Peut-on concilier ces nouveaux concepts avec la perspective classique sur l'induction ?

Il est clair que le principe de chercher une interprétation « économique » des cas, c'est-à-dire prise dans une classe de capacité limitée d'hypothèses, est conforme aux prescriptions de la théorie statistique de l'induction. Ce qui est nouveau en analogie, c'est que l'on cherche en même temps à contraindre au maximum le coût de traduction d'une interprétation dans l'autre $K(C_T|C_S)$, c'est-à-dire, là aussi, à puiser dans un espace de transformations de capacité limitée. En d'autres termes, on est prêt à travailler dans un espace non euclidien, avec un repère dépendant des coordonnées, mais cependant en limitant le « gauchissement »⁶⁹.

Mais que donne la formule d'optimisation 3.13 dans le cas de m sources et de p cibles potentielles ?

$$\begin{aligned} \text{Coût} = & K(C_S) + \sum_{i=1}^m \left[K(C_{S_i}|C_S) + K(\mathbf{x}_i|C_{S_i}) + K(h_{x_i}|C_{S_i}) \right] \\ & + K(C_T|C_S) + \sum_{j=1}^p \left[K(C_{T_j}|C_T) + K(\mathbf{x}_j|C_{T_j}) \right] \end{aligned} \quad (3.14)$$

Cette formule traduit le fait que chaque cas source est interprété grâce à un code local C_{S_i} , qui lui-même est codé en s'aidant d'un code source général C_S . Le même processus de codage a lieu du côté des cibles en s'aidant d'un code cible général C_T qui lui-même est décrit en s'aidant du code source C_S .

Mais si l'on suppose que la fonction de coût K vérifie l'inégalité triangulaire :

$$K(z|x) \leq K(x|y) + K(y|z) \quad \forall x, y, z$$

alors l'optimisation de la formule 3.14 conduit à l'utilisation d'un seul code source et d'un seul code cible :

$$\text{Coût} = K(C_S) + \sum_{i=1}^m \left[K(\mathbf{x}_i|C_S) + K(h_{x_i}|C_S) \right] + K(C_T|C_S) + \sum_{j=1}^p \left[K(\mathbf{x}_j|C_T) \right] \quad (3.15)$$

Si maintenant, on admet, comme c'est fait implicitement dans la théorie de l'induction que le même code est utilisé pour les sources et pour les cibles, l'optimisation porte sur l'expression :

$$\text{Coût} = K(C_S) + \sum_{i=1}^m \left[K(\mathbf{x}_i|C_S) + K(h_{x_i}|C_S) \right] + \sum_{j=1}^p \left[K(\mathbf{x}_j|C_S) \right] \quad (3.16)$$

Il y a maintenant un pas essentiel qui fait la différence entre analogie et induction. En induction, on impose la même hypothèse h sur tout l'espace, d'où la formule :

$$\text{Coût} = K(C_S) + K(h|C_S) + \sum_{i=1}^m \left[K(\mathbf{x}_i|C_S) + K(\mathbf{y}_i|\mathbf{x}_i, h) \right] + \sum_{j=1}^p \left[K(\mathbf{x}_j|C_S) \right] \quad (3.17)$$

69. La *dérivée covariante* en géométrie non euclidienne. Notons qu'ici, comme en géométrie non euclidienne, la dérivée peut dépendre aussi du chemin suivi. (Pour ceux qui sont intéressés, il y a maintenant énormément d'excellents livres d'introduction sur les espaces non euclidiens devenus si essentiels en physique. Je citerais seulement [JS98, D'I99, Tal00, Car04]).

où le terme $K(y_i|x_i, h)$ exprime le fait que si h est imposé alors il est possible que $h(x_i) \neq y_i$ et il faut alors payer le raccord.

On retrouve ici l'idée du principe de minimisation de la longueur de description :

$$h^* = \operatorname{Argmin}_{h \in \mathcal{H}} \left[K(h) + \sum_{i=1}^m K(y_i|x_i, h) \right] \quad (3.18)$$

dans lequel la meilleure hypothèse est celle dont le coût de description ajouté au coût de description des données étant donnée l'hypothèse est minimal.

Le fait d'imposer l'unicité de l'hypothèse sur l'espace \mathcal{X} des données est évidemment une **hypothèse majeure qui fait de l'induction un cas particulier de l'analogie**. Rien ne justifie *a priori* cette hypothèse.

Finalement, comme les sources et les cibles sont les mêmes pour toutes les solutions envisagées, les facteurs $K(x_i|C_S)$ sont communs à toutes et peuvent être ignorés dans l'optimisation.

$$\text{Coût}(h) = K(C_S) + K(h|C_S) + \sum_{i=1}^m K(y_i|x_i, h) \quad (3.19)$$

c'est-à-dire que l'on ignore le coût d'adaptation à chaque point x de l'espace.

On retrouve ici l'expression du principe de minimisation de la longueur de description.

Nous avons ainsi montré que notre formalisation de l'analogie est compatible avec celle classiquement utilisée pour rendre compte de l'induction. Qui plus est, l'induction apparaît comme un cas particulier de l'analogie.

3.4 Publications, projets et stages liés à ces directions de recherche

Les recherches décrites dans ce chapitre ont bénéficié des **collaborations** suivantes :

INDANA (2001-2004) : « Étude du risque cardio-vasculaire »

Responsable : Marie-Christine Jaulet (INSERM)

Participants : Florence d'Alché-Buc (LIP6), Isabelle Colombet (INSERM), Antoine Cornuéjols (LRI), Fabien Torre (LIFL), Rémi Gilleron, (LIFL), Yves Grandvalet (UTC, Compiègne), François Gueyffier (SPC, Lyon), Christophe Marsala (LIP6), Mario Ota (INSERM), Michèle Sebag (LRI).

Collaboration avec l'équipe d'Elena Marchiori de la Vrije Universiteit d'Amsterdam (2003-2004)

Participants : Antoine Cornuéjols (LRI), K. Jong (VUA), Elena Marchiori (VUA), Jérémie Mary (LRI), Michèle Sebag (LRI).

Projet Biogen n°74 (2002-2004) : « Analyse du transcriptome »

Responsable : Marie Dutreix (Institut Curie, Orsay)

Participants : Jean-Paul Comet (Univ. Evry), Antoine Cornuéjols (LRI), Ch. Froidevaux (LRI).

Groupe *Consensus*, Génopole Evry : « Groupe de travail »

Les **stagiaires et doctorants** encadrés ayant travaillé sous ma direction sur ces sujets sont :

- Jacques Ales-Bianchetti (doctorant : 1997 - 2000)
- Jérémie Mary (doctorant : 2002 - 2005)

Les **publications** concernées auxquelles j'ai participé sont :

1. (A-B00) J. Ales-Bianchetti (sous la direction d'A. Cornuéjols et d'Y. Kodratoff). *Le raisonnement par analogie. Une unification des modèles cognitifs et des théories de l'induction pour l'étude du raisonnement par analogie* Thèse de doctorat, Laboratoire de Recherche en Informatique, Université de Paris-Sud, France, Juillet 2000.
2. (Col02) I. Colombet. *Aspects méthodologiques de la prédiction du risque cardiovasculaire : apports de l'apprentissage automatique*, Thèse de doctorat, Paris-5, Juin 2002.
3. (Cor94a) A. Cornuéjols. « Induction from one example and statistics : analogy as a minimization principle », Workshop on *Machine Learning and Statistics* (ECML-94), Catanes, Italy, 1994.
4. (Cor94b) A. Cornuéjols. « Analogy as description minimization principle », Workshop on *Applications of Descriptive Complexity to Inductive, Statistical and Visual Inference* (ICML-COLT-94), Rutgers, USA, 1994.
5. (Cor94c) A. Cornuéjols. « Analogy as a minimization principle », Dagstuhl Seminar *Theory and Praxis of Machine Learning*, Dagstuhl, Germany, 1994.
6. (Cor96a) A. Cornuéjols. « Analogie, principe d'économie et complexité algorithmique », *Journées Francophones d'Apprentissage* (JFA-96), Sète, France, 1996, pp.233-247.
7. (Cor96b) A. Cornuéjols. « Analogy as a minimization of description length », in *Machine Learning and Statistics : The Interface* (Eds. Nakhaeizadeh, G. and Taylor, C.). John Wiley and Sons, 1996, pp. 321-335.
8. (CAB98) A. Cornuéjols and J. Ales-Bianchetti. « Analogy and Induction : which (missing) link ? », Workshop *Advances in Analogy research : Integration of theory ans data from cognitive, computational and neural sciences*, Sofia, Bulgaria, 1998. New Bulgarian University Series (Eds. K. Holyoak, D. Gentner and B. Kokinov), pp. 365-372.
9. (CFM05) A. Cornuéjols, Ch. Froidevaux and J. Mary. « Comparing and combining feature estimation methods for the analysis of microarray data », *JOBIM-05 : Journées Ouvertes Biologie Informatique Mathématiques*, Lyon, France, 2005.
10. (DCCF04) M. Dutreix, J-P. Comet, A. Cornuéjols and Ch. Froidevaux. « Determination of cellular drug targets : searching for functional information in the jungle of microarrays data », *Current Trends in Drug Discovery Research* (CTDDR-04), India, 2004.
11. (JMCMS04) K. Jong, J. Mary, A. Cornuéjols, E. Marchiori and M. Sebag. « Ensemble feature ranking », *Principles of Knowledge Discovery in Databases* (PKDD-04), Pisa, Italy, 2004 Springer-Verlag, LNAI-3202, 267-278.
12. (MMCCFD03) J. Mary, G. Mercier, J-P. Comet, A. Cornuéjols, Ch. Froidevaux and M. Dutreix. « An attribute estimation technique for the analysis of microarray data », Proceedings of the *Dieppe School on Modelling and Simulation of Biological processes in the Context of Genomics*, (Eds. Ph. Amar, F. Képès, V. Norris and P. Tracqui), Publisher Frontier Group, 2003, pp.69-77.

13. (MMCCFD04) J. Mary, G. Mercier, J-P. Comet, A. Cornuéjols, Ch. Froidevaux and M. Dutreix. « Utilisation d'une méthode d'estimation d'attributs pour l'analyse du transcriptome de cellules de levures exposées à de faibles doses de radiation », *Informatique pour l'analyse du transcriptome* (Eds. J-F. Boulicaut and O. Gandrillon), Hermès, 2004, pp.189-205.
14. (MBMACCFD04) G. Mercier, N. Berthault, J. Mary, A. Antoniadis, J-P. Comet, A. Cornuéjols, Ch. Froidevaux and M. Dutreix. « Biological detection of low radiation by combining results of two analysis methods », *Nucleic Acids Research* (NAR), vol.32, No.1, 1-8 (2004).

4

Trajectoires d'apprentissage et circulation d'information

Lorsque j'enseigne l'apprentissage artificiel, il y a toujours un(e) étudiant(e) moins timide que les autres, qui finit par faire part de sa surprise. *Comment ? L'apprentissage serait donc une science de phénomènes statiques ? ! Pas de notions de trajectoire, d'ajustement graduel, de progression ? Pas de mouvement, pas de surprise, d'effet aha, pas de vie ? !* Je ne peux que soupirer. Pas interloqué, non, mais embarrassé.

Et en effet, si j'examine le contenu de mon cours (ou de notre ouvrage [CM02]) qui ne diffère guère de ce qui est enseigné ailleurs, il n'y est presque exclusivement question que de problèmes dans lesquels un système reçoit une collection d'entrées et produit une sortie, et recommence à zéro si un autre jeu de données est disponible. Il est rare que l'on envisage un apprentissage avec des données arrivant au fur et à mesure du temps et traitées en-ligne (l'exception la plus notable étant celle de l'apprentissage par renforcement). Et même dans ces cas là, on suppose que l'environnement reste invariable, c'est-à-dire avec une distribution des exemples et une fonction de coût constante.

Or l'apprentissage naturel est caractérisé par des phénomènes d'adaptation continue, d'effets de facilitation (*Priming effects*, c'est-à-dire d'effets de l'ordre des entrées⁷⁰), ainsi que d'effets de rythme : on sait empiriquement qu'il vaut mieux qu'un certain temps, ni trop court, ni trop long, se passe entre les « leçons ». L'apprentissage naturel est le royaume d'effets dynamiques complexes, et ce d'autant plus que les agents cognitifs sont toujours en situation hors d'équilibre.

Comment peut-on passer à côté de toute cette richesse de comportements ? Comment ne pas essayer de mieux les comprendre, de mieux les prédire et d'en tirer profit ? Comment, à l'horizon, ne pas essayer de voir, par exemple, si les fameux stades piagétiens [Pia37, Pia75, PP79] seraient des étapes nécessaires dans l'apprentissage du monde ?

Il y a là, pour moi, depuis les débuts de ma thèse [Cor89a, Cor89b], motif d'une inquiétude lancinante et stimulante. Un ensemble d'interrogations et d'invitations à tout essayer pour comprendre la dynamique de l'apprentissage.

4.1 Vers une étude de la dynamique de l'apprentissage

On peut étudier la dynamique de l'apprentissage selon plusieurs perspectives, dont le point de vue plus informatique théorique de l'apprentissage en-ligne (voir section 4.1.4), ou celui de

70. Tout enseignant digne de ce nom passe beaucoup de temps à choisir l'ordre de présentation des notions le plus pédagogique.

l'automatique et de la théorie du contrôle⁷¹. Pour un certain nombre de raisons qui vont être exposées, je trouve le cadre de la dynamique en physique également intéressant à explorer.

Dans ce cadre, les notions d'espace d'états et de trajectoire sont centraux.

4.1.1 Apprentissage et dynamique des systèmes

On supposera donc qu'un système d'apprentissage est caractérisé à un instant t par son état noté $e(t)$ appartenant à un espace d'états \mathcal{E} .

Dans cet espace, l'évolution d'un système est décrite par une *trajectoire* γ qui représente la séquence des états traversés par le système.

Un *repère* est un ensemble de fonctions de base sur lequel est défini l'état d'un système. Par exemple, l'état d'un réseau de neurones à n connexions peut être décrit par un vecteur dans un repère constitué de n vecteurs de base, les poids des connexions étant les coordonnées de ce vecteur. La description des états possibles d'un système peut soit s'effectuer dans un repère global, commun à tous les états, soit s'effectuer dans des repères locaux, comme c'est le cas dans le raisonnement par analogie (ou en relativité générale).

Si nous supposons pour le moment que l'espace des états est continu, un changement d'état infinitésimal peut être associé à un vecteur ξ .

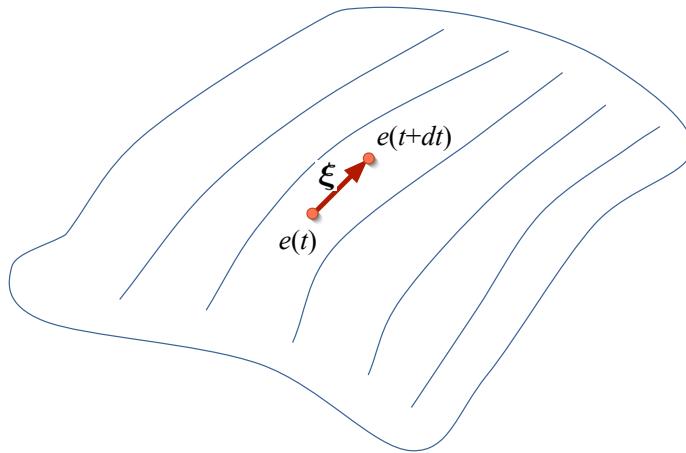


FIGURE 4.1 – Changement d'état infinitésimal dans l'espace des états associé au vecteur ξ .

