1P2-11in

緩い対称性モデルのゲームへの応用

Application of Losse Symmetry Model to game

西村 友伸[*1] Tomonobu Nishimura　齋藤 慧太朗[*1] Keitarou Saitou-　丸山 涼平[*1] Ryouhei Maruyama　大用 庫智[*2] Oyo Kuratomo　高橋 達二[*1] Tatsuji Takahashi

[*1]東京電機大学 Tokyo Denki University　[*2]東京電機大学大学院 Graduate School of Tokyo Denki University

Human reasoning sometimes goes wrong, but the error has a certain trend. The illogicality is the result of environmental adaptation through evolutionary process. If not formally rational, it is adaptively rational as a heuristics under restricted computing resources. It may also be a source of certain properties proper to human. In this article, we apply Shinohara's LS model to game AI and game theoretic framework for constructing human information systems.

1. 序論

本研究では篠原らによって見出された、対称性バイアスと相互排他性バイアスの両方を緩く含んだ因果関係についての期待形成モデル $lossesymmetryModel$ (LS モデル)　[篠原 07] を扱う。既存の研究において LS モデルは 2 本腕バンディット問題において、少ない計算量で高い正解率を記録している。しかしながら、LS モデルについての検証は 2 本腕バンディット問題と若干の問題でしかされておらず、一般的な有効性については明らかではない。そこで本研究は、ゼロ和ゲームであるマッチングペニーゲームと非ゼロ和ゲームである囚人のジレンマ問題、一般的なボードゲームであるリバーシについて勝率などを検証し、LS モデルの一般的な有効性を確かめることを目的とする。

2. LS モデル

ヒトは因果推論において $p \to q$ から $q \to p$ を導く対称性バイアスと、$p \to q$ から $\bar{p} \to \bar{q}$ を導く相互排他性バイアスを持つ。LS モデルはこの二つのバイアスを組み込んだ因果関係を帰納的に推論するモデルである。このモデルは非定常の問題であっても正解率が下がりにくく、少ない計算量で高い正解率を持つことが知られている　[篠原 07]　。

2.1 LS モデルの定義

選択肢 A,B が存在し、その結果として C, D が存在するとき、その組み合わせは 2×2 の分割表で表すこと出来る (表 1)。

表 1: 2×2 分割表

	結果	
	C	D
選択肢 A	a	b
選択肢 B	c	d

このとき $LS(C|A)$ と $LS(C|B)$ は式 (1)、式 (2) で表される。

$$LS(C|A) = \frac{a + \frac{b}{b+d}d}{a + \frac{b}{b+d}d + b + \frac{a}{a+c}c} \tag{1}$$

$$LS(C|B) = \frac{c + \frac{b}{b+d}d}{c + \frac{b}{b+d}d + d + \frac{a}{a+c}c} \tag{2}$$

連絡先: 東京電機大学所属　西村　友伸
e-mail:tomonobu.nishimura[at]gmail.com

3. マッチングペニーゲーム

ゼロ和ゲームにおける LS モデルの有用性を、選択できる行動が 2 種類で結果が 2 種類のマッチングペニーゲームを使い調べる。

3.1 マッチングペニーゲームとは

マッチングペニーゲームとはペニー硬貨を使った単純なゲームで、ゲームのプレイヤーは 2 人、各プレイヤの行動の選択肢は、表を出すか裏を出すかの二択である。ペニー硬貨とは、英国などで使われるポンドの補助通貨のことである。一回限りの勝負では強い弱いは存在しないが、今回の実験では繰り返し対戦をする為勝率などを求められる。また、戦略として相手に行動を読まれない、相手の癖を見抜く等が意思決定において重要となる。勝敗は片方のプレイヤーは相手と同じ手を出すと勝ち、もう片方のプレイヤーは相手違う手を出すと勝ちとする。本実験では表 2 にまとめたように、勝った方に 1 点、負けた方に-1 点を与えることとする。

表 2: マッチングペニーゲームの得点表

		プレイヤー B	
		H'(表)	T'(裏)
プレイヤー A	H(表)	a(1,-1)	b(-1,1)
	T(裏)	c(-1,1)	d(1,-1)

