Vol.2015-GI-33 No.18
2015/3/6

自然な人狼の勝率

西野 順二[1,a)]

概要：多人数不完全情報ゲームの人狼においてその自然な勝率を算出した．人狼は, 代表的な会話にもとづくゲームである．このため, 推論機能を持つ人工知能エージェントによる, 会話と信頼創出のテストベッドとして研究されている．村人の人数に対して人狼側が有利であり, ゲームが拮抗する人数は全体の平方根に従うことが明らかになっている．本研究はこれに対し, 人狼がより自然な戦略を取った場合に人狼側の有利さがさらに高いことを示した．勝率を算出し示すことで, 人工エージェントの能力が有意に高いことを測る具体的な基準値を示した．

1. はじめに

人狼は, ロールプレイングとプレイヤの会話による情報交換をゲームの主体とした多人数による不完全情報ゲーム [1] である．ゲームとしては村人グループと狼グループの 2 プレイヤの対戦であり, 村人からは誰が人狼かが分からず, 人狼グループはそれを知っているという情報の非対称を持つシグナリングゲームの一種である．

ここで行なわれる会話による情報交換を, 高度に知的な情報操作と見立て, 人工エージェントによってプレイを実現することで知能研究を推進する人狼知能プロジェクト [2], [3], [4] も行なわれている．

本発表は, このような人工エージェントの優劣を比較するための勝敗の基準値を具体的に示すことを目的とする．

人狼では, 従来研究により狼グループの有利が知られている [5], [6]．シンプルな仮説と狼と村人にとっての最適な戦略の下においてプレイヤ総人数 R に対して, たかだか $\sqrt{R}$ に比例する狼の数で匹敵する．このモデルによると, たとえば日常的に行なわれる狼 3 人のゲームでは, 総人数が 13 人で狼側の勝率は 0.79 である．この設定で人工エージェントが狼としてプレイしたとき, 統計的には有意水準 0.05 で有意に狼が強いと言えるのは, 100 戦で 87 勝以上が必要であり, 10 戦では 10 勝でも有意とは言えない．ここでの人狼側の最適行動の条件はやや強いものである．既存研究は人狼ゲームの理論的モデル化を行なって, R=1000 以上など実用上は現実と乖離している部分でのモデル化が論じられている．

本研究では狼の最適戦略を自然に見直すことで, 狼の勝率が従来研究より大きいことを示す．さらに狼が 3 人程度, プレイヤ総人数が 30 人程度までの範囲で, その勝率を示す．

[1] 電気通信大学 情報理工学研究科
The University of Electro-Communications
[a)] nishinojunji@uec.ac.jp

2. 人狼

本研究で扱う人狼ゲームのルールについて述べる．人狼は狼と村人に分かれ, 会話による情報交換を主体として, 標準的にはプレイヤ人数は 10 名程度により, グループ相互に戦うテーブルトークゲームである．元来はヨーロッパを発祥とした伝承ゲームであり, 商用化されるにあたって, 市民にまぎれたマフィアと一般市民の戦いを模して Mafia という名前でルール整備がなされた．マフィアを人狼, 市民を村人としているのが人狼 (werewolf game) である．

このようにストーリー設定も含め, 種々のルールバリエーションが存在する．なかには投票排除後の属性開示をするしないといったゲームの本質を変えるルールの違いもある．

2.1 日本での人狼

日本ではイタリアのゲームメーカーダヴィンチによる「タブラの狼」とその派生系が広く行なわれ, 人狼知能やネットゲームでの人狼の基本形となっている．投票による退場時に属性を明かさない, というゲームの性質上大きな特徴がある．既存研究で扱われた Mafia では, 退場時に属性を明かすため, その属性とリンクした範囲でゲームプレイヤ全体に公表された情報が増加する．しかし, タブラの狼ではこれらの情報は不完全なままとなる．

