Annexe :
Contraintes d'un modèle d'apprentissage de concepts

Les paradigmes d'apprentissage

À plusieurs reprises dans le texte de la thèse, nous faisons référence à des méthodes ou à des paradigmes d'apprentissage, tels qu'ils sont proposés dans le domaine de l'apprentissage automatique[1]. Dans cette annexe, nous passons en revue quelques modèles d'apprentissage, avant de les opposer sur la question de l'innéité du biais qu'ils postulent.

L'apprentissage statistique

Les modèles statistiques de l'apprentissage reposent, pour l'essentiel, sur un mécanisme d'extraction de régularités. L'effet de l'apprentissage est de conserver les formes fréquentes en écartant les variations contingentes. Un certain nombre de modèles entrent dans cette catégorie. Citons-en quelques-uns.

- la théorie behavioriste : couplage de stimuli ou association stimulus - réponse.
- les réseaux de neurones artificiels : établissement d'un couplage continu entre une entrée et une sortie, en respectant, dans le cas supervisé, certaines associations imposées.
- l'apprentissage par renforcement : acquisition d'une fonction de coût permettant de mémoriser les meilleures actions dans chaque état.
- les systèmes de regroupement (*clustering*).

Commentons les systèmes de regroupement, car ils sont fréquemment invoqués pour expliquer l'opération de catégorisation perceptuelle, voire conceptuelle. Le principe général des techniques de regroupement consiste à constituer des classes d'objets à partir d'une mesure de similarité, de manière à minimiser la variance à l'intérieur des classes et à maximiser la variance entre les classes. La méthode des nuées dynamiques est un exemple simple d'algorithme de regroupement, dans le cas où la mesure de ressemblance est donnée par une métrique. Elle consiste à choisir *n* objets au hasard, qui consistent la première nuée. On fabrique *n* classes initiales en regroupant chaque objet avec l'élément de la nuée le plus proche. On forme ensuite une nouvelle nuée en prenant les objets les plus proches des centres de gravité des classes. On itère le processus jusqu'à ce que les classes obtenues soient stables. Les éléments de la dernière nuée obtenue peuvent être interprétés comme des prototypes.

Certains systèmes de regroupement statistiques sont de type symbolique. La distance se mesure au nombre d'attributs symboliques que les objets ont en commun. La technique de regroupement reste statistique, mais elle est astreinte à produire des classes qui possèdent une caractérisation simple. Les systèmes de type SBL (*similarity based learning*) fonctionnent typiquement en produisant des classes correspondant à des conjonctions d'attributs.

Les modèles statistiques d'apprentissage nécessitent généralement un grand nombre d'exemples pour pouvoir fonctionner. En cela, ils diffèrent des autres paradigmes d'apprentissage que nous considérons maintenant.

L'apprentissage gestaltiste

Les modèles gestaltistes de l'apprentissage privilégient l'accès aux bonnes formes. Une bonne forme est une forme, concrète ou abstraite, qui reste invariante pour de nombreuses transformations. Dans le domaine perceptif, un cercle ou un carré sont des bonnes formes ;

[1] Le lecteur pourra se reporter à (MITCHELL 1997 [74]) pour une revue des principales techniques d'apprentissage automatique.

elles sont plus symétriques qu'un patatoïde informe et qu'un polygone irrégulier. Pour un mécanisme d'apprentissage gestaltiste, les bonnes formes sont plus faciles à apprendre. On peut même considérer que le résultat de tout apprentissage est une bonne forme.

Il est intéressant de rapprocher la théorie constructiviste (PIAGET 1932 [82]) d'un système gestaltiste. Dans cette théorie, les stades de développement de l'enfant sont caractérisés par l'accès à des structures invariantes pour des groupes de transformations. Ainsi, l'enfant accède à une première version du concept de justice en considérant que toute action qui est bonne envers lui est juste. Ce concept est invariant pour tout changement d'acteur. Lorsqu'il devient capable de se décentrer, l'enfant est en mesure de substituer d'autres patients aux actions. En accédant à ces nouvelles transformations, il forme un nouveau concept de justice, invariant pour toute substitution de personnes dans les rôles d'acteur et de patient : est juste une action qu'il ressentirait comme bonne si elle était dirigée vers lui. La théorie constructiviste explique ainsi que différents enfants, confrontés à des expériences limitées, forment les mêmes concepts de justice (PIAGET 1932 [82]).

L'appariement combinatoire

Le paradigme de l'appariement regroupe un certain nombre de modèles et de techniques, caractérisés par l'existence d'une structure préalable riche. Mentionnons-en quelques-uns.

- la théorie des principes et paramètres.
- l'induction symbolique.
- l'apprentissage par analogie.
- les systèmes sélectifs.

