強化学習を用いた協調学習型 BCI の特性評価

A Characterization of Collaborative Learning Type BCI Using Reinforcement Learning

林　勲[1]*
Isao Hayashi[1]

[1] 関西大学大学院 総合情報学研究科
[1] Graduate School of Informatics, Kansai University

Abstract: Recently, BCI(Brain-computer interface) comes into the research limelight. However, we need an interface model between brain and machine for control and stability. We have already proposed collaborative learning system consisting of reinforcement learning. In this paper, we discuss the usefulness of collaborative learning for BCI using reinforcement learning. We design the collaborative learning system with near-infrared spectroscopy (NIRS), and show the usefulness of the proposed system with a maze problem.

1　はじめに

BCI(Brain-computer interface) [1] では，脳から外部機械に信号を出力するトップダウン処理と外部機械から脳へ信号を入力するボトムアップ処理により，脳と外部機器を相互に接続し外部機器を安定的に制御する必要がある [2]．この内部制御のモデルの一つとして，強化学習 [3] を用いた協調学習型 BCI [4] が提案されている．脳と機器との間に学習型インタフェースモデルを介在させ，外部機器が制御者の意図に合致して制御される (図 1)．また，制御者は外部機器に信号を常時与える必要がなく，効率的な自動制御が実現できる．

本論文では，強化学習を用いた協調学習型 BCI の有用性について検討する．効率的な協調学習に必要な要因を検討し，その効果を迷路探索問題によって示す．具体的には，協調学習の効果を制御課題の精度，制御者の負荷，制御課題の困難性の 3 つの評価値により定義する．制御課題の精度とは，協調学習の制御課題の達成に関する精度であり，制御者の負荷とは，制御者が制御課題の達成に際して強いられる負担量を表す．また，制御課題の困難性とは，与えられた制御課題に対する難易度である．これらの 3 つの評価値を用いて総合評価を定義し，迷路探索問題を例としてその効果を議論する．迷路探索問題では，エージェントが 6×6 の合計 36 マス内の危険地帯を避けるようにスタートからゴールまでの最短経路を探索する．探索過程において，エージェントが危険地帯が近づくと，被験者の避難指示行動の脳信号を近赤外分光法 (NIRS) によりエージェントに与える．エージェントはこの脳示唆を受けて，大きな負の報酬を伴う危険地帯を避けるように効率的に最適経路を学習する．このように，強調学習型 BCI によって，制御者は強化学習を用いて外部機器を安定的に制御することができる．また，脳示唆を強化学習に与えることにより制御者の負担を軽減して外部機器を制御者の意図に沿うよう制御することができる．

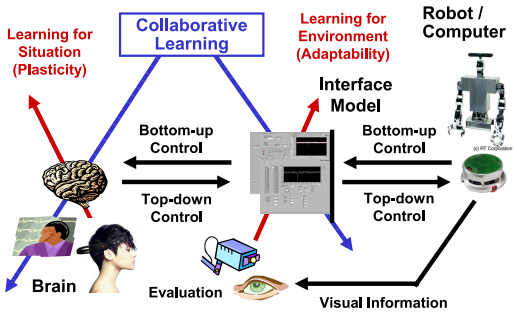

図 1: Concept of Collaborative Learning System

2　強化学習を用いた協調学習型 BCI

学習モデルを強化学習とした場合の協調学習型 BCI の学習過程を図 2 に示す．エージェントは時刻 t に環境の状態 $s(t)$ を観測して行動 $a(t)$ を決定し，それに応

*連絡先：関西大学大学院 総合情報学研究科
大阪府高槻市霊仙寺町 2-1-1
E-mail: ihaya@cbii.kutc.kansai-u.ac.jp

じた報酬 $r(t)$ を得る．協調学習では，同時に，脳信号による示唆 (脳示唆)$su(t)$ が与えられ，この脳示唆を強化学習の教師信号として学習効率を向上させる．

