ベイジアンネットを利用した強化学習エージェントの方策改善

北越大輔[†,☆] 塩谷浩之[††] 栗原正仁[†]

機械学習の１つである強化学習は，報酬を利用して方策を最適化することで，エージェントを環境に適応させることを目的とする．本論文では，強化学習エージェントが得た知識を利用して，方策を改善する手法を提案する．我々はエージェントの知識として確率モデルの１つであるベイジアンネットを用い，その構造は，学習中のエージェントの入出力系列，および報酬をサンプルデータとした情報理論的モデル選択手法によって構築される．本研究において構築されるベイジアンネットは，エージェントの入出力と報酬についての確率的依存関係を表現する．本手法におけるエージェントの方策は，ベイジアンネットの構造（確率的知識）を利用した教師あり学習によって改善される．確率的知識を用いた方策の改善機構を導入することで，強化学習エージェントはより効率的な方策の獲得を可能とする．提案手法の特徴について議論するため，エージェント追跡問題を取り上げて計算機実験を行う．さらに，ベイジアンネットシステムによるエージェントの方策情報表現についても論じる．

An Improvement of Reinforcement Learning Agent's Policy by Using a Bayesian Network

Daisuke Kitakoshi,[†] Hiroyuki Shioya[††]
and Masahito Kurihara[†]

Reinforcement learning is a kind of machine learning. It aims to optimize an agent's policy by adapting the agent to a given environment according to rewards. In this paper, we propose a method for improving policies by using knowledge, in which reinforcement learning agents obtain. We use a Bayesian Network as knowledge of an agent. Its structure is decided by a model selection method based on information theory using series of an agent's input-output and rewards as sample data. A Bayesian Network constructed in our study represents stochastic dependences between input-output and rewards. In our proposed method, policies are improved by supervised learning using the structure of Bayesian Network (i.e. stochastic knowledge). Introducing the mechanism of improving policies makes reinforcement learning agents acquire more effective policies. We carry out simulations in the pursuit problem in order to discuss the characteristics of our proposed method. Furthermore, we discuss the information about agents' policies represented by the Bayesian Network system.

1. はじめに

機械学習の 1 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方策（policy）を最適化することで，対象となる環境に適応することを目的とする手法である．その手法は環境同定型（Q-Learning，sarsa 等）と経験強化型（利益共有法，バケツリレー法等）という 2 つのアプローチに大別され[1)]，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．

これまでに提案された経験強化型の手法は，非マルコフ環境における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)]．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先見的に用いることなく試行錯誤的に学習を行うことから，試行錯誤的な機構は強化学習の特徴を反映したものといえる．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，たとえば利益共有法（Profit Sharing）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

† 北海道大学大学院工学研究科システム情報工学専攻
Division of Systems and Information Engineering, Graduate School of Engineering, Hokkaido University
†† 室蘭工業大学工学部情報工学科
Department of Computer Science and Systems Engineering, Faculty of Engineering, Muroran Institute of Technology
☆ 現在，室蘭工業大学生命ソフトラボラトリ
Presently with Life-Oriented Software Laboratory, Muroran Institute of Technology

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，先に述べた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る．そのような例として，ベイジアンネット（Bayesian Network）を用いた研究があげられる．ベイジアンネットは対象から得られたデータの背景の同時確率構造を，非循環性有向グラフ的に表現する知識表現系モデルであり，文献 3) 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．

ベイジアンネットを他システムに組み込んで利用する場合，そのネットワーク構造を事前情報から設計することが多いが，ネットワークに対応するデータが得られる場合，MDL や AIC 等の情報理論的モデル選択を利用できる[4]．モデル選択のための設定や計算コスト，および，事前に十分なデータを得られない場合が生じるといった問題点が存在する反面，情報理論的モデル選択を実用的なデータマイニングに実装する事例もあることから，ベイジアンネット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論文では，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ベイジアンネット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．先述の文献 3) では，構造が固定されたベイジアンネットによってシステム全体としてのエージェントの方策が表され，これを最適化するための部分的なシステムとして強化学習が用いられたのに対し，本論文では方策学習機構（強化学習）と確率的知識による方策改善機構を併用的に利用することで，より高速な方策学習を可能とするシステムの構築を目指している．本手法では，データ系列と報酬から，環境に対応するエージェントの方策情報を確率的知識として抽出する．抽出された確率的知識によって，エージェントの獲得した方策における改善すべき部分が探索され，その該当部分に関する重み更新によって方策改善がなされる．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，構築されたベイジアンネットの構造を方策改善のための教師信号的な役割として利用する．よって，強化学習における方策の学習に教師あり学習を併用的に導入することとなる．以上の提案手法をエージェント追跡問題に適用した計算機実験の結果から，エージェントにおける方策の学習速度が上昇することを示す．また，ベイジアンネットシステムによるエージェントの方策情報表現についても議論する．

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．強化学習はルールベースの学習システムであり，方策は形式的には，ルールに実数値を与える関数として与えられる．本論文における 1 つのルールは条件部と行動部の対からなり，前者にはエージェントのセンサ入力（もしくは観測状態）についての情報が，後者には実行する行動が記述される．センサ入力 c が条件部と合致したときに行動部に記述される行動 a を選択する “if c then a” というルールを (c,a) と書き，方策 $\boldsymbol{w}$ を以下のように定義する．

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

$\mathcal{C}$，$\mathcal{A}$ はセンサ入力と行動の集合，$\boldsymbol{w}(c,a)$（$c \in \mathcal{C}$,$a \in \mathcal{A}$）の値（>0）をルール (c,a) に対する重みと呼ぶ．エージェントは方策 $\boldsymbol{w}$ のもとで，観測したセンサ入力の集合 C_f（$C_f \subset \mathcal{C}$）に含まれる各要素と条件部が合致するルールの重み $\boldsymbol{w}(c_f,a)$（$\forall c_f \in C_f$）をもとに，以下に示す確率 $P(rule=(c_f,a)\,|C_f,\mathcal{A})$ に従って 1 つのルールを選定し，その行動部に記述された行動を実行する．

$$P(rule=(c_f,a)\,|C_f,\mathcal{A}) = \frac{\boldsymbol{w}(c_f,a)}{\sum_{\bar{c}_f \in C_f, a' \in \mathcal{A}} \boldsymbol{w}(\bar{c}_f,a')}. \tag{2}$$

