1-105 強化学習エージェントの確率的知識を用いた方策改善法に関する研究

A Study on a Method for Improving Policies by Using Stochastic Knowledge of Reinforcement Learning Agents

○北越大輔（北大） 塩谷 浩之（室工大） 栗原 正仁（北大）

Daisuke Kitakoshi, Div. of Syst. and Info. Eng., Hokkaido University, Kita 13 Nishi 8, Kita-ku Sapporo, 060-8628, Japan
Hiroyuki Shioya, Muroran Institute of Technology
Masahito Kurihara, Hokkaido University

Reinforcement learning(RL) is a kind of machine learning, and aims to optimize policies of agents by adapting the agents to a given environment according to rewards. In this paper, we propose a method for improving policies by using stochastic knowledge, in which reinforcement learning agents obtain. We use a Bayesian Network (BN) as the 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. The BN constructed in our study represents stochastic dependences between input-output and rewards. In our proposed method, agents' policies are improved by supervised learning using the structure of BN (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 show the effectiveness of our proposed method.

Key Words: Reinforcement Learning, Bayesian Network, Stochastic Knowledge, Supervised Learning

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．強化学習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[1]，そのうち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ている．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(BN) を用いた研究が挙げられる．BN 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 (2) では，予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．BN を他システムに組み込んで利用する場合，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[3]．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，BN を方策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をもとに構築した BN 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，経験強化機構に利益共有法を適用し，構築された BN 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．提案手法の有効性について検証するため，エージェント追跡問題を適用して計算機実験を実施する．

2. 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．本研究における一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時，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．

$$\boldsymbol{w} : C \times A \to \boldsymbol{R} \tag{1}$$

ここで C と A は，センサ入力と行動の集合，対 $(c,a)(\forall c \in C, \forall a \in A)$ をルールとする．$\boldsymbol{w}(c,a)$ の値 (>0) をルール (c,a) に対する重みと呼ぶ．本研究における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 $C_f(C_f \subset C)$ に含まれる各要素と条件部が合致するルールの重み $\boldsymbol{w}(c_f,a)$ $(\forall c_f \in C_f)$ を基準としたルーレット選択に従って一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．

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，上述の行動選択によって選択されたルール系列 $\{(c_1,a_1),\ldots,(c_G,a_G)\}$ 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．(c_G,a_G) を選択した結果，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．

$$\boldsymbol{w}(c_i,a_i) \leftarrow \boldsymbol{w}(c_i,a_i) + f(i) \tag{2}$$

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

ただし $\gamma \in (0,1]$ とする．

3. Bayesian Network

3.1 Bayesian Network の定義 Bayesian Network (BN) 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[6]．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．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．

$$p(X_1,X_2,\ldots,X_n) = \prod_{i=1}^{n} p(X_i|\pi(X_i)) \tag{4}$$

$\pi(X_i)$ 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．

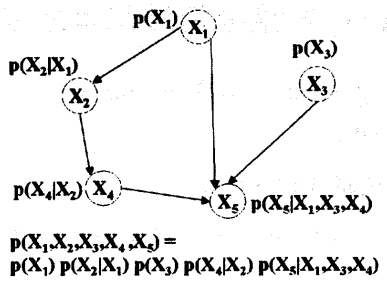

Fig. 1 An example of Bayesian Network

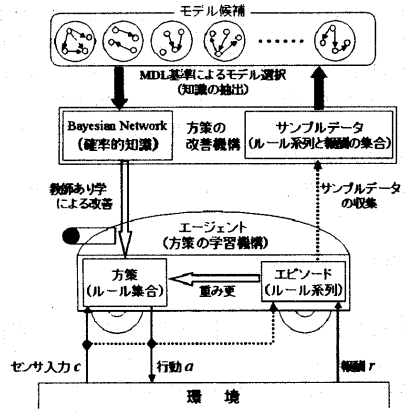

Fig. 2 The framework of a policy improvement system for reinforcement learning agents

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[7]．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，$\boldsymbol{\theta}^i$ 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．

