ILPを用いたワンナイト人狼における特徴的行動の抽出

Extraction of Characteristic Behaviors in One Night Werewolf by Inductive Logic Programming

西崎 絵麻[1]　尾崎 知伸[2]*
Ema Nishizaki[1]　Tomonobu Ozaki[2]

[1] 日本大学大学院 総合基礎科学研究科
[1] Graduate School of Integrated Basic Sciences, Nihon University
[2] 日本大学 文理学部
[2] College of Humanities of Sciences, Nihon University

Abstract: A multiplayer party game named Werewolf game is now widely recognized as a new standard problem for artificial intelligence. In this paper, in order to obtain fundamental knowledge useful for realizing intelligent agents for werewolf games, we conduct a log analysis of a simplified version of werewolf game (One night werewolf). Given background knowledge on utterances and behaviors such as coming out of the role and questions to other players, a logic-based analysis, *i.e.*, Inductive Logic Programming, is applied to the log data of 100 games by 10 subjects to discovery characteristic behaviors by werewolves who falsify seers, those by werewolves in winning games, and those by trustworthy seers. In addition, a role prediction model is also built in the same framework. Through the analysis, we successfully discovered some fundamental but useful behaviors and rules in one night werewolf.

1　はじめに

近年，人狼知能における新たな標準問題として，不完全情報ゲームの一つである人狼を対象とした人狼知能 [1, 2, 3] に関する研究が盛んに行われている．人狼ゲームには，対面式で行うゲームの他に，音声のみで行うゲームやテキスト情報のみで行うゲームなど様々な形式が存在するが，本研究では参加人数を限定した簡易版人狼であるワンナイト人狼を対象とする．詳細なルールは後述するが，ワンナイト人狼では短時間での議論の後，1 回の投票行動で勝敗が決する．このため，各プレイヤには失言を回避しながら，短時間で他者の発言の真否を判断することが要求されることとなり，結果として，無意味で冗長な行動を避け，勝利に直結する人狼ゲームに特有な行為（相手を騙す行為や嘘を見破る行為，信頼を得る行為など）を高い密度で行うと期待できる．

本論文では，ワンナイト人狼における特徴的な行動を獲得するため，初心者 10 名によるワンナイト人狼のプレイログ（100 ゲーム）[4, 5] を対象とした分析を行う．具体的には，論理に基づく分析手法である帰納論理プログラミング (Inductive Logic Programming;ILP)[6] を用い，占い師を騙る人狼や信頼される占い師の特徴的にみられる行動を抽出する．また，役職を推定するためのルール導出も行う．

本論文の構成は以下の通りである．2 章では，準備としてワンナイト人狼のルール説明と，ILP の基本的な枠組みを示す．3 章でデータセットと背景知識について説明した後，4 章で分析結果とそれに対する考察を示す．最後に 5 章でまとめを行い今後の課題を述べる．

*連絡先：日本大学文理学部情報科学科
〒 156-8550 東京都世田谷区桜上水 3-25-40
E-mail: tozaki@chs.nihon-u.ac.jp

2　準備

2.1　ワンナイト人狼のルール

各プレイヤは，ゲーム開始時に自身にのみ知らされる役職に基づき，人間陣営（村人，占い師，怪盗）か人狼陣営（人狼）のどちらかに所属する．その後，人狼プレイヤに対して他プレイヤの所属陣営が知らされ，次いで占い師と怪盗がそれぞれの能力を行使する．占い師には，選択したプレイヤ 1 名が人狼であるか否かが知らされる．また怪盗は，選択したプレイヤ 1 名と役職を交換することができる．なお役職交換を行う場合でも，交換先のプレイヤが持つ情報を得ることはで