主として分類子システム（Classifier System）の枠組みの中で研究されてきた利益共有法[5]は，経験強化型アプローチの 1 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，初期ルール（もしくは報酬獲得時に選択したルール）から次に報酬が得られるまでに，上述の方法に従って選択されたルール系列 $\mathcal{E}=\{(c_1,a_1),\ldots,(c_G,a_G)\}$ をエピソードとして保存する．ここで，系列長 G はエピソード長と呼ばれる．ルール (c_G,a_G) を選択した結果，報酬 r が得られたとすると，(c_i,a_i)（$\in \mathcal{E}$）に対する重み値は，以下に従って更新される．

$$\boldsymbol{w}(c_i,a_i) \leftarrow \boldsymbol{w}(c_i,a_i) + f_i\ , \tag{3}$$

$$f_i = r\gamma^{G-i} \qquad (i=1,\ldots,G)\,. \tag{4}$$

f_i はルールの強化値を決定する強化関数と呼ばれ，$\gamma \in (0,1]$ とする．これにより，報酬を得る直前のルールに対して高い重み値になるよう加算し，強化値がエピソード系列内で単調増加するように設定される．

すなわち，重み更新と方策 $\boldsymbol{w}$ との関係から，部分集合 $C_G = \{c_i\}_{i=1,\ldots,G}(\subset \mathcal{C})$，$A_G = \{a_i\}_{i=1,\ldots,G}(\subset \mathcal{A})$ に関して制限された，関数 $\boldsymbol{w}_G : C_G \times A_G \to \boldsymbol{R}$ のみが変更されることを意味する．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．この問題に対して宮崎ら[6]は，報酬の獲得に不要なルール（無効ルール）の重みを抑制し，報酬分配の合理性を保証するための必要十分条件（合理性定理）を示した（式 (5)）．

$$M\sum_{j=1}^{i-1} f_j < f_i \quad (\forall i = 2,\ldots,G)\,. \tag{5}$$

ここで M は同一のセンサ入力下に存在する有効ルールの最大個数であり，一般的な実装の際には $|\mathcal{A}|-1$ とすれば十分である☆．また文献 7) では，複数エージェントによる協調行動の実現に向けた，マルチエージェント系における報酬分配の合理性定理も示されている．

本論文では，利益共有法を用いた強化学習システムに対して，強化学習エージェントの得た情報を利用した環境に関する知識抽出システムとして，ベイジアンネットを併用的に導入する．次に，基盤となるベイジアンネットについて述べる．

2.2 ベイジアンネット

2.2.1 ベイジアンネットの定義

ベイジアンネットは，同時確率分布 $P(X_1,\ldots,X_n)$ を用い，各確率変数間の依存関係を非循環性有向グラフとして表現した知識表現系ネットワークである[8]．確率変数をノードとし，変数間に確率的依存関係が強いと判断される場合に方向付けられたリンクを設ける（**図 1**）．依存関係を確率的相関と同一視した場合，同時分布 P は，以下のような条件付き確率分布の積で表現される．

$$P(X_1,X_2,\ldots,X_n) = \prod_{i=1}^{n} P(X_i|\pi(X_i))\,. \tag{6}$$

$\pi(X_i)$ は，確率変数 X_i と相関を持つ確率変数のうち，X_i へのリンクを有するもの（親ノード）からなる同時確率変数（$X_{e_1},\ldots,X_{e_{b_i}}$）とする．式 (6) は，ベイジアンネット中の各ノード X_i が $\pi(X_i)$ のみに依存し，子孫（X_i からのリンクをたどって到達できるすべてのノード）を除いた他のノードとは条件付き独立となることを表している．

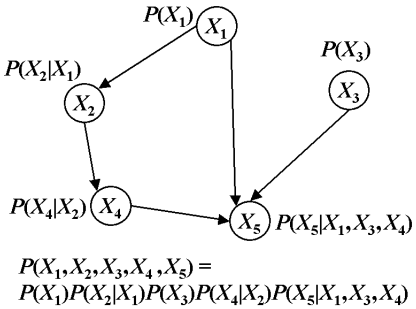

図 1 ベイジアンネットの例
Fig. 1 An example of Bayesian Network.

分布のモデル化について述べる．$P(X_i|\pi(X_i))$ は，パラメータベクトル $\boldsymbol{\theta}^i = (\theta^i_1,\ldots,\theta^i_{d_i})$ によって表現されているものとする[9]．同時確率分布全体のパラメータは，$\boldsymbol{\theta} = (\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^n)$ という，各パラメータベクトル$\boldsymbol{\theta}^i$を合わせたベクトルで表現される．ノード X_i へのリンク数は b_i，パラメータ数は $d_i = dim\,\boldsymbol{\theta}^i$ であり，ネットワーク全体のパラメータ数は $d = \sum_{i=1}^{n} d_i$ となる．これらを決定する手法，つまり確率モデルにおけるモデル選択問題について次に述べる．

2.2.2 ベイジアンネットの構造決定

同時確率分布 P のデータが与えられた場合，ベイジアンネット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，すなわち，確率変数の N 個のサンプルデータ $\mathcal{D} = \{D_1,\ldots,D_N\}$ から結合とパラメータ値を決定することに対応する．ここで，$D_j = (v_{j_1},\ldots,v_{j_n})$ は各確率変数の値を表す多変量データである（$j = 1,\ldots,N$）．

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．MDL 基準は，

$$MDL(\hat{\boldsymbol{\theta}}, d) = -\log P^N_{\hat{\boldsymbol{\theta}}}(\mathcal{D}) + \frac{d\log N}{2}. \tag{7}$$

