変換不変性と内発的動機づけに基づく強化学習

○増山岳人　山下淳　淺間一 (東京大学)

1. 緒言

人の活動環境に内在する社会的場面で運用可能な自律ロボットシステムの開発には多大な社会的発展性が見込まれる．しかし，そのような環境では認識対象の非定常性や局所的な意味規則の発生といった，事前に収集したデータからの解析と分類，適応等では対処が困難な問題が想定される．したがって，ロボット内部に環境との経時的な相互作用を通じた，センサ-モータ系における入出力関係の構造的記述を促す内的基準を設計する必要がある．人間の認識過程においては，知識に基づいて積極的な概念化を行う能動的なトップダウン処理と入力データとその分析結果から得られる特徴に基づく受動的なボトムアップ処理が併用される．本稿では特に，トップダウン処理は知識の力によりボトムアップ処理の不完全な部分を補うものではなく，知識による概念化を積極的に行い，ある種の期待を作り出すものであるという点に着目する [1].

上述の処理過程を強化学習 [2] の枠組みに実装することで，過去の成功体験に関する知識に基づいた主観的な期待によって学習を加速する手法を提案する．つまり，新たな環境における過去の経験の再現性に対する期待が探索空間を縮退させ，センサ情報の概念化を積極的に試みることで環境への適応が促進される．以後，過去に経験したある有限の行動系列と，それに伴って観測されたセンサ情報に関する抽象的知識の組をスキルと呼ぶこととする．ここでの抽象的知識とはスキルの実行に伴って観測されるセンサ情報の時系列を低次元化した量である．それぞれのスキルに対応する抽象的知識は，新たな環境におけるスキルの実行結果の過去の経験に対する再現性を評価するために用いられる．未知環境においてスキルは試行錯誤によって選択され，過去の成功体験に対する高い再現性をもつスキルが内発的に強化される．行動選択過程においては，内発的に動機づけられたスキルの実行がその状態において適切である（あるいは適切でない）というみなしによるバイアスが加えられる．内発的な動機づけはタスクと独立しており，新たな環境において正の外発的報酬を得る保証を与えるものではない．しかし，実際の環境は必ずしも無限定なものではなく，何らかの基準に基づく共通構造を見つけることが可能である．それゆえに，トップダウンな環境に対するみなしが，結果的によりよい方策の学習に寄与し得る．本稿では，過去の成功体験に基づくトップダウンなみなしと経時的なセンサ情報の抽象化によるセンサ - モータ系の分節化が学習を加速することをシミュレーション実験によって示す.

2. 関連研究

漸進的なセンサ-モータ系の構造化は，自律ロボットの設計における重要な論点の一つである．発達ロボティクスなどの分野では，内的な構造化されたデータ系列の表象をロボットに実装するため，内発的に動機づけられた学習手法について盛んに議論されている．Barto らの提案した Intrinsically Motivated Reinforcement Learning (IMRL) [3] では，option による抽象化と，スキル獲得のための salient event に対する内発的報酬が導入されている．Oudeyer らは好奇心によって，適度に学習可能な状況にロボットを駆動させることで効率的な探索を実現した [4]．これらの研究は，タスクと独立な内的な基準にしたがってエージェントを駆動する内発的動機づけが，効率的な探索行動を促進することを示している．しかし，内発的に動機づけられた学習手法のほとんどはセンサ-モータ系のボトムアップ処理に着目しており，トップダウン処理に基づいた構成については議論されていない．情報処理過程におけるトップダウンなバイアスは物体認識の促進と関連付けられる [5]．また，ある種の認知バイアスは動機づけと関連すると言われている [6]．したがって，内発的に動機づけられた，既知の概念を用いた環境認識という方法論についても同様に考慮する必要があると考えられる．すなわち，ロボットが分節化された過去の経験を再現しようとしたとき，それにより現在の行動の帰結が相対化可能となり，概念化された経験の再現性が環境認識の尺度として機能する．このように考えると，内発的に動機づけられた能動的な知識の利用に，効率的な学習過程の促進を期待することができる．