La théorie dite des principes et paramètres a été invoquée pour expliquer l'acquisition par l'enfant de la grammaire de sa langue, sachant qu'il dispose d'un patron de grammaire universel. L'apprentissage se limite à instancier des paramètres binaires du patron au vu de quelques exemples de structures de phrases. Ainsi, un enfant iranien apprendra, en analysant quelques phrases dépourvues de sujet grammatical explicite, que pour le persan, le paramètre *pro□drop* doit être positionné à la valeur `1`, alors qu'un enfant français, en observant des sujets explétifs, fixera le même paramètre à `0`. Ce type d'apprentissage se rapproche du phénomène d'empreinte mis en évidence en éthologie.

Les systèmes d'induction symbolique cherchent à généraliser les exemples qui leur sont soumis. Par exemple, un programme de type ILP (*inductive logic programming*) à qui l'on présente deux instances du concept `MIGNON`, l'un qui est un chat en peluche, l'autre qui est un petit chien en peluche, pourra former la moins générale des généralisations, à savoir qu'un moyen d'être `MIGNON` est d'être un petit animal en peluche. Pour parvenir à ce résultat, le programme utilise une connaissance d'arrière-plan qui lui dit que chiens et chats sont des animaux, et qu'un chat est petit (sans cette connaissance d'arrière-plan, la conclusion aurait été que toute peluche est mignonne, ce qui est excessivement général). La technique utilisée par les programmes ILP est la "résolution inverse". La méthode consiste à partir d'un exemple et d'une règle de la connaissance d'arrière-plan, puis à former une nouvelle règle qui, enchaînée déductivement avec la règle de départ, produit l'exemple.

Les systèmes de type EBL (*explanation based learning*), utilisent également une connaissance d'arrière-plan. Ils mêlent les aspects inductifs et déductifs pour apprendre des concepts qui sont à la fois des généralisations des exemples qui leur sont soumis et des spécialisations des concepts qu'ils connaissent déjà.

Les systèmes comme ILP ou EBL sont capables de produire des représentations structurées. La comparaison entre structures est à la base de l'apprentissage par analogie. Il s'agit d'une forme d'apprentissage dans laquelle on s'intéresse plus au rapports entre exemples qu'aux exemples eux-mêmes. Il est alors possible d'engendrer de nouveaux objets qui entretiennent un rapport connu avec un objet donné, selon les techniques du raisonnement par cas (*case based reasoning*).

Les systèmes d'apprentissage sélectifs utilisent une mesure de la distance entre les formes qu'ils produisent et la forme à obtenir. Ils engendrent de nouvelles formes en ré-assemblant et en modifiant des formes existantes, jusqu'à approcher suffisamment la forme cible. Les algorithmes génétiques offrent un exemple de mécanisme sélectif. Un autre exemple connu d'apprentissage sélectif est donné par la modélisation du système immunitaire. Une population de lymphocytes est pré-assemblée par réarrangement d'un jeu d'éléments génétiques donnés. L'apprentissage d'une molécule étrangère inconnue se fait par sélection du lymphocyte qui présente la plus grande affinité avec la molécule. Le lymphocyte en question se multiplie, engendrant un clone lymphocytaire qui peut perdurer pendant des années. Des mécanismes sélectifs du même type ont pu être proposés pour l'apprentissage conceptuel (CHANGEUX & DEHAENE 1989 [13]).

Tous ces systèmes, que ce soit la théorie des principes et paramètres, l'induction symbolique, l'analogie ou l'apprentissage sélectif, fonctionnent par appariement entre une forme qui sert de stimulus et une forme interne au système. Cette dernière peut résulter d'un assemblage combinatoire, constitué à partir de briques de base connues dès le départ par le système. Dans le cas des systèmes sélectifs, le hasard est seul responsable de cet assemblage. Dans d'autres formes d'appariement combinatoire, l'assemblage est dirigé pour être en adéquation avec le stimulus.

Les modèles d'apprentissage par appariement sont remarquables par le fait qu'ils sont capables d'apprendre de nouvelles formes, de manière non triviale, en une fois. Contrairement aux modèles statistiques, qui nécessitent un échantillon conséquent d'instances de la forme à apprendre, les systèmes d'appariement combinatoire peuvent se satisfaire d'une exposition unique. Cette prouesse est obtenue grâce à la connaissance qui est déjà présente dans le système avant que l'apprentissage ait lieu.