3.2 プレイヤーモデルの紹介

今回のシミュレーションで用いる 2 種類のモデルを紹介する。

・$AgentLS$

LS モデルを用いたプレイヤーモデルを $AgentLS$ と呼び、式 (1) で表を出す価値を、式 (2) で裏を出す価値を求める。このプレイヤーモデルの勝利条件は相手と同じ手 $(H, H')(T, T')$ を出す事とする。

・$AgentLC$

$AgentLS$ と対戦する線形結合のプレイヤーモデル。以下の式に従い行動を選択する。

$$LC(T|H') = \frac{\alpha\ c + \delta\ b}{\alpha\ b + \beta\ d + \gamma\ a + \delta\ c}. \quad (3)$$

$$LC(H|T') = \frac{\alpha\ b + \delta\ c}{\alpha\ b + \beta\ d + \gamma\ a + \delta\ c}. \quad (4)$$

ただし、*AgentLC* は $\alpha, \beta, \gamma, \delta$ の 4 つの変数を持ち、そのうちの 2 つを 1 に固定し、残りの変数を [0,1] の範囲で 0.1 ずつ変化させる。つまり、716 通りの *AgentLC* が出来る。このプレイヤーモデルの勝利条件は相手と違う手 $(R, H')(H, R')$ を出す事とする。ここで、各プレイヤーモデルの式中の a,b,c,d を表 3 にまとめた各事象の共起回数とする。また、H,R,H',R' はそれぞれのプレイヤーモデルが出す手（硬貨の表、裏）とする。

表 3: 各事象と共起確率の表

	H'(表)	T'(裏)
H(表)	a	b
T(裏)	c	d

a:互いに表を出した回数
b:*AgentLS* が表を出し、*AgentLC* が裏を出した回数
c:*AgentLS* が裏を出し、*AgentLC* が表を出した回数
d:互いに裏を出した回数

3.3 シミュレーションの設定

- 各プレイヤーモデルはお互いに相手の戦略を知らず、同時に硬貨の裏表を選択する。
- 1 回の対戦を 1 ステップとする。
- 5, 10, 30, 100, 1000 ステップをそれぞれ 1 試合とし、10 万試合を行う。
- 1 試合ごとに試合に関する情報を初期化し、次の試合への影響はないものとする。
- 10 万回の試合に対する、*AgrntLS* の勝率や勝率の割合を計測する。

3.4 対戦結果

各ステップごとに、*AgentLS* が *AgentLC* に対して、5〜8 割勝ち越すこした割合を表 4 に示す。

表 4: *AgentLS* が勝った *AgentLC* の割合

	5 割	6 割	7 割	8 割
5*step*	61.45%	57.68%	53.49%	39.39%
10*step*	69.41%	56.84%	52.93%	52.37%
30*step*	70.81%	66.34%	64.94%	62.43%
50*step*	70.39%	64.80%	63.97%	63.13%
100*step*	74.58%	65.64%	65.36%	63.53%
1000*step*	74.86%	66.20%	65.22%	64.80%

4. 繰り返し囚人のジレンマ

非ゼロ和ゲームにおける *LS* モデルの有用性を、選択できる行動が 2 種類で結果が 2 種類の囚人のジレンマを使い調べる。

4.1 囚人のジレンマとは

二人の囚人が共犯の容疑で逮捕され、別々の部屋で取り調べを受けそれぞれ自白するか黙秘するかを迫られた。一方だけが自白をすればその囚人は釈放され、黙秘した方は重罪を課せられる。共に自白したなら双方に中程度の刑を、共に黙秘したなら双方に軽い刑を課せることになった。この時、自分は相手が何を選ぼうが自白をすれば重い刑を免れるが、互いに自白してしまうと中程度の刑を科せられることになってしまう。しかし、互いに黙秘をしていれば軽い罪で済む。このように、自分の得になる行動をとったにもかかわらず不利益を被ってしまい、合理的な選択を行うことができない。このような状況のことを囚人のジレンマと呼ぶ。

4.2 ゲームのルール

二人のプレイヤーはそれぞれ、協調か裏切りかを同時に選択し、その結果でそれぞれが得点を得る。互いに協調ならば共に 3 点、一方が裏切りもう一方が協調ならば裏切った方に 5 点協調した方に 0 点、互いに裏切ったなら共に 1 点とする。