そこで，過去の経験の再現性に対する尺度をいかに設計するかが次の問題となる．網羅的な経験の収集と評価はほとんど不可能であるため，ここでは限られた知識からの抽象化が重要な論点となる．Konidaris ら [7] はタスクと独立した Agent space と呼ばれる特徴空間を提案した．しかし，特徴抽出による探索空間の低次元化を考慮しても，複雑な感覚器に対しては膨大な学習コストが必要となってしまう．この問題に対し，本稿では変換不変性を用いた抽象化を提案する．この考え方は群論的な視点から表象の変換不変性について論じた Censi [8] らのものと共通する部分をもつが，筆者らは経時的な表象の構造化に着目しているという点で異なっている．

3. 変換不変性に基づく内発的動機づけ

センサ-モータ系に観測されるパターンの同一性の認識は，自律ロボットに求められる重要な要件の一つである．すなわち，明確な事前知識を用いることなく，データ系列の類似性を評価する尺度が必要となる．そこで，環境認識に関する抽象的な経験の表象として変換不変性を導入する．本稿では，音声認識研究において Qiao らによって提案されたアファイン変換不変性 [9] を用いる．

$\boldsymbol{X}_{t-k_1:t+k_2} = [\boldsymbol{x}_{t-k_1}, ..., \boldsymbol{x}_t, ..., \boldsymbol{x}_{t+k_2}]$ を特徴ベクト

ル $\boldsymbol{x}_t \in \mathbb{R}^d$ の系列とする．$\boldsymbol{x}_t$ へのいかなるアファイン変換

$$\bar{\boldsymbol{x}}_t = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{c} \tag{1}$$

に対しても，$\boldsymbol{X}_{t-k_1:t+k_2}$ に関するアファイン変換不変量 M は $M(\boldsymbol{X}_{t-k_1:t+k_2}) = M(\bar{\boldsymbol{X}}_{t-k_1:t+k_2})$ を満足する．ここで，$\bar{\boldsymbol{X}}_{t-k_1:t+k_2} = [\bar{\boldsymbol{x}}_{t-k_1}, ..., \bar{\boldsymbol{x}}_t, ..., \bar{\boldsymbol{x}}_{t+k_2}]$ である．本稿では，以下の形式のアファイン変換不変量を適用する．

$$M(\boldsymbol{X}_{t-k_1:t+k_2}) = \sqrt{(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^T(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b)^{-1}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)} \tag{2}$$

$\boldsymbol{\mu}_a$，$\boldsymbol{\Sigma}_a$ はそれぞれ $\boldsymbol{X}_{t-k_1:t+k_2}$ の任意の部分列 $\boldsymbol{X}_a := \boldsymbol{X}_{t-k_1:t-1}$ の平均と共分散行列である．

$$\boldsymbol{\mu}_a = \frac{1}{k_1}\sum_{\tau=t-k_1}^{t-1} \boldsymbol{x}_\tau \tag{3}$$

$$\boldsymbol{\Sigma}_a = \frac{1}{k_1}\sum_{\tau=t-k_1}^{t-1} (\boldsymbol{x}_\tau - \boldsymbol{\mu}_a)(\boldsymbol{x}_\tau - \boldsymbol{\mu}_a)^T \tag{4}$$

$\boldsymbol{\mu}_b$，$\boldsymbol{\Sigma}_b$ も部分列 $\boldsymbol{X}_b := \boldsymbol{X}_{t:t+k_2}$ に対して同様に定義する．音声認識研究においては，声道長や収録機器の違いはケプストラムベクトルに対するアファイン変換によって近似的にモデル化できることが知られている．つまり，アファイン変換不変量は大規模データからの話者正規化を行うことなく，ある音声言語特有の構造に対する尺度を提供する．

ロボットのセンサ-モータ系の分節に対しても同様の考え方が成り立つ．つまり，ロボットと環境の相互作用の帰結として，選択された行動系列と環境の特性に依存してセンサ空間に現れるデータ系列の幾何的構造を観測することができるはずである．例として，測距センサを搭載したロボットの壁沿い走行を想定する．このとき，並進運動に対して対称な情報，すなわちロボットのどちら側に壁が存在するのかという情報は重要でない．ここで唯一重要なのは，壁沿い走行動作に付随してセンサ空間に観測される，壁と一定の距離を保てているかといった，センサ空間における幾何的な構造である．交差点での右折や歩行者とのすれ違いといった，より一般的な場面においても同様の議論が成り立つ．つまり，変換不変量は経時的なセンサ情報の幾何的構造の抽象的表象となる．