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–困難な問題となる．このような問題に対しては，多くの場合において確率的探索法が有効となるため，本研究では焼きなまし法による確率的反復改善探索法を用いる．

ネットワーク内のあるノード間にリンクが設定されると，リンクに対応するパラメータとして条件付き確

☆ 文献 6) においては，$L\sum_{j=i}^{W} f_j < f_{i-1}$ $(\forall i = 1,\ldots,W)$ と表されている．ここで L と W はそれぞれ式 (5) における Mと G に対応する．エピソード中のルール系列の並び方が逆で，強化関数の定義が本論文と異なるため，式の形は若干異なるが内容は同一である．

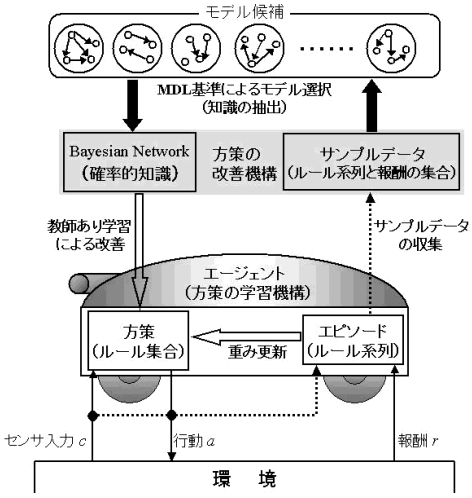

図 2 強化学習エージェントの方策改善システムの枠組み
Fig. 2 The framework of a policy improvement system for reinforcement learning agents.

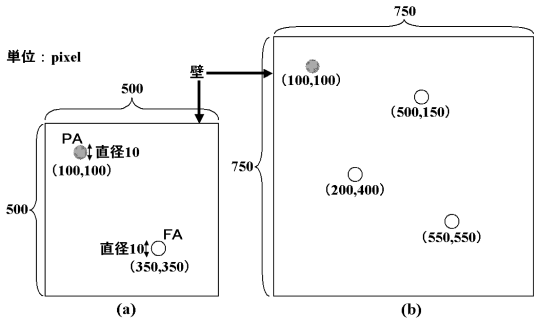

図 3 エージェント追跡問題における環境の例
Fig. 3 Simulation environments and initial positions of agents.

率が割り当てられる．焼きなまし法によって最終的に得られるネットワークのモデルをもとに，そのネットワーク内に設定されたリンクの方向を利用した確率推論が行われる．

3. ベイジアンネット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

システムの枠組みを**図 2** 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みであり，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ベイジアンネットシステムが備えられる．

エージェントのセンサ入力の全体集合 $\mathcal{C}$ の各要素に対応した，センサ状態ノード $X_{c_1},\ldots,X_{c_m}$ を用意する（$m=|\mathcal{C}|$）．センサ c に対応するセンサ状態ノード X_c は，ルール集合 $R_c=\{(c,a)\,|a\in\mathcal{A}\}$ における行動 a に割り当てた整数値（以降これを $\tilde{a}$と表す）を確率変数値とする．また，正の報酬の有無を整数値 $\{1,0\}$ に対応付けた確率変数として報酬ノード X_r を用意する．

まず，提案システムにおける方策改善について，エージェント追跡問題を用いて説明する．エージェント追跡問題は，追跡者エージェント（pursuer，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．**図 3** に示すような環境において，PA がある方策に従う FA を追跡，捕獲するため提案システムを実装する場合について述べる．PA は周囲の壁，および FA の位置情報をセンサ入力 c とし，式 (2) に従って行動 a（移動，停止等）を出力する．PA がエピソードをもとに利益共有法によって方策を学習すると同時に，システムでは学習の過程で得られる報酬 r とルール系列（例：{（壁を上に感知，右に移動）,...,（FA を下に感知，下に移動）}）の対が 1 組のデータとして蓄積される．系列中の各センサ入力がセンサ状態ノードのそれぞれに対応し，その値は行動に割り当てた整数値となる．上の例で，$r>0$ ならば報酬ノード $X_r=1$ となり，センサ状態ノード $X_{\text{「壁を上に感知」}}=$「右に移動」となる．一定時間の利益共有法による学習後，蓄積されたデータを用いてエージェントの確率的知識となるベイジアンネットの構造を決定し，構築されたベイジアンネットをもとに方策改善が行われる．

続いて，具体的な方策改善の手順を以下に示す．

step1-1：利益共有法による方策の学習と同時に，ベイジアンネット構築のためのデータとして，報酬 r とエージェントが選択したルール系列 $\mathcal{L}=\{(c_1,a_1),\ldots,(c_L,a_L)\}$（$L$：系列長）の組 $(r,\mathcal{L})$ を蓄積する．エピソードは，“非 0 の報酬が得られるまでのルール系列” であるのに対し，本手法で蓄積するルール系列には報酬値が 0 の場合のデータも含まれるため，明確に報酬とルールとの確率的依存関係を表現できる．

step1-2：蓄積した報酬とルール系列の集合 $S=\{S_k\}_{k=1,\ldots,N}$（$S_k=(r,\mathcal{L})_k$）の各要素に対して，以下に示す操作を行うことによって，ベイジアンネットの構造決定に利用するサンプルデータ $\mathcal{D}=\{D_k\}_{k=1,\ldots,N}$ の形式を得る．まず，報酬 r の値から報酬ノード X_r のとる値を決定する．次に，(c_i,a_i)（$\in\mathcal{L}$）の条件部 c_i に対応するセンサ状態ノード X_{c_i} の値を $\tilde{a}_i$ とする．$c_i=c_{i+1}$ のように，条件部の内容が重複した場合，$X_{c_i}(=X_{c_{i+1}})$

