*Intellectica*, 1991/2, 12, pp. 139-158

**Laurent BOCHEREAU, Paul BOURGINE et Guillaume DEFFUANT**

# ÉQUIVALENCE ENTRE CLASSIFICATEURS CONNEXIONNISTES ET CLASSIFICATEURS LOGIQUES

## *Introduction*

La proposition d'équivalence comprend deux parties. La proposition directe est bien connue : en effet, la question de l'existence d'un classificateur connexionniste équivalent à une fonction booléenne quelconque a été établie avec les perceptrons multicouches [3]. La proposition inverse est plus délicate, parce qu'il n'est pas toujours aisé de définir, à partir d'une mémoire associative de type connexionniste, un classificateur connexionniste. Cette proposition inverse constitue la question centrale, étudiée par cet article.

La réponse à cette question constitue un enjeu épistémique majeur : elle établit qu'un classificateur connexionniste, quel que soit le support connexionniste avec lequel il réalise son apprentissage, a la possibilité épistémique de représenter son fonctionnement comme un ensemble de règles logiques. Cette possibilité est essentielle pour les besoins de la communication : on imagine mal comment des classificateurs pourraient échanger leur façon de parvenir à des conclusions en communiquant

seulement des pondérations numériques. Les règles logiques constituent au contraire un excellent support pour les explications.

L'équivalence ne se situe pas au niveau de la capacité d'apprentissage : on sait que le choix d'une méthode d'apprentissage, ou de l'architecture d'un réseau, ou encore une division du travail de classification entre un classificateur généraliste et des classificateurs spécialistes peuvent améliorer de façon essentielle les performances de l'apprentissage, en termes de complexité du calcul ou de taux de réussite. L'équivalence se situe au niveau fonctionnel, et la représentation du classificateur sous forme de clauses logiques constitue, en quelque sorte, la forme “confluente” de l'apprentissage, qui permet à plusieurs machines connexionnistes d'échanger des “raisonnements”. Le classificateur connexionniste est un support approprié pour la cognition et l'apprentissage, et le classificateur logique est un support approprié pour la communication et les explications.

Cet article a un autre but, celui d'entrouvrir une des voies possibles pour l'explicitation de l'expertise. De nombreux auteurs reconnaissent que l'expert, dans les nombreux domaines où l'expérience joue un rôle majeur, a compilé ses connaissances. La décompilation des connaissances, dans un microdomaine représentable par quelques centaines de règles, est une opération difficile et coûteuse lorsqu'elle est réalisée par interview des experts ; une telle méthode d'explicitation de l'expertise semble tout à fait irréaliste pour des domaines de connaissances, tel celui d'un maître international aux échecs (ou encore la médecine interne, voire la grammaire de la langue française), où les estimations du nombre de règles atteignent un ordre de grandeur d'environ 100.000 règles. L'extraction automatique des règles logiques équivalentes à un classificateur connexionniste offre en principe une réponse à l'explicitation de l'expertise.

L'extraction automatique de règles n'est pas une question facile et rencontre un certain nombre de difficultés, qui font l'objet des différentes parties de l'article :

- Un réseau connexionniste généralise à partir d'exemples, et ses conclusions sont hésitantes ; il faut appliquer à ces conclusions une fonction de décision, qui n'est pas toujours facile à définir. C'est la composition entre la mémoire associative et une fonction de décision, qui produit un classificateur connexionniste. Une conséquence directe de ceci est que l'ensemble des règles logiques dépend du choix de la fonction de décision : il y a autant de classificateurs logiques qu'il y a de fonctions de décision possibles (§ 2).

- Deux réseaux connexionnistes ne généralisent pas de la même manière. Il peut être utile de construire plusieurs classificateurs connexionnistes (éventuellement à l'aide de plusieurs méthodes, tout en se plaçant dans le cadre de la théorie de Valiant) et d'étudier leur comportement “moyen”. Quelques propositions sur le fonctionnement d'un “congrès” de classificateurs figurent au § 3 et conduisent à la définition d'un domaine de validité : il est en effet essentiel qu'un classificateur (ou un ensemble de classificateurs) puisse répondre “je ne suis pas compétent”, là où les exemples sont rares et ne contraignent pas suffisamment leur réponse.

- L'extraction est, dans le cas général, un problème NP-Complet ; le § 4 expose quelques considérations heuristiques, qui permettent d'améliorer les temps de calcul.

## *I. Classificateurs connexionnistes et classificateurs logiques*

### *1. Réseaux connexionnistes et fonctions de décision*

Nous considérons des réseaux connexionnistes dont les unités de sortie sont booléennes et peuvent correspondre à des classes, des prototypes ou des hypothèses. Deux modèles de ces réseaux seront présentés : le *modèle*

*de Hopfield* et le *perceptron multi-couches*. Les *fonctions de décision* sont des fonctions qui ont pour résultat un vecteur booléen d'hypothèses.