De la nature des biais d'apprentissage

Les théories de l'apprentissage, en sciences cognitives, ont souvent été opposées du point de vue des hypothèses qu'elles faisaient à propos des connaissances *a priori* prêtées au sujet apprenant, autrement dit le biais d'apprentissage. Dans le cas du langage, on a pu opposer les théories behavioristes, les théories constructivistes et les théories innéistes (PIATTELLI-PALMARINI 1979 [84]). Nous mentionnons ici une classification qui s'attache davantage aux propriétés de symétrie du biais qu'à une simple mesure quantitative de ce biais. L'intérêt de cette classification pour les sciences cognitives est de permettre de caractériser le type d'apprentissage à l'œuvre à partir des propriétés des formes apprises.

Certains modèles d'apprentissage possèdent les propriétés d'isotropie et de relativité. Un mécanisme d'apprentissage est isotrope (respectivement relatif) si le fait de faire subir une rotation (respectivement une translation) à toutes ses entrées, ce qui revient à un changement de repère, ne modifie pas son comportement global après apprentissage (DESSALLES 1998 [29]). Ces propriétés signifient que le système d'apprentissage n'utilise au plus que les distances relatives entre les données. En d'autres termes, le système ne possède aucune connaissance absolue lui permettant de privilégier une direction particulière dans les données auxquelles il est exposé.

Il est intéressant de constater que, parmi les modèles couramment proposés pour rendre compte de l'apprentissage animal et humain, ces propriétés d'isotropie et de relativité distinguent clairement les systèmes statistiques et les systèmes gestaltistes d'un côté, des systèmes d'appariement combinatoire de l'autre (DESSALLES 1998 [29]). Pour prendre un exemple, un perceptron multi-couches est insensible à une permutation systématique de ses neurones d'entrée (rotation) ou à une complémentation partielle des signaux qui lui sont présentés (translation). La classification acquise par le système est rigoureusement la même lorsque les données d'apprentissage et les données de test subissent la même transformation systématique, sachant qu'il s'agit d'une isométrie (rotation ou translation) n'affectant pas la distance, et donc la ressemblance, entre les données.

Le résultat principal concernant l'isotropie des mécanismes d'apprentissage est qu'un système isotrope et relatif est nécessairement astreint à une forme de gestaltisme (DESSALLES 1998 [29]). Plus précisément, pour une classification apprise par un tel système, le produit de son harmonie (nombre d'isométries laissant les classes invariantes) par sa variété (nombre de classifications différentes qui auraient été obtenues sur des données transformées) est constant. La conséquence de ce résultat est importante pour la modélisation cognitive. Le tableau qui suit résume l'idée principale : il compare les trois paradigmes d'apprentissage mentionnés ci-dessus au regard de trois propriétés, la convergence des résultats de l'apprentissage, l'isotropie et la relativité du mécanisme considéré, et l'harmonie des formes apprises.

	convergence	*isotropie/relativité*	*harmonie*
statistique	non	oui	non
gestaltiste	oui	oui	oui
Appariement	oui	non	non

La convergence est la propriété selon laquelle des individus confrontés à des données différentes apprennent les mêmes formes. Cette propriété distingue d'un côté les systèmes statistiques, qui peuvent apprendre toute forme de régularité, et d'un autre côté les systèmes gestaltistes, dans lesquelles les formes apprises se comportent comme des attracteurs, et les systèmes par appariement, capables de produire des résultats fiables dans des environnements variés (il s'agit de l'argument classique de la pauvreté du stimulus). Certaines formes d'apprentissage conceptuel sont considérées comme convergentes. Ainsi, les individus d'une culture donnée acquièrent des concepts comparables, ce qui leur permet de communiquer. Lorsque l'on s'intéresse aux catégories fondamentales de la cognition que tout individu humain est supposé posséder, et qu'on les considère comme acquises, la convergence est encore plus manifeste.

Le résultat théorique mentionné plus haut contraint tout modèle de l'apprentissage cognitif à choisir entre les trois lignes du tableau. Or ce choix est fortement contraint. Si l'on s'intéresse à des acquisitions comme celle du concept de justice, il est parfaitement possible d'opter pour un système gestaltiste, puisque la forme apprise présente un certain nombre de symétries. En revanche, si la forme apprise ne peut être considérée harmonieuse, *id est* comme une bonne forme, ce qui est sans doute le cas pour la plupart des acquisitions conceptuelles quotidiennes (rappelons qu'un enfant acquiert pendant plusieurs années une dizaine de mots nouveaux par jour, avec les concepts associés), il faut choisir entre un apprentissage statistique et un mécanisme d'appariement. Or, le choix statistique, en l'absence de la propriété de convergence, n'est acceptable que si l'on est en mesure de montrer que les différents individus ont été confrontés aux mêmes expériences. Si l'on considère une telle contrainte comme inacceptable, la seule option qui demeure ouverte est celle de l'appariement combinatoire.