きず，また交換されたプレイヤにはその情報が通知されないままゲームが進行する．能力行使後，プレイヤ全員による議論が開始される．議論の目的は人狼プレイヤの特定であり，人狼プレイヤはそれを邪魔するために嘘をつくなどの行動を行う．議論終了後，各プレイヤが最大 1 回，他プレイヤを選択し質問を行う．議論と質問結果に基づき，全プレイヤが人狼と疑わしきプレイヤへ投票を行い，最多票のプレイヤが処刑される．結果として，処刑されたプレイヤの所属陣営が敗北，非所属陣営が勝利となる．

2.2 帰納論理プログラミング

帰納論理プログラミング (Inductive Logic Programming; ILP)[6] は述語論理上で帰納推論を行う枠組みであり，背景知識（領域知識）を伴い事例を一般化することで，事例を説明する仮説（ルール）を獲得する．形式的には，述語論理で表現される目標概念に対する正例集合 E^+ と負例集合 E^-，事例に関する背景知識 B に対し，条件

$$B \cup H \models E^+ \quad \wedge \quad B \cup H \not\models E^-$$

を満たす，すなわち背景知識と共に正例集合のみを説明し，負例集合を説明しない仮説 H を獲得する．

代表的な ILP システム Progol[7] や Aleph[8] では，正事例 $e \in E^+$ に対する逆伴意関係

$$B \cup \{\neg e\} \models \{\neg h_e\}$$

に基づき，e を説明する最も特殊な仮説を底とする探索空間（束）を構築し，その中をトップダウンに探索することで単一節仮説 h_e を獲得する．ここで条件を満たす仮説は複数考えられるが，事例集合 $E^+ \cup E^-$ と背景知識 B,（すでに得られている）仮説 H に対し，ある評価基準で最適となる仮説 h_e を選択する．一方で節集合仮説 H は，集合被覆アルゴリズム [7] に基づき，$e \in E^+$ により得られる h_e を仮説集合 H に追加するとともに，H に説明される正例を事例集合から取り除く，すなわち更新

$$H \;=\; H \cup \{h_e\}, \quad E^+ = E^+ \setminus \{e \mid B \cup H \models e\}$$

を正例集合が空になるまで繰り返すことで抽出される．

ILP システム Aleph には，集合被覆アルゴリズムを拡張した二つのアルゴリズム induce_cover 及び induce_max が準備されている．induce_cover では，単一仮説を評価するための事例集合を $E' = E^+$ と初期化し，単一仮説 h_e の評価を（E^+ ではなく）E' を用いて行う．一方 induce_max は，各正例 $e \in E^+$ それぞれに対して h_e を求めることで仮説 $H = \{h_e \mid e \in E^+\}$ を生成する，

3 データセットと背景知識

本研究では，ワンナイト人狼 100 ゲームのプレイログ [4, 5] を分析の対象とする．このデータセットは，ワンナイト人狼の初心者である大学生 10 名を被験者として収集されたものであり，各ゲームにおける役職割り当ての上限は，村人 2 名，占い師 1 名，怪盗 2 名，人狼 2 名となっている．データの収集方法や内容の詳細に関しては，文献 [4] を参照されたい．

本論文では，表 1 に示す 12 の述語を用い，ログデータに含まれる各プレイヤの行動とその順序関係を表現する．またこれらは，ILP システムへ与える背景知識となる．なお述語 `order/3` と `lteq/3` はルール形式の背景知識であり，`co_ord/3` を用いてそれぞれ次のように定義される．

```
order(G,P1,P2):- co_ord(G,P1,O1),
                 co_ord(G,P2,O1), O1 < O2.
lteq(G,P1,ORD):- member(ORD,[1,2,3,4]),
                 co_ord(G,P1,X), X =< ORD.
```