従来研究 [5] では, 投票によって退場するときにその属性を明らかにするルールでモデルを構築している．

2.2 人狼のルール

人狼のルールと進行をタブラの狼に即して述べる．

- プレイヤには互いに分からないように属性として, 人狼と一般の村人が割り振られる.
- 全員が顔を伏せ, 人狼だけお互いに顔を上げて互いに認識する. 村人は誰が人狼で誰が村人かは分からないままである.
- ゲームは 1 ターンごとに夜と昼を交互に繰り返し, 夜は人狼が秘密に選んだプレイヤが一人ゲームから外され, 昼は全員の投票によって選んだプレイヤが一人ゲームから外される. いったん外れたプレイヤはそのゲームには参加できず, 発言やジェスチャーによる情報伝達などは禁止される.
- 外されたプレイヤが人狼であったか村人であったかは開示されない.
- 人狼と一般の村人で最終的に外されずに残ったプレイヤのチームの勝利となる. 実際には人狼が村人と同数以上になれば人狼の勝利となる.
- また, 村人や人狼の中に特殊な役職を割り振り, これも本人以外の全員に秘密である. 様々なルールバリエーションがあり, 基本となる役職は占い師 (Mafia では Detectives) である. 占い師は毎夜に 1 ターンに 1 名の属性を知る事ができる.

タブラの狼では, ゲームから外れる時を含め, 役職や属性を明らかにすることがない. このため, 占い師が自己申告したとしてもその真偽を確かめる論理的な方法は存在しない. 通常のゲームでは, 雰囲気の怪しさや, 会話の矛盾などからこれを心理的に推測する.

3. 自然な人狼の勝率モデル

現在の人狼の数を w , 村の人数を v とする. 投票によって人狼が外される確率を $f(w,v)$ で表すとする.

ゲームターンの進展により, 昼の投票によって 1 人, 夜の人狼によって村人が 1 人ずつ単調に減っていく. これを人数の状態遷移としてとらえ人狼が勝利する確率 $ww(w,v)$ を漸化式で表すと式 (1) となる.

$$ww(w,v) = f(w,v) \times ww(w-1,v-1) + (1-f(w,v)) \times ww(w,v-2) \quad (1)$$

人狼の勝率は, 昼間の投票によって人狼が外される確率に依存する.

以下では, いずれも占い師の居ないゲームを対象とする.

3.1 既存モデル

Braverman らの結果 [5] では, 昼間の投票行動において人狼を含む全員がランダムに投票することが最良戦略であるとしている. 占い師がいないとき, 人狼と村人の勝敗が拮抗するとき, 人狼の人数が総人数の平方根に比例する.

これは,

(1) 村人は人狼を判別する確かな情報を持っていない. 会話の不自然さなど心理的な判断は, 情報量を増やさない.

(2) 人狼は村人と異なる行動をすると検出されるため村人と同じ行動をする.

という前提にもとづいている.

人狼の行動のうち村人に可観測なのは投票行動だけであり, 不審な投票を行なうことは人狼であることが検出されることにつながるので村人と人狼は同一の投票行動を取る, という信念にもとづいている. Braverman らは, さらにこのランダム投票を可能にするための, 多人数意思決定アルゴリズムも同時に提案している.

誰に投票するかの基準はランダムであるなら, その結果としてあるプレイヤに投票が集まる可能性も均一となる.

そこで v 人村の中で w 人の人狼が投票によって排除される確率は式 (2) に示すように単純な人数比となる.

$$f(w,v) = \frac{w}{v} \quad (2)$$

3.2 投票モデル

ランダムな投票によって外すプレイヤを決定するとき, 仲間を互いに知る人狼が, 仲間に投票しないことは自然である. 本研究では, この自然な行動モデルについて検討する.

まず, 仲間に投票をしないことにより人狼であると検出される危険性を考える.

ランダムに投票したとき, 村全体が v 人の状態から n ターン (昼夜) の間に, ある特定の m 人が互いに投票しないことが観測される確率 p は式 (3) となる.

$$p = \prod_{i=0}^{n-1} \frac{v-m-2i}{v-2i} \quad (3)$$

例えば,13 人中人狼が 2 人いて 3 昼夜の間互いに投票しないことは, ランダムに投票したとしても 0.53 の確率で発生する. さらにここで, 特定の二人の組は 76 組あることから, 数十組が同時にこの条件に当てはまり区別できず, 実質的な検出は不可能とみなせる. 全員がランダムに投票をする, という合意の下で, 人狼同士が互いに投票をしなくても, 検出されるおそれは無いと言える.

このため, 人狼は次の戦略が自然である.

- ランダムな投票において, 人狼同士は投票しない.

この戦略の元で, 人狼が外される確率 $f(w,v)$ を考える.

