2002 年 ベイジアンネットセミナー
Tutorial on Bayesian Networks (BN2002)
Tokyo, Japan, Sep. 1-2, 2002.

強化学習エージェントの確率的知識表現を用いた方策改善

An Improvement of Reinforcement Learning Agents' Policy by Using a Representation of Stochastic Knowledge

北越 大輔* 栗原 正仁† 塩谷 浩之‡
北海道大学大学院 工学研究科 室蘭工業大学 工学部

Abstract: 強化学習エージェントは、置かれている環境についての情報を用いることなく、試行錯誤的に方策を学習する。本稿では、強化学習エージェントの試行錯誤の過程で得られるデータ系列と報酬をもとに、情報理論的モデル選択手法によって構築した Bayesian Network を用いた方策改善手法を提案する。計算機実験の結果、Bayesian Network による方策改善機構の導入によって、より効率的な方策の獲得が可能であること、および、構築されたネットワークが、エージェントの置かれた環境における確率的な知識表現となっていることを示す。

1 はじめに

機械学習の一つである強化学習 (Reinforcement Learning) は，報酬という外界からの入力を手がかりに，対象となる環境に適応する手法であり，方策（policy）を最適化することを目的としている．その手法は環境同定型と経験強化型という二つのアプローチに大別され，前者は主にマルコフ決定過程の環境への適応を，後者は非マルコフ決定過程の環境への適応を目的としている [1].

経験強化型の手法は，エージェントの行動決定のための方策の学習によく用いられる [2][3]．その際，強化学習エージェントは，置かれている環境についての情報を用いることなく，試行錯誤的に学習を行う．強化学習エージェントが報酬を得る過程において，状態と行動という組のデータが生成されるが，経験強化型アプローチである利益共有法では，報酬を利用したデータ系列の重み値の更新により方策の学習を行う．ここで，観測したデータ系列と報酬を蓄えて別の形で利用することで，経験強化型学習システムの外部から方策の改善を行う方式も有効となり，そのような例として Bayesian Network を用いた研究が挙げられる．文献 [4] では，あらかじめ設計された方策モデルとして利用した場合の有効性が報告されている．Bayesian Network を方策の改善に利用することに加え，環境に対応する知識ベース構築の基礎となる確率的知識ネットワークシステムとして強化学習機構の上に組み込むことで，さらなる有効性が期待される．

これらの考えをもとに本稿では，強化学習エージェントのデータ系列と報酬から，情報理論的モデル選択手法を用いて構築した Bayesian Network システムを確率的知識として利用した方策改善法を提案する．具体的には，母体となる経験強化機構に利益共有法を適用し，構築された Bayesian Network システムを用いた確率推論によって，方策改善のための教師信号を生成する．よって，強化学習における方策の改善に教師あり学習を併用的に導入することとなる．提案手法の有効性について検証するため，計算機実験を実施する．また，Bayesian Network システムによるエージェントの環境情報表現をも視野に入れた適用の一例についても述べる．

2 準備

2.1 利益共有法による強化学習

強化学習エージェントは，報酬という外界からの入力を用いて方策を最適化することで学習を行う．方策は，ルールと呼ばれる観測状態と行動の対に実数値を与える関数$\boldsymbol{w}$として次式で与えられる．

$$\boldsymbol{w} : \mathcal{S} \times \mathcal{A} \rightarrow \boldsymbol{R} \tag{1}$$

ここで$\mathcal{S}$と$\mathcal{A}$は，エージェントが取り得る状態と行動の集合，対$(s,a)(\forall s \in \mathcal{S}, \forall a \in \mathcal{A})$をルールとする．$\boldsymbol{w}(s,a)$の値をルール$(s,a)$に対する重みと呼ぶ．エージェントは，方策$\boldsymbol{w}$に現在の観測状態$s(s \in \mathcal{S})$を取り入れることで定まる，各ルールの重みを基準としたルーレット選

*〒 060-8628 札幌市北区北 13 条西 8 丁目, tel: 011-706-6861, e-mail: kitakosi@main.eng.hokudai.ac.jp, URL: http://aibm4.main.eng.hokudai.ac.jp/ kitakosi/
†〒 060-8628 札幌市北区北 13 条西 8 丁目, e-mail: kurihara@main.eng.hokudai.ac.jp,
‡〒 050-8585 北海道室蘭市水元町 27-1 tel: 0143-46-5436, e-mail shioya@csse.muroran-it.ac.jp,