### *1.1. Modèles de Hopfield*

Le modèle de Hopfield est un modèle connexionniste où tous les neurones sont booléens et symétriquement interconnectés [1, 11, 15]. Hopfield a prouvé qu'un tel modèle était capable de mémoriser des formes, encore appelées *prototypes*, comme des attracteurs. Lorsque le réseau est placé dans un état initial donné, il est capable d'évoluer librement jusqu'à se stabiliser dans l'un des états précédemment mémorisés.

Nous noterons B = {0,1} et I = [0,1] dans la suite de l'article. Les neurones ont des états d'activation booléens X $B^m$ ; les *n* prototypes sont donnés comme $X^j = (x_1^j, .. x_m^j)$. T est la fonction de transition définie par :

$T(x_i) = 1$ si $x_i = \Sigma_j w_{ij}x_j \geq 0$ où $w_{ij}$ est le poids de la connexion
$T(x_i) = 0$ si $x_i = \Sigma_j w_{ij}x_j < 0$ entre les neurones i et j.

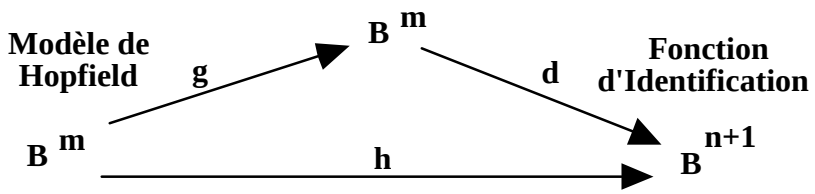


*figure 1 : Construction d'un modèle de Hopfield suivi par une fonction d'identification*

Pour chaque prototype $X^j$, il est possible de définir un bassin d'attraction $B^j$ par :

$$B^j = \{X \quad B^m / \text{ il existe } k \quad N, \ T^k(X) = X^j \}$$

Nous pouvons maintenant définir l'espace de sortie $B^{n+1}$, avec n+1 booléens exclusifs :

- $(X \; B^j)$ pour $j \leq n$,
- “unknown” dont la valeur est vraie si tous les $(X \; B^j)$ sont faux.

La fonction de décision liée à un modèle de Hopfield est donc simple : la fonction compare l'état du réseau avec les prototypes mémorisés. Si l'état du réseau correspond à l'un des prototypes, la fonction renvoie le nom du prototype ; dans le cas contraire, la réponse est “unknown”.

Avec un modèle de Hopfield non déterministe, on peut obtenir un vecteur de probabilité donnant la fréquence des différents prototypes lorsque l'on répète le calcul. La fonction de décision est alors plus compliquée et la discussion est exactement la même qu'au paragraphe précédent.

### *1.2. Perceptrons multi-couches*

Les perceptrons multi-couches sont des réseaux de neurones multi-couches complètement connectés entre couches successives [9,10]. Le nombre de couches cachées entre la couche d'entrée et la couche de sortie est arbitraire. La relation d'entrée-sortie pour chaque unité (excepté pour ceux appartenant à la couche d'entrée) est donnée par :

$$y = f(z) \qquad \text{où } z = \sum_{i=1}^{n} w_i x_i - \theta$$

où y représente l'état d'activation de l'unité, n le nombre d'entrées $x_i$, $w_i$ les poids des connexions et $\theta$ le seuil ; f est la fonction sigmoïde pour chaque unité. Un ensemble d'exemples $\{ X^k, Y^k \}$ est donné, où $X^k$ prends ses valeurs dans l'hypercube $B^m$ et $Y^k$ dans l'hypercube $B^n$. Le réseau calcule une transformation **g** (figure 1) de $B^m$ dans $I^n$ par l'application de l'algorithme de rétropropagation du gradient [14,16].

Les perceptrons multi-couches fournissent une réponse continue pour chaque hypothèse, qui peut être considérée comme une mesure

d'acceptation de l'hypothèse. Ceci n'est pas, à l'évidence, un mécanisme pour prendre la décision d'éliminer ou d'accepter des hypothèses. Afin d'obtenir un classificateur ou une machine décisionnelle, il est nécessaire d'envoyer le vecteur continu de sortie du réseau connexionniste dans une fonction de décision qui sera alors capable de sélectionner certaines hypothèses booléennes (cf. figure 2).

