Catégorisation adaptative de données sensori-motrices pour un système d'apprentissage par renforcement

Rémi Munos*
CEMAGREF, LISC, Parc de Tourvoie,
BP 121, 92185 Antony Cedex, FRANCE.
Tel : 01 40 96 61 21, Poste 64 14. Fax : 01 40 96 60 80
E-mail : Remi.Munos@cemagref.fr

Résumé :

Cet article propose d'utiliser le formalisme de l'Apprentissage par Renforcement pour mettre en évidence certains processus cognitifs mis en jeu lors de la prise de décisions d'un système en interaction avec son environnement. Nous pensons que la capacité d'apprendre par l'expérience nécessite une coopération entre les deux processus dynamiques d'apprentissage et de catégorisation. Nous définissons la fonction valeur, fonction construite par les techniques d'apprentissage par renforcement, qui anticipe les renforcements à venir et permet de déterminer le comportement adéquat. Enfin nous interprétons la fonction valeur comme une perception émotionnelle d'états corporels et pensons que la capacité de ressentir des émotions est utilisée dans les processus de prise de décisions. Aussi, nous proposons l'idée que le sentiment d'identité corporel pourrait naître des capacités d'anticipation et d'action sur l'environnement donc de la bonne adéquation de la valeur émotionnelle perçue en fonction de la situation vécue et des dynamiques d'interaction système-environnement.

1 Introduction

Dans cet article, nous souhaitons mettre en évidence certains processus cognitifs permettant à un système agissant au sein d'un environnement d'acquérir, à

*DASSAULT-AVIATION, DGT-DTN-EL-Et.Avancées, 78 quai Marcel Dassault, 92214 Saint-Cloud

partir de l'expérience de ses interactions sensori-motrices, un ensemble de compétences ou d'aptitudes qui lui permettent d'agir de manière adéquate. Nous illustrons les mécanismes nécessaires à ces capacités d'apprentissage en proposant des modèles algorithmiques que l'on souhaite être les plus simples possibles et dont les simulations sur ordinateur rendent compte de tels comportements.

Un courant de l'Intelligence Artificielle s'intéresse à modéliser ces processus cognitifs par des algorithmes dits "d'*apprentissage par renforcement*". Il s'agit de considérer un *système artificiel*, ou *agent*, "plongé" dans un environnement, dont le comportement est défini par des séquences d'actions (ou décisions ou contrôles). L'information dont dispose le système pour améliorer ses performances est le *renforcement*, une rétribution occasionnelle et a posteriori de son comportement, qui représente une émotion ou une perception affective. C'est un scalaire positif lorsqu'il simule un encouragement (ou un bienfait, par exemple la découverte de nourriture) ou négatif lorsqu'il s'agit d'une sanction (ou un comportement néfaste, par exemple le fait de tomber lorsque l'on apprend à marcher).

Cette rétribution permet au système d'évaluer les prises de décisions passées, lui permettant de rectifier son comportement en renforçant les séquences d'actions qui mènent à des renforcements avantageux.

Nous pensons que de telles capacités d'apprentissage peuvent avoir lieu si le système contient les deux processus dynamiques suivants :

- **Une dynamique d'apprentissage** qui permet d'anticiper les renforcements à venir en construisant itérativement une fonction d'évaluation, qui associe à chaque donnée sensori-motrice (c'est à dire à chaque couple perception-action) une estimation du renforcement à venir.
- **Une dynamique de catégorisation** qui permet d'établir des catégories sensori-motrices homogènes selon le critère de l'estimation des renforcements à obtenir.

A partir de données neurophysiologiques concernant l'observation de modifications comportementales chez des personnes ayant subi certaines lésions cérébrales, nous tenterons de donner une interprétation biologique de ces deux dynamiques.

Nous avancerons l'idée que le sentiment d'identité, tout au moins corporel, nécessite un mécanisme qui permet de percevoir une émotion relative à l'"état corporel" du système, à partir de l'appréciation des dynamiques d'interactions avec son environnement, afin que l'émotion ressentie soit cohérente avec la situation vécue. Nous pensons qu'une telle capacité de ressentir des émotions est utilisée dans les processus de prise de décisions et permet au système d'engendrer des stratégies d'action sur son environnement.