À chaque état e peuvent être associées un ensemble de fonctions scalaires correspondant à une fonction d'erreur ou de risque pour cet état. Par exemple, on pourrait associer à un état le risque empirique mesuré en cet état sur un échantillon d'apprentissage \mathcal{S} , que nous noterons $\Phi_{\mathcal{S}}(e)$ (c'est, dans le cas classique de l'induction, l'équivalent du risque empirique $R_{\text{Emp}}(h)$). Mais, dans un contexte d'apprentissage en-ligne, on pourrait être intéressé par la fonction de perte instantanée $\Phi_z(e)$ où z est l'entrée instantanée, qui peut être un couple (x, u) où x est un exemple et u son étiquette, ou sa réponse désirée.

On a alors $d\Phi$ qui est une 1-forme (ici une fonction *gradient*). Combinée à un vecteur ξ dans un produit scalaire $d\phi \cdot \xi$, on obtient la quantité dont varie Φ selon le vecteur ξ .

La **dynamique** d'un système d'apprentissage vient alors du fait que le système calcule son déplacement instantané ξ en fonction du gradient $d\Phi$, ce qui en retour modifie son état et donc

⁷¹. Qui commence à prendre sa place dans les mathématiques financières et les études d'évolution de cours de bourse.

son gradient, etc.

L'étude de la dynamique des apprentissages est donc liée à celle des trajectoires du système dans l'espace des états.

Il est possible d'envisager que les trajectoires sont dues à l'exercice de forces en chaque point (état) qui sont liées au gradient de la fonction d'erreur Φ . Un autre point de vue est celui introduit par Lagrange (1736-1813) selon lequel à chaque trajectoire peut être associée une *action*. La trajectoire effectivement suivie par le système étant alors celle qui rend l'action stationnaire (minimale classiquement, d'où le nom de *principe de moindre action*).

On remplace alors une description locale de la dynamique par une description globale, au niveau des trajectoires directement. Nous verrons ce que cela peut apporter. Dans la formulation lagrangienne, en supposant que l'état d'un système dépende de n degrés de liberté, on représente le système dans un espace à $2n$ coordonnées en ajoutant aux n coordonnées de position, n coordonnées de vitesse ou dérivées par rapport au temps. La trajectoire du système est alors considérée dans cet « espace de phase ».

L'expression de l'action dans le cas de la mécanique du point matériel dans un champ conservatif U fait intervenir une fonction appelée *Lagrangien*⁷² donnée dans ce cas par la formule : $\mathcal{L}(e) = E(e) - U(e)$, où $E(e)$ l'énergie cinétique du point en e , et $U(e)$ est la valeur du potentiel en e . L'action d'une trajectoire entre deux points e_1 et e_2 dans l'espace des phases est : $\mathcal{A} = \int_{e_1}^{e_2} \mathcal{L}(e) de$.

Les équations d'Euler-Lagrange et celles d'Hamilton expriment l'équivalence entre la description locale du mouvement à l'aide du concept de force et la description globale faisant appel au Lagrangien.

4.1.2 Symétries et propagation d'information

Ce qui est intéressant, c'est que le Lagrangien associé à un système doit obéir à toutes les propriétés d'invariance dont on pense qu'elles sont satisfaites par le système⁷³.

L'un des grands mérites de la formulation lagrangienne est de permettre d'établir mathématiquement⁷⁴, grâce au théorème d'Emmy Noether (1882-1935), la correspondance entre *symétrie* du mouvement (par rapport à des transformations des équations) et lois de *conservation de quantité physique*⁷⁵.

La puissance du raisonnement sous-jacent à la formulation lagrangienne, la capacité de cette formulation à prendre en compte les propriétés de symétrie, permettent de généraliser le principe de moindre action à une grande variété de systèmes dynamiques⁷⁶.

72. Attention, ce Lagrangien n'est pas le même que celui qui intervient dans les problèmes d'optimisation sous contraintes et qui est devenu familier depuis l'émergence des méthodes à noyau dont les Séparateurs à Vastes Marges.

73. Ainsi dans le cas de la mécanique classique, les équations sont invariantes par un changement d'origine (les lois de la physique sont les mêmes dans mon labo et dans celui de mon collègue à Rio de Janeiro), par un changement de l'origine du temps, et par une rotation du référentiel.

74. Sous des conditions de différentiabilité qui ne s'appliquent malheureusement pas aux systèmes discrets que l'on retrouve souvent en apprentissage.

75. Ainsi, aux invariances des équations de la mécanique classique par rapport respectivement aux changements d'*origine du temps*, changements d'*origine de l'espace* et *rotations*, correspondent respectivement les quantités conservées d'*énergie*, de *quantité de mouvement* et de *moment cinétique*.

76. L'une d'entre elles concerne l'électromagnétisme. On peut en effet montrer que les équations de Maxwell découlent d'un principe de moindre action. On peut écrire un Lagrangien de Maxwell en fonction du champ électromagnétique qui est traité comme un système dynamique dépendant d'un nombre infini de degrés de liberté, qui fait intervenir une « énergie cinétique » de propagation du champ (le potentiel vecteur A).

Il est également possible, grâce à cette formulation, de relier des symétries locales du Lagrangien, lorsqu'on lui fait subir des transformations localement et non plus globalement sur tout l'espace, à des couplages entre grandeurs physique. C'est ainsi que ces théories de jauge locale sont devenues centrales en physique (en particulier en physique des particules) et qu'elles permettent de rendre compte des couplages par émission/réception de particules qui sont de cette manière prédictes par la théorie (parfois longtemps avant d'avoir été observées).

Devant le succès de ce type d'approche et devant sa puissance de prédiction, on peut se demander si il ne serait pas possible de chercher à l'adapter au cas de l'étude des systèmes d'apprentissage et de leur dynamique.

Avant de passer à l'étude d'un type de symétrie qui concerne l'ordre des entrées en apprentissage en-ligne, il est important de réaliser que la découverte d'un Lagrangien permettant de décrire un système n'est pas facile. Pour le moment, personne ne connaît de recette ou de démarche systématique pour ce faire, et il s'agit beaucoup d'un art dans lequel l'intuition éduquée de l'expert est cruciale. Un petit exemple en donnera l'illustration.

Exemple 4.1

Pour des raisons de simplicité, nous nous plaçons ici dans le cas discret. Soit un système (d'apprentissage) minimal qui calcule en-ligne la moyenne des nombres x_i qui lui sont fournis séquentiellement. On a alors les équations :

$$\mu_1 = x_1 \tag{4.1}$$

$$\mu_i = \frac{(i-1)\mu_{i-1} + x_i}{i} \quad (\forall i > 1) \tag{4.2}$$

Il s'agit d'un système très simple qui présente la propriété d'être indépendant de l'ordre dans lequel lui sont présentées les entrées : le résultat final est invariant par rapport au groupe des permutations des entrées. Mais cela se fait au prix d'avoir à conserver en mémoire à la fois la moyenne courante μ_i , mais surtout le nombre i d'entrées déjà prises en compte.

Or concernant ce système si simple, deux questions sont non résolues :

1. *Quel est un lagrangien qui permettrait de rendre compte de sa trajectoire (e.g. sachant que la séquence d'entrées a été $\langle 3, 5, 7, 5 \rangle$, pouvoir dire que la trajectoire suivie est définie par la séquence des moyennes $[3, 4, 5, 4]$) ?*
2. *Comment montrer qu'il faut au moins passer une information scalaire d'un état à l'autre (ici le nombre i) pour assurer l'invariance par permutation de l'ordre des entrées ?*

Notons que l'on peut facilement traduire ce système dans un cas continu pour lequel on ne connaît pas plus les réponses à ces questions.

4.1.3 Le cas des systèmes d'apprentissage indépendants de l'ordre des entrées

Jusqu'à ces dernières années, très peu d'études ont été consacrées à l'apprentissage en-ligne ou incrémental. La section 4.1.4 présentera le point de vue de la théorie de l'apprentissage concernant ce cadre, mais en gros l'idée était, jusque récemment, que si un système était prêt à faire face au pire cas, il n'était pas besoin d'étudier des stratégies face à d'autres possibilités⁷⁷. Très peu d'algorithmes authentiquement incrémentaux ont été décrits, du moins en apprentissage supervisé, et souvent, le premier soin des concepteurs était de montrer que leur algorithme

⁷⁷. Réflexion meurrière d'Avrim Blum quand je lui parlais en 1997 de mon intérêt pour étudier des stratégies pour un enseignement optimal.

était indépendant de l'ordre des entrées (par exemple voir [Utg89] qui regrette que l'algorithme ID4 soit dépendant de l'ordre et qui propose ID5 [Utg94] pour avoir un équivalent incrémental de ID3). Ainsi donc la communauté en apprentissage, dans sa grande majorité, ignorait, voire évitait, une caractéristique fondamentale de l'apprentissage en-ligne et se concentrat sur l'apprentissage « batch » dans lequel les données sont disponibles d'un seul coup⁷⁸.

Pourtant, plusieurs raisons militent pour une étude de l'apprentissage en-ligne, et plus particulièrement des effets de l'ordre des entrées sur l'apprentissage :

- D'un point de vue pratique, l'apprentissage est de plus en plus vu comme un module devant s'intégrer dans des systèmes de « longue vie » dans lesquels il devient absurde et coûteux de redémarrer l'apprentissage à zéro à chaque fois que de nouvelles données sont disponibles.
- Par ailleurs, il est difficile de garantir qu'un apprentissage incrémental est équivalent à un apprentissage batch, c'est-à-dire donne le même résultat sur tout jeu de données d'apprentissage, quelque soit l'ordre dans lequel ces données sont présentées.
- Il devient donc intéressant d'étudier les phénomènes qui peuvent se produire en apprentissage incrémental, y compris les cas les pires ou les plus favorables pour un objectif donné.
- En fait, on peut même considérer que si l'ordre d'exposition des données influe sur le résultat de l'apprentissage, c'est qu'il correspond à une information supplémentaire fournie au système. Il est intéressant de se demander en quoi cette information est comparable à celle qui est fournie par les exemples.
- Finalement, il n'est pas interdit d'imaginer que ce type d'étude puisse contribuer à la modélisation d'apprentissages naturels et particulièrement celui des élèves en situation scolaire, ce qui a des implications évidentes.

Avant d'étudier toute la richesse de l'apprentissage incrémental, il est intéressant d'**étudier le cas limite dans lequel cet apprentissage est indépendant de l'ordre**.

Comme mentionnée plus haut, la symétrie par rapport aux permutations de l'ordre des entrées s'accompagne certainement d'une circulation d'information entre les états. Une illustration peut rendre cela plus clair. Supposons que l'on représente l'état du système à un instant donné par un point dans l'espace des états. L'effet d'une entrée z sur le système dans un état e conduit le système dans un état e' , ce qui peut se représenter comme un petit morceau de trajectoire $e - (z) \rightarrow e'$ entre e et e' .

Supposons alors deux trajectoires :

$$\gamma_1 : e_1 - (z_1) \rightarrow e_2 \dots - e_{i-1} - (z_i) \rightarrow e_i \dots - e_{j-1} - (z_j) \rightarrow e_j \dots - (z_{m-1}) \rightarrow e_m$$

$$\gamma_2 : e_1 - (z_1) \rightarrow e_2 \dots - e_{i-1} - (z_j) \rightarrow e'_i \dots - e'_{j-1} - (z_i) \rightarrow e_j \dots - (z_{m-1}) \rightarrow e_m$$

qui diffèrent par l'ordre des entrées : z_i et z_j étant intervertis (voir figure 4.2).

Si le système d'apprentissage est indépendant de l'ordre, alors, n'importe où sur la trajectoire γ_1 , après l'entrée z_i , il suffit de présenter l'entrée z_j pour que le système retrouve la fin de la trajectoire γ_2 . Cela n'est absolument pas trivial. D'un point de vue informel, cela signifie que le système conserve l'information suffisante pour que, après bifurcation, les deux trajectoires restent corrélées et conservent une « différence » juste égale à l'entrée z_j .

Dans un papier présenté à ECML-93 [Cor93a], j'attribuais les effets d'ordre à deux causes possibles, chacune suffisante, mais qui peuvent jouer simultanément :

78. Une exception à cet ostracisme a été celle du projet européen Learning in Humans and Machines (LHM) financé par l'European Science Foundation entre 1994 et 1997 et dont un des cinq groupes de travail avait pour but d'étudier précisément les effets de séquence en apprentissage (voir [LR06])

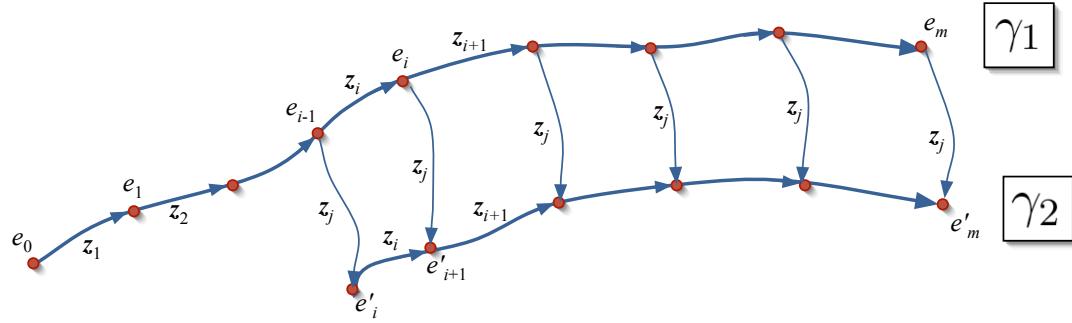


FIGURE 4.2 – Deux trajectoires d'apprentissage dans lesquelles on intervertit l'ordre des entrées z_i et z_j .

1. La *non optimalité* instantanée, c'est-à-dire le fait que le système ne soit pas capable de trouver l'optimum global de la fonction à optimiser à chaque instant. Par exemple, dans le cas d'optima multiples et d'une procédure d'optimisation par gradient, l'optimum trouvé à chaque instant dépend de l'histoire passée du système.
2. *L'oubli* d'information dépendant de l'histoire passée du système.

Me concentrant sur le deuxième point, j'étudiais la question de déterminer quelle information un système peut oublier sans introduire d'effets d'ordre.

Dans [Cor93a], je me plaçais dans le cas de l'apprentissage de concepts par des exemples positifs et négatifs. On a alors la notion d'espace des versions (voir la section 2.1), c'est-à-dire d'ensemble d'hypothèses cohérentes avec les exemples d'apprentissage. En supposant que cet espace est doté d'une relation de généralité, on peut représenter l'état du système par les bornes inférieures et supérieures du treillis des hypothèses cohérentes avec les données : le S -set et le G -set (voir [Mit82]). L'apprentissage consiste alors à mettre à jour ces bornes au fur et à mesure que les exemples sont disponibles. Il est crucial de noter que cet algorithme jouit d'une propriété de commutativité : son résultat est indépendant de l'ordre de présentation des données.

L'oubli d'information, dans ce cadre, peut naturellement être associé à l'oubli d'éléments du S -set et/ou du G -set, par exemple pour des raisons de trop grande complexité en espace⁷⁹. La question que je me posais devenait alors : quelle stratégie d'oubli, exercée au cours de l'histoire du système, peut garantir une indépendance sur l'ordre de présentation des données ? On notera qu'on ne demande pas ici que le résultat de l'apprentissage soit le même que celui d'un apprentissage sans oubli, mais juste qu'il soit indépendant de l'ordre de la séquence d'apprentissage, quelle que soit celle-ci.

La démonstration passe par un lemme intéressant, et capital ici, qui est que l'oubli d'éléments des bornes de l'espace des versions est équivalent à la fourniture d'exemples supplémentaires. Ainsi donc, paradoxalement à première vue, **l'oubli est équivalent à disposer de plus d'information**. De ce lemme découle le fait que l'on peut remplacer le problème difficile de l'étude d'une stratégie d'oubli dépendante de l'histoire du système par l'ajout d'exemples à l'algorithme qui, lui, est indépendant de l'ordre des exemples. De la sorte, la question sur la stratégie d'oubli est remplacée par celle du choix des exemples à ajouter, n'importe quand, dans la séquence d'apprentissage.