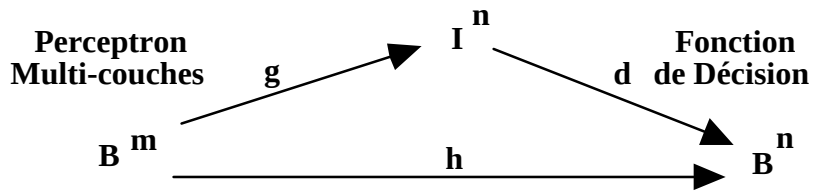


*figure 2 : Construction d'un perceptron multi-couches suivi d'une fonction de décision*

Il existe un grand nombre de fonctions de décision possibles ; nous présentons ci-dessous un certain nombre d'exemples :

(i) La règle du maximum permet de ne garder qu'une hypothèse parmi un ensemble d'hypothèses exclusives.
(ii) La règle du seuil permet de sélectionner un nombre variable d'hypothèses.
(iii) La règle des p meilleurs, dans l'ordre ou le désordre.

La classe des machines décisionnelles peut être étendue en introduisant les risques de seconde espèce $c_{ij}$ qui sont définis comme le coût de choisir l'hypothèse i lorsque j est la réponse correcte. Ceci peut être illustré en considérant un médecin qui doit faire un diagnostic et qui hésite entre une tumeur maligne ou bénigne. Dans ce cas, les coûts $c_{ij}$ et $c_{ji}$ peuvent être très différents. Une stratégie permet alors au médecin de limiter le risque attaché à sa décision :

(iv) choisir l'hypothèse i minimisant $\Sigma_j y_j c_{ij}$.

Cette expression devient un coût moyen si les $y_j$ sont des probabilités. Ceci peut être obtenu si, pendant la période d'apprentissage, un terme de pénalité est introduit pour contraindre le réseau avec la relation $\Sigma_i y_i = 1$.

*2. Classificateurs connexionnistes comme machines abductives*

Un *classificateur connexionniste* est une machine décisionnelle définie par la composition entre un réseau connexionniste **g** et une fonction de décision **d** (figure 3).

Puisque **g** envoie l'hypercube $B^m$ dans E et que **d** envoie E dans l'hypercube $B^n$, un classificateur connexionniste est donc une transformation entre l'hypercube $B^m$ et l'hypercube $B^n$.

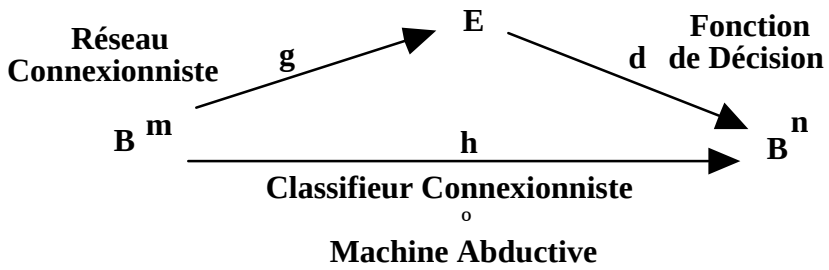


*figure 3 : Construction d'un classificateur connexionniste*

Peirce[1] définit l'*abduction* comme la capacité à se focaliser sur un petit nombre d'hypothèses à soumettre à l'expérience en laissant de côté la vaste majorité des hypothèses possibles. Selon cette définition, le classificateur connexionniste se comporte comme une *machine abductive* : il est en effet capable de sélectionner un nombre réduit d'hypothèses et d'éliminer les hypothèses restantes.

Nous avons montré que deux modèles de réseaux connexionnistes (le perceptron multi-couches et le modèle de Hopfield) sont conformes au

[1] Peirce, *Collected Papers*, 6-530.

schéma de la figure 3. Cependant, nous pensons que le cadre précédant peut être étendu à d'autres types de réseaux connexionnistes [8, 13]. Les fonctions de décision prennent leurs arguments dans un ensemble d'hypothèses et ont pour image un vecteur booléen de classes.

### *3. Classificateurs logiques*

Nous définissons maintenant un certain nombre de notations qui seront utiles par la suite. $X = [x_i]_{i\ [1,m]}$ représente le vecteur d'entrées et $Y = [y_j]_{j\ [1,n]}$ le vecteur des classes ou des hypothèses.

Un *classificateur logique* est défini comme un ensemble de clauses possédant une variable booléenne de sortie positive unique.

$$( x_{i1} \ 3 \ x_{i2} \ 3 \ x_{i3} \ 3 \ ... \ \ y_i ) \ 3 \ ( x_{j1} \ 3 \ ... \ \ y_j ) \ 3 \ ( x_{k1} \ 3 \ ... \ \ y_k ) \ 3 \ ...$$

où $x_{ij}$ X sont des variables booléennes d'entrée,

$y_i$ Y sont des variables booléennes de sortie positives.