ここでは，協調学習の総合評価として，与えられた制御課題の精度と被験者の負荷，及び，制御課題の困難性の 3 種類の評価から総合的に定義する．

F_1：制御課題の精度

F_2：制御者の負担

F_3：制御課題の困難性

F_1, F_2, F_3 の 3 評価の総合評価を F とし，評価式を次のように定義する．

$$F = w_1F_1 + w_2F_2 + w_3F_3$$

ただし，w_1, w_2, w_3 は重み係数である．

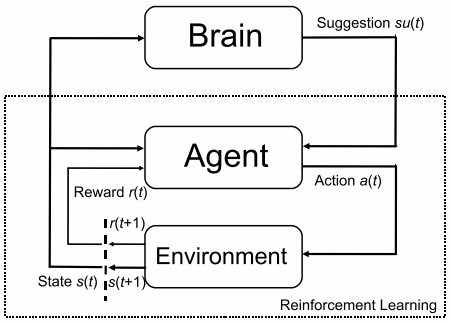

図 2: Proposed Collaborative Learning System

3 迷路探索問題

協調学習システムの例として迷路探索問題を取り上げる．エージェントの制御者は 20 歳代の 3 名である．制御者は 150cm 先に表示された迷路のモニターを固視し，強化学習によるエージェントの探索に NIRS 計測機器により意図を介入する．

図 3 に迷路を示す．迷路は 6×6 の合計 36 マスから構成され，脳示唆を与えない場合には，エージェントは，[上，下，左，右] の 4 方向から 1 方向を選択して行動する Q 学習によりスタート (S) からゴール (G) までの経路を探索する．ただし，ゴールに到達した場合には報酬「10」を得るが，壁に衝突した場合には「-1」，危険地帯に侵入した場合には「-10」を得て，1 ステップ前の位置から再探索する．制御者の脳信号は，近赤外分光法 (NIRS) 計測機器を用いて，国際 10-20 法に

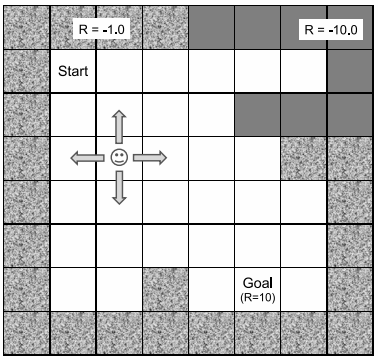

図 3: Maze

よる前頭葉の $FP1$ と $FP2$ の酸化ヘモグロビン変化量 (ox) と還元ヘモグロビン変化量 $(deox)$ として得られる．探索前に制御者の脳示唆を規範意図として計測し，各マスでの制御者の脳信号が，この規範意図に合致する制御者の [行使行動] か，あるいは，エージェントの ε-greedy 法の Q 学習を採用する [採用行動] かを推定する．制御者の脳信号は実験中に常時計測されており，エージェントへの指示行動として解釈する．

行使行動：制御者の意図行動

採用行動：エージェントの Q 学習行動

実験前の規範行動の観測では，制御者の行使行動と採用行動のそれぞれの脳信号を 10sec 間で 5 回計測し平均値として算出した．また，迷路探索中の脳示唆では，制御者の $FP1$ と $FP2$ の酸化ヘモグロビン変化量 (ox) と還元ヘモグロビン変化量 $(deox)$ をそれぞれ 10sec 間計測した．被験者の [行使行動] か [採用行動] は次式で判定した．

行使行動：

$$\sum_{h=\{ox,deox\}} |E(h) - E(e)| \leq \sum_{h=\{ox,deox\}} |E(h) - E(a)|$$

採用行動：

$$\sum_{h=\{ox,deox\}} |E(h) - E(e)| > \sum_{h=\{ox,deox\}} |E(h) - E(a)|$$

ここで，$E(h)$ は探索中に観測した酸化ヘモグロビン変化量 (ox)，及び，還元ヘモグロビン変化量 $(deox)$ の平均値であり，$E(e)$，$E(a)$ は，それぞれ，規範行動の [行使行動 (e)] と [採用行動 (e)] を示す．