La *partie 2* fournit une introduction aux techniques d'apprentissage par renforcement et décrit le formalisme employé.

La *partie 3* traite des deux dynamiques d'apprentissage et de catégorisation et fait référence à deux réalisations informatiques mettant en application ces mécanismes.

La *partie 4* fournit une interprétation biologique de la fonction d'évaluation et des deux dynamiques précédentes.

2 Introduction aux méthodes d'apprentissage par renforcement

Les modèles d'apprentissage par renforcement sont issus des travaux en psychologie expérimentale (voir par exemple, [RW72]) et en neurophysiologie. Ils sont essentiellement utilisés en *théorie du contrôle* comme méthode adaptative pour déterminer le contrôle optimisant le gain (voir [Bar90] par exemple). Ils fournissent également des algorithmes d'adaptation (par exemple l'algorithme du "Bucket Brigade" utilisé dans les systèmes de règles, voir [Gol89]) pour des systèmes artificiels dits *animats* (modèles informatiques et robotiques qui s'inspirent de comportements animaux ; voir les conférences "Simulation of Adaptive Behavior", par exemple [CHMW94]).

L'apprentissage par renforcement s'intéresse à l'acquisition de compétences d'un système en interaction avec un environnement. Il s'agit davantage de l'acquisition d'un "savoir-faire" ou d'une habileté à réagir en situation (par exemple les coordinations sensori-motrices nécessaires pour tenir en équilibre sur une bicyclette ou pour jouer au tennis), qu'à l'apprentissage de connaissances symboliques (au sens représentationnel ou linguistique).

Lors de son développement, l'individu acquiert, conditionné par les contraintes et possibilités de son être physique en interaction dans un environnement, des compétences motrices qui lui permettent d'agir de manière adéquate en fonction de ses propres capacités et limites.

L'apprentissage par renforcement, tente modestement de modéliser le processus de prise de décisions qui permet de générer des comportements pertinents afin d'optimiser la satisfaction générale, définie assez librement en termes de renforcement, d'un système. En effet, la récompense peut provenir directement de l'environnement (guidage par encouragement ou sanction, méthodes de conditionnement, de dressage) mais elle peut aussi modéliser une motivation interne, propre au système (par exemple selon un principe de plaisir), ce qui lui octroye une certaine autonomie, dans les limites des contraintes externes. En général, les modélisations informatiques prennent en compte un renforcement (interne ou externe) qui ne depend que de l'état présent.

La forme d'apprentissage par renforcement n'exclut pas l'utilisation d'autres formes d'apprentissage (symbolique, supervisé, par analogie, etc.) ni la notion de connaissance a priori, lorsqu'elles sont disponibles. Voici le formalisme mathématique généralement employé.

2.1 Formalisme utilisé

2.1.1 Positionnement du problème

A l'instant t, le système reçoit une certaine perception (de l'environnement et de lui-même) qui définit son *état* $x(t) \in X$ (où X est l'espace d'état). En fonction des compétences actuelles Q (notation détaillée plus tard), le système choisit ou sélectionne une action (ou un contrôle ou un comportement) $a(t) \in A$ ou A est l'ensemble des actions possibles.

Suite à cette action, en fonction de la *dynamique d'interaction agent-environnement*, le système perçoit un nouvel état $x(t+1)$ et peut éventuellement reçevoir un renforcement $r(x(t), a(t))$. Selon les processus que l'on cherche à contrôler, on utilise un modèle du temps discret ou continu :

- **Modèle discret**. Si la dynamique est stochastique, l'état résultant $x(t+1)$ est défini selon une distribution de probabilité qui dépend de $x(t)$ et $a(t)$. Si la dynamique est déterministe, l'état résultant s'exprime comme une fonction successeur : $x(t+1) = succ(x(t), a(t))$.
- **Modèle continu**. Dans le cas déterministe, la dynamique s'exprime comme une équation différentielle dépendant de l'action a :

$$\frac{dx}{dt} = f(x(t), a(t))$$

Dans la suite, nous nous limiterons à étudier les dynamiques déterministes sachant que les conclusions auxquelles nous aboutirons seront généralisables au cas stochastique.