## *II. Équivalence entre un classificateur connexionniste et un classificateur logique*

Dans cette partie, nous montrons l'équivalence entre un classificateur connexionniste et un classificateur logique. Le diagramme suivant illustre cette équivalence :

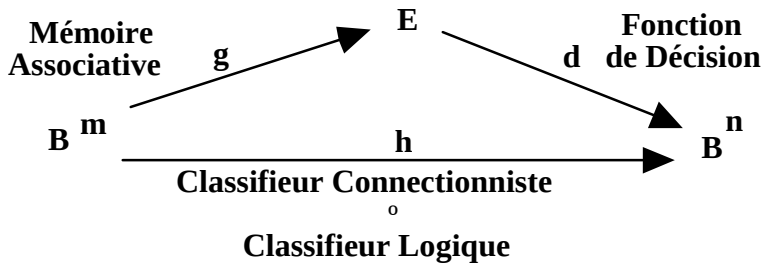


*figure 4 : Diagramme d'équivalence*

*Theorème* : Un classificateur connexionniste est fonctionnellement équivalent à un classificateur logique.

*Démonstration* : Chaque variable booléenne de $B^n$ est une fonction de variables booléennes de $B^m$. Cette variable peut donc s'exprimer comme une disjonction de conjonctions :

$$[ ( x_{i1} \; 3 \; x_{i2} \; 3 \; x_{i3} \; 3 \; ...) \; £ \; ( x_{j1} \; \# \; x_{j2} \; 3 \; x_{j3} \; 3... ) \; ... \; ] \quad y_j$$

et donc,

$$[ ( x_{i1} \; 3 \; x_{i2} \; 3 \; x_{i3} \; 3 \; ... \quad y_j ) \; 3 \; ( x_{j1} \; 3 \; x_{j2} \; 3 \; x_{j3} \; 3... \quad y_j ) \; 3 \; .. ]$$

Puisque ceci est vrai pour toute variable booléenne de $B^n$, il s'ensuit qu'un classificateur connexionniste est équivalent à un classificateur logique.

Ce résultat nous permet de regarder le classificateur sous deux angles qui sont fonctionnellement équivalents, soit comme la composition d'un réseau connexionniste et d'une fonction de décision, soit comme une machine logique. Le premier point de vue est intéressant du point de vue de l'apprentissage. Le deuxième point de vue convient mieux pour les explications.

Le choix d'une fonction de décision conduit à un ensemble de règles logiques donné. Changer la fonction de décision (c'est-à-dire modifier **d** en **d'**) entraîne un changement de l'ensemble des clauses logiques équivalentes.

Dans le cas où le classificateur connexionniste est construit comme la composition d'un perceptron multi-couches et d'une fonction de décision, ses performances ont tendance à se dégrader lorsqu'on lui présente des cas éloignés de la base d'apprentissage. Les règles logiques extraites correspondant à de tels cas ne seront pas pertinentes. Afin de surmonter ce problème, nous proposons par la suite de considérer, non pas un classificateur connexionniste unique, mais un ensemble de classificateurs connexionnistes entraînés sur la même base d'apprentissage, et de comparer leurs résultats.

## *III. Construction d'un domaine de validité*

Le concept de domaine de validité est lié à l'idée qu'il existe un voisinage de la base d'apprentissage pour lequel la généralisation est bonne [5]. La construction d'un tel domaine de validité peut être obtenue par différentes méthodes en fonction du problème considéré et de la probabilité d'occurrence de chaque exemple de la base d'apprentissage [2].

### *1. Borne supérieure du domaine de validité*

Une borne supérieure D du domaine de validité peut être obtenue, soit à l'aide de *connaissances a priori*, liées au codage des données ou formulées par un expert du domaine considéré, soit à partir de *régularités statistiques* observées sur la base d'apprentissage. On obtiendra alors des contraintes sur l'hypercube d'entrée $B^m$. Ceci permettra également de réduire la cardinalité initiale du domaine ($2^m$).

Si l'on suppose que le vecteur d'entrée est un vecteur de booléens, *les contraintes sur le domaine peuvent alors s'exprimer comme des formules logico-mathématiques*. Nous pouvons aussi interpréter l'ensemble de booléens comme la fonction caractéristique d'un sous-ensemble et obtenir des contraintes comme propriétés de ces sous-ensembles, par exemple leur cardinalité. Nous décrivons ci-dessous un certain nombre d'exemples de restriction de domaine avec les règles de calcul de cardinalité associées :

a) Une méthode simple consiste à utiliser la distance de Hamming par rapport à $\{X_1^k\}_{k\ K}$ où $X_1$ est un sous-ensemble de X. Nous pouvons alors définir une mesure d'appartenance au domaine de validité par la définition suivante : un exemple x est dit appartenir au domaine de validité D du classificateur connexionniste avec un coefficient de similitude $c = 1 - s/m$ $[0,1]$ si :

il existe y $\{X_1^k\}_{k\ K}$ / x hypersphère de centre y et de rayon s

b) D est formé de booléens $(a_1, ..., a_q)$ parmi lesquels p sont VRAI et $p_1 \leq p \leq p_2$,

$D = \{ A' \subseteq A \ / \ p_1 \leq \text{card}(A') \leq p_2 \}$ avec $\text{card}(D) = \Sigma_{p \ [p_1,p_2]} C_{(q,p)}$

b') Quand $p_1=p_2=p$, D est formé de booléens $(a_1, ..., a_q)$ parmi lesquels p sont VRAI.