また，制御者の負担はエージェントへの脳示唆の頻度とし，エージェントの探索ステップを 15, 30, 45, 60, 75, 90, 105, 120, 200, 300, 400, 500, 600 回で変化させた場合に，探索ステップごとに与える脳示唆の負担とした．

4　探索結果

4.1　総合評価 F

実験では，3 名の被験者に対して合計 7 回の計測を行った．F_1, F_2, F_3 の各評価は，それぞれ，探索効率，脳示唆行動回数，危険地帯到達回数として $[0,1]$ に変換した．各評価を次のように定義し，総合評価 F を算出する．

$$F_1 = \frac{P_1 + P_2 + P_3}{3} \quad (1)$$

$$P_1 = \frac{x_1 - min(x_1)}{max(x_1) - min(x_1)} \quad (2)$$

$$P_i = \frac{max(x_i) - x_i}{max(x_i) - min(x_i)}, \quad i = 2,3 \quad (3)$$

$$F_i = \frac{max(x_{i+2}) - x_{i+2}}{max(x_{i+2}) - min(x_{i+2})}, \quad i = 2,3 \quad (4)$$

ただし，x_1 は収益，x_2 は試行数，x_3 はステップ数，x_4 は脳試行回数，x_5 は危険地域到達回数とし，F_1 は，評価 P_1, P_2, P_3 から成り立つとする．

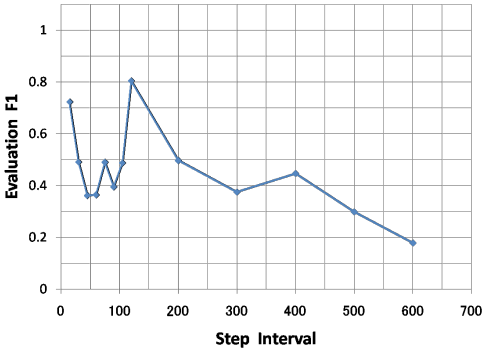

図 4: Estimation of Search Efficiency

探索効率 F_1 の結果を図 4 に示す．脳示唆ステップ間隔が上昇すると，探索効率 F_1 が降下している．脳示唆を与えない (強化学習のみ) による評価値は 0.43 として得られた．したがって，脳示唆ステップ間隔が 250 回以内では，協調学習が強化学習よりも効率が良く，250 回以上では，強化学習のみの方が効率的であるといえる．脳示唆行動回数 F_2 の結果を図 5 に示す．脳示唆ステップ間隔が上昇すると，脳示唆行動回数 F_2 が上昇し，ステップ間隔が 400 回以上では，ほぼ 1.0 を満足している．危険地帯到達回数 F_3 の結果を図 6 に示す．脳示唆ステップ間隔が上昇すると，制御者が介入する機会が増えるので，危険地帯到達回数 F_3 が降下している．強化学習のみによる評価値は 0.51 として得られた．脳示唆ステップ間隔が 200 回以内では，協調学習が強化