Ainsi, pour tout état initial x, le choix d'une séquence d'actions $a(t)$ génère une séquence d'états $x(t)$ appelée *trajectoire*.

Pour définir la valeur, ou le gain d'une trajectoire, on peut comptabiliser la somme des récompenses obtenues au cours de la trajectoire pondérée par un taux d'actualisation $\gamma < 1$ (notion courante en économie). On définit ainsi le *gain* :

$$J(x; a(.)) = \sum_{0}^{\infty} \gamma^t r(x(t), a(t)) \text{ ou } J(x; a(.)) = \int_0^{\infty} \gamma^t r(x(t), a(t)) dt$$

selon le formalisme employé. Le facteur γ permet d'établir une préférence pour une récompense immédiate plutôt que pour une récompense de même valeur mais différée dans le temps. On peut alors formuler l'objectif de l'apprentissage par renforcement ainsi :

> Trouver pour chaque état x, la séquence d'actions optimale $a^*(t)$ qui permet de maximiser le gain $J(x, a(t))$.

2.1.2 Méthode de la Programmation Dynamique

Le principe de la Programmation Dynamique (voir [Ber87]) consiste à remplacer le problème d'optimisation global précédent (dont l'espace de recherche est l'espace énorme de tous les comportements possibles) par une condition locale (simple) à satisfaire en tous points de l'espace d'état.

Pour cela, on construit une fonction d'anticipation, appelée la *fonction valeur* V, qui représente l'espérance maximale de gain, c'est à dire le gain obtenu lorsque le système effectue le comportement optimal à chaque instant. Ainsi, la fonction valeur est définie par :

$$V(x) = \sup_{a(.)} J(x; a(.))$$

La condition locale est appelée équation de la Programmation Dynamique (formalisme discret) ou équation de Hamilton-Jacobi-Bellman (formalisme continu) et porte sur la fonction valeur :

- **Equation de la Programmation Dynamique :**

$$V(x) = \sup_{a \in A} [r(x,a) + \gamma.V(succ(x,a))]$$

- **Equation de Hamilton-Jacobi-Bellman :**

$$V(x) \ln \gamma + \sup_{a \in A} [r(x,a) + DV(x).f(x,a)] = 0$$

où $DV(x)$ représente le gradient de V en x.

Ainsi, nous avons une équation de "propagation" de la fonction valeur le long de la dynamique d'interaction agent-environnement, et en certains états nous connaissons les renforcements correspondants. Ainsi, en partant des renforcements et en les propageant de proche en proche dans tous l'espace d'etat, le système peut construire le paysage de l'espérance de gain à venir.

L'approximation de la fonction valeur permet d'anticiper et de prévoir les renforcements à venir. Nous l'interpréterons (voir la *partie* 4) comme le fait de "ressentir émotionnellement" la valeur d'une situation ; par exemple si elle est négative, elle peut correspondre à une peur à l'approche d'un danger.

Cependant, pour un état donné, la seule information de la fonction valeur ne suffit pas pour déterminer quelle est la meilleure action à choisir. Pour cela on utilise la *fonction qualité* $Q(x,a)$ définie pour chaque couple (état, action) et qui permet de déterminer l'action optimal en chaque état. Cette fonction attribue à chaque donnée sensori-motrice (état x, action a) une estimation du meilleur gain lorsque que le système perçoit l'état x et choisit l'action a. Considérons ici le cas discret, l'équation de la programmation dynamique s'écrit :

$$\begin{aligned} Q(x,a) &= r(x,a) + \gamma.V(succ(x,a)) \\ \text{avec } V(x) &= \sup_{a \in A} Q(x,a) \end{aligned}$$