$D = \{ A' \subseteq A \ / \ \text{card}(A') = p \}$ avec $\text{card}(D) = C_{(q,p)}$

b'') Pour p = 1, D est formé de booléens exclusifs $\{a_1, ..., a_q\}$ et card(D) = q

c) D est égal au produit cartésien de domaines $D_1, ..., D_q$,

$D = D_1 \times ... \times D_q$ avec $\text{card}(D) = \text{card}(D_1) \times ... \times \text{card}(D_q)$

Les règles a, b, b' and b'' sont des descriptions de domaines feuilles. La règle c appelle récursivement les règles a, b, b', b'' and c. Ainsi, un langage récursif peut être utilisé pour décrire les contraintes sur le domaine de validité du classificateur connexionniste. Ce langage peut être étendu en considérant d'autres constructions [6].

### *2. Construction par m-stabilité*

Supposons que l'on construise de manière indépendante un grand nombre de classificateurs connexionnistes PAC (*Probably Almost Correct*), entraînés sur une même base d'apprentissage caractérisée par une distribution de probabilité donnée. En utilisant le cadre de Valiant [17], ces classificateurs connexionnistes $\mathbf{h_i}$ (chacun défini par un couple $(\mathbf{g_i}, \mathbf{d_i})$) peuvent classer correctement, avec une probabilité de $1-\delta$, une proportion d'au moins $(1-\varepsilon)$ des exemples d'une base de test K (possédant la même distribution) [2]. En considérant que la fonction de décision **d** est la règle du maximum, nous pouvons écrire :

Probabilité [ $E < \varepsilon$ ] > $1 - \delta$

où $E = (1/\text{ card }(K)) \Sigma_{k \in K} d_{Hamming} [Y^k, \mathbf{h} (X^k)]$

La définition de classificateurs connexionnistes PAC vaut pour des dépendances fonctionnelles strictes. On peut l'étendre au contexte de la classification statistique : dans ce cas, on a sait que la probabilité minimale d'erreur de classification ($E_0$) est obtenue par la loi de Bayes et Kanaya [12] définit un réseau qui donne une *généralisation valide* à partir d'une base d'apprentissage de taille m comme un réseau tel qu'il existe $0 < \eta, \varepsilon < 1$ satisfaisant :

Probabilité [ $E < E_0 + \varepsilon$ ] > $1 - \delta$

où $E = (1/\text{ card }(K)) \Sigma_{k \in K} d_{Hamming} [Y^k, \mathbf{h} (X^k)]$
et $E_0$ est l'erreur minimale obtenue par la règle de Bayes.

Dans le cadre de Valiant, la dépendance fonctionnelle est stricte : on a alors $E_0 = 0$, et on retrouve bien la formule précédente.

Supposons maintenant que l'on construise de manière indépendante un grand nombre (supérieur à 20) de classificateurs connexionnistes $\mathbf{h_i} = \mathbf{dog_i}$, i = 1, ..., p, entraînés sur la même base d'apprentissage.

Pour chaque point de l'espace d'entrée $B^m$, nous pouvons considérer

$$m_p (x) = (1/p) \Sigma_{i \in [1,p]} \mathbf{h_i} (x)$$

Nous pouvons alors définir le *classificateur connexionniste moyen* ***h*** par :

$$\boldsymbol{h} : B^m \varnothing [0, 1]^n$$
$$x \varnothing m_p (x)$$

Pour chaque $x \in B^m$, la $j^{ème}$ composante de $\boldsymbol{h} (x)$ est notée $[\boldsymbol{h} (x)]_j$

Le classificateur connexionniste moyen ***h*** est dit *m-stable* au point x $B^m$, s'il existe j [1,n] tel que :

$$[\boldsymbol{h}(x)]_j \geq \mu \qquad \mu \in [0, 1]$$

Quand p devient grand, chaque composante $[\boldsymbol{h}(x)]_i$ peut être considérée comme une mesure de probabilité de l'hypothèse j dans l'espace des classificateurs possibles.

Nous pouvons maintenant définir le domaine de *m-validité* comme l'ensemble des points dans D où le classificateur connexionniste p-moyenné est μ-stable.

$$D_\mu = \{ x \in D / \boldsymbol{h} \text{ est } \mu\text{-stable au point } x \}$$

Les règles logiques *m-pertinentes* sont alors définies comme les règles logiques extraites du domaine de μ-validité d'un classificateur connexionniste p-moyenné. Cette idée sera reprise plus loin.