3.2 Bayesian Network の構造決定　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 $\boldsymbol{D}$ 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[3]．MDL 基準は，

$$MDL(\hat{\boldsymbol{\theta}}, d) = -\log P^N_{\hat{\boldsymbol{\theta}}}(\boldsymbol{D}) + \frac{d \log N}{2} \qquad (5)$$

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．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てモデル選択を行う．

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

4.1 システムとその設定の詳細　システムの枠組を図 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であり，その上層に，サンプルデータから確率的知識を抽出すべく BN システムが備えられる．

センサ入力の全体集合 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 ($|C| = m$)．状態 c に対応するセンサ状態ノード X_c は，ルール集合 $R_c = \{(c, a)|a \in A\}$ 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 $\{1, 0\}$ に対応する

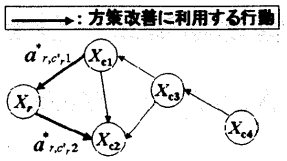

Fig. 3 An example of the action ($a^*_{r,c'_r j}$) used for improving the agent's policy

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学習と同時に，BN 構築のためのサンプルデータとして，エージェントが選択したルール系列 $\{(c_1, a_1), \ldots, (c_L, a_L)\}$ と報酬 r の組を蓄積する．

step2: 一定時間の利益共有法による学習の後，蓄積されたデータを用いて BN 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 BN 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，全状態と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，全サンプルデータ $\boldsymbol{D}$ から，$X_r = 1$ 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 についてのサンプルデータ $\boldsymbol{D'}$ を取り出し，これらに対応するセンサ状態ノードと報酬ノード X_r に関する同時確率分布に関して MDL 学習を行う．

step3: 構築された BN 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 $X_{c'_r 1}, \ldots, X_{c'_r a}$ とおく．BN 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 $a^*_{r,c'_r j}$ を選択する．

$$a^*_{r,c'_r j} = \arg\max_a p(X_r = 1 | X_{c'_r j} = \tilde{a}) \qquad (6)$$

$a^*_{r,c'_r j}$ よりルールの重みを次式に従って更新する．

$$w(c'_r j, a^*_{r,c'_r j}) \leftarrow (1 + r_{im}) \cdot w(c'_r j, a^*_{r,c'_r j}) \qquad (7)$$

ただし，r_{im} 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．

4.2 システムの特性　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，BN 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 BN 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．

本研究では，(6) 式で求められる $a^*_{r,c'_r j}$ 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，強化学習システムの上層に位置する確率的知識（BN）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 3）．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．

5. 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，方策改善システムの特性について検討する．エージェント追跡問題は，追跡者エージェント（PA）が逃亡者エージェント（FA）を捕獲する問題であり，多様な設定が可能である．実験の実施に

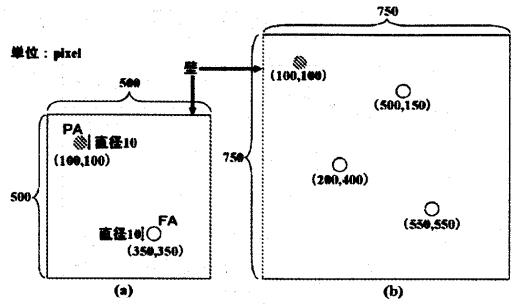

Fig. 4 Simulation environments and initial positions of agents

Table 1 Settings of variables

variable	value	variable	value
E_0	3	r_p	100
E_{stay}	1	r_n	-100
E_o	2000	w_0	200
Vr	140	w_{min}	1
γ	0.1	w_{max}	20000
r_{im}	0.1	G, L	5

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，(**p1**) 常に停止，(**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**) 常にランダムに行動選択，の 3 種類を採用する．

実験環境は，図 4 に示す 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，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環境情報の抽出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，および捕獲した時に $r_p(>0)$，FA が Vr 外に逃れた時，PA が壁に接触した時に $r_n(<0)$ 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取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取り続け[8]，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，学習を収束させるため，PA が連続して FA を捕獲するにつれて，報酬値を 0 へと減少させる．実験で用いる変数の設定を表 1 に示す．