Ainsi, la fonction qualité donne une évaluation des couples de données perception-action et permet de définir l'action optimale $a^*(x)$ en chaque état x :

$$a^*(x) = \arg\sup_{a \in A} Q(x,a)$$

Cependant, afin de permettre une exploration de son environnement (notamment en début d'apprentissage), le système peut avoir intérêt à choisir des actions sous-optimales, le choix pouvant se faire, par exemple, selon une distribution de Boltzmann :

$$\text{L'action } a \text{ est choisi selon la probabilité } p(x,a) = \frac{e^{\frac{1}{T}Q(x,a)}}{\sum_{a' \in A} e^{\frac{1}{T}Q(x,a')}}$$

qui permet de générer un bon compromis (en faisant varier la valeur de la "température" T) entre l'exploration de l'environnement (T élévée) et l'exploitation en fonction des connaissances actuelles (T proche de 0). Voir [Meu96] pour une étude sur le dilemme exploration/exploitation dans les systèmes d'apprentissage par renforcement.

Le fait que le choix des actions $a^*(x)$ dépend de l'évaluation $Q(x,a)$ des situations-actions sera rapproché (à la *partie 4)* à la corrélation, observée chez des personnes ayant subi des lésions cérébrales, entre une déficience de leur capacité de prendre des bonnes décisions et une absence de perception émotionnelle correcte des situations.

Nous allons maintenant décrire un exemple d'algorithme d'apprentissage par renforcement, c'est à dire une méthode qui permet au système en interaction avec l'environnement de construire la fonction d'évaluation.

2.2 Un exemple d'algorithme : le Q-learning

Un *algorithme d'apprentissage* est une loi de modification itérative des compétences du système ; en particulier l'algorithme du Q-learning (créé par [Wat89]) fournit la règle suivante portant sur les qualités :

Lorsque le système se trouve dans l'état x, choisit l'action a et se retrouve dans l'état résultant y en percevant éventuellement une récompense $r(x,a)$ alors la qualité $Q(x,a)$ est modifiée selon :

$$\Delta Q(x,a) = \alpha \left[r(x,a) + \gamma . V(y) - Q(x,a) \right]$$

(avec α un coefficient d'apprentissage décroissant au cours du temps). Cet algorithme est intéressant car on connaît les hypothèses pour lesquelles l'apprentissage converge vers la solution optimale.

Ainsi, un algorithme d'apprentissage par renforcement construit *itérativement, à partir des successions des perceptions-actions vécues,* une fonction d'anticipation qui attribue à chaque donnée sensori-motrice une estimation des rétributions à venir.

3 Une double dynamique d'adaptation

3.1 Nécessité d'une catégorisation

Dans le cas général, l'espace des perceptions-actions est immense, ce qui entraîne un problème de représentation sur un ordinateur de la fonction qualité ainsi qu'une divergence du nombre d'itérations à effectuer pour que les valeurs convergent. Il est alors tentant de chercher les régularités de la fonction qualité et d'utiliser ces propriétés pour représenter la fonction avec un nombre limité de paramètres.

Ainsi, on peut essayer de regrouper des perceptions-actions différentes ayant des qualités similaires, donc créer des sous-ensembles de l'espace des perceptions-actions appelés *catégories sensori-motrices*. L'apprentissage consiste alors à déterminer la qualité de chaque catégorie sensori-motrice.

Ainsi, une possibilité serait de faire coopérer les deux processus dynamiques suivants (voir la figure 1 pour une illustration) :

- *Une dynamique d'apprentissage* qui évalue, via le renforcement, les qualités des catégories sensori-motrices existantes.
- *Une dynamique de catégorisation* qui partitionne l'espace des données sensori-motrices en catégories contenant des états dont les gains espérés sont homogènes.