Comme l'illustre la figure 5, le domaine de μ-validité décroît lorsque μ croît. Parallèlement, les performances du classificateur connexionniste p-moyenné mesurées sur le domaine de μ-validité augmentent lorsque μ augmente.

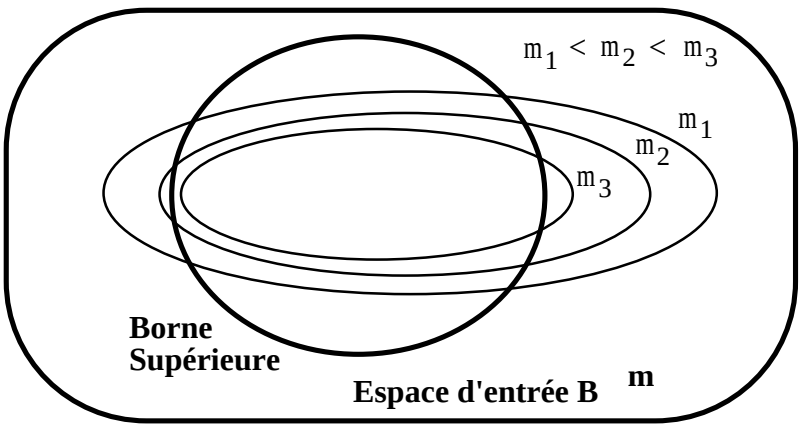

*figure 5 : Construction du domaine de validité*

Lorsque $\mu = 1$, le taux d'erreur $\varepsilon$ du classificateur connexionniste p-moyenné n'est pas nécessairement nul, car tous les classificateurs connexionnistes peuvent se tromper pour certains points du domaine de 1-validité. D'autre part, il existe une borne inférieure non nulle pour $\mu$, en dessous de laquelle le domaine de $\mu$-validité est égal à D (par exemple, lorsque **d** est la règle du maximum, la borne inférieure est 1/n).

Afin de construire le domaine de validité, nous proposons de lier le choix de $\mu$ au niveau de performance désiré ; Le niveau de performance peut être estimé sur la base de test K par :

$$E(\mu) = (1/N) \Sigma_{k \; K \leftrightarrow D_\mu} \; d_{Hamming} [Y^k, \; \mathbf{h}(X^k)]$$
$$\text{où} \quad N = \text{card}(K \leftrightarrow D_\mu)$$

Le niveau de performance désiré dépend du problème considéré ; cependant, $E_0 < E(\mu) < E_0 + \varepsilon$, et le choix de E permet de déterminer $\mu$.

Le problème de l'extraction de règles équivalentes est NP-complet [4]. En choisissant $\mu$ proche de 1, nous améliorons la pertinence et la reproductibilité des règles et nous réduisons la complexité du calcul. D'autres méthodes permettant de réduire la complexité du calcul vont maintenant être présentées.

## *IV. Obtention d'un classificateur logique à partir d'un classificateur connexionniste*

Les classificateurs connexionnistes considérés dans cette partie sont des perceptrons multi-couches composés avec des fonctions de décision. Néanmoins, l'extraction de règles logiques d'un modèle de Hopfield est également possible [18]. Des techniques pour l'obtention d'un

classificateur logique à partir d'un classificateur connexionniste sont maintenant décrites.

*1. Construction du domaine de m-stabilité*

Cette idée a déjà été présentée dans la troisième partie. Le domaine de validité est restreint en calculant dans D le domaine de μ-stabilité. Pour des réseaux de neurones multi-couches, la différence de construction des réseaux est obtenue par le caractère aléatoire du tirage initial des poids et du tirage de la suite des exemples à chaque cycle d'apprentissage. Cette méthode est complémentaire de celle décrite ci-dessous et améliore le temps de calcul ainsi que la pertinence des règles.

*2. Détection d'invariance*

Nous considérons maintenant des transformations d'un neurone $f[w_0+\Sigma_j w_j x_j]$ qui laissent sa fonction de transfert invariante, mais qui favorisent de meilleures propriétés sémantiques.

*2.1. Moyennage des poids*

Quand le nombre des $x_j$'s est grand, il est possible de réaliser un histogramme des poids $w_j$ et de répartir les poids en plusieurs classes. Quand les classes sont difficiles à distinguer, il est toujours possible d'avoir recours à des méthodes statistiques telles que l'analyse discriminante. En choisissant un coefficient moyen $w_k$ pour chaque classe, nous pouvons obtenir un neurone équivalent $f[w_0+\Sigma_k w_k(\Sigma_{jk} x_{jk})]$ où $x_{jk}$ appartient à la classe k.

Ceci peut également être utilisé pendant la période d'apprentissage. Quand les poids évoluent dans la même direction, ils peuvent être moyennés, puis contraints. En réduisant le nombre de degrés de liberté, cette procédure force le réseau à extraire des traits sémantiques et permet d'accélérer la procédure d'apprentissage.