択によって，一つに選定されたルールに従い行動を実行する．

利益共有法は，経験強化型アプローチの一つとして知られており，エピソードと呼ばれるルール系列を利用して方策 $\boldsymbol{w}$ を更新する手法である．エージェントは現在の方策の下で，初期状態（もしくは報酬が得られた状態）から次に報酬が得られるまでに，上述の行動選択によって選択されたルール系列 $\{(s_1, a_1), \ldots, (s_C, a_C)\}$ をエピソードとして保存する．ここで，系列長 C はエピソード長と呼ばれる．状態 s_C で行動 a_C を実行した結果，報酬 r が得られたとすると，エピソード内の各ルールに対する重み値は，以下に従って更新される．

$$\boldsymbol{w}(s_i, a_i) \leftarrow \boldsymbol{w}(s_i, a_i) + f(i) \quad (2)$$

$$f(i) = r\gamma^{C-i} \quad (3)$$

ただし $\gamma \in (0, 1]$ とする．

マルチエージェント問題等の，複雑な問題に対して有効とされる利益共有法であるが，エージェントの置かれている環境の推測，もしくは同定が困難であるため，最適方策の獲得は保証されない．本研究では，利益共有法自体に環境に関する情報抽出機能を補完するのではなく，強化学習エージェントの情報を用いた，環境情報に関する知識抽出のための併用システムを導入する．そこで次に，基盤となる Bayesian Network について述べる．

2.2 Bayesian Network の構造決定

本稿では，Bayesian Network の定義については省略し，得られたデータからネットワークの構造を決定する具体的方法について紹介する．サンプルデータから同時確率分布 P が得られた場合，ネットワークの構造を決定することは，データを表現するために最も適切な結合とパラメータ値を決定することである．すなわち，確率変数の N 個のサンプルデータ Db から結合とパラメータを決定することである．

本研究では，情報理論的妥当性がある MDL 基準を用いたモデル選択を採用する [6]．MDL 基準は，

$$MDL(\hat{\boldsymbol{\theta}}, d) = -\log P^N_{\hat{\boldsymbol{\theta}}}(\boldsymbol{D}) + \frac{d \log N}{2} \quad (4)$$

と定義され，この情報量が最小となるモデルを選択する．ここで，パラメータ$\hat{\boldsymbol{\theta}}$は最尤法により得られたものである．このモデル選択は，MDL 基準が最小となるネットワークの結合配置を求めるもので，NP 問題となり，確率的探索法が必要となる．そのため本研究では，焼きなまし法を用いたモデル選択を行う．

構造決定に利用するサンプルデータ，および，提案する方策改善システムの詳細について次節で述べる．

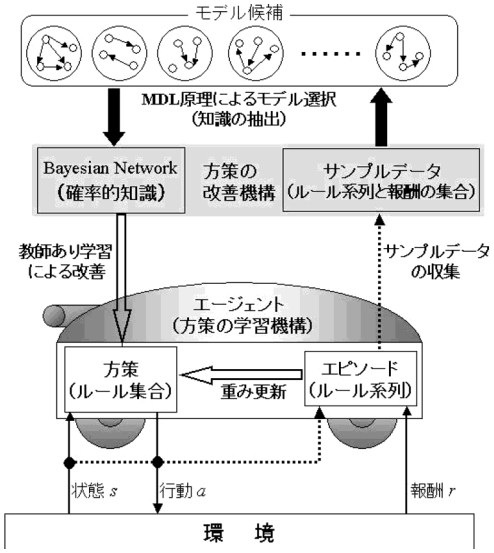

図 1: 強化学習エージェントの方策改善システムの枠組み

3 Bayesian Network を用いた確率的知識による方策改善システム

3.1 システムとその設定の詳細

システムの枠組を図 1 に示す．環境および方策学習機構の部分は従来の強化学習の枠組であり，その上層に，ルール系列と報酬に関するデータから確率的知識を抽出すべく Bayesian Network システムが備えられる．

観測状態の全体集合 $\mathcal{S}$ の各要素に対応した，観測状態ノード $X_{s_1}, \ldots, X_{s_m}$ を用意する ($|\mathcal{S}| = m$)．状態 s に対応する観測状態ノード X_s は，ルール集合 $R_s = \{(s, a)|a \in \mathcal{A}\}$ における行動 a に割り当てた整数値を確率変数値としてとるものとする．また，正の報酬の有無を $\{1, 0\}$ に対応する確率変数として報酬ノード X_r を用意する．方策の改善は以下のようにして行う．

step1:利益共有法による方策の学習と同時に，Bayesian Network 構築のためのデータとして，エージェントが選択したルール系列 $\{(s_1, a_1), \ldots, (s_L, a_L)\}$ (L : 系列長) と報酬 r の組をサンプルデータとして蓄積する．

step2:一定時間の学習後，蓄積されたデータを用いて Bayesian Network の構造を学習する．強化学習エージェントは試行錯誤的に方策を学習するため，ルール系列をサンプルデータとする Bayesian Network のシステムが，全状態のデータを得られる保証はない．サンプルデータが不完全である場合，全状態，および報酬に関する同時確率分布が未知となるため，ネットワークの適切な構造決定は困難である．従って本研究では，得られた全サンプルデータ D1 から，共通に含まれる観測状態を選択し，