実際のロボットの運用においては，しばしば高次元のセンサ空間が想定される．したがって，センサ情報から直接アファイン変換不変量を計算するのは計算コストの面から望ましくない．加えて，アファイン変換不変量は入力ベクトルの系列がなす幾何形状に関する変換不変量であるため，センサ情報の誤差は不変量に対し大きな影響を及ぼす可能性がある．そこで前処理として，センサ情報を特徴空間に写像し低次元化を施すこととする．センサ入力 $\boldsymbol{s} \in \mathbb{R}^d$ の特徴空間への写像を $\boldsymbol{\xi} : \mathbb{R}^d \to \mathbb{R}^n$ とする．ここで $d \geq n$ である．ただし，特徴抽出による $\boldsymbol{\xi}$ のコーディングはロボットの主観的な環境認識に直接的に関わるものであり，その設計の方法論に関しては今後の十分な議論が必要である．センサ情報から得られる特徴ベクトルの系列間の幾何的変換に関する類似性を測る尺度として変換不変量を導入することで，時間，空間的なスケールは異なるが同一の特性をもつ情報を評価することが可能となる．

4. 内発的に動機づけられた強化学習

アファイン変換不変量をスキル実行に関する環境認識の尺度とする，内発的に動機づけられた強化学習手法を提案する．ここで S をセンサ空間，A を行動空間とする．提案手法は Q-learning [10] の枠組みと並行に Semi-Markov Decision Process (SMDP) における学習則が適用される．本稿ではスキル Λ は過去の経験において獲得された最適方策に基づく行動系列と，その実行に伴って観測されるアファイン変換不変量の組として定義される．以後，いくつかのスキルがそれぞれ異なる環境において事前に獲得されていることとして議論を進める．

4·1 スキルの再現性と内発的動機づけ

実行されたスキルを評価する一つの方法は，通常の強化学習と同様に外発的報酬である．本稿では，さらに新たな環境におけるスキル実行の結果得られる，過去の成功体験の再現性に基づく内発的報酬による評価を与える．環境との経時的な相互作用により得られるセンサ情報の履歴は，スキル実行の帰結として解釈し得る．スキルは過去の成功体験に基づいて構築されるため，現在のスキル実行及び過去の経験において得られたそれぞれのアファイン変換不変量で評価される状況の再現性によってスキルは動機づけられる．

ここで，内発的報酬 $r^{int} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ を以下の通りに定義する．現在の環境においてスキル Λ を実行した結果得られる変換不変量を M，Λ のもつ変換不変量を M_Λ とすると，

$$r^{int}(M, M_\Lambda) = r_p^{int} \exp\left(-\frac{|M - M_\Lambda|}{\tau}\right) - r_n^{int} \tag{5}$$

$r_p^{int} \in \mathbb{R}_+$，$r_n^{int} \in \mathbb{R}_+$，$\tau \in \mathbb{R}_+$ は正のパラメータである．この定式化によりアファイン変換不変量の意味で高い再現性を示したスキル程強く動機づけられることになる．次に，以下では強化学習の枠組みにおいて，上述の内発的報酬をスキル選択過程に組み込んでいる．ここで，内発的報酬それ自体はタスクとは独立であることに注意されたい．つまり，内発的報酬に基づくスキルの強化はタスクに対する効率性に寄与する保証を何ら与えるものではない．

4·2 行動価値及びスキル価値

提案手法では行動価値関数 $Q(s,a)$ とスキル価値関数 $Q_\Lambda(s,\Lambda)$ が並行に学習される．ここで $s \in S$ は状態，$a \in A$ は行動である．行動価値は Q-learning の学習則により更新される．

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\left(r_{t+1}^{ext} + \gamma \max_{a'} Q(s',a')\right) \tag{6}$$

$\alpha \in (0,1]$ は学習率，$\gamma \in [0,1]$ は割引率である．$r_{t+1}^{ext} \in \mathbb{R}$ は外発的報酬，$s' \in S$ は a の実行により遷移する次ステップにおける状態である．