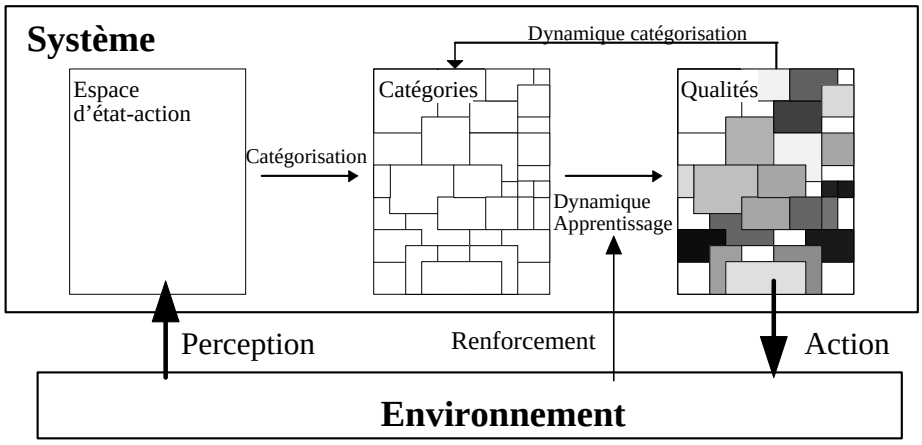

Figure 1: *Le processus de catégorisation des données sensori-motrices via le renforcement.* L'ensemble des perceptions et actions possibles définit l'espace d'état-action. Cet espace est partitionné en régions appelées catégories. La dynamique d'apprentissage permet d'évaluer les qualités (en gris) des catégories. La dynamique de catégorisation modifie le partitionnement pour avoir des catégories homogènes.

Remarque : la modélisation du processus de catégorisation décrite ici est outrageusement simplifiée. En effet, en général, la simple perception de l'état du système met en jeu des processus complexes (détection de traits, discrimination, segmentation, etc.) qui est déjà un mécanisme de catégorisation. Nous n'étudierons ici, pour des raisons de simplicité, qu'un processus de catégorisation qui ne fait que regrouper des états (i.e. les catégories sont des parties de l'espace d'état-action).

3.2 Modèles informatiques

Nous avons étudié les deux réalisations suivantes :

- Le *Q-learning partitionné* (défini dans [MP94]) utilise un formalisme discret et définit une fonction qualité constante à l'intérieur des catégories. Celles-ci sont représentées sous forme de règles et utilisent le formalisme des "Classifiers Systems". La dynamique d'apprentissage consiste à modifier les qualités des règles sélectionnées en fonction du renforcement obtenu. Le processus de catégorisation permet de remplacer une règle générale (règle mère) par une série de règles plus spécifiques (règles filles).

- Le *Finite-Element Reinforcement Learning* avec triangulation adaptative (voir [Mun96] et [Mun97]) utilise un formalisme continu et approche la fonction valeur par des méthodes aux Eléments Finis. Pour une triangulation donnée de l'espace d'état, la fonction qualité est approchés par une fonction linéaire à l'intérieur des triangles (qui définissent les catégories) et la dynamique d'apprentissage consiste à modifier les valeurs aux sommets de la triangulation. Le processus de catégorisation consiste à ajouter des points à l'ensemble des sommets et ainsi permet de construire une nouvelle triangulation plus fine.

Pour ces deux réalisations, la dynamique de catégorisation est identique. Il s'agit d'un processus de spécialisation (descendant) des connaissances, c'est à dire que le système est initialisé avec un petit nombre de "grandes" catégories (ce qui représente une connaissance potentielle grossière et de faible pouvoir discriminant, i.e. connaissance du novice) et le processus de catégorisation consiste à subdiviser certaines catégories (selon un critère estimant l'hétérogénéité de l'espérance de gain dans chaque catégorie) en plusieurs catégories plus fines. Pour déterminer quelles sont les catégories qu'il est nécessaire de subdiviser, on utilise un critère local représentant l'erreur d'approximation de l'espérance de gain dans une catégorie : l'ecart-type des modifications de la qualités de cette catégorie. Ainsi, si une même catégorie mène à des renforcements différents, cela signifie qu'elle contient des états dont l'estimation de gain est variable, sa qualité sera modifiée constamment et elle sera subdivisée.

Ainsi, la connaissance acquise après la dynamique d'apprentissage devient plus précise et pertinente. Ce processus modélise l'acquisition de connaissances expertes.

Cependant ces réalisations souffrent d'une absence de processus ascendants pour lesquels des catégories spécialisées pourraient se regrouper en une catégorie globale afin d'établir un mécanisme de généralisation.