Il est alors facile de montrer que pour assurer l'indépendance sur l'ordre de présentation des

⁷⁹. Haussler en 1989, [Hau89] a montré que le S -set pouvait croître de manière exponentielle avec le nombre d'exemples positifs.

exemples, il faut que le système d'apprentissage, en choisissant les exemples supplémentaires, se comporte comme s'il connaissait à l'avance l'espace des versions à obtenir finalement. Il en découle que, dans ce cadre, **il n'existe pas de stratégie d'oubli permettant un apprentissage indépendant de l'ordre**, sauf les deux stratégies extrêmes de non oubli ou de tout oubli.

Pour résumer, dans cette étude on a pris un algorithme d'apprentissage incrémental intrinsèquement indépendant de l'ordre des entrées, l'algorithme d'élimination des candidats de Tom Mitchell, et on l'a modifié en supposant qu'à chaque instant, en fonction de l'état du système, une heuristique d'oubli pouvait choisir quelles informations oublier. Et nous avons vu que n'importe quel oubli dans ce cas est équivalent à la prise en compte d'exemples supplémentaires. On considère alors toutes les trajectoires d'apprentissage correspondant aux $m!$ permutations d'une séquence d'apprentissage donnée $\langle z_1, z_2, \dots, z_m \rangle$. Nous avons montré que pour que ces trajectoires se terminent toutes dans le même état final e_m , il est nécessaire et suffisant que l'effet de l'heuristique d'oubli sur chaque trajectoire soit équivalent à l'ajout du *même ensemble d'exemples*. Cela signifie donc la nécessité d'une corrélation très forte entre ces trajectoires, c'est-à-dire entre les choix instantanés de l'heuristique d'oubli.

On peut également interpréter ce résultat comme le fait que sur chaque trajectoire d'apprentissage le système partage une information commune implicite qui correspond à un même ensemble d'exemples supplémentaires.

Il est donc clair que l'indépendance sur l'ordre des exemples impose des conditions fortes sur le système apprenant. La démonstration vaut pour le cadre d'apprentissage de concepts avec élimination des hypothèses non cohérentes avec les données. Mais elle peut certainement s'étendre à d'autres cadres. L'un de mes objectifs est d'étendre l'étude au cadre continu de la régression.

Avant de passer à l'analyse de méthodes d'apprentissage actif (section 4.2), il est utile d'examiner rapidement l'état de l'art dans l'étude de l'apprentissage en-ligne.

4.1.4 Théorie de l'apprentissage en-ligne

Dans l'apprentissage en-ligne (*on-line learning*), on suppose que le processus se décompose en épisodes. Dans chacun de ceux-ci, un nouvel exemple est fourni à l'apprenant qui, selon le modèle étudié, doit soit prédire son étiquette, soit émettre une hypothèse. Il reçoit alors une pénalité qui est fonction de l'écart entre la réponse produite et la réponse désirée. Un cas typique est celui dans lequel chaque erreur vaut 1. On évalue l'apprentissage par le nombre d'erreurs produites par l'apprenant avant d'avoir identifié le concept cible. C'est le modèle *mistake-bound* [Blu94].

Deux problèmes liés se posent pour définir ce cadre. D'une part, de quel lien suppose-t-on l'existence entre le passé et l'avenir ? D'autre part, comment évalue-t-on la qualité d'un apprenant ?

Dans la théorie statistique de l'apprentissage, on suppose qu'une même distribution fixe régit les exemples d'apprentissage et les exemples à venir. Si on conserve cette hypothèse dans le cadre de l'apprentissage en-ligne, alors il n'y a aucune différence fondamentale entre l'apprentissage hors-ligne ou *batch* et l'apprentissage en-ligne. La distribution sur l'espace des exemples induit une distribution sur les séquences d'apprentissage possibles, qui, par exemple, rend très peu probables les séquences dans lesquelles tous les exemples négatifs seraient fournies avant les exemples positifs.

Mais les théoriciens formés à l'informatique fondamentale sont obsédés par les études de pire cas. Ils souhaiteraient donc caractériser un apprenant par son comportement dans le cas

des séquences d'apprentissage les plus défavorables. Or il est facile de montrer que si l'ordre des exemples est choisi par un adversaire tenant compte de l'état de l'apprenant, alors cet adversaire peut forcer des performances de l'ordre de celles qui seraient obtenues en répondant au hasard. Il faut donc avoir recours à des *critères de performances relatifs* et des bornes associées (*relative loss bounds*).

L'une des possibilités est de comparer l'apprenant en-ligne à un apprenant hors-ligne qui aurait d'emblée toutes les données d'apprentissage. On cherche alors à quantifier le désavantage lié à l'ignorance des données d'apprentissage à venir [BDKM97]. Une autre possibilité est de comparer l'apprenant à un ensemble d'« experts ». Un bon apprenant, dans ce cadre, est celui qui se comporte pas beaucoup plus mal que le meilleur expert de cet ensemble. Ce cadre a donné naissance à un modèle dérivé : l'apprentissage à partir d'avis d'experts *learning from expert advice*.

Dans ce cadre, on suppose donc l'existence d'un ensemble de n d'experts (ou d'hypothèses). Pour chaque nouvel exemple \mathbf{x} , chaque expert fait une prédiction y . L'algorithme d'apprentissage \mathcal{A} n'a accès qu'à ces « avis » des experts pour former sa propre prédiction $y_{\mathcal{A}}$. Il apprend alors la bonne prédiction y^* . On cherche à ce que l'algorithme \mathcal{A} obtienne une performance presque aussi bonne que le meilleur expert dans l'ensemble. Si $L(e)$ est la perte de l'expert e , et $l(\mathcal{A})$ la perte de l'algorithme d'apprentissage, il est souvent possible de prouver :

$$L(\mathcal{A}) \leq \min_e L(e) + \log(n) \quad (4.3)$$

sur toutes les séquences d'apprentissage possibles.

Il faut noter le caractère nouveau de ce type d'étude. Il n'y a plus d'hypothèse selon laquelle les données sont indépendamment et identiquement distribuées (i.i.d.), et il n'y a pas de supposition sur la valeur des experts : ceux-ci peuvent être médiocres. L'inégalité 4.3 ne tient pas en probabilité seulement, elle tient toujours, quelque soit la séquence d'apprentissage rencontrée.

Ce cadre, si il se révèle intéressant pour toutes les applications de gestion de portefeuille par exemple, dans lesquelles il est possible de combiner les avis d'experts et de quantifier la performance à chaque épisode, est cependant limité. Les questions que l'on aimera étudier concernant l'apprentissage en-ligne sont :

- L'algorithme utilise-t-il bien les données qui lui parviennent séquentiellement ?
- Comment qualifier et quantifier l'information fournie par l'organisation de la séquence (ordre et vitesse de présentation). Il est évident qu'il y a là une information importante pour guider l'apprentissage.
- La mesure de la qualité d'un apprenant par rapport à une professeur. Est-il capable de parvenir rapidement à satisfaire l'objectif du professeur : concept cible ou performance donnée ? Évidemment, cela ramène au débat sur les possibilités de collusion entre professeur et élève.

Il reste à trouver un cadre théorique permettant d'examiner ces questions.

4.2 L'apprentissage actif

4.2.1 L'existant en apprentissage actif

L'étude de l'apprentissage actif est motivée en grande partie par la réalisation que, dans de nombreuses applications, il est aisément d'obtenir un grand nombre de données non étiquetées (par exemple des signaux vocaux ou des pages web) mais qu'il est coûteux de les faire étiqueter par un « expert ». L'idée est alors de chercher à réduire ce coût d'étiquetage par des questions

intelligemment posées. Un cadre un peu différent est celui dans lequel il n'existe pas *a priori* d'ensemble d'exemples non étiquetés, mais où l'agent a l'initiative de chercher les exemples qui lui semblent les plus informatifs.

Voici un exemple simple.

Exemple 4.2

On suppose que les données sont disposées sur la droite des réels suivant une distribution \mathcal{D} , et que l'espace des hypothèses est constitué des fonctions seuil, $\mathcal{H} = \{h_w\}$:

$$h_w(x) = 1 \text{ si } x > w, \text{ et } 0 \text{ autrement.}$$

La théorie statistique de l'apprentissage prédit que si la fonction cible appartient à \mathcal{H} , alors il suffit de tirer un échantillon aléatoire de données, suivant la distribution \mathcal{D} , de taille $\mathcal{O}(1/\varepsilon)$ pour obtenir une hypothèse de taux d'erreur d'au plus ε (suivant \mathcal{D}).

Supposons maintenant que m points non étiquetés soient tirés suivant la distribution \mathcal{D} . Sur la droite des réels, leurs étiquettes inconnues forment une séquence de '0' suivie d'une séquence de '1'. Le but est de trouver un point w pour lequel se produit la transition entre les deux séquences. Une simple recherche binaire permet de trouver un tel point en demandant l'étiquette de seulement $\mathcal{O}(\log m)$ exemples. L'apprentissage actif peut donc fournir un avantage exponentiel sur le nombre d'étiquettes nécessaires.

Malheureusement, peu de choses sont connues en dehors de cet exemple jouet. L'un des rares résultats théoriques à ce jour est celui relatif à l'étude de l'algorithme *query-by-committee* (QBC) [SOS92, FSST97]. Dans ce modèle, l'apprenant observe une séquence de données non étiquetées et peut à chaque instant décider de demander l'étiquette de l'exemple courant. Il est alors prouvé que si les exemples de la séquence sont tirés uniformément sur la surface d'une hypersphère unité de dimension d et que le concept cible correspond à un séparateur linéaire passant par l'origine, il est possible d'apprendre une hypothèse d'erreur ε après avoir vu $\mathcal{O}(d/\varepsilon)$ exemples et demandé l'étiquette de $\mathcal{O}(d \log 1/\varepsilon)$ d'entre eux, ce qui représente un avantage exponentiel sur la taille $\mathcal{O}(d/\varepsilon)$ requise pour apprendre un séparateur linéaire dans le cadre classique de l'apprentissage supervisé passif. Il s'agit d'un résultat remarquable, obtenu cependant au prix d'un algorithme complexe qui doit calculer les volumes des espaces des versions intermédiaires.

En dehors de ces travaux théoriques, plusieurs heuristiques ont été proposées pour accélérer l'apprentissage [AM98, CAL94, CGJ96, IAZ00, MM04, Opp99, PF91, PRZ02, RM01, STP04, Ton01]. Il s'agit essentiellement d'augmenter la probabilité d'échantillonnage dans les régions de forte incertitude, ce qui est très lié aux méthodes d'*importance sampling* dans les techniques de Monte-Carlo (voir par exemple [Gla04]).

4.2.2 Le cas d'une stratégie active de sélection d'attributs

Toutes les stratégies d'apprentissage actif évoquées portent sur le choix d'exemples d'apprentissage et donc sur une modification de la distribution des exemples afin de faciliter l'apprentissage ou de conduire à de meilleures performances en généralisation avec moins de données. Un autre objectif pourrait être, non pas de jouer sur les exemples et leur distribution, mais sur les attributs de description. On revient là à des problèmes de sélection d'attribut, mais cette fois-ci dans le cadre d'une démarche incrémentale active.

Ce problème nous a été suggéré, à Jérémie Mary et moi-même, par des applications en bioinformatique. L'analyse du transcriptome, sur des puces à ADN, implique des coûts qui sont en relation avec le nombre de «spots» testés sur la puce. Ainsi, l'un des objectifs de l'étude

sur la mise en évidence des faibles radiations (section 3.2.1) supervisée par Marie Dutreix est l'identification d'un petit nombre de gènes discriminants permettant la mise au point de puces à ADN dédiées de faible coût (testant par exemple 100 gènes) et donc utilisables dans des campagnes de dépistage à grande échelle.

En supposant que le coût d'une puce à ADN est proportionnel au nombre de gènes testés⁸⁰, et que le budget alloué à une étude soit fixé, la question est alors d'étudier si il est possible de déterminer un protocole expérimental permettant de maximiser la précision et/ou le rappel sur les attributs discriminants (les gènes pertinents).

On suppose que le nombre total d'attributs est de d_0 dont p_0 sont des attributs pertinents et $n_0 = d_0 - p_0$ sont donc non pertinents. On suppose par ailleurs que le coût unitaire de test d'un attribut sur un exemple est de c et que le budget disponible total est de M .

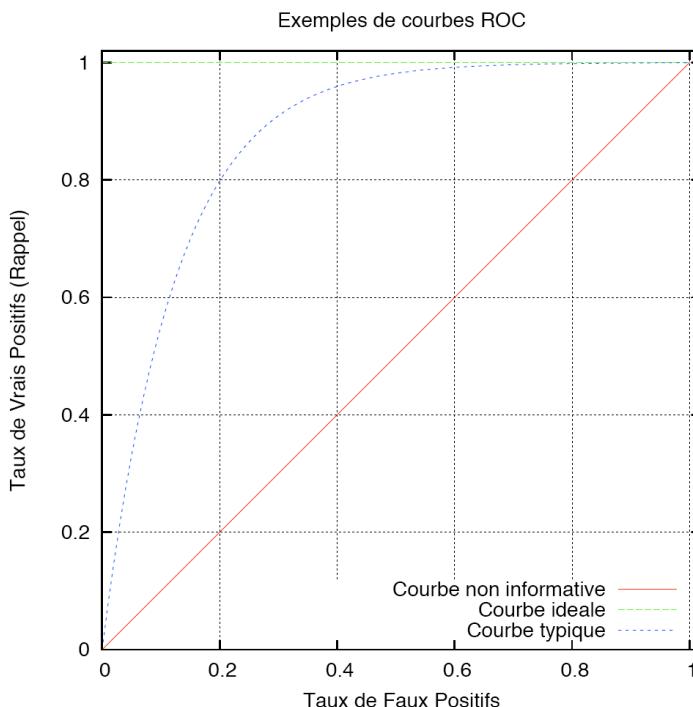


FIGURE 4.3 – Exemples de courbes ROC correspondant à des fonctions de décision plus ou moins informative sur la vraie classe des attributs. (Merci à Jérémie pour cette figure).

Une démarche de sélection active pourrait alors être la suivante. Dans une première étape, on investit m_1 tests sur les d_0 attributs. Cela correspond à m_1/d_0 exemples (ou lames dans le cas du transcriptome). On utilise une méthode d'évaluation d'attributs (e.g. RELIEF) pour établir un classement des attributs. Selon un seuil à fixer, on élimine les attributs les moins prometteurs, et on investit m_2 tests supplémentaires sur les d_1 attributs restants. On peut répéter cette procédure n fois jusqu'à ce que la somme des investissements $m_1 + m_2 + \dots + m_n$ soit égale au budget alloué M .

La question est alors de déterminer la séquence optimale des investissements m_i et des seuils d'élimination σ_i ($1 \leq i \leq n$) afin de maximiser un critère objectif.

Avec Jérémie, nous avons analysé deux critères. L'un est la *maximisation de l'aire sous la*

80. Une hypothèse simplificatrice pour commencer, mais qu'il faudrait adapter pour être plus réaliste.

courbe ROC (correspondant au taux de vrais positifs en fonction du taux de faux positifs sur le choix des attributs). L'autre est la *maximisation de la précision pour un rappel souhaité*, ou vice-versa.