スキル価値は明示的には行動価値に影響を及ぼさず，行動選択過程を通して行動価値と関係づけられる．

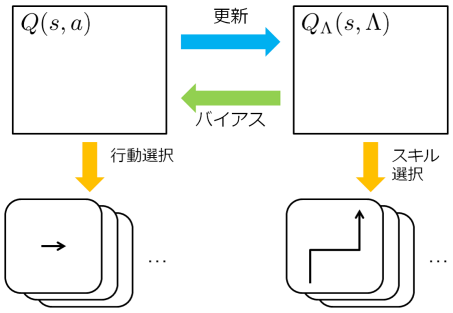

図 1 行動価値とスキル価値の関係性

$Q_\Lambda(s,\Lambda)$ には SMDP に拡張された Q-learning [11] と類似した更新則を適用する．また，$Q_\Lambda(s,\Lambda)$ は各スキルの行動系列の終端，もしくは目標状態に到達した場合にのみ更新される．

$$Q_\Lambda(s_\Lambda,\Lambda) \leftarrow (1-\alpha)Q_\Lambda(s_\Lambda,\Lambda) + \alpha\left(r^{int} + R_\Lambda + \gamma^{T_\Lambda}\max_{a'} Q(s',a')\right) \quad (7)$$

ここで，$s_\Lambda \in S$ はスキルが開始された状態，$T_\Lambda \in \mathbb{R}_+$ はスキルの実行開始からの経過時間を表す．$R_\Lambda \in \mathbb{R}$ はスキル実行過程において累積される割り引かれた外発的報酬の総和であり，以下で求められる．

$$R_\Lambda = \sum_{t_\Lambda=1}^{T_\lambda} \gamma^{t_\Lambda-1} r^{ext}_{t+t_\Lambda} \quad (8)$$

スキル実行中の即時報酬は割り引かれ，累積され，スキル実行が完了した時点で更新に利用される．

4·3 スキルによる行動へのバイアス

スキル価値は行動価値の更新過程に明示的に作用しないため，ϵ-greedy やソフトマックス行動選択といった，Q-learning に適用される任意の行動選択方策が利用可能である．しかしながら，行動価値はスキル価値によるバイアスを加えられ，スキルは間接的に行動価値に埋め込まれる．つまり，スキルの実行に沿って，各時刻でスキルが指定する行動価値はスキル価値によるバイアスを受ける．ある時刻において実行中のスキルの指定する行動の価値は以下の通り変更される．

$$Q(s,a_\Lambda) \leftarrow Q(s,a_\Lambda) + \beta\gamma^{-(t_\Lambda-1)}Q_\Lambda(s_\Lambda,\Lambda) \quad (9)$$

t_Λ はスキルが選択されてからの経過時間，$a_\Lambda \in A$ は t_Λ においてスキルの指定する行動，$s_\Lambda \in S$ はスキルの実行が開始されたときの状態である．$\beta \in \mathbb{R}$ は a_Λ に対するバイアスの強度を決定するパラメータである．バイアスは各時刻において割増されたスキル価値によって計算される．式 (9) は更新則ではなく，バイアスは一時的な加算であり，バイアスが加えられた $Q(s,a_\Lambda)$ はその時刻における行動が選択された時点で元の値に戻される．

以上が提案手法の枠組みであり，$Q(s,a)$ と $Q_\Lambda(s,\Lambda)$ の関係性は図 1 の通りである．スキル価値は式 7 にしたがって，過去の経験の再現性に基づく内発的報酬，外発的報酬及び行動価値によって更新される．他方，行動価値の更新にスキル価値は用いられない．しかし，スキルは式 9 によるバイアスによって行動価値上に埋め込まれる．