4 Interprétation biologique

4.1 Faits neurophysiologiques

A. Damasio (voir [Dam95]) rapporte le cas de certains patients qui, suite à des lésions cérébrales (causées par des maladies neurologiques ou des accidents ayant détruit des régions corticales), ont subitement été handicapés dans l'aptitude à prendre des décisions. Pourtant, ces personnes semblaient tout à fait normaux par ailleurs ; leurs capacités linguistiques, de mémorisation, d'attention, de raisonnement logique étaient intactes. Seul un symptôme accompagnait cette déficience à agir de manière adaptée : une incapacité à ressentir la moindre émotion.

Ainsi, des lésions cérébrales à différents niveaux :

- régions ventro-médianes du cortex préfrontal,
- aires corticales somatosensorielles dans l'hémisphère droit,
- système limbique (notamment l'amygdale)

entraînent une déficience dans la prise des bonnes décisions (au niveau social, personnel ou de la motricité) ainsi qu'un affaiblissement de la perception des émotions.

Ainsi, on peut considérer cette incapacité à agir de manière pertinente dans leur environnement comme une impossibilité d'attribuer des poids différents aux diverses solutions possibles de sorte que le paysage émotionnel où s'effectue les prises de décisions est plat ou pire, ne correspond pas à la situation vécue.

4.2 Interprétation

Considérons l'"*état corporel*" : ensemble des perceptions internes provenant de l'état du corps, des viscères, des muscles, de la peau, des organes sensoriels, etc.

- Nous identifions l'état du système x à cet état corporel.
- Nous identifions le renforcement $r(x, a)$ à une *émotion primaire* ressentie de façon réflexe et préprogrammée en réponse à la détection de certains traits de l'état corporel, par exemple, de certains sons (rugissement d'un lion), de certains types de mouvements (reptiles), de certains contacts physiques avec des corps extérieurs (douleur, plaisir), etc.

Selon Damasio, la perception d'une émotion correspond à l'information sensorielle provenant d'une partie de l'état corporel. Ainsi,

- nous identifions la fonction valeur $V(x)$ (resp. la fonction qualité $Q(x,a)$) à la perception émotionnelle de l'état corporel x (resp. de l'état x et de l'action a).

Il s'agit d'une *émotion secondaire* apprise par le système et qui résulte d'une propagation des émotions primaires (les renforcements) selon sa dynamique d'interaction avec l'environnement. Ainsi une peur sera apprise à l'approche d'un danger. Un appétit à l'approche d'un repas.

En poursuivant l'analogie, les deux dynamiques peuvent être interprétées de la manière suivante :

- La dynamique d'apprentissage permet d'attibuer à chaque catégorie corporelle la valeur émotionnelle adaptée à l'état actuel, c'est à dire celle qui satisfait une bonne adéquation du système en interaction avec l'environnement. Ainsi, un algorithme d'apprentissage par renforcement (par exemple le Q-learning) permet de construire itérativement cette perception émotionnelle.

- La dynamique de catégorisation permet de regrouper les états corporels en fonction du seul critère de similarité de leur valeur émotionnelle. Ainsi, des états seront perçus comme étant différents uniquement s'ils ont des influences différentes sur le renforcement à venir.

Il nous semble que le sentiment d'identité corporel peut apparaître lorsqu'un système développe des capacités d'anticipation et d'action sur l'environnement lui permettant d'assurer une certaine autonomie. D'après les faits neurophysiologiques précédemment cités, il apparaît que la capacité de ressentir des émotions est utilisée dans les processus de prise de décisions à différents niveaux. Ainsi, nous proposons l'idée selon laquelle l'apparition de l'identité corporelle d'un système nécessite l'existence d'un mécanisme de perception émotionnelle ainsi qu'un processus d'acquisition par l'expérience de ces émotions (par les deux dynamiques d'apprentissage et de catégorisation) afin d'assurer une bonne adéquation de la valeur émotionnelle perçue par rapport à la situation vécue.

5 Conclusion

Nous avons succinctement proposé une modélisation possible des mécanismes cognitifs mis en jeu dans l'apprentissage par l'expérience en distinguant deux processus complémentaires. La dynamique d'apprentissage permet de construire pour chaque catégorie sensori-motrice une évaluation émotionnel afin de permettre des prises de décisions et comportements satisfaisants en fonction du renforcement à venir. La dynamique de catégorisation regroupe des données sensori-motrices en fonction du seul critère de similarité de leur espérance de renforcement.