4.3 シミュレーションの設定

- 各モデルはお互いに相手の戦略を知らず、同時に協調か裏切りを選ぶ。
- 10 万回勝負を行い得点を記録する。
- どのモデルも勝負回数は得られないものとする。
- 自身を含む 14 種類のモデルによる総当り戦を行う。

4.4 各プレイヤーモデルの紹介

LS モデルは表 2 の表と裏を協調と裏切りに変えたものをもとに協調か裏切りかの価値を求める。今回使用する *LS* モデル以外の 13 種類のモデルとその行動を表 5 に示す。

表 5: 各モデルの紹介

モデル名	行動
allC	全て協調。
allD	全て裏切り。
PerCD	*CDCD* と繰り返す。
PerDC	*DCDC* と繰り返す。
PerCDD	*CDDCDD* と繰り返す。
PerDCC	*DCCDCC* と繰り返す。
TFT	初回は協調、次回からは前回の相手の真似をする。しっぺ返し。
TFTT	初回は協調、相手が過去に 2 回裏切った場合裏切り返す。
Joss	初回は協力、相手が裏切ったら裏切り返し、相手が協調ならば 10%の確率で裏切る。
Davis	最初の 5 回は協調、その時に裏切られなければずっと協力。
Friedman	初回は協調、一度でも裏切られたら裏切り続ける。
*Friedman*80	初回は協調、一度でも裏切られたら 80%の確率で裏切る。
Random	50%の確率で協調し、50%の確率で裏切る。

4.5 対戦結果

それぞれのモデルの総当り戦の結果を表 6 に示す。結果として、*LS* モデルは 14 位中 4 位の成績だった。

5. リバーシ

前節では、ゼロ和、非ゼロ和ゲームにおいて、LS モデルの有効性を論じて来た。以降の章ではリバーシ (商標オセロ) に N 本腕バンディット問題を発展させたモンテカルロ計画法を実装 [Sylvain 06, 大用 09] し、*LS* の一般的な有効性を検証する。リバーシとは二人用の対戦ボードゲームである。交互に白と黒の石をボードに置き、相手の石を上下左右または斜め何れかの方向で自分の石を使い挟む事で相手の石を自分の石とする事が出来る。最終的に自分の色の石が多い場合勝利となるゲームである。

表 6: 総当たり戦の結果

	LS	$allC$	$allD$	CD
平均得点	255624.29	222860.36	205646.57	219637.64
順位	4	7	10	8
	DC、	CDD	DCC	TFT
平均得点	209009.00	200720.79	200677.64	249414.14
順位	9	12	13	5
	$TFTT$	$Joss$	$Davis$	FR
平均得点	225446.64	202939.71	283555.14	271464.36
順位	6	11	1	2
	$FR80$	$random$		
平均得点	262821.79	199952.21		
順位	3	14		

5.1 N 本腕バンディット問題

N 本腕バンディット問題とは N 個のスロットマシンの中から最も報酬が得られるマシンを探索し、総獲得報酬の最大化を目的とした機械学習の問題である。バンディット問題では各マシンの当たり確率は知る事が出来ず、実際にマシンに対して試行を行い確かめる必要がある。これは最良のマシンをプレイし最大の報酬を得ることと、最良のマシンを探索する行動とをうまく配分しなければならないことを示している。また最良のマシンを探すためには、各マシンを平均的にプレイしなければならないがそれは平均的なマシンでプレイしていることと変わりがない。この、既知の最大の報酬をもたらすマシンを選択することと、より良い可能性のある他のマシンを探索することの両立の難しさを「報酬と探索のジレンマ」と呼ぶ。

5.1.1 $UCB1$ アルゴリズム

バンディット問題の腕選アルゴリズムやモデルは数多く存在するが、囲碁 AI に実装されたことで有名になった $UCB1$ が知られている [Auer 02]。このアルゴリズムは、損失の上限が保証されいるため多くの探索が許されていれば正解の選択肢を選択する事が出来る。$UCB1(x_i)$ は式 (5) で定義され、x_i は i 番目のマシン、$\bar{X}_i$ は i 番目のマシンの勝率、n は全体のプレイ数、n_i は i 番目のマシンのプレイ数をそれぞれ表している。

$$UCB1(x_i) = \bar{X}_i + \sqrt{\frac{2\log n}{n_i}} \tag{5}$$

5.1.2 LS モデルの N 本腕バンディット問題への適用

LS モデルは 2×2 分割表で定義されている。よって 2 本腕バンディット問題までしか扱えず着手可能手が 3 手以上ある盤面では最良の手を導くことが出来ない。本研究ではこの問題に対応するため着手可能手を 2 つずつの組に分け、その組を 2 本腕バンディット問題と見なすことで対応する。この組をトーナメント形式にし、最後に残った一つの手をその盤面での最良の手と考える。また着手可能手が奇数個であった場合には、余った一つの手をシードとする。