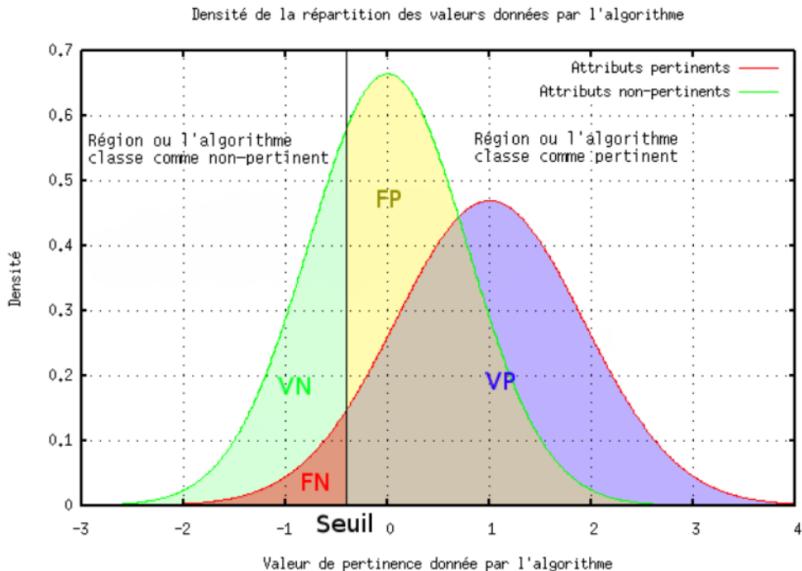


FIGURE 4.4 – *Visualisation du positionnement de VP, FP, VN, FN en fonction du seuil de tolérance. Pour une séparation parfaite, on voudrait FN = FP = 0. Plus le seuil est faible -sélection faible- plus on réduit FN (mais plus FP grandit) et plus le seuil est élevé -sélection forte- plus on réduit FP (mais plus FN grandit).* (Merci à Jérémie pour cette courbe).

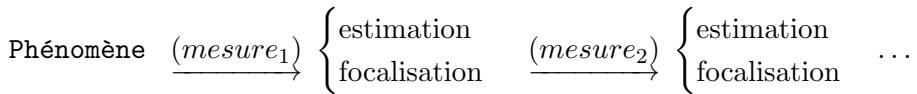
Évidemment, il y a un compromis à régler car plus on investit à une étape, meilleure est la séparation entre les attributs pertinents et les non pertinents (voir la courbe 4.4), mais moins il reste à investir pour les étapes suivantes qui se focalisent sur les attributs apparemment les plus prometteurs.

Le problème est de manière assez surprenante assez difficile à formuler (e.g. une courbe ROC qui est une combinaison de courbes ROC ne portant pas sur les mêmes attributs) et doit à la fin faire l'objet d'une procédure d'optimisation numérique lourde (le nombre de paramètres interdépendants à déterminer est grand). Nous ne doutions pas de pouvoir montrer que dans une très large classe de situations, cette démarche active pouvait conduire à une amélioration des performances par rapport à la démarche classique consistant à tester tous les attributs en une étape unique. À notre surprise, c'est l'inverse qui se produit. Nos expériences avec des modèles artificiels exprimant le pouvoir discriminant des méthodes d'évaluation d'attributs montrent que si ce pouvoir est trop faible ou trop fort, alors la méthode classique en une étape est plus performante. Avant donc de pouvoir publier sur une idée qui paraissait de prime abord si évidemment bénéfique, il nous faut caractériser les classes de situations dans lesquelles elle conduit à des améliorations. Ce travail porte en particulier sur des conditions impliquant la borne de Cramer-Rao liant, d'une part, l'information qui peut être gagnée sur la valeur des paramètres d'un modèle à estimer à chaque nouvelle observation, et d'autre part, l'information de Fisher⁸¹.

81. Sorte de mesure du pouvoir de résolution d'une expérience. C'est-à-dire, en supposant deux valeurs de

Ce travail montre d'abord que l'apprentissage actif ne conduit pas nécessairement à de meilleures performances que l'apprentissage « batch » avec un jeu de données prédéterminé. Ici, la meilleure estimation de la valeur de certains attributs (ceux qui sont sélectionnés jusqu'aux dernières étapes) ne compense pas la perte d'information sur les attributs éliminés dans les premières étapes.

Ensuite, le mécanisme que nous proposons peut être vu comme une métaphore du mécanisme d'attention. Étant donné un phénomène dans l'environnement de l'agent, celui-ci y prête une première attention légère, puis se focalise graduellement sur quelques aspects particuliers de ce phénomène.



Cette attention activement dirigée est-elle nécessairement avantageuse par rapport à une mesure uniforme sur l'environnement ? Sous quelles conditions y a-t-il bénéfice à appliquer cette stratégie ? Notre travail, jusqu'à présent, constitue une étude exploratoire de ces questions sous un angle qui, je crois, n'est pas celui des études actuelles en « perception située » [HB88, MT03]. Peut-être y a t-il là un point de vue nouveau et intéressant. Nous le souhaitons et allons continuer à travailler dans cette direction.

4.3 L'effet tunnel cognitif : comment construire un nouveau domaine conceptuel

« Il n'y a rien de neuf sous le soleil » [Ecclesiastes 1 :9]. Cela est peut-être littéralement vrai. On peut aussi argumenter que les anneaux de Saturne n'ont pris d'existence qu'avec l'invention de la lunette astronomique, et qu'il en est de même de bien d'autres phénomènes. Il n'empêche que cette citation souligne le rôle de la créativité qui consiste en particulier à savoir réinterpréter ce qui est observé⁸².

D'une manière abstraite et très réductrice, on peut considérer ce problème comme celui de la construction d'un nouveau référentiel pour interpréter une situation, là où d'autres référentiels existaient déjà, qui permettaient des interprétations, sans doute partielles, imparfaites, mais néanmoins opérationnelles.

On peut alors se poser deux questions :

1. Comment se construit un nouveau référentiel pour rendre compte de situations pour lesquelles, éventuellement, existaient déjà des interprétations ?
2. Comment peut-on aider ou forcer la construction d'une nouvelle interprétation et du référentiel associé ?

Cette dernière question est évidemment d'une grande importance potentielle pour l'enseignement, en particulier l'enseignement des sciences.

L'étude de ce genre de questions, passionnantes, est extrêmement délicate. Un problème essentiel est qu'en effet, alors que nous sommes sans doute créatifs en permanence, à plus ou moins grande échelle, il est très difficile de réaliser des expériences contrôlées de créativité. Les

paramètre θ et θ' , quelle doit être la différence $|\theta - \theta'|$ minimale pour que l'expérience permette de les distinguer (Voir par exemple [Kul68, Jay03]).

82. Voir aussi savoir construire des sortes de simulations ou expériences de pensée qui font jouer les ressorts essentiels d'un modèle du monde [Gan02].

4.3. L'effet tunnel cognitif : comment construire un nouveau domaine conceptuel

chercheurs en sont donc soit réduits à proposer des modèles de processus cognitifs (et éventuellement sociaux) dont ils essaient de montrer la validité sur des compte-rendus historiques plus ou moins fidèles de découvertes scientifiques (ou artistiques) (voir [Koe60, Koe64] qui décrit le processus de *bissolement*, [Hol73] sur les schémas, superstructures à l'œuvre dans les interprétations scientifiques, ...), soit amenés à étudier des « micro » situation-problèmes qui stimulent des processus cognitifs créatifs (voir les études sur l'analogie (section 3.3), ou celles sur le *blending* [Fau85, Fau97, FT98]).

Les recherches en didactique des sciences, sans évoquer explicitement la créativité, mettent en scène et organisent des ré-interprétations, des coordinations de domaines conceptuels et des constructions de nouveaux domaines conceptuels.

Depuis toujours intéressé par le raisonnement inventif (par exemple, voir [Pol57]) et le processus de découverte scientifique (voir par exemple [SL90, VP95, MNT99]), je n'aurais peut-être jamais été y voir de près si je n'avais eu la chance de rencontrer Andrée Tiberghien, spécialiste de didactique de la physique, en 1994, tandis qu'elle cherchait à mettre sur pied un programme pluridisciplinaire d'étude de l'apprentissage de connaissances complexes. Nous avons alors entamé une collaboration de plusieurs années, dans une entente parfaite, et dans un esprit de rigueur et d'ouverture exemplaire sur le plan scientifique⁸³.

Le problème qu'Andrée proposait d'étudier était séduisant à deux égards. D'une part, il concernait des activités d'apprentissage et de découverte, donc le vaste champ de l'articulation, de la coordination et de la construction de domaines conceptuels, d'autre part, il se fondait sur des expériences très précises, très circonscrites, très bien documentées, d'apprentissage du concept d'énergie en physique par des élèves de lycée.

Andrée disposait d'enregistrements de l'activité de dyades d'élèves en train d'essayer de rendre compte de petits montages expérimentaux à l'aide d'une modélisation en termes de transfert, de transformation et de stockage d'énergie [MT95, Meg95, Tib94, Tib96]. Les élèves disposaient d'une proto-théorie sur l'énergie consistant essentiellement en quelques définitions et contraintes sémantiques (e.g. l'énergie se conserve) et en un langage d'expression du modèle à construire ainsi que de quelques contraintes syntaxiques (e.g. une chaîne énergétique commence et se termine par un réservoir, le réservoir initial et le réservoir final devant être différents) (voir figure 4.5).

Les élèves devaient donc construire une interprétation explicite d'une situation-problème en respectant un ensemble de contraintes sémantiques et syntaxiques correspondant au domaine conceptuel en grande partie nouveau pour eux.

Dans une des tâches en particulier, visant à rendre compte d'un petit montage électrique (pile-ampoule), ce qui était remarquable, c'est que toutes les dyades (7) produisaient à un moment donné clé de leur séquence de raisonnements un même modèle erroné de la situation-problème (voir figure 4.6). Un tel cas est intéressant car il signale qu'une « erreur » de raisonnement correspond à un mécanisme cognitif puissant. Il importe alors de le comprendre, soit pour le combattre, soit pour en tirer profit.

Deux questions se posaient. D'une part, *pourquoi et comment ce modèle erroné se construisait et était un « attracteur » de l'activité interprétative ?* D'autre part, *quelles étaient les conséquences de ce processus cognitif ?* Pouvaient-elles conduire à des activités d'apprentissage spécifiques ?

La première question peut paraître élémentaire. Il est évident que dans le cas du montage pile-ampoule, les élèves produisent un modèle inspiré par leurs connaissances du domaine de

83. Evelyne Cauzinille-Marmèche, alors à Paris 5 et maintenant au CNRS dans l'Université de Provence, et Gérard Collet, de l'IUFM de Grenoble, ont été d'autres grands acteurs de cette aventure.

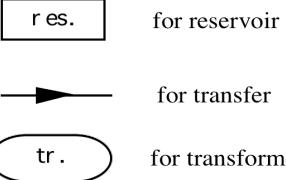
Theory (seed)	Model (seed)
<p>Energy can be characterized by:</p> <ul style="list-style-type: none"> * its properties : <ul style="list-style-type: none"> - Storage - Transformation - Transfer <ul style="list-style-type: none"> - by work : mechanical or electrical - by heat, - by radiation. * a fundamental principle of conservation The energy is conserved whatever the transformations, transfer and forms of storage 	<p>* Symbols to be used:</p>  <ul style="list-style-type: none"> res. for reservoir → for transfer tr. for transformer <p>* Under the constraints:</p> <ul style="list-style-type: none"> - a complete energy chain starts and ends with a reservoir; - the initial reservoir is different from the final reservoir.

FIGURE 4.5 – Version simplifiée de la proto-théorie du domaine de l'énergie fournie aux élèves dans les expériences d'Andrée Tiberghien. La partie droite fournit les symboles et les règles syntaxiques à utiliser pour construire un modèle dans le monde de l'énergie.

l'électricité. Reste cependant à rendre compte de manière plus formelle des mécanismes qui conduisent cette interprétation à être produite malgré les consignes de la tâche et à se « déguiser » en modèle appartenant au domaine de l'énergie.

Inspirés en partie par les travaux de Gentilhomme en linguistique [Gen94], Gérard Collet, Andrée et moi-même avons proposé l'idée selon laquelle l'activité interprétative lorsqu'elle vise un domaine conceptuel mal maîtrisé a recours à un *niveau notionnel* des connaissances plutôt qu'au niveau conceptuel utilisé par les experts du domaine. Ce niveau notionnel se caractérise par des primitives très prégnantes (comme transport, fabrication), liées davantage aux sensations (comme chaud), et très « accommodantes », c'est-à-dire informelles et se prêtant facilement aux changements de sens et aux dérives métaphoriques (voir [CTC98, CTC, Col00] pour une description détaillée). Par exemple, la notion d'énergie inclut des croyances telles que : l'énergie est un fluide ou une substance qui peut circuler, qui peut être stockée (e.g. « je suis plein d'énergie aujourd'hui ») et qui est souvent liée à la causalité.

Ce niveau notionnel permet des appariements aisés entre entités appartenant à des domaines conceptuels distincts. Il permet en particulier de mobiliser des entités interprétatives dans des domaines mal connus, entités qui pourront être ultérieurement spécialisées et définies de manière formelle pour donner naissance aux concepts du domaine en construction. Une autre vertu des *entités-Janus* ainsi constituées à la jonction entre deux domaines est qu'elles permettent de transporter des propriétés et des mécanismes d'inférence d'un domaine à l'autre. (Ainsi, après avoir associé l'entité **Réservoir** et le poids dans un petit montage physique, un élève a pu dire en levant le poids : « tu vois, je le remplis », ce qui dénote l'importation d'une propriété initialement absente de poids).

Nous avons qualifié d'*effet tunnel* le mécanisme cognitif à l'œuvre dans ces tâches d'apprentissage en vertu d'une deuxième observation remarquable faite sur les élèves au cours de leurs raisonnements. En effet, une fois parvenus à leur modèle erroné issu, en grande partie, d'une interprétation dans le domaine conceptuel source (e.g. le domaine électrique), les élèves utilisent ce modèle pour faire des inférences dans le domaine cible en construction (ce qui leur permet de

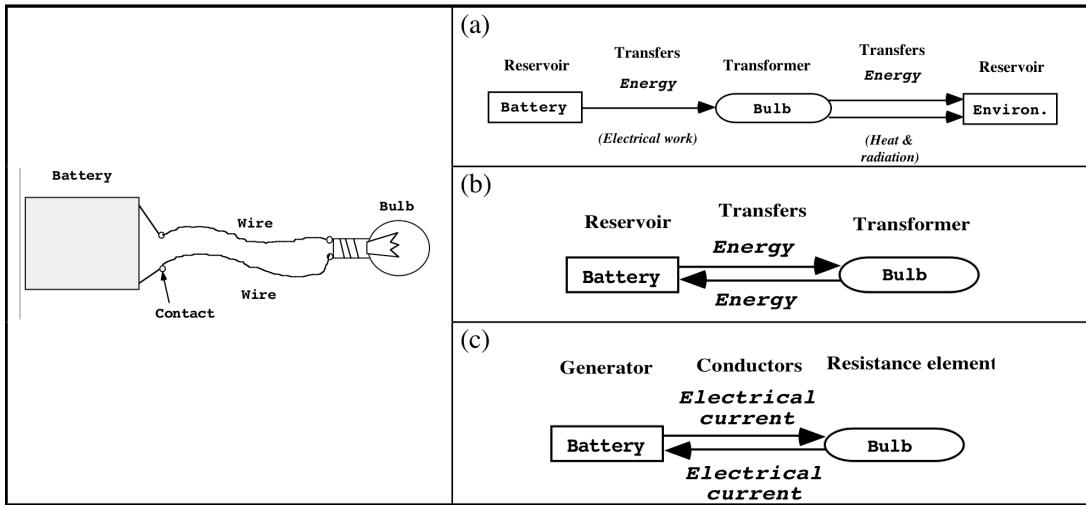


FIGURE 4.6 – Trois modèles d'une situation-problème correspondant à un montage pile-ampoule. Le modèle (a) est le modèle cible, le modèle (b) correspond à un modèle erroné produit par les élèves, le modèle (c) à un modèle du montage dans le domaine d'interprétation de l'électricité.

prédir que la pile ne s'usera pas, ou que la borne '+' est le Réservoir_initial et la borne '-' le Réservoir_final, etc.).