*2.2. Apparition de variables numériques intermédiaires*

Si les neurones $x_{jk}$s de la même classe k sont — ou se comportent comme — des unités booléennes (c'est toujours le cas sur la couche d'entrée), alors la somme $\Sigma_{jk}\, x_{jk}$ peut être interprétée comme le nombre de $x_{jk}$ égal à 1. Nous pouvons alors introduire une variable numérique intermédiaire définie comme le nombre d'éléments de la classe k. Cette méthode contribue à une meilleure interprétation et simplifie la tâche d'extraction. Des contraintes telles que “le nombre d'éléments de la classe k est compris entre $n_1$ et $n_2$" peuvent alors être exprimées avec la distance de Hamming :

$$n_1 \leq d_{Hamming}[X,X_0] \leq n_2$$

où $X_0$ est le vecteur égal à X, sauf pour les éléments $x_{jk}$ de classe k $[x_1, ..., x_{jk}]$ qui ont pour valeur 0.

De telles contraintes peuvent être facilement prises en compte par les techniques d'énumération implicite utilisables pour l'extraction des règles logiques.

*3. Extraction de règles logiques*

Il est toujours possible, *par énumération*, d'extraire une forme normale disjonctive. Nous pouvons nous permettre d'utiliser un *algorithme d'énumération explicite lorsque le nombre de variables a été nettement réduit.* Mais, il est toujours préférable de recourir à une *énumération implicite en utilisant des méthodes heuristiques* telles que les techniques de propagation et d'extraction de contraintes [6].

L'*introduction d'un domaine de validité* pour le classificateur connexionniste (§ 3) permet d'accélérer le mécanisme d'extraction de

règles : les contraintes décrivant le domaine de validité sont faciles à introduire dans les méthodes de propagation et d'extraction de contraintes.

Les méthodes présentées ci-dessus ont été testées avec des réseaux dont les couches cachées comprenaient moins de 10 unités.

## *V. Conclusion*

Il n'est pas toujours utile de produire les clauses logiques équivalentes à un classificateur connexionniste : pour tous les domaines qui relèvent du sens commun, les explications sont généralement inutiles, puisqu'il s'agit de connaissances largement partagées, qui peuvent rester implicites.

Lorsqu'il est nécessaire de produire des explications à partir d'un classificateur connexionniste, l'extraction des clauses est un problème NP complet. Cette extraction nous semble praticable pour quelques milliers de clauses. Mais il devient nécessaire de comparer alors le temps de calcul global et les performances (construction du réseau + extraction) avec celui des méthodes qui produisent directement des règles. C'est le cas de l'apprentissage symbolique. C'est aussi le cas des classificateurs génétiques.

Quelle que soit la méthode utilisée pour l'extraction de règles, la possibilité effective de construire de grandes bases de règles permettrait de les comparer aux règles obtenues à l'aide d'interview d'experts. On pourrait vérifier ainsi l'hypothèse plausible que les experts communiquent (et enseignent aux novices) en priorité les “grosses” règles. D'autres protocoles de psychologie cognitive pourraient être envisagés. Il serait possible, par exemple, d'approfondir trois questions délicates, rarement évoquées dans la construction des systèmes experts, qui ont toutes une influence importante sur la base de règles censée représenter l'expertise. La première concerne le choix de la fonction de décision, en présence de risques éventuels de deuxième espèce ; la seconde a trait à la distribution des exemples et au domaine de validité ; la troisième prend en compte la

difficulté d'obtenir d'excellents classificateurs (bayésiens). Nous discutons pour terminer ces trois derniers points.

Tout d'abord, il faut apporter une attention particulière à la fonction de décision à composer avec un réseau connexionniste, pour obtenir un classificateur. Le choix de cette fonction de décision peut être délicat et fournit autant de classificateurs différents. Son choix dépend du domaine, des risques de deuxième espèce liés à une décision erronée, du rôle de l'expertise (doit-elle fournir une seule hypothèse, un doublé, un tiercé dans l'ordre ou le désordre ?). Ces questions deviennent particulièrement intéressantes lorsque l'on cherche à représenter une suite d'étapes de raisonnement comme dans le domaine médical : symptômes, syndromes, maladie, thérapeutique, effets des médicaments, nouvelle thérapeutique... Dans une telle suite, les fonctions de décision jouent un rôle très différent : parfois, il est nécessaire de faire un choix ; ailleurs, on peut garder les hésitations du réseau sans forcer les sorties à être booléennes.