5.2 モンテカルロ計画法

ランダムなサンプリングを用いてシュミレーションをおこない、近似解を得る手法としてモンテカルロ法が知られている。原始モンテカルロ法は着手可能手に対して、決められた回数分だけプレイアウトを行い着手を決定する方法である。プレイアウトとは、ある局面からランダムな選択をしながら終盤まで選択を行うことをいう。モンテカルロ法では評価関数の変わりにプレイアウトの結果 (勝ち、負け) を用いる。着手可能手をバンディット問題の腕として扱いサンプリングを工夫するアルゴリズムはモンテカルロ計画法と呼ばれている。

5.2.1 局面のゲーム木化

現在の局面を一つのノードとして考えたとき、次の局面は現在の局面の子ノードとして考えることができ、現在の局面と先の局面の関係は木構造で表される。これは一般にゲーム木と呼ばれている。本研究ではモンテカルロ計画法とゲーム木を組み合わせたモンテカルロ木探索を行う。木の成長はサンプリング数が閾値に達したならば、$UCB1$ または LS が最良と判断するノードからゲーム木を成長させた局面を次のノードとする。 このゲーム木の作成法では自分の手番においては最良の結果が期待でき、相手の手番では相手が最良となる結果が選択される。このようなゲーム木の作り方では、AI は常に両者において最良の結果を想定しながらプレイを行うことができる。一般にこの考え方をミニマックスアルゴリズムと呼ぶ。図 1 はゲーム木の構成を表す。この場合現在の局面が根であり、二手先の局面の 2 つと一手先の局面の内、中央と右のノードが葉ノードである。図 2 はモンテカルロ計画法によりプレイアウトが行われる葉ノードの選択方法を表す。前述の通り、全ての葉ノードの集まりを 2 つずつの組に分けトーナメント式に最良の葉ノードを選ぶ。図 2 では 2 つ目の葉ノードが最良のノードとして選ばれる。

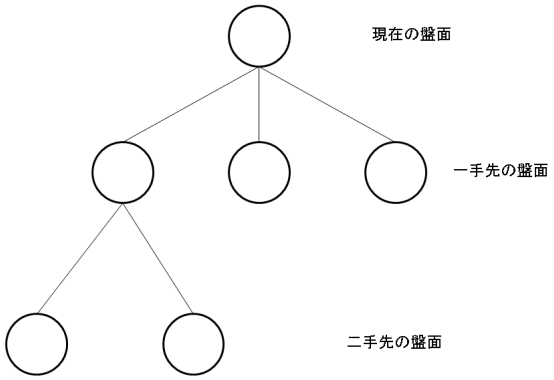

図 1: ゲーム木

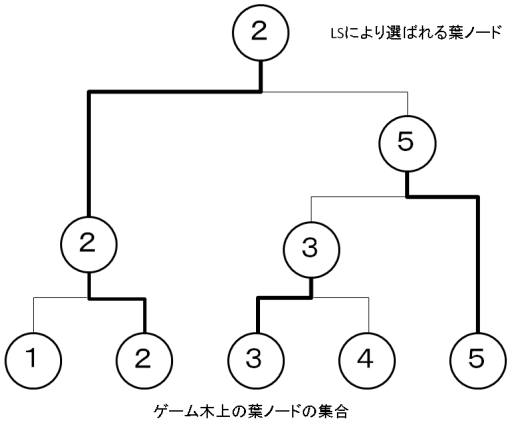

図 2: LS による葉ノードのトーナメント

5.3 シミュレーション

本シミュレーションではゲーム木が成長する閾値は 100 とする。LS モデルと $UCB1$ を実装した AI を先攻、後攻を入れ替えそれぞれ合計 1 万回対局させる。このときプレイアウト回数は {2, 3, 5, 10, 15, 20, 30, 50, 100, 150, 300, 500, 1000} とする。