このとき村人 $A, B, \cdots$ と人狼 $W, X, \cdots$ の投票行動確率 P は行列 (4) で表す事ができる.

$$P = \begin{pmatrix} 0 & p_{AB} & \cdots & p_{AW} & \cdots & p_{AZ} \\ p_{BA} & 0 & \cdots & p_{BW} & \cdots & p_{BZ} \\ & & \cdots & & & \\ p_{WA} & p_{WB} & \cdots & 0 & 0\cdots & 0 \\ & & \cdots & & & \\ p_{ZA} & p_{ZB} & \cdots & 0 & 0\cdots & 0 \end{pmatrix} \quad (4)$$

対角が 0 となるのは自分には投票しないことを表し, 右下部は人狼同士は互いに投票をしないことを表す.

村人 A は全員に均一に投票するため, プレイヤ i に投票する確率 p_{Ai} は

$$p_{Ai} = \frac{1}{v-1} \tag{5}$$

また, 他の村人も同じ確率分布で行動するものとする. 人狼は人狼を除いて均一に投票するため, プレイヤ i に投票する確率 p_{Wi} は

$$p_{Wi} = \frac{1}{v-w} \tag{6}$$

となる. なお, 人狼が一人 ($w = 1$) のときは人狼がグールプとならず式 (5) と式 (6) は一致する.

このとき, 村人 A が過半数以上を取る確率 p_v は村人 A の得票数を g としてそれが過半数以上であるすべての場合の確率の総和であり, 式 (7) で与えられる.

$$p_v = \sum_{g=\lfloor (v+1)/2 \rfloor}^{v} \{ \sum_{h=ws}^{w} {}_{v-w}C_{g-h} \ {}_wC_h \ p_{BA}^{(g-h)} \ p_{WA}^{h} \} \tag{7}$$

ここで $ws = min(g-(v-w), 0)$ である.

同様に人狼が過半数を取る確率を p_w とすると, 式 (8) となる.

$$p_w = \sum_{g=\lfloor (v+1)/2 \rfloor}^{v} {}_{v-w}C_g p_{BA}^{g} \tag{8}$$

人狼がこのため, 人狼の誰かが投票で外される確率は一人あたり確率と人数から, 式 (9) となる.

$$f(w, v) = \frac{wp_w}{wp_w + (v-w)p_v} \tag{9}$$

これらを元に, 村の人数 30 人まで, 人狼 3 人までのゲームでの, 人狼チームの勝率, 人狼同士が互いに投票していない事象と区別の付かない確率を計算し付表に示した.

4. おわりに

村人にとってランダムな投票は適切な行動であるが, このとき人狼が共謀して互いに投票しなかったとしても, それを検出することは事実上むずかしいことを示した. このことから, 人狼は互いに投票しない行動を取るとすると, 従来のモデルより, さらに大幅に勝率が上がることが分かった.

とくに人狼が 2 人居る場合には 30 人のゲームでも 0.99 の勝率となった. これは, 占い師の行動や心理的な推論を考慮していないため人狼に有利な値となっていると考えられる.

もしも, 村人が互いに認識できれば, 今回の人狼相互が投票しないのと同様な行動によって村人に有利な投票ができることも示している.

占い師の効果は, 人狼の発見にまして, このような確実な村人グループを構成することにある. Braverman らの提案する占い師の最適戦略でも, できるかぎりホワイトリストを作って残す, ことが示されている.

占い師についても不確定ながら一定の信頼度によって投票行動に偏りのあるモデルとして取り入れる必要がある.

参考文献

[1] Chittaranjan, G. and Hung, H.: Are you awerewolf? detecting deceptive roles and outcomes in a conversational role-playing game, *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, IEEE, pp. 5334–5337 (2010).
[2] 篠田孝祐，鳥海不二夫，片上大輔，大澤博隆，稲葉通将：汎用人工知能の標準問題としての人狼ゲーム，人工知能学会全国大会論文集， Vol. 28, pp. 1–3 (2014).
[3] 稲葉通将，鳥海不二夫，大澤博隆，片上大輔，篠田孝祐，西野順二：同調と反駁に着目した人狼ゲームの分析，人工知能学会全国大会論文集， Vol. 28, pp. 1–4 (2014).
[4] Katagami, D., Takaku, S., Inaba, M., Osawa, H., Shinoda, K., Nishino, J. and Toriumi, F.: Investigation of the effects of nonverbal information on werewolf, *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*, IEEE, pp. 982–987 (2014).
[5] Braverman, M., Etesami, O., Mossel, E. et al.: Mafia: A theoretical study of players and coalitions in a partial information environment, *The Annals of Applied Probability*, Vol. 18, No. 3, pp. 825–846 (2008).
[6] Migdał, P.: A mathematical model of the Mafia game, *arXiv preprint arXiv:1009.1031* (2010).