La distribution des exemples joue bien sûr un rôle majeur sur les clauses obtenues par extraction. Une forte densité d'exemples le long des frontières permet aux clauses d'opérer des distinctions fines entre les différentes classes. De faibles densités impliquent au contraire davantage de degrés de liberté le long des frontières, et des classificateurs, même s'ils constituent des solutions de Bayes pour la distribution des exemples, peuvent générer des clauses équivalentes divergentes. La distribution des exemples a aussi un impact sur la détermination du domaine de validité. Mais les risques de deuxième espèce ont aussi un rôle sur la détermination du domaine de validité : si les risques associés à une erreur sont graves, la réponse de non-compétence doit se faire plus fréquente ; cette réponse doit être soigneusement distinguée d'une réponse d'hésitation entre deux classes le long de leurs frontières.

À partir d'une même base d'exemples et des mêmes risques de deuxième espèce, les méthodes de calcul produisent alors, au mieux, une approximation des solutions de Bayes (par définition même de ces

solutions). Bien souvent, les classificateurs obtenus s'en éloignent de façon importante et peuvent très bien diverger dans un grand nombre de cas. Les règles associées divergent de la même façon.

Toutes ces observations relatives aux classificateurs connexionnistes conduisent à s'interroger sur l'objectivité des règles équivalentes : ces règles associées ont, en fait, le caractère de règles pour soi. Cette constatation est celle de l'intrication entre l'objectivité et la subjectivité de tout savoir construit à partir de cas acquis par l'expérience. Certes, les besoins de la communication conduisent à échanger ce savoir sous une forme plus ou moins prédicative et logique. Mais il peut être utile de se rappeler qu'il s'agit de règles pour soi et qu'elles peuvent coexister avec d'autres règles pour soi.

Laurent BOCHEREAU, Paul BOURGINE
et Guillaume DEFFUANT
CEMAGREF, BP 121, 92185 Antony

## *Références*

[1] Amit D. J., Gutfreud H. et Sompolinsky H., (1985) “Storing infinite numbers of patterns in a spin-glass model of neural network”, *Phys. Rev. Lett.*, vol. 50, pp. 1110-1112.

[2] Baum E.B. (1990) “ Are k-nearest neighbor and back propagation accurate for feasible sized sets of examples?”, Lecture Notes in *Computer Science*, 412, pp. 2-26, Springer Verlag.

[3] Bochereau L. et Bourgine P., “Implémentation et extraction de traits sémantiques sur un réseau neuromimétique”, *Actes de Neuro-Nîmes 89*, pp. 125-143.

[4] Bochereau L. et Bourgine P., “Extraction of semantic features and logical rules from a multilayer neural network”, *Proceedings IJCNN 90*, Washington DC, 15-19/01/90.

[5] Bochereau L. et Bourgine P., “Validity domain and extraction of rules on a multilayer neural network”, *Proceedings IJCNN 90*, San Diego, 17-21/06/90.

[6] Bourgine P., “PROMAT language and his virtual machine”, *Actes de RFIA 89*, AFCET/INRIA, Paris.

[7] Clancey W. (1985) “Heuristic Classification”, *Artificial Intelligence*, 27, 289-350.

[8] Deffuant G., “Neuron units recruitment algorithm for generation of decision trees”, *Proceedings IJCNN 90*, San Diego, 17-21/06/90.

[9] Feldman S.E. et Ballard D.H. (1982) “Connectionist models and their properties”, *Cognitive Science*, 6, pp. 205-254.

[10] Fogelman-Soulié F., Gallinari P., Lecun Y., Thiria S. (1987) “Automata networks and artificial intelligence” dans *Automata Networks in Computer Science*, F. Fogelman-Soulié, Y. Robert, M. Tchiente, éd., Manchester University Press, pp. 133-186.

[11] Hopfield J.J., “Neural networks and physical systems with emergent collective computational abilities”, *Proceedings of the National Academy of Sciences,* 1982, 79, pp. 2554-2558.

[12] Kanaya F., “On the Bayes Statistical Behavior and Valid Generalization of Pattern Classifying Neural Networks”, *Cognitiva 90*, Madrid, 20-23/11/90.

[13] Kohonen T. (1987) *Self-Organization and Associative Memory*, Springer-Verlag, Berlin.

[14] Le Cun Y. (1985) “A learning procedure for asymmetric networks”, *Actes de Cognitiva 85*, Paris, pp. 599-604.

[15] Personnaz L., Guyon I. et Dreyfus G. (1986) “Collective computational properties of neural networks : new learning mechanisms”, *Phys. Rev. A*, Vol. 34, pp. 4217-4228.

[16] Rumelhart D.E., Hinton G.E., Williams R.J. (1986) “Learning internal representations by error propagation” in *Parallel Distributed Processing*, D.E. Rumelhart & J.L. MacClelland, éd., Cambridge (Mass.), MIT Press, Vol. 1, pp. 318-369.

[17] Valiant L.G. (1984) “A theory of the learnable”, *Communications of the ACM V27*, n° 11, pp. 1134-1142.

[18] Victorri B. (1988) “Modéliser la polysémie”, *TA Informations*, Vol. 29, 1-2, pp. 21-42.