$UCB1$ に対して LS モデルの勝率を図 3 に示す。プレイアウトの回数を x 軸、勝率を y 軸で表す。

5.4 実験結果

実験結果を図 3 に示す。プレイアウトの回数を x 軸、勝率を y 軸で表す。

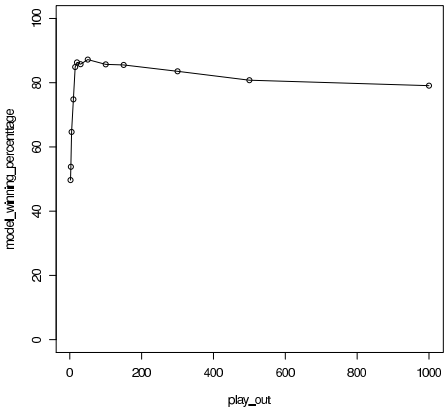

図 3: 対戦結果

6. 考察

いずれの実験においても、LS モデルは、少ない計算量で高い成績を示し、今回の三つの問題構造にたいして柔軟に対応することができた。それぞれの実験結果について以下に記す。

6.1 マッチングペニーゲーム

このゲームでは、相手の戦略を素早く見極める事が重要であるが、表 4 に上げた勝率を見ると、5 ステップという短い期間の戦いにおいて、716 通りの $AgentLC$ のうち、440 通りの $AgentLC$ に勝つことができた。コレは LS のもつ少ない計算量で高い正答率を得る特徴がうかがえる。また、8 割以上の勝率を収める割合を見ると、5 ステップでは 40%程にとどまるがステップ数が増えるごとに増えていることが分かる。これは、勝負の回数が増えればより正確に相手の戦略を見極めることが出来るようになることを示している。

6.2 囚人のジレンマ

総当たり戦の結果 14 モデル中 4 位という平均得点を獲得し、全て協調や全て裏切りといった極端な相手 ($allC$ や $allD$ など) に対する戦略を観察したところ、ひとつの手に注目するだけではなく常に他の選択肢についても考慮していることがわかった。また、手をたまに変えるという変則的な戦略をとっている相手（$Joss$ や CDD、DCC など）にも対応することができた。このことから LS は機械的な硬直性を超えた柔軟な思考を行えるモデルであると考えられる。4 位という結果になってしまったのは、総当り戦の得点の平均から順位付けをしたため、上位のモデルにとって有利な相手が多かったためと考えられる。

6.3 リバーシ

実験の結果、サンプリング数が少ない段階で高い勝率を出し、サンプリング数が 15 回を超えた段階で 80% を上回った。その後サンプリング数が 300 までは勝率は約 85% が続き, サンプリング数が 500 回、1000 回と増えると勝率は約 80% と若干低下した。この結果は、木の成長のない少ないサンプリング数では先行研究　[大用 09]　と一致し、木の成長が発生するサンプリング数が 100 回以上の場合先行研究を若干上回っている。特にサンプリング数が１０００回の時、先行研究では勝率が 75% 程度であったが、木構造を導入することにより約 5% の上昇がみられた。これは木の成長が発生することにより新たな探索が発生し、それが LS の持つ少ない計算量で良い結果を出す性質に有効に作用した結果であると考えられる。

参考文献

[Auer 02] Auer, P, Cesa-Bianchi, N., Fischer, P., Finite-time analysis of the multi-armed bandit problem, $MachineLearning$, 47, 235-256 (2002)

[Sylvain 06] Sylvain Gelly, Yizao Wang, Remi Munos., , Teytaud, Modification of UCT with Pattens in Monte-Carlo Go (2006)

[大用 09] 大用庫智, ヒト認知バイアスのモンテカルロ法への応用, 2009 年度　情報科学科卒業論文 (2009)

[篠原 07] 篠原修二，田口 亮，桂田浩一，新田恒雄：因果性に基づく信念形成モデルと N 本腕バンディット問題への適用 ，人工知能学会論文誌，Vol.22，No.1，pp.58-68 (2007).

[星 04] 星正明,. リバーシのアルゴリズム　C++& Java 対応, 工学社.

[丸山 11] 丸山涼平, マッチングペニーゲーム、囚人のジレンマに対するゆるい対称性モデルの有効性, 2010 年度　情報システムデザイン学系卒業論文 (2011).