の値は系列 $\mathcal{L}$ 中のより新しいルールの行動部に対応した値をとるものとする．したがってこの場合，$X_{c_i} = \tilde{a}_{i+1}$ となる．ここで，$\mathcal{L}$ 中の各条件部に対応するセンサ状態ノードからなる同時確率変数を $X_{\mathcal{F}} = (X_{f_1}, \ldots, X_{f_F})$ $(1 \leq F \leq L)$，それ以外のセンサ状態ノードからなる同時確率変数を $X_{\mathcal{N}} = (X_{n_1}, \ldots, X_{n_{(m-F)}})$ とおき，S_k について上述の操作によって得られる同時確率変数 $(X_r, X_{\mathcal{F}}, X_{\mathcal{N}})$ の実現値を D_k とする．ただし，$X_{n_j} (j = 1, \ldots, m-F)$ に対応するセンサ入力は観測されていないため，それらの値は存在しない．

step2：一定時間の利益共有法による学習後，蓄積されたサンプルデータを用いてベイジアンネットの構造を学習する．同時分布 $P(X_r, X_{\mathcal{F}}, X_{\mathcal{N}})$ は，$\mathcal{D}$ の頻度分布により求められる．強化学習システムでは，報酬をもとにして試行錯誤的に方策を学習するため，**step1-2** に示したとおり，ルール系列をサンプルデータとするベイジアンネットのシステムが全状態のデータを得られる保証はない．上述の不完全なサンプルデータは，全センサ状態と報酬に対応する同時確率分布 $P(X_r, X_{\mathcal{F}}, X_{\mathcal{N}})$ において，ルール系列に含まれないセンサ状態に関する周辺分布 $P(X_r, X_{\mathcal{F}}) = \sum_{X_{\mathcal{N}}} P(X_r, X_{\mathcal{F}}, X_{\mathcal{N}})$ から生成されたデータであると見なせる．サンプルデータが不完全である場合，$P(X_r, X_{\mathcal{F}}, X_{\mathcal{N}})$ は未知であるため，ネットワークの適切な構造決定は困難である．したがって本研究では，全サンプルデータ $\mathcal{D}$ から，報酬ノード $X_r = 1$ においてルール系列に含まれる確率の高いセンサ状態 c'_i $(i = 1, \ldots, s)$ に対応するデータ $\mathcal{D}'$ を取り出し，これらに対応するセンサ状態ノード $\{X_{c'_i}\}_{i=1,\ldots,s}$ $(\subset \{X_{c_j}\}_{j=1,\ldots,m})$ の全要素と報酬ノード X_r に関する同時確率分布について MDL 学習を行う．よって，扱われるベイジアンネットのノード数は $s+1$ となる．センサ状態ノード群 $\{X_{c'_i}\}_{i=1,\ldots,s}$ の要素数 s は，ルール系列に含まれる確率の高い上位何番目までを MDL 学習の対象とするかを示しており，その値は問題設定に応じて決定されるものとする．

step3：ベイジアンネットにおける条件付き独立の観点から，直接リンクされるノードどうしの関係に着目する．構築されたネットワークにおいて報酬ノードとのリンクを有するセンサ状態ノードを $X_{c'_r 1}, \ldots, X_{c'_r l}$ とおき，そのそれぞれについて以下の式を満たす行動 $a^*_{r,c'_r j}$ を選択する．

$$a^*_{r,c'_r j} = \arg\max_a P(X_r = 1 | X_{c'_r j} = \tilde{a}) . \quad (8)$$

$a^*_{r,c'_r j}$ よりルールの重みを次式に従って更新する．

$$\boldsymbol{w}(c'_r j, a^*_{r,c'_r j}) \leftarrow (1 + r_{im}) \cdot \boldsymbol{w}(c'_r j, a^*_{r,c'_r j}) . \quad (9)$$

ただし，r_{im} は更新の割合で定数とする．

エージェント追跡問題の例に置き換えれば，方策改善の対象となるルール $(c'_r j, a^*_{r,c'_r j})$ の選択は，PA が蓄積したデータから構築したベイジアンネットによって，報酬との依存関係の強いセンサ入力（例：$c'_r j =$「壁を上に感知」）を抽出し，その入力が得られた際に正の報酬が得られる確率が最も高い行動（例：$a^*_{r,c'_r j} =$「下に移動」）を選択することに対応する．

以上によりエージェントの方策を改善し，方策改善後は利益共有法による学習を再開する．

全センサ状態ノードについての構造決定

不完全データをもとにすべてのセンサ状態についての構造決定を行うためには，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における $\{X_{c'_i}\}_{i=1,\ldots,s}$ の全要素と，**step2** では扱わなかった各センサ状態ノードへの結合を，$\mathcal{D}$ から取り出した完全データをもとに推定する．本論文においては，**step2** で扱わなかった各センサ状態ノード，すなわち追加的学習を行う各ノード間は独立であると見なし，これら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．加えて，ネットワーク全体の構造を一括して学習する場合，ノード数の増加にともない，MDL 基準に関する計算時間，および焼きなまし法によるネットワークの適切な結合状態の探索時間が著しく増大するのに対し，追加的学習ではもとのネットワークに対して追加する各ノードに関する結合のみを学習するため，構造学習に要する時間を短縮可能である．その一方で，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を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がある．

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学習する．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ベイジアンネットにより表現する．環境全体についての確率的知識を表現する場合，環境のすべ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，データ収集の時間やデータ量が莫大にな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学習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．したがって，本研究で構築されるベイジアンネットは環境全体についての確率的知識を表現するものではなく，正の報酬を得るために必要な，環境に対応する方策情報についての確率的知識表現となる．

ベイジアンネットを，エージェントの確率的知識を表現するために用いる場合，センサ入力，行動，および報酬のそれぞれにノードを割り当てることも可能である．この方法は実装が容易である反面，ルールを構成する 2 つの要素が別のノードとなるため，報酬とルールとの依存関係を直観的に表現することが困難であり，提案する方策改善法へ適用する場合には問題がある．本研究では，各センサ入力におけるルール集合をセンサ状態ノードに割り当てることによって，ルールどうし，および報酬とルールとの依存関係を表現可能としている．