上述の設定によって構築される BN 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，$X_r = 1$ においてルール系列中に含まれる確率の高い 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 BN を構築する．

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．

6. 結 果 と 考 察

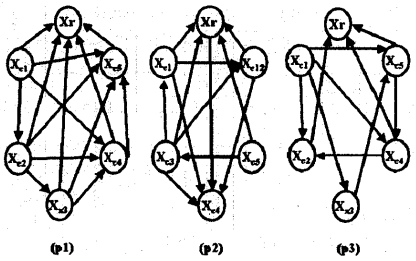

Fig. 5 Typical examples of the structures of constructed Bayesian Networks

Table 2 A reward and sensory inputs corresponding to nodes

報酬，およびセンサ入力	ノード
報酬	X_r
何も感知していない	X_{c1}
壁を左方向に感知	X_{c2}
壁を下方向に感知	X_{c3}
壁を右方向に感知	X_{c4}
壁を上方向に感知	X_{c5}
エージェントを右上に感知	X_{c12}

Table 3 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	0.06	0.11	0.21

6.1 BN による環境情報表現 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 BN 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）．各ノードが表す報酬とセンサ入力を表 2 に示す．また，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 3）．表中の値は 10 個のネットワークにおける平均値である．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，二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．

続いて，同時エントロピーの平均値について比較する．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 3 より，同時エントロピーの大小関係は環境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，本稿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わかる．

今回実施した実験では，ルール系列中に含まれる確率の高いセンサ状態ノードを対象に BN 構築を行った．不完全データをもとに全てのセンサ状態についての構造を決定するためには，ノードの追加的学習を行う必要がある．追加的学習では，4.1 節の **step2** で構築したネットワークの各センサ状態ノードと，選択されなかったノードそれぞれとの結合を，$\boldsymbol{D}$ から取り出した完全データをもとに推定する．この方法により，**step2** で構築し

Table 4 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	98.25	206.23	198.63

Table 5 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	76.46	194.70	203.62

Table 6 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	80.99	226.24	225.45

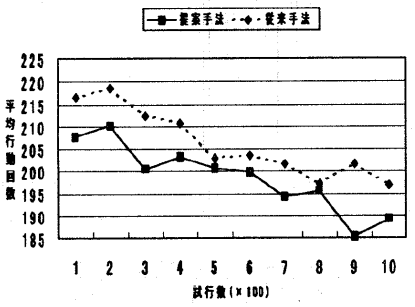

Fig. 6 A transition of the average number of actions in the second half of the trials in (p1)

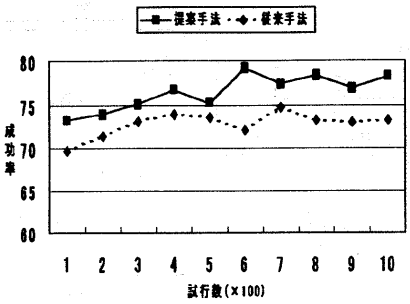

Fig. 7 A transition of success rate in the second half of the trials in (p2)

Table 7 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	72.19	20.84	24.18	59.19	59.34

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．しかしながら，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，その場合モデル選択が適切に行われないことも予想される．

6.2 方策改善法の特徴，有効性 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），及び全試行に占める成功試行数の割合（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，有効性について考察する．(p1) ～ (p3) における結果を表 4~7 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．表 4 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 4 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，従来手法よりも効率的な方策が獲得できたことがわかる．(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 5）．同様の結果は (p3) でも観察できる（表 6）．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，その値は常に従来手法よりも大きい．表 3 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 6 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ている．方策 (p3) では，報酬の与えられ方が行動の取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．

最後に，大規模な環境（b）における成功率を表 7 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．

7. おわりに

本稿では，BN 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．提案手法の有効性について検証するため，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，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の複雑さ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．

今後の課題としては，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挙げられる．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．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なども課題として挙げられる．

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) 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.

(3) 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.

(4) 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.

(5) 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.

(6) 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.

(7) David Heckerman, “A Tutorial on Learning With Bayesian Networks”, *Technical Report (Microsoft Research Advanced Technology Division)*, 1995.

(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.