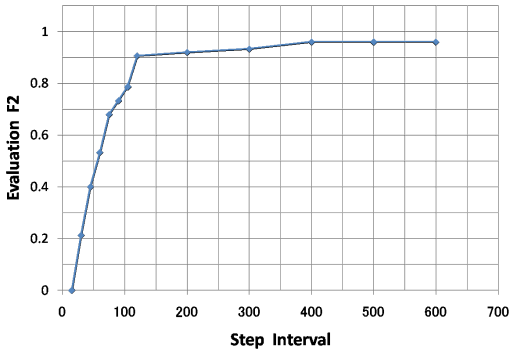

図 5: Estimation of Supervisor Instruction

学習よりも危険地帯に到達する回数が少なく，脳示唆ステップ間隔を 200 回以上にすると，危険地帯に到達する回数が強化学習の場合よりも多くなる．

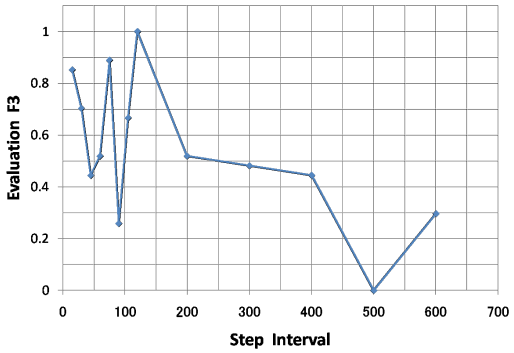

図 6: Estimation of Reaching Dangerous Area

F_1 から F_3 の評価値を $w_i = 1/3$, $i = 1,2,3$ として加重平均し，総合評価 F の結果を図 7 に示す．脳示唆ステップ間隔が 120 回の探索のとき総合評価 F は最大評価値を示している．すなわち，協調学習は脳示唆ステップ回数が 120 回の場合に，最も効率の良い探索を行っているといえる．

4.2　壁衝突回数の比較

脳示唆ステップ回数が 120 回の場合の協調学習と強化学習とを壁衝突回数で比較した．図 8 に各試行の壁衝突回数の比較を示す．協調学習の壁衝突回数を実線で，強化学習の壁衝突回数を点線で示す．協調学習は，3 名の被験者に対して，脳示唆ステップ回数を 120 回として 15 回の連続試行を行った．強化学習は，Q 値を

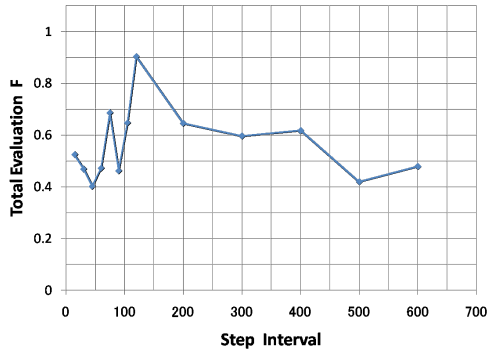

図 7: Total Evaluation

保持しながら 15 回の連続試行を行い，5 回の実験の平均値である．協調学習は強化学習と比較して，ほとんどの試行で強化学習よりも低い衝突回数を示し，さらに試行を繰り返すことによって，衝突回数が低下していることがわかる．

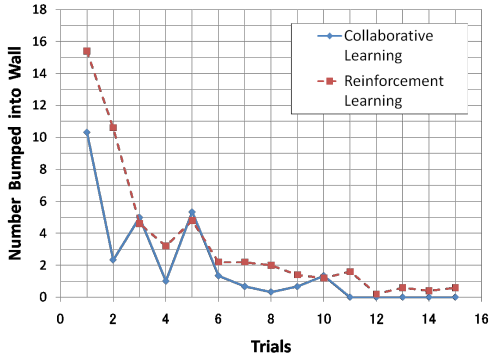

図 8: Number Bumped Wall

4.3 Q 値の比較

3 名の被験者の協調学習と強化学習との Q 値を比較した．結果を図 9 に示す．協調学習は，脳示唆ステップ間隔を 120 回とし，15 回の連続試行の Q 値の平均値である．強化学習は，15 回の連続試行で 5 回実験を行った Q 値の平均値である．各セルでは，上移動 (左上に表示)，下移動 (右下に表示)，左移動 (左下に表示)，右移動 (右上に表示) を表し，上部の数値が協調学習の Q 値であり，下部の数値が強化学習の Q 値である．協調学習は，危険地帯付近の $D1$, $D2$, $E3$ で危険地帯を避ける方向に Q 値が更新されていることがわかる．