本研究では，式 (8) で求められる $a^*_{r,c'_r j}$ を用いて方策を改善する．これは，ニューラルネットワークの学習等で利用される，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，強化学習システムの上層に位置する確率的知識（ベイジアンネット）から一意に生成された “最も望ましい出力” という意味で，強化学習システムにとっての教師信号的な役割を果たしているといえる（**図 4**）．$a^*_{r,c'_r j}$ を選択することは，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．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，方策改善の対象となるルール重みの更新回数は他のルールよりも多くなり，効率的な方策をより高速に学習可能となることが期待される．

4. 適用例に関する計算機実験

3.1 節で紹介した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．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 の方策として以下に示す 3 種類を採用する．

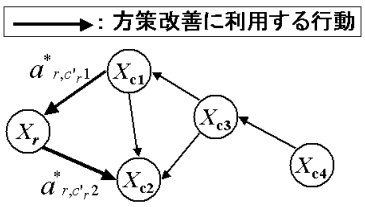

図 4 方策改善に利用する行動（$a^*_{r,c'_r j}$）の例
Fig. 4 An example of the action ($a^*_{r,c'_r j}$) used for the improvement of policy.

（**p1**） つねに停止．
（**p2**） PA を感知するまで停止し続け，感知後は PA から遠ざかるように，かつ壁に接触しないように行動選択．
（**p3**） つねにランダムに行動選択．

実験環境は，図 3 に示した 2 種類を採用する．エージェントは，周囲 Vr 内のエージェントと壁の位置情報をセンサ入力とし，上，右上，$\cdots$，左，左上の 8 方向への移動（移動量は 3pixel），および停止という計 9 種類の行動の 1 つを出力とする．マルチエージェント環境におけるエージェントの状態遷移が，他のエージェントの行動等に依存する場合，状態遷移に不確実性が生じて方策の学習が困難となることから，環境は複雑であると見なせる．上述の方策を FA に実装した場合，エージェントの得られる入力情報の範囲に制限があるため，（**p1**）〜（**p3**）の順に環境の複雑さは増大する．PA は初期値 E_0 のエネルギーを有し，壁への接触時，移動時に E_-，停止時に E_{stay} のエネルギーを失う．また，FA の PA による捕獲を，PA と FA の接触に対応させる．PA が FA を捕獲した試行を成功試行，PA のエネルギーが 0 になった試行を失敗試行と呼び，各試行後におけるエージェントの位置，エネルギーは初期値に再設定される．

報酬は，PA が FA を感知，および捕獲したときに $r_p(>0)$，FA が Vr 外に逃れたとき，PA が壁に接触したときに $r_n(<0)$ を与える．重みの初期値，最小値，最大値をそれぞれ w_0，w_{min}，w_{max} とし，重みの更新には，宮崎らの合理性定理を満たす公比 0.1 の等比減少関数を使用する．エピソード長 G，ルール系列長 L の最大値を 5 に固定し，5 個以上のルールを保存する際は，古いものから削除する．また，各センサ入力において適切な行動をとる方策を学習するため，エージェントはセンサ入力の集合 C_f に変化が生じるまで同じ行動をとり続け[10]，状態変化が生じた際にルールおよびエピソードを蓄積する．さらに，学習した方策を固定させるため，PA が連続して FA を捕獲するにつれて，報酬値を 0 へと減少させ，重み更

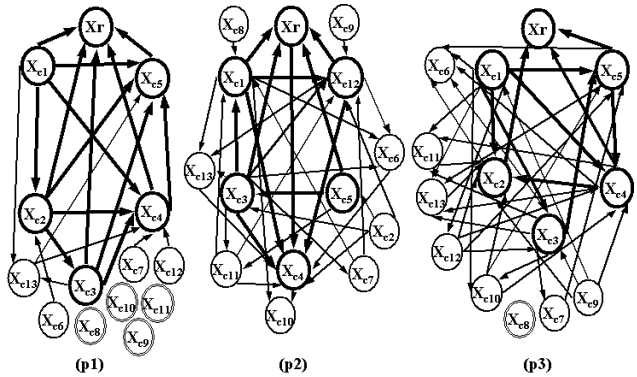

図 5 構築されたベイジアンネットの典型的な例
Fig. 5 Typical examples of the structures of constructed Bayesian Networks.

表 1 各種変数の設定
Table 1 Settings of variables.

変数	値	変数	値	変数	値
E	3	s	5	G	5
E_{stay}	1	r_p	100	L	5
E_O	2000	r_n	−100	T_0	10
Vr	140	W_O	200	σ	0.95
γ	0.1	W_{min}	1	τ	10
r_{im}	0.1	W_{max}	20000		

表 2 ネットワーク中のノードに対応する報酬，センサ入力
Table 2 A reward and sensory inputs corresponding to nodes in the network.

報酬，およびセンサ入力	ノード	報酬，およびセンサ入力	ノード
報酬	Xr	エージェントを左に感知	X_{c7}
何も感知していない	X_{c1}	エージェントを左下に感知	X_{c8}
壁を左方向に感知	X_{c2}	エージェントを下に感知	X_{c9}
壁を下方向に感知	X_{c3}	エージェントを右下に感知	X_{c10}
壁を右方向に感知	X_{c4}	エージェントを右に感知	X_{c11}
壁を上方向に感知	X_{c5}	エージェントを右上に感知	X_{c12}
エージェントを左上に感知	X_{c6}	エージェントを上に感知	X_{c13}

新を抑制する．上述の設定によって構築されるベイジアンネットは，報酬ノードと 13 のセンサ状態ノードを有する．実験では $s = 5$ として，5 つのセンサ状態ノードに報酬ノードを加えた 6 ノードのベイジアンネットを構築する．その後，$\{X_{c_i'}\}_{i=1,\ldots,s}$ に含まれる各センサ状態ノードと，残りの 8 ノードとの結合について追加的学習を行う．ネットワークの構造決定に用いる焼きなまし法の初期温度は T_0 とし，温度の更新には温度減少率 σ の指数型アニーリングを使用する．また，ネットワーク構造の探索は評価関数（MDL 基準）の値が τ 回変化しなかった場合に終了させる．実験で用いる変数の設定を**表 1** に示す．