Il y a donc là un phénomène tout à fait étonnant, dans lequel un modèle qui n'aurait pas pu être construit si les élèves avaient effectivement joué le jeu et travaillé entièrement dans le domaine conceptuel cible (e.g. le domaine de l'énergie) se trouve importé depuis le domaine source, emmenant avec lui, en contrebande, des inférences (comme celle portant sur le sens du transfert de l'énergie décalqué sur celui du courant électrique). Il y a bien réinterprétation du modèle exprimé, qui devient support d'inférence, dans le domaine cible.

Il est notable que ce phénomène observé chez les élèves est certainement à l'œuvre dans des travaux scientifiques célèbres, comme par exemple quand Carnot construit le modèle du cycle thermodynamique en travaillant en fait avec la notion de calorique en tête. Ce modèle permettra, plus tard, à Joule et à Clausius de construire le concept d'entropie (comme une fonction d'état). Un exemple de réinterprétation réussie. Dans le cas des élèves, la réinterprétation dans le domaine cible conduit à des contradictions, soit avec leurs connaissances sur le monde (e.g. la pile s'use), soit avec des contraintes du domaine de l'énergie (e.g. le réservoir initial doit être différent du réservoir final).

L'examen de ces contradictions focalise l'attention des élèves sur les aspects problématiques de leurs modèles et les conduit à ré-examiner les entités-Janus mises en jeu (e.g. le transfert d'énergie n'est pas directement appariable avec le courant électrique) et à spécialiser ces entités qui deviendront progressivement des concepts autonomes dans le nouveau champ conceptuel (voir [CMCCT97]).

L'effet tunnel cognitif repose donc sur :

- la construction d'un modèle exprimé dans le langage formel du domaine cible (e.g. avec des flèches, des boîtes et des mots comme énergie, réservoir, transformateur)
- à partir des connaissances d'un domaine source
- grâce à des associations au niveau notionnel qui produisent des entités-Janus

- modèle qui est ensuite utilisé comme source d'inférence dans le domaine cible en construction (réinterprétation)
- ce qui conduit éventuellement à des contradictions soit sémantiques (avec les connaissances *a priori* sur le monde, ou des connaissances sur le domaine cible) soit syntaxiques (avec les contraintes imposées sur les modèles valides dans le domaine cible).

Nous avons montré dans nos publications que ce mécanisme permettait d'expliquer de nombreuses productions chez les étudiants, de même qu'il donne des pistes pour rendre compte de découvertes scientifiques comme celles de Carnot, de Maxwell sur l'électromagnétisme ou de Planck sur les quanta. Il est par ailleurs distinct du raisonnement par analogie. Il ne porte que sur une situation-problème et non deux, et implique un domaine conceptuel cible en gestation, incapable de fournir les interprétations sémantiques organisées et hiérarchisés que demandent les modèles classiquement proposés de l'analogie [FFG89, Gen83, Gen89].

Nous pensons donc avoir mis en évidence et analysé un nouveau mécanisme de coordination et de construction de domaines conceptuels, reposant sur la mise en jeu d'un niveau notionnel de la connaissance, et impliquant l'utilisation d'un support symbolique se prêtant à des interprétations dans plusieurs domaines. Il reste à montrer la pertinence de cette modélisation sur une large classe de protocoles expérimentaux, et surtout à en donner une spécification assez précise pour que l'on puisse la coder en vue de simulations informatiques dans des domaines circonscrits. Il s'agit là d'investissements en recherche, à la frontière de disciplines, et pour des modélisations cognitives, qui ne sont pas dans l'air du temps.

Selon un point de vue purement informatique sur les techniques d'exploration d'espaces de modèles pour l'optimisation d'un critère donné, l'effet tunnel (dans une version très basique) peut être considéré comme apparenté aux techniques d'évolution simulée. Dans celles-ci, le principe est de combiner des morceaux des hypothèses candidates les meilleures pour construire de nouvelles hypothèses, dont les meilleures seront à leur tour décomposées, etc. Dans l'effet tunnel, dans une première étape, on construit des hypothèses dans un langage cible à partir de connaissances d'un domaine source différent, puis on utilise des connaissances du domaine cible pour modifier incrémentalement les hypothèses candidates.

4.4 Publications, projets et stages liés à ces directions de recherche

Les recherches décrites dans ce chapitre ont bénéficié des **collaborations** suivantes :

European Science Foundation (Learning in Humans and Machines) (1994-1997) :
« Sequence effects in learning »
Responsable : Erno Lehtinen (Helsinki)
Participants : 14 participants.

GIS - CNA-47 (Cognition Naturelle et Artificielle) (1997-1998) : « Structuration des connaissances dans l'apprentissage de connaissances complexes »
Responsables : Andrée Tiberghien (Equipe COAST, ENS-Lyon) et Antoine Cornuéjols (LRI)
Participants : Evelyne Cauzinille (Paris-5), Gérard Collet (Grenoble), ...

GIS - CNA-II (Cognition Naturelle et Artificielle) (1999-2000) : « La notion de changements de repères en sciences cognitives »
Responsable : Antoine Cornuéjols (LRI)

4.4. Publications, projets et stages liés à ces directions de recherche

Participants : Jacques Ales-Bianchetti (LRI), Louis Bourelly (CREPCO), Gérard Collet (Grenoble), André Didierjean (CREPCO), Thierry Ripoll (CREPCO), Andrée Tiberghien (Lyon-2), Bruno Vivicorsi (CREPCO), Jean-Daniel Zucker (ACASA, Paris-6)

Groupe de Travail Ecole (2001-2002) : « Des connaissances naïves aux savoirs scientifiques »

Responsable : Andrée Tiberghien (Lyon-2)

Participants : Christian Buty (Lyon-2), Marylin Coquidé (Orsay), Françoise Cordier (Univ. Poitiers), Antoine Cornuéjols (LRI), Colette Laborde (IMAG, Grenoble), Janine Rogalski (Paris-8), Lorenza Saitta (Univ. Turin), Laurent Veillard (Lyon-2).

Projet ROBEA (2003-2004) : « Contrôleur et représentation pour un robot auto-nome : Agir, Anticiper et s'Adapter »

Responsables : Marc Schoenauer et Michèle Sebag

Participants :

Les **stagiaires et doctorants** encadrés ayant travaillé sous ma direction sur ces sujets sont :

- Jérémie Mary (doctorant : 2002 - 2005)
- Jérémy Barbay (stagiaire DEA : 1998) : « Étude théorique des capacités d'adaptation des algorithmes d'évolution simulée »
- Yann Merour (stagiaire DEA : 1998) : « Modélisation de l'acquisition du concept d'énergie par des lycéens »

Les **publications** concernées auxquelles j'ai participé sont :

1. (BC02) C. Buty and A. Cornuéjols. « Évolution des connaissances chez l'apprenant », in *Des connaissances au savoir scientifique* (Ed. A. Tiberghien), ACI Ecole et Sciences Cognitives, pp.41-66, 2002.
2. (Cor89a) A. Cornuéjols. *De l'apprentissage incrémental par adaptation dynamique : le système INFLUENCE*, Thèse de doctorat, Université d'Orsay, 1989.
3. (Cor89b) A. Cornuéjols. « An exploration into incremental learning : the INFLUENCE system », *International Conference on Machine Learning (ICML-89)*, Cornell, USA, 1989. ACM Press, pp. 383-386.
4. (Cor93a) A. Cornuéjols (Ed.) *Training Issues in Incremental Learning*, AAAI Press, 1993.
5. (Cor93b) A. Cornuéjols. « Getting Order Independence in Incremental Learning », *European Conference on Machine Lerarning (ECML-93)*, (Ed. P. Brazdil), Vienna, Austria (1993). Springer-Verlag, LNAI-667, pp. 196-212.
6. (CMCCT97) E. Cauzinille-Marmèche, G. Collet, A. Cornuéjols and A. Tiberghien A. « Co-Adaptation of Students' Knowledge Domains when Interpreting a Physical Situation in Terms of a New Theory », in Proc. of the *2nd European Conference on Cognitive Science (ECCS-97)*, Manchester, April 9-12, (1997), pp. 107-112.
7. (CTC98) A. Cornuéjols, A. Tiberghien and G. Collet « Decomposing the scientific discovery process using multiple interpretations of notions », *1st Conference on Model-Based Reasoning (MBR-98)*, Pavia, Italy, (1998).

8. (CTC99a) A. Cornuéjols, A. Tiberghien and G. Collet « Tunnel effects in cognition : A transfer mechanism from known conceptual domains to new ones », in *Conference on Artificial Intelligence and Simulation of Behaviour (AISB-99)*, Edinburgh, United Kingdom, April 6-9, 1999. (pp. 9-18 in vol. on Scientific Creativity).
9. (CTC99b) A. Cornuéjols, A. Tiberghien and G. Collet « L'effet tunnel en cognition : un mécanisme de transfert entre domaines conceptuels », in Actes de la 1ère Conférence d'Apprentissage (CAP-98), Palaiseau, France, 16-19 juin, 1999, pp. 243-250.
10. (CTC00) A. Cornuéjols, A. Tiberghien and G. Collet « A new mechanism for transfer between conceptual domains in scientific discovery and education ». *Foundations of Science*, vol.5, No.2, (2000), 129-155.
11. (Cor06) A. Cornuéjols. « Machine Learning : The Necessity of Order », *In Order to Learn : How ordering processes and sequencing effects in machines illuminate human learning and vice-versa*, E. Lehtinen and F. Richter (Eds.), Cambridge University press, 2006.
12. (Mar05) J. Mary. *Étude de l'apprentissage actif*. Thèse de doctorat, Université de Paris-Sud, Orsay, 2005.

5

Perspectives

Dans un contexte où l'informatique en général et l'apprentissage artificiel en particulier peuvent parfois être conçus comme des disciplines de service (seulement) destinées à fournir des outils aux autres disciplines et secteurs d'activité, je veux affirmer que pour moi l'objectif central est d'étudier l'apprentissage en tant qu'objet à part entière. Ce qui signifie, d'une part, étudier ses faits propres, observables tant dans les systèmes artificiels que dans les organismes naturels, ses règles de cohérence interne et ses lois, et, d'autre part, façonner des outils conceptuels et des techniques qui découlent de cette étude et qui peuvent trouver leur utilité en dehors du champ de l'apprentissage.

Pour cela, je crois à la valeur d'une double démarche consistant, d'un côté, à chercher à **comprendre les fondements par une modélisation théorique** et le test sur des micro-mondes, et, de l'autre, à **se confronter à des problèmes réels** afin d'y trouver la source d'interrogations et le banc d'essai où mettre à l'épreuve les concepts et les techniques inventés.

Or, pour ce qui concerne l'étude de l'apprentissage, nous sommes loin d'une théorie finale du grand tout. Il reste à faire pour notre génération et pour les suivantes.

À ses débuts, l'apprentissage artificiel était intégré dans l'étude de l'intelligence en général et se donnait pour objectif de modéliser les *processus* d'apprentissage. À partir des années quatre-vingt et jusqu'à maintenant, l'arrivée concomitante du connexionnisme, donc des mathématiques du continu, de l'approximation et des modèles semi-paramétrés, et de l'approche statistique pour théoriser l'apprentissage a eu un double résultat. D'abord, l'apprentissage qui était jusque là perçu comme le problème de l'interprétation des entrées et de leur intégration dans un modèle du monde en construction, a soudain été conçu comme un problème d'analyse de données où modéliser ou apprendre signifient trouver une fonction (simple) s'ajustant aux données. Ensuite, en raison des outils mathématiques disponibles, il n'est possible dans ce paradigme que de rendre compte d'échantillons de données indépendamment et identiquement distribués (i.i.d.). Aucune information ne peut être extraite, selon ce point de vue, de l'ordre et de l'organisation de la séquence des données. Fondamentalement, la théorie statistique de l'apprentissage est incapable de rendre compte de telles structures dans les données.

Cette vision très purifiée de l'apprentissage a conduit à des progrès considérables dans la compréhension des facteurs en jeu dans l'induction d'hypothèses paramétrées ou semi-paramétrées à partir de petits échantillons de données et en supposant un critère de performance simple, fondé sur une mesure d'erreur ou d'écart. Il a aussi permis de concevoir de nouveaux algorithmes très intéressants comme le boosting ou les méthodes à noyau. Ce cadre est par ailleurs opérant dans plusieurs domaines d'application comme par exemple, certaines parties de la bioinformatique ou la fouille de bases de données en général.

Mais, selon cette perspective, l'apprentissage est une *science de phénomènes statiques et non de processus dynamiques*. Toutes les données sont supposées être disponibles d'un seul coup, et l'ordre dans lequel elles peuvent être prises en compte n'est pas censé changer le résultat final. De plus, l'apprentissage est considéré comme un outil de description, non un moyen de découvrir et de comprendre. Il n'y a pas de place pour l'idée d'interaction avec les connaissances préalables et les théories rivales déjà opérationnelles dans le système, ni pour examiner la possibilité d'une recherche active de nouvelles données par un système cherchant à comprendre le monde.

Même en ignorant un instant les sévères limitations imposées par ce paradigme à notre compréhension de l'apprentissage et à la prise en compte de sa nature dynamique, les conséquences en termes d'application sont embarrassantes. En effet, de nombreuses tâches d'apprentissage, comme la fouille de texte, l'analyse de molécules et de processus biologiques ou la robotique, impliquent des effets de séquences et plusieurs niveaux d'organisation dans les données. Toutes bénéficiaient de théories de l'apprentissage et de méthodes permettant de sortir du cadre i.i.d., toutes auraient besoin que puissent être considérés des critères de performance qui aillent au-delà du simple taux d'erreur et qui prennent également en compte l'intégration de la connaissance produite avec les autres formes de connaissances et de théories disponibles, à la fois dans le système et dans la communauté des utilisateurs humains.

Il faut donc résolument chercher de nouveaux cadres conceptuels qui nous permettent de *sortir du cadre limité* à une distribution i.i.d. des données, des critères inductifs simples comme la minimisation du risque empirique et ses variantes, et des données et des espaces d'hypothèses supposés sans structure.

Pour cela, le programme que j'envisage est le suivant : se concentrer sur trois axes de recherche.

1. **Étude de l'apprentissage actif et de l'apprentissage guidé.** La plupart des activités d'apprentissage impliquent la recherche active de données, et non une réception passive. Certains travaux en cognition située montrent même qu'il est impossible de percevoir et d'apprendre sans autonomie et action. Dans le cadre de l'apprentissage artificiel, l'apprentissage actif peut être utile soit lorsque les données sont coûteuses à obtenir, soit au contraire lorsqu'elles sont trop abondantes et qu'il faut être capable de diriger son attention vers les informations pertinentes. Jusqu'à présent, les travaux, en nombre limité, portant sur ce sujet adoptent essentiellement un point de vue proche de l'«importance sampling» connu dans les méthodes de Monte-Carlo. L'idée est de concentrer l'échantillonnage sur les régions dans lesquelles les données peuvent faire varier le plus les hypothèses produites par le système. Si ce point de vue permet de sortir du cadre i.i.d. classique, c'est de manière très limitée. La distribution des exemples d'apprentissage est maintenant différente de la distribution en test, mais le critère de performance reste le taux d'erreur et il n'est pas question de tenir compte de l'organisation de l'espace des hypothèses ou de la structure de celles-ci (qui pourrait conduire à privilégier d'abord l'apprentissage de sous-concepts par exemple).