5. シミュレーション実験

提案手法の有効性を検証するため，グリッドワールド環境における実験を行った．エージェントの行動は上下左右方向のいずれかに 1 セル移動する 4 種とした．環境は壁に囲まれた正方形領域とし，1 セルを占有する障害物を複数設置した．ただし，壁と隣接するセルには障害物は設置しない．エージェントの初期状態を $(1,1)$，目標状態 $s_d = (15,15)$ と設定した．s_d においてエージェントは 1 の外発的報酬を受け取り，障害物に衝突した場合は-1 の報酬を受ける．また，各行動には-0.1 の移動コストがかかることとした．$Q(s,a)$ 及び $Q(s,\Lambda)$ は初期状態で $[0,1]$ の間の一様分布から各状態についてランダムな値をとった．$\tau = 10$，$r_p^{int} = 1$，$r_n^{int} = 0.5$，$\alpha = 0.2$，$\gamma = 0.95$，$\beta = 0.5$，エピソードの終了条件は s_d への到達，もしくは 300 ステップの経過とした．方策は ϵ-greedy を適用し，現在のエピソード e について $\epsilon = 0.5\exp(-e/80)$ とした．特徴ベクトル $\boldsymbol{\xi}$ はエージェントに隣接する障害物の数及び s_d と現在の状態のユークリッドノルムを要素にもつ 2 次元ベクトルとした．アファイン変換不変量の計算において，$\boldsymbol{\xi}$ の系列は中間地点で分割した．スキルは各々の新規の環境について構成し直した．スキルを獲得する環境における s_d は $(5,5)$ で，5 個の障害物を設置し，Q-learning によって最適方策を求めた．本実験ではスキル数は 5 に設定した．以上の設定で，Q-learning，内発的報酬を $r^{int} \equiv 0$ とした提案手法，提案手法について実験を行った．障害物の配置，スキルのそれぞれ異なる 30 試行の実験の平均を図 2，3 に示す．それぞれの試行について，300 エピソードの実験を 30 回行った結果の平均を用いている．

図 2 はそれぞれの手法の外発的報酬に関する学習曲線を示している．縦軸はエピソード毎の累積報酬，横軸はエピソードである．行動価値の更新則自体は 3 つの手法で同一であるが，学習曲線には顕著な違いがみられる．学習の初期段階において，提案手法と内発的報酬を与えない提案手法は Q-learning より速い立ち上がりを示した．しかし，30 エピソード程で内発的報酬を与えない場合は学習が停滞をはじめ，120 エピソード程で Q-learning の方が高い累積報酬を獲得するようになった．一方で内発的報酬を導入した場合は学習が進行し続けた．この差は当然，内発的報酬の有無の影響を反映したものである．スキルは少なくとも一つの特定の環境における最適方策としての整合性をもつため，スキルの導入により一定の水準までは学習を加速することは可能である．しかしながら，スキルによるバイアスは学習を阻害することもある．そのため，内発的報酬，すなわち変換不変性に基づく環境認識によるスキルの評価と，それによる行動選択過程に対するバイアスの調整が重要となる．スキルがその環境に対して適合するものでなかったとしても，内発的報酬によるトップダウンなみなしは，より高い再現性を得るようスキルを調整し，期待される形に環境の概念化を積極的に推し進める．その結果，スキルは新たな環境

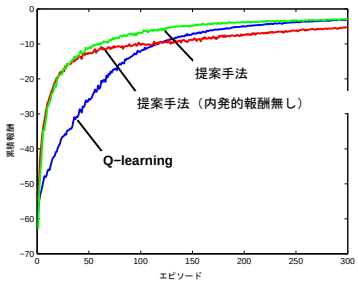

図 2 外発的報酬に対する学習曲線

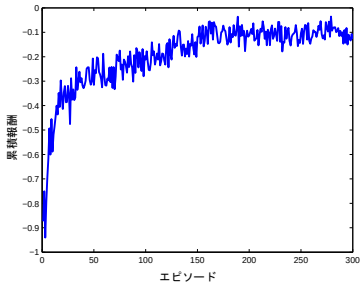

図 3 内発的報酬に対する学習曲線

において，新たな形で学習し直されることになる．

図 3 は内発的報酬の遷移を示しおり，縦軸はエピソード毎の累積報酬，横軸はエピソードである．内発的報酬はタスクとは独立しているため利用できるスキルと環境に敏感な応答を示すが，学習の進行に伴って高い再現性をもつスキルが選択されるという傾向がみられた．ただし，高い再現性をもつスキルが必ず優先的に実行されるようになるわけではなく，エピソードを通じて内発的報酬の累積値が減少するケースも観測された．これはスキル群の環境全体に対する適合性の問題である．つまり，スキルの実行によりパフォーマンスが損なわれる場合もあり得るが，その場合スキル価値によってスキルの指定する行動に負のバイアスを加えることで，この場合もスキルを否定的な概念として扱うことで学習が効率化される．