	A	B	C	D	E	F
1	-0.79 0.0 0.02 **S** -0.06 -0.83 0.01 0.01 -0.1	-0.67 0.0 0.06 -0.18 0.0 -0.79 0.0 -0.03	-0.81 0.0 -0.02 -0.22 0.0 0.01 0.0 -0.25	-6.67 0.0 1.36 0.0 0.0 0.0 -0.02 -0.37	-7.4 0.0 0.65 0.0 0.0 -5.83 0.0 0.42	-7.5 -8.75 -0.09 -2.5 0.0 -6.25 0.0 0.18
2	0.0 -0.67 0.0 0.12 -0.67 0.08 0.09 -0.26	**Wall**	0.0 0.0 -0.01 0.0 -0.67 0.21 -0.02 -0.5	0.0 -6.67 0.0 0.18 0.0 0.0 -0.01 -0.83	**Wall**	
3	0.0 -0.5 -0.02 0.34 -0.63 0.38 0.15 -0.45		0.0 0.0 -0.03 -1.06 -0.5 0.81 0.19 -0.27	0.0 0.0 0.0 -0.77 0.0 0.0 0.0 -2.39	0.0 -0.5 5.86 0.08 0.0 0.34 -0.43 -0.86	
4	0.0 0.97 -0.01 -0.22 -0.58 0.0 0.05 -0.39	-0.58 0.0 0.08 -1.66 0.0 1.62 -0.01 1.33	0.0 0.0 0.0 -2.84 0.0 2.57 0.0 0.74	0.0 0.0 -0.66 -1.74 0.0 1.27 -0.51 -2.92	0.0 0.0 -0.08 0.0 0.46 2.25 -0.67 -0.99	-0.5 -0.5 0.1 0.18 0.0 0.0 -2.48 0.0
5	0.0 0.0 -0.02 -0.71 0.0 0.0 0.71 0.0	0.44 2.59 0.44 0.65 0.0 0.0 0.0 -0.06	0.42 6.41 -0.03 1.93 0.0 -0.75 0.0 -0.08	0.0 8.53 -0.6 6.90 1.20 0.0 0.81 -6.4	0.0 0.0 -2.27 -0.99 0.0 9.93 0.0 1.71	0.0 -0.5 -0.23 0.0 1.13 0.75 0.48 -1.22
6	0.0 0.0 0.0 0.0 -0.5 -0.5 0.08 0.10	0.0 -0.5 -0.46 0.24 0.0 0.0 0.0 0.63	**Wall**	3.61 5.0 2.52 -4.67 0.0 -0.5 0.53 0.09	**Goal**	0.0 0.0 0.0 0.5 7.5 0.0 0.89 0.5

図 9: Comparison of Q Values

5 おわりに

本論文では，迷路探索問題を例として，強化学習を用いた協調学習型 BCI の有用性を制御課題の精度，制御者の負荷，制御課題の困難性の 3 つの評価値により総合的に議論した．今後，迷路探索問題だけではなく他の多くの応用問題に適用して，協調学習の有用性を議論する必要がある．

なお，本研究の一部は「文部科学省私立大学戦略的研究基盤形成支援事業 (平成 20 年度〜平成 24 年度)」によって行われた．

参考文献

[1] M.A.Lebedev, J.M.Carmera, J.E.O'Doherty, M.Zacksenhouse, C.S.Henriquez, J.C.Principe, and M.A.L.Nicolelis: Cortical ensemble adaptation to represent velocity of an artificial actuator controlled by a brain-machine interface, *Journal of Neuroscience*, Vol.25, No.19, pp.4681-4693 (2005).

[2] 林，徳田，清原，田口，工藤：生体表現システム：ファジィインタフェースを用いた培養神経細胞とロボットとの相互接合，知能と情報，Vol.23, No.5, pp.761-772 (2011)

[3] R.S.Sutton, A.G.Barto(著), 三上，皆川 (訳)：強化学習，森北出版 (2000)

[4] 林，三輪，仙浪：強化学習と脳信号による BCI 協調学習の基礎的研究，第 *25* 回ファジィシステムシンポジウム講演論文集，1B1-02 (2009)