Ce critère de regroupement est le plus faible si l'on considère le seul objectif de l'optimisation des rétributions à venir.

En effet, un critère plus fort consisterait à regrouper des perceptions-actions menant à des situations similaires, mais ceci nécessiterait de construire au cours de l'apprentissage, un modèle de l'environnement (voir par exemple [Sut90]), ce qui serait utile, mais coûteux d'un point de vue cognitif (puisque cela utiliserait un processus de création de cartes mentales ainsi qu'un processus de planification).

Sous le formalisme étudié ici, la faculté de créer un modèle de l'environnement n'est pas utile car la modélisation de la dynamique agent-environnement ($r(x,a)$, $succ(x,a)$ et $f(x,a)$) ne dépend que de la perception et de l'action à l'instant présent (ce sont des modèles markoviens) ce qui est une simplification abusive pour la plupart des cas réels. Dans les cas où il y a une dépendance du passé ou lorsque la perception est imparfaite (voir les travaux sur les processus partiellement observables de [LCK95] et sur les variables cachées de [McC95]), il semble nécessaire de construire (au moins partiellement) un modèle de l'environnement.

Ainsi, il apparaît que les dynamiques d'apprentissage et de catégorisation représentent les capacités cognitives les plus simples qui rendent compte de comportements d'apprentissage par renforcement et qu'il faut leur associer, selon la complexité de l'environnement dans lequel le système évolue, divers "modules" cognitifs.

Nous avons proposé une interprétation de la fonction valeur comme étant une perception émotionnelle d'un état corporel et soutenu que les processus de prise de décisions dépendaient de l'adéquation de l'émotion perçue avec la situation vécue, relativement à la dynamique d'interaction système- environnement, c'est à dire du bon fonctionnement des dynamiques d'apprentissage et de catégorisation.

References

[Bar90] Andrew G. Barto. *Neural networks for control.* W.T. Miller, R.S.Sutton, P.J. Werbos editors. MIT press, Cambridge, Massachussetts, 1990.

[Ber87] Dimitri P. Bertsekas. *Dynamic Programming : Deterministic and Stochastic Models.* Prentice Hall, 1987.

[CHMW94] D. Cliff, P. Husbands, J.A. Meyer, and S.W. Wilson. *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior.* The MIT Press/Bradford Books, 1994.

[Dam95] A. Damasio. *L'erreur de Descartes, La raison des émotions.* Odile Jacob, 1995.

[Gol89] D.E. Goldberg. *Genetic Algorithms in search, optimization and machine learning.* Addisson Wesley, 1989.

[LCK95] M. Littman, A. Cassandra, and L.P. Kaelbling. Learning policies for partially observable environments: Scaling up. *International Conference on Machine Learning*, 1995.

[McC95] R.A. McCallum. Instance-based utile distinctions for reinforcement learning with hidden state. *International Conference on Machine Learning*, 1995.

[Meu96] N. Meuleau. *Le dilemme Exploration/Exploitation dans les systèmes d'apprentissage par renforcement.* PhD thesis, Université de Caen, 1996.

[MP94] R. Munos and J. Patinel. Reinforcement learning with dynamic covering of the state-action space : Partitioning q-learning. *Simulation of Adaptive Behavior*, 1994.

[Mun96] Rémi Munos. A convergent reinforcement learning algorithm in the continuous case : the finite-element reinforcement learning. *International Conference on Machine Learning*, 1996.

[Mun97] Rémi Munos. Finite-element methods with local triangulation refinement for continuous reinforcement learning problems. *European Conference on Machine Learning*, 1997.

[RW72] R.A. Rescorla and A.R. Wagner. A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 1972.

[Sut90] R.S. Sutton. Integrated architectures for learning, planning and reacting based on approximating dynamic programming. *International Conference on Machine Learning*, 1990.

[Wat89] Christopher J.C.H. Watkins. *Learning from delayed reward.* PhD thesis, Cambridge University, 1989.