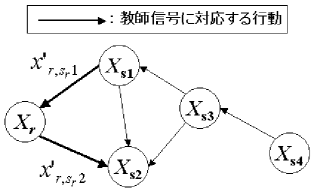

図 2: Bayesian Network を利用した教師信号の生成例

その状態群と報酬に関する同時確率分布に関して MDL 学習を行う．すなわち，不完全データの中から共通に含まれる観測状態に関する完全データを取り出して，ネットワーク構造を決定する．

step3:構築された Bayesian Network において報酬ノードとのリンクを有する観測状態ノードを $X_{s_r1}, \ldots, X_{s_ra}$ とおく．Bayesian Network における条件付独立の観点から，直接リンクされるノード同士の関係に着目し，以下の式を満たす行動 x'_{r,s_rj} を選択する．

$$x'_{r,s_rj} = \arg\max_{x \in \mathcal{A}} p(X_r = 1 | X_{s_rj} = x) \qquad (5)$$

x'_{r,s_rj} よりルールの重みを次式に従って更新する．

$$w(s_rj, x'_{r,s_rj}) \leftarrow (1 + r_j) \cdot w(s_rj, x'_{r,s_rj}) \qquad (6)$$

ただし，r_j は更新の割合で定数とする．以上により，エージェントの方策を改善し，利益共有法による学習を再開する．

不完全データをもとに全ての観測状態について構造推定を行うためには，ノードの追加的学習を行う必要がある．追加的学習では，**step2** で構築したネットワークの各観測状態ノードと，選択されなかったノードそれぞれとの結合を，D1 から取り出した完全データをもとに推定する．この時，追加的学習を行った観測状態ノード間の結合に関する学習は行わない．この方法により，**step2** で構築したネットワーク内のリンクとパラメータに影響を与えず，残りの各ノードに対する結合について追加的な学習を行うことができるが，追加的学習を行う部分については，十分な量のサンプルデータが得られないことが多く，モデル選択が適切に行われない可能性も高い．追加的学習を行ったノードは，方策の改善には関与しない．

3.2 システムの特性

強化学習エージェントは報酬獲得の直前に選択したルール，あるいはエピソードに含まれるルール系列を対象として局所的に方策を改善する．利益共有法の場合，エージェントは学習の進行に従い，特定のルール系列を選択する傾向にあるが，これは頻繁に報酬が与えられるルールが強化されることに起因する．提案手法では，強化学習エージェントの行動によって得られるルール系列と報酬の組をもとに，エージェントが置かれた環境におけるルールと報酬についての確率的依存関係を，Bayesian Network により表現する．環境全体についての確率的知識を表現する場合，環境の全てにおいて一様にサンプルデータを収集する必要があり，データの収集に莫大な時間を要することが予想されるのに対し，提案手法の場合，強化学習による方策の更新と同時に，方策改善に必要なサンプルデータの収集が可能である．このため，本研究で構築される Bayesian Network は環境全体を表現するものではなく，あくまで，正の報酬を得るために必要な環境情報についての確率的知識表現となる．従って，確率的知識から教師信号を生成することは，報酬獲得に関与するルール集合の全体から，改善することが望ましいルールを探索することに対応する．

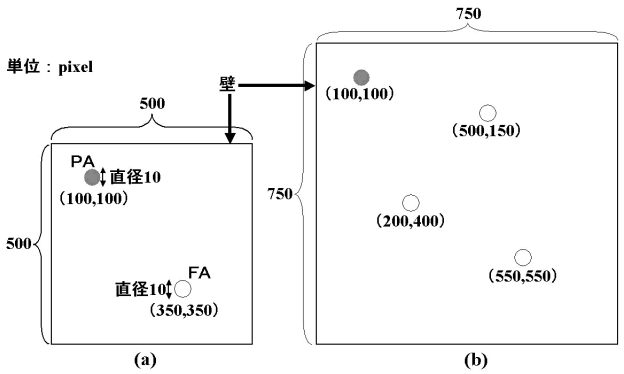

図 3: 実験環境とエージェントの初期位置

本研究では，上記手順（**step3**）における x'_{r,s_rj} を教師信号と表現する（図 2）．教師信号は，ニューラルネットワークの学習などで利用される，外部から与えられる正答例を用いた教師信号とは異なるが，Bayesian Network システムを利用して一意に生成される “最も望ましい出力” という意味で,「教師信号」としている．強化学習を用いた局所的な方策学習機構に加え，提案手法を利用した全体的な観点からの方策改善機構を導入することで，より効率的な方策の獲得が期待される．