6. 結言

本稿では過去の成功体験としてのスキルと，その新たな環境における再現性に基づいた内発的動機づけを用いた強化学習手法を提案した．成功体験の再現性に駆動されて，強化学習エージェントはスキルを選択する．スキルの選択がその環境において妥当である保証は与えられないにも関わらず，エージェントが過去の経験の再現性を認識したとき，環境の概念化を積極的に促進することで学習が加速し得ることをシミュレーション実験によって示した．提案手法の基盤は内発的動機づけと抽象的な環境認識にある．高い再現性を示したスキルを強化する内発的報酬は，行動系列の帰結としてのセンサ情報の経時的な抽象化によって生成される．スキルに関する抽象的知識としてはアファイン変換不変量を導入した．センサ空間に現れる幾何的構造の変換不変性を利用することでスキルの再現性を評価し，タスクと独立した動機付けを行った．新たな環境においては行動価値とスキル価値が同時に学習され，また，スキルは行動価値に対する一時的なバイアスによって行動価値に埋め込まれる．したがって，内発的報酬系はバイアスを制御し，スキルの環境への適応を実現する．

トップダウンなみなしは環境の概念化を推し進める強力なツールになり得るが，ボトムアップ処理もまた適応的かつ自律的なロボットの構成に重要である [12]．例えば，本来スキルの獲得はエージェントの発達過程において自律的に獲得されることが望ましい．感覚運動系における時系列の分節化やモデルに基づく制御のための予測器の構築においてボトムアップ処理が必要となる．したがって，トップダウン処理とボトムアップ処理が相補的に機能する包括的な発達的システムの構築は今後の課題である．そのためには行動空間における抽象化によって，2 つの抽象的表象間の関係性を論ずる必要が生じるが，これに関しては統計的な観点からのアプローチを考えている．

参考文献

[1] 甘利俊一, 中川聖一, 鹿野清宏, 東倉洋一: 音声・聴覚と神経回路網モデル. オーム社, 1990.
[2] R. S. Sutton, A. G. Barto: Reinforcement Learning:An Introduction. Cambridge, MA, MIT Press, 1998.
[3] S. Singh, A. G. Barto, N. Chentanez: “Intrinsically Motivated Reinforcement Learning”, Proceedings of Advances in Neural Information Processing Systems 17, pp. 1281-1288, 2005.
[4] P. -Y. Oudeyer, F. Kaplan, V. V. Hafner: “Intrinsic Motivation Systems for Autonomous Mental Development”, IEEE Transactions on Evolutionary Computation, vol. 11, No. 2, pp. 265-286, 2007.
[5] C. Summerfield, T. Egner: “Expectation (and Attention) in Visual Cognition”, Trends in Cognitive Sciences, vol. 13, No. 9, pp. 403-409, 2009.
[6] Z. Kunda: “The Case for Motivated Reasoning”, Psychological Bulletin, vol. 103, No. 3, pp. 480-498, 1990.
[7] G. Konidaris, A. G. Barto: “Building Portable Options: Skill Transfer in Reinforcement Learning”, Proceedings of the 20th International Joint Conference on Artificial Intelligence, vol. 2, pp. 895-900, 2007.
[8] A. Censi, R. M. Murray: “Uncertain Semantics, Representation Nuisances, and Necessary Invariance Properties of Bootstrapping agents”, Proceedings of IEEE International Conference on Development and Learning, vol. 2, pp. 1-8, 2011.
[9] Y. Qiao, M. Suzuki, N. Minematsu: “Affine Invariant Features and Their Application to Speech Recognition”, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp.4629 - 4632, 2009.
[10] C. J. C. H. Watkins, P. Dayan: “Q-learning”, Machine Learning, vol. 8, pp. 279-292, 1992.
[11] A. G. Barto, S. Mahadevan: “Recent Advances in Hierarchical Reinforcement Learning”, Discrete Event Dynamical Systems: Theory and Applications, vol. 13, pp. 341-379, 2003.
[12] A. Bonarini, A. Lazaric, M. Restelli, P. Vitali: “Self-Development Framework for Reinforcement Learning Agents”, Proceedings of 5th International Conference on Development and Learning, 2006.