実験は，利益共有法によって前後半 1000 試行ずつ行い，前半終了後に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）との結果を比較する．3 種類の方策について，各手法をそれぞれ 10 回適用して実験を行う．

5. 結果と考察

5.1 ベイジアンネットによる方策情報表現

はじめに，小規模な環境（a）における FA の方策 3 種類に対して構築したベイジアンネットの典型的な例を示し，それらの特徴について比較する（**図 5**）．図中の太い丸で示された 6 個のノード，およびそれらの結合（太矢印）は，3.1 節の **step2** で構築するネットワークに対応し，残りの 8 個のノードに関する結合は追加的学習によって決定したものである．また，2 重丸で表されるノードは他のノードと独立な（リンクの存在しない）ノードである．各ノードが表す報酬とセンサ入力を**表 2** に，方策の改善に利用する，センサ状態ノードから報酬ノードへのリンクを条件付き確率表として**表 3** に示す．表の各項目における “( )” 中の数字は，それぞれの条件において $X_r = 1$ となった該当データ数を示している．データ数のばらつきは FA のとる方策の違いから生じていると予測されるが，それぞれのデータはエージェントの試行錯誤による方策学習の過程によって得られたものであるため，データ数の多少によらず，条件付き確率表は各環境に対応する PA の方策の特徴を表現できていると考えられる．図 5 より，太い丸のノードは（**p2**）の X_{c12} を除いたすべてが壁の位置情報を表すノードであり，これは PA が，FA を感知する頻度より壁を感知する頻度が多いことを示している．各方策における，報酬ノードとセンサ状態ノードとのリンク数，総リンク数，およ

表 3 条件付き確率表
Table 3 A conditional probability table.

方策	$P(Xr=1\|X_{On})$	X_{On}の値(行動)								
		0	1	2	3	4	5	6	7	8
(p1)	$P(Xr=1\|X_{O1})$	0.00	0.50(2)	0.33(1)	0.30(3)	0.10(1)	0.17(6)	0.00	0.00	0.00
	$P(Xr=1\|X_{O2})$	0.00	0.00	0.00	0.23(3)	0.26(10)	0.00	0.00	0.00	0.00
	$P(Xr=1\|X_{O3})$	0.00	0.21(3)	0.17(2)	0.21(7)	0.00	0.00	0.00	0.00	0.33(1)
	$P(Xr=1\|X_{O4})$	0.00	0.27(10)	0.33(1)	0.00	0.00	0.00	0.00	0.22(2)	0.00
	$P(Xr=1\|X_{O5})$	0.00	0.00	0.00	0.00	0.15(3)	0.21(6)	0.20	0.00	0.00
(p2)	$P(Xr=1\|X_{O1})$	0.00	0.13(18)	0.06(25)	0.11(1)	0.25(1)	0.06(2)	0.00	0.25(1)	0.00
	$P(Xr=1\|X_{O3})$	0.00	0.09(34)	0.06(14)	0.00	0.00	0.00	0.00	0.00	0.00
	$P(Xr=1\|X_{O5})$	0.00	0.00(1)	0.00	0.08(40)	0.08(7)	0.00	0.00	0.00	0.00
	$P(Xr=1\|X_{O12})$	0.08(1)	0.00	0.07(42)	0.00	0.33(1)	0.40(2)	0.50(2)	0.00	0.00
(p3)	$P(Xr=1\|X_{O2})$	0.00	0.00	0.06(15)	0.04(7)	0.06(3)	0.00	0.00	0.00	0.00
	$P(Xr=1\|X_{O4})$	0.04(3)	0.00	0.00	0.00	0.00	0.05(5)	0.05(7)	0.05(11)	0.04(4)
	$P(Xr=1\|X_{O5})$	0.01(1)	0.00	0.00	0.05(11)	0.06(9)	0.00	0.04(4)	0.05(5)	0.00

表 4 リンク数および同時エントロピーの平均値の比較
Table 4 A comparison of the average number of links and average value of joint entropies in three policies.

FAの方策	(p1)	(p2)	(p3)
報酬ノードとの平均リンク数	4.1	4.6	3.6
ネットワークの平均総リンク数	13.0	12.5	11.9
同時エントロピーの平均値	3.36	4.28	5.14

びネットワーク内の全ノード（確率変数）の同時エントロピー値を示す（**表 4**）．表中の値は追加的学習前の 10 個のネットワークにおける平均値である．また，（**p1**）〜（**p3**）において，追加的学習前のネットワーク構築に要した焼きなまし法の反復回数（平均値）はそれぞれ 41.3，39.1，33.6 となった．表における報酬ノードとの平均リンク数の値は，環境における報酬と行動の依存関係に対応付けることができる．（**p3**）に従う FA は，センサ状態に依存せずランダムに行動を選択するが，2 つのノード間に依存関係が存在しなければノード間のリンクも存在しないため，リンク数が最小値を示していると考えられる．逆に，報酬ノードとのリンク数が最も多い（**p2**）では，PA が選択する行動と得られる報酬に，より顕著な確率的依存関係が存在しているといえる．さらに，状態遷移の不確実性の影響に注目して，平均総リンク数について比較すると，不確実性の影響の少ない単純な方策である（**p1**）のリンク数が最も多く，影響の大きな（**p3**）のリンク数が最小となっている．

一方で，追加的学習後のネットワーク全体のリンク数について比較すると，図 5 から明らかに（**p3**）におけるリンク数が最も多い．本実験では，エージェントのルール系列と報酬をサンプルデータとしているため，状態遷移の不確実性の影響が大きい（環境が複雑である）場合，適切なモデル選択を可能とするために必要な量のサンプルデータを得ることは困難である．3.1 節でも述べたとおり，この問題は追加的学習を行うノードについて特に顕著に表れることが予想される．その結果，不確実性の影響が増大するほどこれらのノードに関するリンク数が増大し，表 4 と相反する結果となった．ネットワーク全体として適切な構造を決定するためには，特に追加的学習の対象となるノードについて十分な量のデータを収集する必要がある．