Une approche intéressante est de *renverser le problème et de se demander comment un professeur pourrait aider ou guider l'apprenant*. Les études théoriques sur cette question sont actuellement paralysées par le fait qu'il semble impossible de trouver un protocole à la fois raisonnable et qui empêche une collusion condamnable entre le professeur et l'apprenant. Mais c'est jeter le bébé avec l'eau du bain. En attendant de trouver une solution à ce problème, on peut sortir du cadre du PAC-apprentissage. Par exemple, nos travaux sur la transition de phase pouvant affecter la variation du taux de couverture (section 2.3) pointe vers l'idée intéressante qu'à chaque instant l'algorithme d'apprentis-

sage n'a accès qu'à un sous-espace limité de l'espace des hypothèses. Or ce sous espace peut avoir des caractéristiques particulières vis-à-vis de la couverture des exemples. Une possibilité est alors pour le professeur de fournir à chaque instant les exemples qui maximisent le gain d'information possible, c'est-à-dire ici qui optimisent le gradient local du taux de couverture.

Plus généralement, la solution n'est peut-être pas de chercher à circonscrire ou à éliminer la collusion entre le professeur et l'apprenant. Après tout, c'est au contraire cette collusion que nous cherchons à mettre en place dans les salles de classe. Ce qui en limite l'efficacité c'est qu'une connaissance ne s'acquiert jamais indépendamment d'autres connaissances et de plusieurs champs interprétatifs et de mise en applications. On peut fournir à l'étudiant une définition encyclopédique, encore lui reste-t-il à l'intégrer, à l'articuler, avec ses savoirs et ses connaissances courantes [Tib02]. Ma proposition serait donc de ne pas chercher à limiter la collusion, mais à enrichir les critères de performances pour faire place à l'apprentissage incrémental de connaissances complexes. Un bon apprenant est alors un agent capable d'intégrer les nouveaux concepts au sein de ses connaissances actuelles et d'en préparer l'interprétation, la traduction, l'expansion, dans de nouveaux domaines. Il faut alors à tout le moins être capable de mesurer des distances entre connaissances structurées.

Les apprentissages guidés par la connaissance de la structure des concepts à apprendre sont aussi à étudier.

Les *applications* se prêtant bien à ces recherches sont en particulier la recherche d'information sur le web et plus généralement la recherche d'information dans d'énormes bases de données (certaines applications de bioinformatique, données provenant des grandes expériences de physique des particules ou d'observation des rayons cosmiques, ...) ou dans des flux importants de données (e.g. vision en robotique), ainsi que lorsque les données sont coûteuses à acquérir (certaines applications de la bioinformatique, société d'agents aux ressources très limitées comme en téléphonie mobile ou autres applications du même type).

2. **Étude de l'apprentissage en-ligne et à long terme.** De plus en plus d'applications impliquent un apprentissage en-ligne, d'une part parce que les données ne deviennent souvent disponibles que progressivement, avec éventuellement une dérive de leurs caractéristiques, d'autre part parce qu'on cherche de plus en plus à avoir des algorithmes « any-time » capables de produire un résultat, même imparfait, n'importe quand. Il devient donc de plus en plus inacceptable d'être obligé de recommencer tout l'apprentissage à chaque fois que de nouvelles données arrivent, et il faut disposer d'algorithmes capables d'apprendre incrémentalement, modifiant leur modèle du monde au fur et à mesure de l'arrivée des entrées.

Il est crucial d'analyser les propriétés de ces apprentissages incrémentaux, en particulier en regard des effets de séquence, c'est-à-dire des variations de l'ordre et du rythme de présentation des entrées. J'ai argumenté largement dans ce mémoire l'intérêt de cette étude pour mieux comprendre les liens entre apprentissage et circulation de l'information. Il est aussi intéressant de pouvoir définir à quel type d'information est équivalent le choix de la séquence d'apprentissage si il a un effet sur le résultat final.

Mais généralement une meilleure connaissance des propriétés des apprentissages incrémentaux permettrait de définir de nouvelles stratégies pour l'apprentissage actif et, de manière duale, pour l'enseignement.

Comme je l'ai montré dans ce document, je crois à l'intérêt de passer par une étude des

trajectoires dans un espace de phase en utilisant des techniques issues de la physique des systèmes dynamiques et de la théorie du contrôle. Cependant, ces approches reposent sur des hypothèses fortes de continuité de la trajectoire qui ne conviennent pas à la plupart des protocoles d'apprentissage existants (sauf éventuellement en régression). Il est donc probable qu'il faille aller voir du côté des travaux sur les marches aléatoires et les flots stochastiques qui permettent de considérer des trajectoires discontinues, mais qui sont intrinsèquement très liées au cadre i.i.d. Il est clair qu'il existe là un champ théorique très important pour dépasser les limites des cadres existants.

Les *domaines d'application* concernés incluent les apprentissages à partir de séquences (temporelles, ou spatiales comme en bioinformatique), et les apprentissages en présence de flux de données, comme en robotique, ou en recherche d'information sur le web.

3. **Étude de l'apprentissage avec d'autres critères de performances.** Après s'être concentré pendant plus de deux décennies sur des critères de performances directement liés à des mesures d'erreur entre hypothèses apprises et hypothèses cibles, la communauté des chercheurs en apprentissage commence à examiner d'autres critères. Par exemple :

- L'aire sous la courbe ROC qui intègre l'espérance d'erreur en fonction des seuils de décision possibles et tend à rendre l'apprentissage plus robuste vis-à-vis des variations de l'environnement.
- La qualité d'un tri ou d'un classement.
- Le coût d'acquisition des données
- ...

Ces nouveaux critères de performance renouvellent les questions et approches théoriques, et motivent la recherche d'autres algorithmes d'apprentissage.

Toujours dans l'optique de l'apprentissage en-ligne et de l'apprentissage guidé, dans lesquels l'intégration des nouvelles informations dans les modèles et théories existant dans le système est au cœur de l'apprentissage, il serait intéressant d'envisager aussi d'autres critères mesurant le coût de cette intégration, son caractère continu ou non, la possibilité d'interaction avec des experts, artificiels ou humains.

À côté des nombreuses questions et directions de recherche mentionnées au cours de ce mémoire et qui constituent mon agenda à court terme, il existe donc un programme de recherche excitant à moyen terme. Aux Etats-Unis, il est d'usage lors des séminaires d'embauche des nouveaux professeurs de leur demander ce qu'ils feraient si ils avaient 1 million de dollars (ou deux ...) et 5 ans de liberté pour la recherche. Pour ma part, je parierais sur les directions esquissées ci-dessus. J'espère que d'autres chercheurs se retrouveront aussi dans ce programme ou un programme proche, et que j'aurais encore la chance de connaître des collaborations aussi heureuses et fructueuses que celles qui ont jalonné ma carrière jusqu'ici. Merci à tous ceux avec qui j'ai pu travaillé, et bienvenue à ceux à venir.

A

(Publications sélectionnées)

On trouvera en annexe à ce document les papiers suivants :

1. (PCS05b) N. Pernot, A. Cornuéjols and M. Sebag, « Phase transition within grammatical inference », *Int. Joint Conf. on Artificial Intelligence (IJCAI-05)*, Edinburgh, UK, 2005, (Ed. L. P. Kaelbling), pp.811-816
2. (JMCMS04) K. Jong, J. Mary, A. Cornuéjols, E. Marchiori and M. Sebag. « Ensemble feature ranking », *Principles of Knowledge Discovery in Databases (PKDD-04)*, Pisa, Italy, 2004 Springer-Verlag, LNAI-3202, 267-278.
3. (CFM05) A. Cornuéjols, Ch. Froidevaux and J. Mary. « Comparing and combining feature estimation methods for the analysis of microarray data », *JOBIM-05 : Journées Ouvertes Biologie Informatique Mathématiques*, Lyon, France, 2005.
4. (Cor96a) A. Cornuéjols. « Analogie, principe d'économie et complexité algorithmique », *Journées Francophones d'Apprentissage (JFA-96)*, Sète, France, 1996, pp.233-247.
5. (Cor93b) A. Cornuéjols. « Getting Order Independence in Incremental Learning », *European Conference on Machine Lerarning (ECML-93)*, (Ed. P. Brazdil), Vienna, Austria (1993). Springer-Verlag, LNAI-667, pp. 196-212.
6. (CTC00) A. Cornuéjols, A. Tiberghien and G. Collet « A new mechanism for transfer between conceptual domains in scientific discovery and education ». *Foundations of Science*, vol.5, No.2, (2000), 129-155.

Annexe A. (Publications sélectionnées)

B
(Curriculum Vitae)

Annexe B. (Curriculum Vitae)

C
(Liste de publications)

Annexe C. (Liste de publications)

Index

Voici un index

alphabet auxiliaire d'une grammaire, 30
alphabet terminal d'une grammaire, 30

Couverture

d'une hypothèse, 18
taux de, 18

grammaire formelle, 30
grammaire hors-contexte, 31
grammaire régulière, 31

langage engendré par une grammaire, 31
langage hors-contexte, 31
langage régulier, 31

mot engendré par une grammaire, 31
mots dérivés selon une grammaire, 31

règle de production d'une grammaire, 30

Bibliographie

- [AB92] M. Anthony and N. Biggs. *Computational Learning Theory*. Cambridge University Press, 1992.
- [AB99] M. Anthony and P. Bartlett. *Neural network learning : theoretical foundations*. Cambridge University Press, 1999.
- [AB00] J. Ales-Bianchetti. *Le raisonnement par analogie. Une unification des modèles cognitifs et des théories de l'induction pour l'étude du raisonnement par analogie*. Ph.d., Université de Paris-Sud, Orsay, 2000.
- [ABNK⁺87] S.-I. Amari, O. Barndorff-Nielsen, R. Kass, S. Lauritzen, and C. Rao, editors. *Differential geometry in statistical inference*, volume 10. Institute of Mathematical Statistics, Hayward, California, 1987.
- [AKKK01] D. Achlioptas, L. Kirousis, E. Kranakis, and D. Krizanc. Rigorous results for (2+p)-sat (with l.m. kirousis, e. kranakis, d. krizanc) theoretical computer science, 265 (1-2), (2001), p.109-129. *Theoretical Computer Science*, 265(1-2) :109–129, 2001.
- [AM98] N. Abe and H. Mamitsuka. Query learning strategies using boosting and bagging. In *International Conference on Machine Learning (ICML-98)*, pages 1–9, Madison, Ill., 1998.
- [AM02a] E. Alphonse and S. Matwin. Feature subset selection and inductive logic programming. In *International Conference on Machine Learning (ICML-02)*, Sidney, Australia, 2002. Morgan Kaufmann.
- [AM02b] C. Ambroise and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences, USA*, 99(10) :6562–6566, 2002.
- [Ang82] D. Angluin. Inference of reversible langauges. *Communications of the ACM*, 29 :741–765, 1982.
- [AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Very Large Data Bases (VLDB-94)*, pages 487–499, Santiago, Chile, 1994.
- [AU72] A. Aho and J. Ullman. *The theory of Parsing, Translation and Compiling, Vol 1 : Parsing*. Prentice-Hall, 1972.
- [Bad90] A. Baddeley. *Human Memory. Theory and practice*. Allyn and Bacon, 1990.
- [BBL05] O. Bousquet, S. Boucheron, and G. Lugosi. Theory of classification : A survey of recent advances. *ESAIM : Probability and Statistics*, 2005. Une synthèse remarquable des résultats récents sur la théorie statistique de l'apprentissage.
- [BCM01] G. Biroli, S. Cocco, and R. Monasson. Transitions de phases et complexité en informatique : le temps d'un choix. In *Images de la Physique*. CNRS-Editions, 2001.

Bibliographie

- [BDKM97] S. Ben-David, E. Kushilevitz, and Y. Mansour. Online learning versus offline learning. *Machine Learning Journal*, 29 :45–63, 1997.
- [BE02] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2 :499–526, 2002.
- [BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM*, (36) :929–965, 1989.
- [BGSS99] M. Botta, A. Giordana, L. Saitta, and M. Sebag. Relational learning : Hard problems and phase transition. In *AIIA '99*, pages 178–189. Springer-Verlag, 1999.
- [BGSS03] M. Botta, A. Giordana, L. Saitta, and M. Sebag. Relational learning as search in a critical region. *Journal of Machine Learning Research*, 4 :431–463, 2003.
- [BK92] E. Baum and L. Kevin. Query learning can work poorly when a human oracle is used. In *Int. Joint Conf. in Neural Networks*, Beijing, China, 1992.
- [BL97] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence journal*, (97) :245–271, 1997.
- [Blu94] A. Blum. Separating distribution-free and mistake-bound learning models over the boolean domain. *SIAM Journal on Computing*, 23(5) :990–1000, 1994.
- [BLV03] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4 (Special issue on learning theory) :861–894, 2003.
- [Bot03] L. Bottou. Stochastic learning. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advance lectures on machine learning. ML summer school 2003*, volume LNAI-3176, pages 146–168. Springer-Verlag, 2003.
- [Bou03] O. Bousquet. New approaches to statistical learning theory. *Annals of the Institute of Statistical Mathematics*, 55(2) :371–389, 2003.
- [Bou05] O. Bousquet. Svm, noyaux, régularisation. quelques idées récentes en machine learning, Tutoriel à CAP-05, Nice, 30-05-05 2005.
- [Bre98] L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3) :801–849, 1998.
- [Bre02] N. Bredèche. *Ancrage de lexique et perceptions : changements de représentation et apprentissage dans le cadre d'un agent situé et mobile*. Ph.d., Université de Paris-Sud, Orsay, 2002.
- [BRS02] J. A. Bianchetti, C. Rouveiro, and M. Sebag. Inductive logic programming out of phase transition : A bottom-up constraint-based approach. In *International Conference on Machine Learning (ICML-02)*, pages 35–42, The University of New South Wales, Sydney, Australia, 2002.
- [BS95] A. Bell and T. Sejnowski. Blind separation and blind deconvolution : an information-theoretic approach. In *ICASSP*, Detroit, USA, 1995.
- [BSZar] N. Bredèche, Z. Shi, and J.-D. Zucker. Perceptual learning and abstraction in machine learning : an application to autonomous robotics. *IEEE Transactions on Systems, Man, and Cybernetics, Part C : Applications and Reviews*, To appear.
- [BTC96] A. Brunie-Taton and A. Cornuéjols. Classification en programmation génétique. In *11ème Journées Françaises d'Apprentissage (JFA-96)*, pages 303–316, Sètes, France, 1996.
- [Bur95] B. Burns. Fluid concepts and creative analogy : A review. *AI Magazine*, 16(3) :81–83, 1995.

-
- [CAB98] A. Cornuéjols and J. Ales-Bianchetti. Analogy and induction : which (missing) link ? In Keith Holyoak, Dedre Gentner, and Boicho Kokinov, editors, *Workshop "Advances in Analogy research : Integration of theory ans data from cognitive, computational and neural sciences*, pages 365–372, Sofia, Bulgaria, 1998. New Bulgarian University.
- [CAL94] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning Journal*, 15(2) :201–221, 1994.
- [Car04] S. Carroll. *An introduction to general relativity. Spacetime and geometry*. Pearson / Addison-Wesley, 2004.
- [CC87] J.-P. Cassou and A. Cornuéjols. Statistical filtering of motion field from image sequences. In *GRETSI-87*, Nice, France, 1987.
- [CFM05] A. Cornuéjols, C. Froidevaux, and J. Mary. Comparing and combining feature estimation methods for the analysis of microarray data. In *JOBIM-05 : Journées Ouvertes Biologie Informatique Mathématiques*, Lyon, France, 2005.
- [CG91] C. Clement and D. Gentner. Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, 15 :89–132, 1991. Not owned.
- [CG98] J. Castro and D. Guijarro. Query, pacs and simple-pac learning. Technical report, Dept Llengatges i Sistemes Informatics, 1998.
- [CGJ96] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4 :129–145, 1996.
- [CM02] A. Cornuéjols and Laurent Miclet. *Apprentissage Artificiel. Concepts et Méthodes*. Eyrolles, 2002.
- [CMCCT97] E. Cauzinille-Marmèche, G. Collet, A. Cornuéjols, and A. Tiberghien. Co-adaptation of students' knowledge domains when interpreting a physical situation in terms of a new theory. In *2nd European Conference on Cognitive Sciences (ECCS'97)*, pages 107–112, Manchester, 1997.
- [Col00] G. Collet. *Langage et modelisation scientifique : Le verbe, levier de l'apprentissage*. Editions du CNRS, 2000.
- [Col02] I. Colombet. *Aspects méthodologiques de la prédiction du risque cardiovasculaire : apports de l'apprentissage automatique*. Ph.d., Paris-5, 2002.
- [Com01] F. Comte. Quelques contributions à l'étude de la dépendance en statistiques et en économétrie. Technical report, Paris 6, Laboratoire de Probabilités et Modèles aléatoires, Mars 2001 2001.
- [Cor89a] A. Cornuéjols. *De l'apprentissage incrémental par adaptation dynamique : le système INFLUENCE*. Ph.d, Paris-Sud, Orsay, 1989.
- [Cor89b] A. Cornuéjols. An exploration into incremental learning : the influence system. In *International Conference on Machine Learning (ICML-89)*, pages 383–386, Cornell, USA, 1989. ACM Press.
- [Cor93a] A. Cornuéjols. Getting order independence in incremental learning. In Pavel Brazdil, editor, *European Conference on Machine Lerarning (ECML-93)*, volume LNAI-667, pages 196–212, Vienna, Austria, 1993. Springer-Verlag.
- [Cor93b] A. Cornuéjols, editor. *Training Issues in Incremental Learning*. AAAI Press, Stanford, USA, 1993.