付表:狼の勝率等一覧

4.1 人狼勝率表

表 1 人狼勝率表 1

人狼 (w)	村人数 (v)	従来の人狼勝率	提案モデルの勝率
1	3	0.667	
1	4	0.750	
1	5	0.533	
1	6	0.625	
1	7	0.457	
1	8	0.547	
1	9	0.406	
1	10	0.492	
1	11	0.369	
1	12	0.451	
1	13	0.341	
1	14	0.419	
1	15	0.318	
1	16	0.393	同左
1	17	0.300	
1	18	0.371	
1	19	0.284	
1	20	0.352	
1	21	0.270	
1	22	0.336	
1	23	0.259	
1	24	0.322	
1	25	0.248	
1	26	0.310	
1	27	0.239	
1	28	0.299	
1	29	0.231	
1	30	0.289	

表 2 人狼勝率表 2

人狼 (w)	村人数 (v)	従来の人狼勝率	提案モデルの勝率
2	5	0.867	0.995
2	6	0.917	0.998
2	7	0.771	0.987
2	8	0.844	0.995
2	9	0.702	0.978
2	10	0.784	0.991
2	11	0.648	0.969
2	12	0.736	0.985
2	13	0.605	0.960
2	14	0.695	0.980
2	15	0.570	0.952
2	16	0.661	0.975
2	17	0.540	0.944
2	18	0.631	0.970
2	19	0.515	0.937
2	20	0.605	0.964
2	21	0.493	0.930
2	22	0.582	0.959
2	23	0.474	0.923
2	24	0.561	0.955
2	25	0.456	0.917
2	26	0.543	0.950
2	27	0.441	0.911
2	28	0.526	0.945
2	29	0.427	0.905
2	30	0.511	0.941

表 3 人狼の勝率表 3

人狼 (w)	村人数 (v)	従来の人狼勝率	提案モデルの勝率
3	7	0.943	1.000
3	8	0.969	1.000
3	9	0.886	1.000
3	10	0.931	1.000
3	11	0.835	0.999
3	12	0.895	1.000
3	13	0.792	0.999
3	14	0.860	1.000
3	15	0.755	0.998
3	16	0.829	0.999
3	17	0.722	0.998
3	18	0.801	0.999
3	19	0.693	0.997
3	20	0.776	0.999
3	21	0.668	0.996
3	22	0.752	0.999
3	23	0.645	0.996
3	24	0.731	0.998
3	25	0.625	0.995
3	26	0.712	0.998
3	27	0.606	0.994
3	28	0.693	0.997
3	29	0.589	0.993
3	30	0.677	0.997

4.2 あるグループ同士が投票しない事象の自然発生確率

3 ターン続けてある n 人のグループが互いに投票しないことが自然に起こる確率.

表 4 同時非投票発生率 1

グループ人数	村人数	自然発生率
2	6	0.000
2	7	0.143
2	8	0.250
2	9	0.333
2	10	0.400
2	11	0.455
2	12	0.500
2	13	0.538
2	14	0.571
2	15	0.600
2	16	0.625
2	17	0.647
2	18	0.667
2	19	0.684
2	20	0.700
2	21	0.714
2	22	0.727
2	23	0.739
2	24	0.750
2	25	0.760
2	26	0.769
2	27	0.778
2	28	0.786
2	29	0.793
2	30	0.800

表 5 同時非投票発生率 2

グループ人数	村人数	自然発生率
3	7	0.000
3	8	0.078
3	9	0.152
3	10	0.219
3	11	0.277
3	12	0.328
3	13	0.373
3	14	0.413
3	15	0.448
3	16	0.479
3	17	0.507
3	18	0.532
3	19	0.555
3	20	0.576
3	21	0.594
3	22	0.612
3	23	0.628
3	24	0.642
3	25	0.656
3	26	0.668
3	27	0.680
3	28	0.691
3	29	0.701
3	30	0.711

表 6 同時非投票発生率 3

グループ人数	村人数	自然発生率
4	8	0.000
4	9	0.048
4	10	0.100
4	11	0.152
4	12	0.200
4	13	0.245
4	14	0.286
4	15	0.323
4	16	0.357
4	17	0.388
4	18	0.417
4	19	0.443
4	20	0.467
4	21	0.489
4	22	0.509
4	23	0.528
4	24	0.545
4	25	0.562
4	26	0.577
4	27	0.591
4	28	0.604
4	29	0.617
4	30	0.629