続いて，同時エントロピーの平均値について比較する．本実験における同時エントロピーは，ネットワーク内の総リンク数が 0 で，各ノードが一様分布となるときに最大値（11.679）をとる．ここで，同時エントロピーの大小関係を比較することによって，FA の方策を含めた環境に対応する PA の方策の複雑さ，および相違について議論することが可能である．表 4 より，同時エントロピーの大小関係は方策の複雑さと対応付けられ，状態遷移の不確実性の影響が増大するに従い同時エントロピーの値も増加している．

これらの結果から，本論文で構築されるネットワークの構造は，FA の 3 種類の方策を含む環境に対応する PA の方策の複雑さを間接的に表現し，環境の相違がリンク数や同時エントロピーの差に反映されていることが分かる．

5.2 方策改善法の特徴

小規模な環境（a）における，FA の捕獲に要した行動数の平均（平均行動回数），および全試行に占める成

表 5 （**p1**）における結果の比較
Table 5 A comparison of two methods in (**p1**).

	従来手法	提案手法	従来手法	提案手法
	成功率		平均行動回数	
1〜1000試行	95.92		234.86	
1001〜2000試行	98.45(94)	98.25(93)	206.23(157)	198.63(144)
標準偏差	0.56	0.79	15.99	16.26

表 6 （**p2**）に対する結果の比較
Table 6 A comparison of two methods in (**p2**).

	従来手法	提案手法	従来手法	提案手法
	成功率		平均行動回数	
1〜1000試行	69.99		193.01	
1001〜2000試行	72.81(52)	76.46(60)	194.70(151)	203.62(147)
標準偏差	6.92	5.66	15.83	13.96

表 7 （**p3**）に対する結果の比較
Table 7 A comparison of two methods in (**p3**).

	従来手法	提案手法	従来手法	提案手法
	成功率		平均行動回数	
1〜1000試行	76.05		222.81	
1001〜2000試行	80.82(67)	80.99(70)	226.24(171)	225.45(179)
標準偏差	4.26	4.21	9.66	7.68

功試行数の割合（成功率）について提案手法と従来手法を比較し，提案手法の特徴について考察する.（**p1**）〜（**p3**）における結果を**表 5**，**表 6**，**表 7** に示す．表中の値は，10 回の実験の平均値であり，平均行動回数は成功試行のみについて計算した値である．“( )” 中の値は，実験における結果の最小値である．また，表中には方策改善後の成功率，平均行動回数の標準偏差も示しているが,（**p1**),（**p3**）における成功率，および（**p1**）における平均行動回数の標準偏差は両手法とも同程度の値となっており，残りの部分については提案手法が従来手法より若干小さな値を示している．標準偏差が同程度となった原因は，これら 2 つの手法に適用した利益共有法の設定が同一であったためと考えられる．また，方策改善の結果，成功率，平均行動回数のばらつきが小さくなったため，残りの部分に関する提案手法の標準偏差が小さくなったと予測される．表 5 より,（**p1**）における成功率は，両手法の前後半とも 100%に近い結果を示しており，PA は試行開始直後から FA を捕獲する方策を獲得していると考えられる． また，表 5 および**図 6** から，方策改善後（後半）における提案手法の平均行動回数の値，およびその最小値は従来手法より少なく，確率的知識を利用した教師あり学習によって効率的に FA を捕獲するように方策が改善されていることが分かる．その一方で，後半終了後の平均行動回数の値はこの実験設定における最小値（約 80）に達しておらず，利益共有法の特徴をよく表した結果が得られた.（**p2**）においては，双方の手法における後半の成功率，平均行動回数が前半より大きな値を示している（表 6）．同様の結果は（**p3**）でも観察できる（表 7）．平均行動回数が上昇した原因は，PA が FA を追跡する方策を獲得した結果，多くの行動を費すことによって，FA を捕獲可能となった試行が増加したためと考えられる．また,（**p2**）における方策改善後の成功率の推移（**図 7**）から，提案手法は方策改善の直後から成功率が上昇しており，その値はつねに従来手法よりも大きい．加えて，表 6 から，方策改善後の提案手法における最小成功率も従来手法より大きな値となっていることが分かる．表 4 で示したように,（**p2**）における報酬ノードとのリンク数は 3 つの方策のうち最も多く，報酬ノードとの依存関係を有するこれらのノードを用いた方策改善が有効に作用しているといえる．

一方，表 7 より,（**p3**）の後半における提案手法の成功率，平均行動回数は従来手法とほぼ同じ値を示し

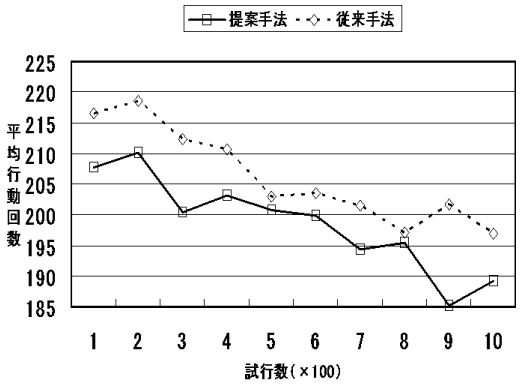

図 6 （**p1**）の後半 1000 試行における平均行動回数の推移
Fig. 6 A transition of the average number of actions in the second half of the trials in (**p1**).

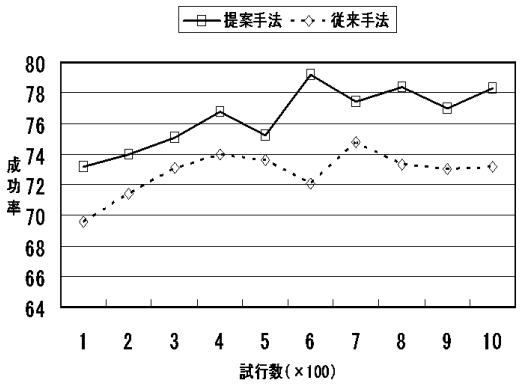

図 7 （**p2**）の後半 1000 試行における成功率の推移
Fig. 7 A transition of success rate in the second half of the trials in (**p2**).