Bibliographie

- [Cor94a] A. Cornuéjols. Analogy as a minimization principle. In *Dagstuhl Seminar "Theory and Praxis of Machine Learning"*, Dagstuhl, Germany, 1994.
- [Cor94b] A. Cornuéjols. Analogy as description minimization principle. In *Workshop on "Applications of Descriptive Complexity to Inductive, Statistical and Visual Inference" ICML-COLT-94*, Rutgers, USA, 1994.
- [Cor94c] A. Cornuéjols. Induction from one example and statistics : analogy as a minimization principle. In *Workshop on "Machine Learning and Statistics" (ECML-94)*, Catanes, Italy, 1994.
- [Cor96a] A. Cornuéjols. Analogie, principe d'économie et complexité algorithmique. In *Journées Francophones d'Apprentissage (JFA-96)*, pages 233–247, Sètes, France, 1996.
- [Cor96b] A. Cornuéjols. Analogy as a minimization of description length. In G. Nakhaeizadeh and C. Taylor, editors, *Machine Learning and Statistics : The Interface*. John Wiley and Sons, 1996.
- [CS05] A. Cornuéjols and M. Sebag. A note on phase transitions and computational pitfalls of learning from sequences. In *Second Franco-Japanese Workshop on Information Search, Integration and Personalization (ISIP-05)*, Lyon, France, 2005.
- [CSM04] A. Cornuéjols, M. Sebag, and J. Mary. Classification d'images à l'aide d'un codage par motifs fréquents. In *Workshop sur la fouille d'images (RFIA-04)*, Toulouse (France), 2004.
- [CTC] A. Cornuéjols, A. Tiberghien, and G. Collet. Tunnel effects in cognition : A new mechanism for scientific discovery and education. *International Journal on Computer-Human Studies (submitted to)*.
- [CTC98] A. Cornuéjols, A. Tiberghien, and G. Collet. Decomposing the scientific discovery process using multiple interpretations of notions. In *1st Conference on Model-Based Reasoning (MBR-98)*, Pavia, Italy, 1998.
- [CTC99a] A. Cornuéjols, A. Tiberghien, and G. Collet. L'effet tunnel en cognition : un mécanisme de transfert entre domaines conceptuels. In Michèle Sebag, editor, *Conférence d'Apprentissage (CAP-99)*, pages 243–250, Palaiseau, France, 1999.
- [CTC99b] A. Cornuéjols, A. Tiberghien, and G. Collet. Tunnel effects in cognition : A transfer mechanism from known conceptual domains to new ones. In *Artificial Intelligence and Simulation of Behaviour (AISB-99)*, pages 9–18, Edinburgh, Great-Britain, 1999.
- [CZ02] S. Climer and W. Zhang. Searching for backbones abd fat : A limit-crossing approach with applications. In *National Conf. on Artificial Intelligence (AAAI-02)*, 2002.
- [DB95] T. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2 :263–286, 1995.
- [DCCF04] M. Dutreix, J.-P. Comet, A. Cornuéjols, and C. Froidevaux. Determination of cellular drug targets : searching for functional information in the jungle of microarrays data. In *Current Trends in Drug Discovery Research (CTDDR-04)*, India, 2004.
- [DD01] O. Dubois and G. Dequen. A backbone-search heuristic for efficient solving of hard 3-sat formulae. In *Int. Conf. on Artificial Intelligence (IJCAI-01)*, pages 248–253, 2001.

-
- [DDG96] F. Denis, C. D’Halluin, and R. Gilleron. Pac learning with simple examples. In *13th Annual Symposium on Theoretical Aspects of Computer Science*, volume LNCS-1046, pages 231–242, Grenoble, France, 1996. Springer.
- [Del94] J.-P. Delahaye. *Information, complexité et hasard*. Hermès, 1994.
- [DFLS04] P. Duchon, P. Flajolet, G. Louchard, and G. Schaeffer. Boltzmann samplers for the random generation of combinatorial structures. *Combinatorics, Probability, and Computing*, 13(4-5 (Special issue on Analysis of Algorithms)) :577–625, 2004.
- [DG97] F. Denis and R. Gilleron. Pac learning under helpful distributions. In *Workshop on Algorithmic Learning Theory (ALT-97)*, volume LNAI-1316, pages 132–145, Berlin, 1997. Springer.
- [DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer Verlag, 1996.
- [DGS97] F. Denis, R. Gilleron, and J. Simon. Apprentissage pac avec enseignant. In *JFA-97*, pages 175–186, 1997.
- [D’I99] R. D’Inverno. *Introducing Einstein’s relativity*. Oxford University Press, 1999.
- [dlH96] C. de la Higuera. Ensembles caractéristiques en inférence grammaticale. *Rapport interne : LIRMM*, 1996.
- [dlH97] C. de la Higuera. Characteristic sets for polynomial grammatical inference. *Machine Learning*, (27) :125–138, 1997. Kluwer Academic Publishers. Manufactured in Netherland.
- [dlHOV96] C. de la Higuera, J. Oncina, and E. Vidal. Identification of dfa : data-dependent versus data-independant algorithms. *ICGI’96*, 1996.
- [DM98] P. Dupont and L. Miclet. Inférence grammaticale régulière : fondements théoriques et principaux algorithmes. Technical Report 3449, INRIA, 1998.
- [Dod03] Y. Dodge. *Premier pas en statistique*. Springer-Verlag, 2003.
- [DT03] N. Denquive and P. Tarroux. Codages fréquentiels et catégorisation de scènes visuelle. Technical report, CNRS-LIMSI, 2003.
- [Dup94] P. Dupont. Regular grammatical inference from positive and negatives samples by genetic search : the gig method. In *Grammatical Inference and Applications*, volume LNAI-862, pages 236–245. Springer Verlag, 1994.
- [EdB01] A. Engel and C. Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.
- [EHW96] T. Hogg (Eds), B. Huberman, and C. Williams. Special volume on frontiers in problem solving : phase transitions and complexity. *Artificial Intelligence journal*, 81(1-2), 1996.
- [Eng96] T. English. Evaluation of evolutionary and genetic optimizers : No free lunch. In L.J. Fogel, P.J. Angeline, and T. Bäck, editors, *Evolutionary Programming V : Proc. of the Fifth Annual Conf. on Evolutionary Programming*, pages 163–169. MIT Press, 1996.
- [FA85] Y. Fu and P. Anderson. Application of statistical mechanics to np-complete problems in combinatorial optimisation. *Journal of Physics A : Mathematics and General*, 19(9) :1605–1620, 1985.
- [Fau85] G. Fauconnier. *Mental spaces*. University of Chicago Press, 1985.

Bibliographie

- [Fau97] G. Fauconnier. *Mappings in thought and language*. Cambridge University Press, Cambridge, 1997.
- [FFG89] B. Falkenhainer, K. Forbus, and D. Gentner. The structure-mapping engine : algorithm and examples. *Artificial Intelligence journal*, 41 :1–63, 1989.
- [FKGL94] Forbus, Kenneth, D. Gentner, and K. Law. Mac/fac : A model of similarity-based retrieval. *Cognitive Science*, 19 :141–205, 1994.
- [Fre95] R. French. *The subtlety of sameness*. The MIT Press, 1995. Not owned.
- [FRPT99] C. Fonlupt, D. Robillard, P. Preux, and E.-G. Talbi. Fitness landscape and performance of metaheuristics. In *Meta-heuristics - Advances and Trends in Local Search Paradigms for Optimization*, pages 255–266. Kluwer Academic Press, 1999.
- [FSST97] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning Journal*, 28 :133–168, 1997. Apparemment (d’après mon survol) très important. 12-03-02.
- [FT98] G. Fauconnier and M. Turner. Conceptual integration networks. *Cognitive Science*, 22(2) :133–187, 1998.
- [Gan02] J.-G. Ganascia. *Gédéon ou les aventures extravagantes d'un expérimentateur en chambre*. Le Pommier (série "Romans & plus"), 2002.
- [GBS99] A. Giordana, M. Botta, and L. Saitta. An experimental study of phase transitions in matching. In *Int. Joint Conf. on Artificial Intelligence (IJCAI-99)*, pages 1198–1203, Stockholm, Sweden, 1999.
- [GE03] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3 :1157–1182, 2003.
- [Gen83] D. Gentner. Structure-mapping : A theoretical framework for analogy. *Cognitive Science*, 7 :155–170, 1983.
- [Gen89] D. Gentner. The mechanisms of analogical reasoning. In Vosnadiou and Ortony, editors, *Similarity and analogical reasoning*. Cambridge University Press, 1989.
- [Gen94] Y. Gentilhomme. L’éclatement du signifié dans les discours technoscientifiques. *Cahiers de lexicologie*, 64, 1994.
- [GH04] M. Gitterman and V. Halpern. *Phase transitions. A brief account with modern applications*. World Scientific, 2004.
- [Gla04] P. Glasserman. *Monte Carlo methods in financial engineering*. Springer, 2004.
- [GMP] P. Grünwald, I. J. Myung, and M. Pitt, editors. *Advances in Minimum Description Length : Theory and Applications*. MIT Press.
- [Gor02] M. Gordon. An introduction to statistical physics. Technical report, Leibniz-IMAG, 2002.
- [GS00] A. Giordana and L. Saitta. Phase transitions in relational learning. *Machine Learning Journal*, 41 :217–251, 2000.
- [GS05] C. Gomes and B. Selman. Can get satisfaction. *Nature*, 435 :751–752, 2005.
- [GSSB00a] A. Giordana, L. Saitta, M. Sebag, and M. Botta. Analyzing relational learning in the phase transition framework. In *Int. Conf. on Machine Learning (ICML-00)*, pages 311–318, Stanford, USA, 2000.
- [GSSB00b] A. Giordana, L. Saitta, M. Sebag, and M. Botta. La programmation logique induc-tive à la lumière de la transition de phase. In Colin de la Higuera, editor, *Conférence d'Apprentissage (CAP-00)*, pages 157–172, Saint-Etienne, France, 2000. Hermès.

-
- [GV03] P. Grünwald and P. Vitanyi. Kolmogorov complexity and information theory. *Journal of Logic, Language and Information*, 12 :497–529, 2003.
- [Hau89] D. Haussler. Learning conjunctive concepts in structural domains. *Machine Learning*, 4(1) :7–40, 1989.
- [Hay97] B. Hayes. Can't get no satisfaction. *American Scientist*, 85(2) :108–112, 1997.
- [HB88] D. Horswill and R. Brooks. Situated vision in a dynamic world : Chasing objects. In *AAAI-98*, pages 796–800, St. Paul, MN, USA, 1988.
- [HGK] K. Holyoak, D. Gentner, and B. Kokinov, editors. *Advances in Analogy research : Integration of theory ans data from cognitive, computational and neural sciences*. New Bulgarian University.
- [HKO01] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.
- [HKS94] D. Haussler, M. Kearns, and R. Schapire. Bounds on the sample complexity of bayesian learning using information theory and the vc dimension. *Machine Learning Journal*, 14 :83–113, 1994.
- [HKT89] Holyoak, Keith, and Thagard. Analogical mapping by constraint satisfaction. *Cognitive Science*, 13 :295–355, 1989.
- [HO00] A. Hyvärinen and E. Oja. Independent component analysis : Algorithms and applications. *Neural Networks*, 13 :411–430, 2000.
- [Hof85] D. Hofstadter. Analogies and roles in human and machine thinking. In Douglas Hofstadter, editor, *Metamagical themas*, page Cha.24. Bantam Books, 1985.
- [Hof95] D. Hofstadter. *Fluid concepts and creative analogies*. Basic Books, 1995.
- [Hol73] G. Holton. *Thematic origins of scientific thought. Kepler to Einstein*. Harvard University Press, 1973.
- [Hyv99] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2 :94–128, 1999.
- [IAZ00] V. Iyengar, C. Apte, and T. Zhang. Active learning using adaptive resampling. In *ACM SIGKDD-2000*, 2000.
- [Jay03] E. Jaynes. *Probability theory : the logic of science*. Cambridge University Press, 2003.
- [JCS⁺03] S. Jouteau, A. Cornuéjols, M. Sebag, P. Tarroux, and J.-S. Liénard. Nouveaux résultats en classification à l'aide d'un codage par motifs fréquents. *Revue d'Intelligence Artificielle (Proc. of the EGC-03 Conf.)*, 17(1-3) :521–532, 2003.
- [JMC⁺04] K. Jong, J. Mary, A. Cornuéjols, E. Marchiori, and M. Sebag. Ensemble feature ranking. In *Principles of Knowledge Discovery in Databases (PKDD-04)*, volume LNAI-3202, pages 267–278, Pisa, Italy, 2004. Springer-Verlag.
- [Jou02] S. Jouteau. *Reconnaissance de scènes naturelles*. Master's thesis, Paris-6, 2002.
- [JS98] J. José and E. Saletan. *Classical dynamics. A contemporary approach*. Cambridge University Press, 1998.
- [JS04] J.-M. Jolian and I. Simand. Représentation d'images par des chaînes de symboles : application à l'indexation. In *Compression et Représentation des Signaux Audiovisuel (CORESA '04)*, Lille, France, 2004.