4 適用例に関する計算機実験

エージェント追跡問題を例に計算機実験を行うことで，これまで述べた併用的方策改善システムの特性について改めて検討する．エージェント追跡問題は，追跡者エージェント（persuer，以降 PA）が逃亡者エージェント（fugitive，以降 FA）を捕獲する問題であり，多様な設定が可能である．実験環境は，図 3 に示す 2 種類を採

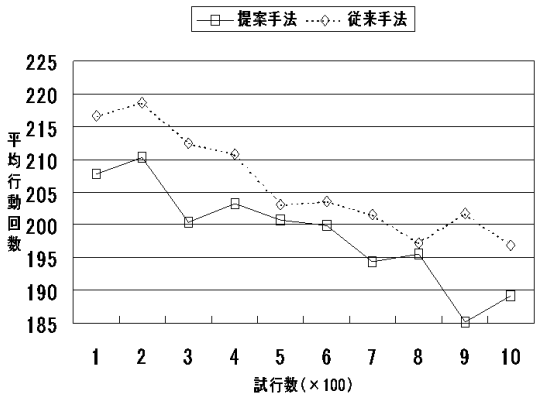

図 4: (**p1**) の方策改善後における平均行動回数の推移

用する．実験の実施にあたり，PA には方策学習のための利益共有法を適用し，FA には 3 種類の異なる方策を用意する．

一回の試行は，PA が FA を捕獲した時，もしくは PA の有するエネルギー（初期値 E_0，行動選択ごとに減少）が 0 となった時に終了し，前者を成功試行，後者を失敗試行と呼ぶ．一回の実験では，PA が利益共有法によって前後半 1000 試行づつ方策を学習し，前半終了後に確率的知識による方策改善を行った場合（提案手法）と行わなかった場合（従来手法）について比較し，提案手法の特徴，有効性について検証する．FA の 3 種類の方策それぞれについて，各手法を 10 回適用して，その結果をもとに考察を行う．具体的には，提案手法において構築された Bayesian Network が表現する環境情報についての分析に加え，双方の手法における，FA の捕獲に要した行動数（平均行動回数）や，全試行に占める成功試行数の割合（成功率）の比較から，方策改善法の効果について議論する．

5　結果・考察および今後の指針

現在まで，提案手法によって構築された Bayesian Network が，FA の方策を含む環境の複雑さを間接的に表現し，環境の相違がネットワーク構造の差に反映されることを確認した．しかしながら，追加的学習を行った観測状態ノードについては 3.1 節に述べた通り，必要なサンプルデータ数の不足等によって，適切なモデル選択は困難であった．上に挙げた問題は，環境の複雑さにも起因していると考えられる．

また，提案手法を適用した PA は確率的知識を利用した方策の改善によって，より効率的な方策を獲得したことを，平均行動数，成功率から示し，より大規模な環境においても同様の結果を得た．得られた結果の一例として，小規模環境 (a) における，(**p1**) に従う FA に関する平均行動回数の推移を示す（図 4）．グラフは方策改善直後からの 100 試行ごとの値をプロットしており，改善直後から，提案手法を適用した PA（実線）が，従来手法（点線）を適用したものより少ない行動回数で FA を捕獲していることが示されている．

現在，他の強化学習法，特に環境同定型アプローチに対して本手法を適用した場合における有効性について検証すべく計算機実験を行っている．

参考文献

[1] 山村 雅幸, 宮崎 和光, 小林重信: エージェントの学習, 人工知能学会誌, Vol. 10, No. 5, pp. 683–689 (1995).

[2] 堀内 匡, 藤野 昭典, 片井 修, 椹木 哲夫: 経験強化を考慮した Q-Learning の提案とその応用, 計測自動制御学会論文集, Vol. 35, No. 5, pp. 645–653 (1999).

[3] D. Kitakoshi, H. Nonaka and T. Da-te: An application of action selection networks adjusting their structures, *Proceedings of MS'2000 International Conference on Modelling and Simulation*, pp. 173–180 (2000).

[4] 山村 雅幸: Bayesian Network 上の強化学習, 第 24 回知能システムシンポジウム (1997).

[5] 本村 陽一, 赤穂 昭太郎, 麻生 英樹: ベイジアンネット学習の知能システムへの応用, 計測と制御, Vol. 38, No. 7, pp. 468–473 (1999).

[6] 山西 健司: 統計的モデル選択と機械学習, 計測と制御, Vol. 38, No. 7, pp. 420–426 (1999).

[7] 宮崎 和光, 山村 雅幸, 小林 重信: 強化学習における報酬割り当ての理論的考察, 人工知能学会誌, Vol. 9, No. 4, pp. 580–587 (1994).

[8] 浅田 稔: 強化学習の実ロボットへの応用とその課題, 人工知能学会誌, Vol. 12, No. 6, pp. 831–836 (1997).