表 8 大規模環境（b）における成功率
Table 8 Success rate on a large-scale environment (b).

試行数＼方策	(p1)		(p2)		(p3)	
	従来手法	提案手法	従来手法	提案手法	従来手法	提案手法
1〜1000	64.20		14.59		53.36	
1001〜2000	69.23(53)	72.19(58)	20.84(2)	24.18(3)	59.19(46)	59.34(47)
標準偏差	4.76	4.77	8.39	8.16	3.25	3.45

表 9 学習速度の比
Table 9 The ratio of learning speed.

環境＼FAの方策	(p1)	(p2)
(a)	1.4*	2.0
(b)	1.9	1.9

ている．5.1 節でも述べたとおり，方策（**p3**）では，報酬の与えられ方が行動のとり方に依存しないため，提案した確率的依存関係を利用する方策改善法が効率的に作用しなかったと考えられる．加えて，強化学習による方策の学習が困難であるような，状態遷移の不確実性の大きい環境においては，ネットワーク構築の際のモデル選択が適切に行われないことが予測され，上と同様，効率的な方策改善は期待できない．

最後に，大規模な環境（b）における成功率を**表 8** に示す．表より，成功率の値は全般的に小規模環境のものより低いが，（**p1**），（**p2**）においては提案手法が従来手法より高い値を示している．したがって，提案した学習システムは，大規模な問題設定においてもその構成を変更することなく効率的に方策を改善可能であるといえる．また，両手法の学習速度について比較するため，後半 1000 試行における提案手法の平均成功率と同程度の値を示すまでに，従来手法が要した試行数を比（従来手法の試行数/1000）によって示す（**表 9**）．小規模な環境（a）における方策（**p1**）（表中の*）に関しては，両手法ともほぼ 100%の成功率であったため，平均行動回数についての比を示した．なお，方策改善が有効に作用しなかった（**p3**）についての比較は行っていない．表に示されるとおり，不確実性の影響が大きい（**p2**）や大規模な環境（b）において，従来手法は提案手法のおよそ 2 倍の試行数を費やさなければ，同程度に良い結果を得られないことが分かる．

提案手法の従来手法に対する十分な優位性を確認するためにはさらなる検証が必要であるが，今回得られた結果から，方策改善法はエージェントの効率的な方策獲得，および方策学習速度の改善に寄与していると考えられる．

6. おわりに

本論文では，ルール系列と報酬をもとに構築したベイジアンネットシステムを確率的知識として利用した，強化学習エージェントの方策改善法を提案した．提案手法の特徴について考察するため，エージェント追跡問題を取り上げ計算機実験を実施した．実験の結果として，構築されたネットワークが逃亡者エージェントの方策を含む環境に対応する，追跡者エージェントの方策についての確率的知識表現となっていることを，リンク数，および同時エントロピーを比較することによって確認した．また，追跡者エージェントの方策は，構築されたネットワークを利用した教師あり学習によって効率的に改善可能であること，および，大規模な問題設定に対してもシステム構成を変更することなく，同様の結果を得られることを示した．さらに，提案手法の適用が困難な問題設定について考察を行った．

今後の課題としては，より多くの実験結果をもとに統計的検定等を含めた様々な手法によって，提案手法の有効性について検証することや，他の強化学習法，特に環境同定型アプローチに本手法を適用した場合の有効性の検証があげられる．Q-learning の場合，ルール重みと Q 値との整合性の考慮や，Q-Learning の最適性に沿った本手法の修正が必要となる．ベイジアンネットの学習の高速化，オンライン化については，情報論的学習におけるベイジアンネットの学習法に関する個別的研究成果を，本研究における提案手法にそのまま適用することで改善が期待される．また，強化学習への適用によるベイジアンネットの扱いの特殊性を考慮した改善等も課題としてあげられる．

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) 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資料 (1997).
4) 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).
5) Sen, S. and Sekaran, M.: Multiagent coordination with learning classifier systems, *Working Notes of the IJCAI-95 Workshop on Adaptation and Learning in Multiagent Systems*, pp.84–89 (1995).
6) 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).
7) 宮崎和光，荒井幸代，小林重信：Profit Sharing を用いたマルチエージェント強化学習における報酬配分の理論的考察，人工知能学会誌，Vol.14, No.6, pp.1156–1164 (1999).
8) 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).
9) Heckerman, D.: A Tutorial on Learning With Bayesian Networks, Technical Report (Microsoft Research Advanced Technology Division) (1995).
10) 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).

(平成 14 年 10 月 7 日受付)
(平成 15 年 9 月 5 日採録)

北越　大輔

1975 年生．1998 年北海道大学工学部情報工学科卒業．2000 年同大学院工学研究科システム情報工学専攻修士課程修了．2003 年同博士後期課程修了．現在，室蘭工業大学生命ソフトラボラトリ非常勤研究員．博士（工学）．機械学習等の研究に従事．

塩谷　浩之（正会員）

1964 年生．1990 年北海道大学理学部数学科卒業．1992 年同大学院工学研究科情報工学専攻修士課程修了．1995 年同博士後期課程修了．現在，室蘭工業大学工学部情報工学科助教授．博士（工学）．数理情報工学の研究に従事．電子情報通信学会，日本神経回路学会各会員．

栗原　正仁（正会員）

1955 年生．1978 年北海道大学工学部電気工学科卒業．1980 年同大学院工学研究科情報工学専攻修士課程修了．現在，同大学院工学研究科システム情報工学専攻教授．工学博士．人工知能およびソフトウェア科学における数理モデル等の研究に従事．電子情報通信学会，人工知能学会，米国人工知能学会各会員．