- [KJ97] R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence journal*, pages 273–324, 1997.
- [KLD94] M. Keane, T. Ledgeway, and S. Duff. Constraints on analogical mapping : A comparison of three models. *Cognitive Science*, 18 :387–438, 1994.
- [KMF04] S. Kiritchenko, S. Matwin, and A. Fazel Famil. Functional annotation of genes using hierarchical text categorization. In *BioLINK SIG : Linking Literature, Information and Knowledge for Biology*, pages 1–4, Detroit, Michigan, 2004.
- [Koe60] A. Koestler. *Les somnambules (The Sleepwalkers : A History of Man's Changing Vision of the Univers)*. Calman-Lévy, 1960.
- [Koe64] A. Koestler. *Le cri d'Archimède (The act of creation)*. 1964.
- [KR92] K. Kira and L. Rendell. A practical approach to feature selection. In *Int. Conf. on Machine Learning (ICML-92)*, pages 249–256. Morgan Kaufmann, 1992.
- [KS] H. Krautz and B. Selman. The state of sat. *Discrete and Applied Math*, ((to appear)).
- [Kul68] S. Kullback. *Information theory and statistics*. Dover, 2nd (first, 1959) edition, 1968.
- [KV94] M. Kearns and U. Vazirani. *An introduction to computational learning theory*. MIT Press, 1994.
- [Lan92] K. Lang. Random dfa's can be approximately learned from sparse uniform examples. *5th ACM workshop on Computation Learning Theory*, pages 45–52, 1992.
- [LB87] K. Van Lehn and W. Ball. A version space approach to learning context-free grammars. *Machine Learning Journal*, 2(1) :39–74, 1987.
- [LG94] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *International ACM-SIGIR Conf. on Research and Development in Information Retrieval*, pages 3–12, 1994.
- [LL03] M. Laguës and A. Lesne. *Invariances d'échelle. des changements d'états à la turbulence*. Belin, 2003.
- [LPP98] K. J. Lang, B. A. Pearlmuter, and R. A. Price. Results of the abbadingo one DFA learning competition and a new evidence-driven state merging algorithm. *Lecture Notes in Computer Science*, 1433 :1–12, 1998.
- [LR06] E. Lehtinen and F. Ritter, editors. *In Order to Learn : How ordering processes and sequencing effects in machines illuminate human learning and vice-versa*. Cambridge University Press, 2006.
- [LS99] D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 :788–791, 1999.
- [LV94] J.-H. Lin and J. S. Vitter. A theory for memory-based learning. *Machine Learning Journal*, 17 :1–26, 1994.
- [LV97] M. Li and P. Vitanyi. *An introduction to Kolmogorov complexity and its applications (2nd ed.)*. Springer-Verlag, 1997.
- [Mac04] D. MacKay. *Information theory, Inference, and Learning Algorithms*. Cambridge University Press, 2004.

-
- [MBD05] L. Miclet, S. Bayoudh, and A. Delhay. Définitions et premières expériences en apprentissage par analogie dans les séquences. In *CAP-05 (Conférence d'Apprentissage)*, pages 31–48, Nice, France, 2005. PUG.
- [MBM⁺04] G. Mercier, N. Berthault, J. Mary, A. Antoniadis, J.-P. Comet, A. Cornuéjols, C. Froidevaux, and M. Dutreix. Biological detection of low radiation by combining results of two analysis methods. *Nucleic Acids Research (NAR)*, 32(1) :1–8, 2004.
- [Meg95] O. Megalakaki. Expériene marianne : corpus de dialogue de trois groupes d’élèves résolvant une séquence de problèmes mettant en jeu une activité de modélisation. Technical Report CR-11/95, CNRS-COAST, 1995.
- [Mic80] L. Miclet. Regular inference with a tail-clustering method. *IEEE Transactions on SMC*, SMC-10 :737–743, 1980.
- [Mit82] T. Mitchell. Generalization as search. *Artificial Intelligence*, 18 :203–226, 1982.
- [Mit93] M. Mitchell. *Analogy-making as perception*. MIT Press, 1993.
- [Mit97] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.
- [MM04] P. Melville and R. Mooney. Diverse ensembles for active learning. In *Int. Conf. on Machine Learning (ICML-04)*, 2004.
- [MMC⁺03] J. Mary, G. Mercier, J.-P. Comet, A. Cornuéjols, C. Froidevaux, and M. Dutreix. An attribute estimation technique for the analysis of microarray data. In Philippe Amar, François Képès, Victor Norris, and P. Tracqui, editors, *Proceedings of the Dieppe School on Modelling and Simulation of Biological processes in the Context of Genomics*, pages 69–77. Publisher Frontier Group, 2003.
- [MMC⁺04] J. Mary, G. Mercier, J.-P. Comet, A. Cornuéjols, C. Froidevaux, and M. Dutreix. Utilisation d’une méthode d’estimation d’attributs pour l’analyse du transcriptome de cellules de levures exposées à de faibles doses de radiation. In Jean-François Boulicaut and Olivier Gandrillon, editors, *Informatique pour l’analyse du transcriptome*, pages 189–205. Hermès, 2004.
- [MMR01] O. Martin, R. Monasson, and Z. Riccardo. Statistical mechanics methods and phase transitions in optimization problems. *Theoretical Computer Science*, 265 :3–67, 2001.
- [MNT99] L. Magnani, N. Nersessian, and P. Thagard, editors. *Model-Based Reasoning in Scientific Discovery*. Kluwer Academic / Plenum Publishers, New York, 1999.
- [Mon02] R. Monasson. Threshold phenomena and complexity : a statistical physics analysis of the random satisfiability problem. In *School on complexity, CIRM, Marseille, 28 janvier - 1 février 2002*. (<http://www.lpt.ens.fr/~monasson>), 2002.
- [MPV87] M. Mézard, G. Parisi, and G. Virasoro. *Spin glass theory and beyond*. World Scientific, Singapore, 1987.
- [MR93] M. Murray and J. Rice. *Differential geometry and statistics*. Chapman and Hall, 1993.
- [MRK⁺99] R. Monasson, Z. Riccardo, S. Kirkpatrick, B. Selman, and L. Troyansky. Determining computational complexity from characteristic ‘phase transitions’. *Nature*, 400 :133–137, 1999.
- [MS04] J. Maloberti and M. Sebag. Fast theta-subsumption with constraint satisfaction algorithms. *Machine Learning Journal*, 55 :137–174, 2004.

Bibliographie

- [MT95] O. Megalakaki and A. Tiberghien. Corpus de dialogues de trois groupes d'élèves résolvant trois problèmes mettant en jeu une activité de modélisation. Technical Report CR-10/95, CNRS-COAST, 1995.
- [MT03] J. Machrouh and P. Tarroux. Perceptual agents : a situated framework for image analysis. In *WAPCV 03. International Workshop on Attention and Performance in Computer Vision*, page 6 pages, Graz, Autriche, 2003.
- [MU83] T. Mitchell and P. Utgoff. Learning by experimentation : Acquiring and redening problem solving heuristics. In Tom Mitchell, Ryszard Michalski, and Jaime Carbonell, editors, *Machine Learning : An Articial Intelligence Approach*, pages 137–162. Morgan Kaufmann, 1983.
- [OF96] B. Olshausen and D. Field. Sparse coding with an overcomplete basis set : A strategy employed by v1 ? *Vision Research*, 37(23) :3311–3325, 1996.
- [OG92] J. Oncina and P. García. Inferring regular languages in polynomial updated time. In *Pattern Recognition and Image Analysis : Selected papers from the IVth Spanish Symposium*, pages 49–61. World Scientific, 1992.
- [Opp99] M. Opper. A bayesian approach to online learning. In David Saad, editor, *On-line learning in neural networks*, pages 363–378. Cambridge University Press, 1999.
- [PCS05a] N. Pernot, A. Cornuéjols, and M. Sebag. Phase transition in grammatical inference. In François Denis, editor, *CAP-05 (Conférence Francophone d'Apprentissage)*, pages 49–60, Nice, France, 2005. PUG.
- [PCS05b] N. Pernot, A. Cornuéjols, and M. Sebag. Phase transition within grammatical inference. In Leslie Pack Kaelbling, editor, *Int. Joint Conf. on Artificial Intelligence (IJCAI-05)*, pages 811–816, Edinburgh, UK, 2005.
- [Per04] N. Pernot. *Exploration d'espaces d'hypothèses en présence de transition de phase*. Master degree, Université de Paris-Sud, Orsay et CNRS UMR-8623, 2004.
- [PF91] S. Porat and J. Feldman. Learning automata from ordered examples. *Machine Learning*, 7 :109–138, 1991.
- [PH97] R. Parekh and V. Hanovar. Learning dfa from simple examples. In Ming Li and Akira Maruoka, editors, *Workshop on Algorithmic Learning Theory (ALT-97)*, volume LNAI-1316, pages 116–131, Berlin, Germany, 1997. Springer.
- [Pia37] J. Piaget. *The construction of reality in the child*. New-York, Basic Books (1952), 1937.
- [Pia75] J. Piaget. *The development of thought : Equilibration of cognitive structures*. New-York, Viking Press (1977), 1975.
- [Pin01] S. Pinto. Etude du phénomène de transition de phase dans l'induction supervisée. Technical Report 18122, L.R.I., Université de Paris-Sud, Orsay et CNRS UMR-8623, 1-07-2001 2001.
- [PlS03] *Les illusions des sens (Numéro Spécial)*, volume 39. Pour la Science, Avril 2003.
- [Pol57] G. Polya. *How to Solve It (2nd ed.)*. Princeton University Press, 1957.
- [PP79] M. Piattelli-Palmarinio, editor. *Théories du langage. Théories de l'apprentissage. le débat entre Jean Piaget et Noam Chomsky*. Seuil, France., 1979.
- [Pro96] P. Prosser. An empirical study of phase transition in binary constraint satisfaction problem. *Artificial Intelligence journal*, (81) :81–109, 1996.

-
- [PRZ02] J.-M. Park, J. Reed, and Q. Zhou. Active feature selection in optic nerve data using support vector machines. In *Int. Joint Conf. on Neural Networks (IJNN-02)*, 2002.
 - [RGG01] S. Risau-Gusman and M. Gordon. Statistical mechanics of learning with soft margin classifiers. *Phys. Rev. E*, 64, 2001.
 - [RLJS05] J. Ros, C. Laurent, J.-M. Jolion, and I. Simand. Comparing string representations and distances in a natural images classification task. In Brun (Eds) and M. Vento, editors, *Graph-Based Representations in Pattern Recognition (GbRPR 2005)*, volume LNCS 3434, pages 72–81, Poitiers, France, 2005. Springer-Verlag.
 - [RM01] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Int. Conf. on Machine Learning (ICML-01)*, 2001.
 - [Ros05] R. Ros. *Transition de phase en apprentissage artificiel. Pistes pour sa mise en évidence en robotique*. Master’s thesis, Université de Paris-Sud, Orsay, et C.N.R.S. UMR-8623, 2005.
 - [RSK03] M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning Journal*, 53 :23, 2003.
 - [Rus87] S. Russell. Analogy and single-instance generalization. In *Fourth International Workshop on Machine Learning (IWML-87)*, 1987.
 - [Rus89] S. Russell. *The use of knowledge in analogy and induction*. Pitman Publishing / Morgan Kaufmann, 1989.
 - [SAL04] M. Sebag, J. Azé, and N. Lucas. Roc-based evolutionary learning : Application to medical data-mining. In *Artificial Evolution VI*, volume LNCS-2936, pages 384–396. Springer-Verlag, 2004.
 - [SBS99] B. Schölkopf, C. Burges, and A. Smola. *Advances in Kernel Methods. Support Vector Learning*. MIT Press, 1999.
 - [SBSS00] A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans. *Advances in large margin classifiers*. MIT Press, 2000.
 - [SDDO03] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar. Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research*, 3 :1399–1414, 2003.
 - [Sel95] B. Selman. Stochastic search and phase transition : Ai meets physics. In *International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 998–1002, 1995.
 - [SGS01] A. Serra, A. Giordana, and L. Saitta. Learning on the phase transition edge. In *Int. Joint Conf. on Artificial Intelligence (IJCAI-01)*, pages 921–926, Seattle, USA, 2001.
 - [SL90] J. Shrager and P. Langley, editors. *Computational models of discovery and theory formation*. Morgan Kaufman, 1990. Not owned.
 - [SM90] A. Shinohara and S. Miyano. Teachability in computational learning. In *ALT’90*, pages 246–257, Tokyo, 1990. Springer Verlag.
 - [SOS92] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT’92*, pages 287–294, Pittsburgh, PA, USA, 1992. ACM Press.
 - [SPBJ04] I. Simand, D. Pellerin, S. Bress, and J.-M. Jolion. Spatio-temporal signature for video copy detection, advanced concepts for intelligent vision. In *Advanced Concepts for Intelligent Vision Systems (ACIVS’04)*, pages 421–427, Brussels, Belgium, 2004.

Bibliographie

- [SS02] B. Schölkopf and A. Smola. *Learning with kernels. Support vector machines, regularization, optimization, and beyond.* MIT Press, 2002.
- [STC04] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis.* Cambridge University Press, 2004.
- [STP04] M. Saar-Tsechansky and F. Provost. Active sampling for class probability estimation and ranking. *Machine Learning Journal*, (54) :153–178, 2004.
- [SW01] J. Slaney and T. Walsh. Backbones in optimization and approximation. In *17th Int. Conf. on Artificial Intelligence (IJCAI-01)*, 2001.
- [SY05] N. Stroppa and F. Yvon. Apprentissage par analogie et rapports de proportion : contributions méthodologiques et expérimentales. In *CAP-05 (Conférence d'Apprentissage)*, pages 61–62, Nice, France, 2005. PUG.
- [SZ00] L. Saitta and J.-D. Zucker. Abstraction and phase transitions. In *Symposium on Abstraction, Reformulation and Approximation (SARA '00)*, volume LNAI-1864, pages 291–302, HorseshoeBay (Lake LBJ), Texas, 2000. Springer-Verlag.
- [Tal00] R. Talman. *Geometric mechanics.* Wiley Inter-Science, 2000.
- [Tas04] P. Tassi. *Méthods statistiques.* Economica, 2004.
- [Tib94] A. Tiberghien. Modelling as a basis for analysing teaching-learning situations. *Learning and Instructions*, 4(1) :71–87, 1994.
- [Tib96] A. Tiberghien. Construction of prototypical situations in teaching the concept of energy. In G. Welford, J. Osborne, and P. Scott, editors, *Research in science and education in Europe*, pages 100–114. Falmer Press, London, 1996.
- [Tib02] A. Tiberghien, editor. *Des connaissances naïves au savoir scientifique.* ACI Ecole et Sciences Cognitives, 2002.
- [Ton01] S. Tong. *Active learning : theory and applications.* Phd., Stanford University, 2001.
- [TTC01] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences, USA*, 98(9) :5116–5121, 2001.
- [Utg89] P. Utgoff. Incremental induction of decision trees. *Machine Learning Journal*, 4 :161–186, 1989.
- [Utg94] P. Utgoff. An improved algorithm for incremental induction of decision trees. In *Int. Conf. on Machine Learning (ICML-94)*, pages 318–325. Morgan Kaufmann, 1994.
- [Val84] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11) :1134–1142, 1984.
- [Vap82] V. Vapnik. *Estimation of dependencies based on empirical data.* Springer-Verlag, 1982.
- [Vap95] V. Vapnik. *The nature of statistical learning theory.* Springer-Verlag, 1995.
- [VC71] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2) :264–280, 1971.
- [Vid03] M. Vidyasagar. *Learning and Generalization. With applications to neural networks. (2nd ed.).* Springer-Verlag, 2003.

-
- [VP95] R. Valdez-Perez. Systematic methods of scientific discovery : Papers from the 1995 spring symposium. Technical report, AAAI, 1995.
 - [Wol92] D. Wolpert. On the connection between in-sample testing and generalization error. *Complex Systems*, 6 :47–94, 1992.
 - [Wol96] D. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7) :1341–1390, 1996.
 - [XJK01] E. Xing, M. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray data. In *Int. Conf. on Machine Learning (ICML-01)*, pages 601–608, 2001.
 - [ZBS02] J.-D. Zucker, N. Bredèche, and L. Saitta. Abstracting visual percepts to learn concepts. In *Symposium on Abstraction, Reformulation and Approximation (SARA-2002)*, pages 256–273, Kananaskis, Alberta, Canada, 2002.
 - [ZG96] J.-D. Zucker and J.-G. Ganascia. Changes of representation for efficient learning in structural domains. In *International Conference on Machine Learning (ICML'96)*, Bari, Italy, 1996. Morgan Kaufmann.
 - [Zha01] W. Zhang. Phase transitions and backbones of 3-sat and maximum 3-sat. In *7th Int. Conf. on Principles and Practice of Constraint Programming (CP-2001)*. Springer, 2001.
 - [Zuc96] J.-D. Zucker. *Appariements et Changements de Représentation pour l’Apprentissage Symbolique*. Ph.d, Univ. Paris 6, 1996.