一方，`qst_ctgr/4` における質問内容カテゴリ `CTGR` として，`q0`：投票先を尋ねる，`q1`：役職を尋ねる，`q2`：誰が（人狼／村人／占い師／怪盗）だと思うか尋ねる，`q3`：能力者の行動に対する内容を尋ねる，`q4`：役職構成について尋ねる，`q5`：役職表明に関連した内容について尋ねる，`q6`：疑いを持っているプレイヤに尋ねる，`q7`：自分を人狼と推測した理由を尋ねる，`q8`：その他，の 9 つを準備した．また `act_cnt/3` における行動頻度カテゴリ `CNT` は，各プレイヤが行った推測，同調，反駁行動回数の合計値を離散化したものであり，`c0`：0 回，`c1`：1–2 回，`c2`：3 回，`c3`：4 回，`c4`：5 回以上，の 5 つを準備した．

4 実験と考察

実験では，

1. 占い師を騙る人狼 (`s_deceive/2`)
2. 勝利ゲームにおける人狼 (`winner/2`)
3. 同調された占い師 (`trust/2`)
4. プレイヤの役職 (`p_role/3`)

をそれぞれ目標概念と設定し，ILP システム Aleph[8] に実装されている集合被覆アルゴリズム (induce) と induce_cover，induce_max の 3 アルゴリズムを用いてルール抽出を行った．なお，単一節仮説が説明する正例の下限数を 5，負例の上限数（許容するノイズ数）を

表 1: 背景知識

述語名	内容（G はゲーム，P1，P2 はプレイヤを表す））	回数
comingout(G, P1, R)	G において，P1 が自身の役職を R であると表明した	500
co_ord(G, P1, ORD)	G において，P1 は ORD 番目に役職表明した	500
estimate(G, P1, P2, R)	G において，P1 が P2 の役職を R と推測した	973
agree(G, P1, P2)	G において，P1 は P2 に同調した	305
disagree(G, P1, P2)	G において，P1 は P2 に反駁した	194
inspect(G, P1, P2, R)	G において，P1 が P2 の占い結果を R であると報告した	92
change_co(G, P1, P2, R)	G において，P1 が P2 の怪盗結果を R であると報告した	72
sqt_ctrg(G,P1,P2,CTGR)	G において，P1 による P2 への質問内容カテゴリは CTGR である	275
act_cnt(G, P1, CNT)	G における，P1 の行動頻度のカテゴリは CNT である	500
team(G, P1, P2, T)	G において，P1 は P2 の所属陣営 T を知っている	625
order(G, P1, P2)	G において，P1 より後に P2 が役職表明を行った	
lteq(G, P1,ORD)	G において，P1 は ORD 番目までに役職表明を行った	

表 2: 抽出された単一仮説（ルール）の数

	induce	i_cover	i_max	Union
s_deceive/2	2	6	8	8
winner/2	6	18	23	27
trust/2	4	9	10	12
p_role/3	15	43	56	67

Union は重複するルールを除いた 3 アルゴリズムの結果の合計

2（役職推定のときは 100）と設定している．表 2 に得られたルール数を示す．

以下，それぞれの目標概念に対し，その詳細を説明すると共に，得られた特徴的ルールの例を示す．

4.1 占い師を騙る人狼

嘘つき行為に関する特有な行動を抽出することを目的に，目標概念「ゲーム G において人狼プレイヤ P1 が占い師を騙った (s_deceive(G,P1))」を学習対象とした．なお，正例は占い師を騙る人狼（29 事例），負例は本物の占い師（79 事例）とした．また，背景知識 team/4 は正例（人狼）にのみ有効な述語であるため，学習時には使用を避けている．

学習により全部で 8 のルールが抽出されたが，そのうち 6 ルールが述語 estimate(G, P1, P2, werewolf) を含む結果となった．これとは対照的に，占い師との推測（estimate(G, P1, P2, seer)）を含むルールは抽出されなかった．これは初心者を対象としたことにより，占い師を騙る知識がない被験者が多く，人狼と推測される述語が多く現れた結果であると推測できる．

以下に得られたルールの例を示す．

ルール 1（正例：19，負例：0）

```
s_deceive(G, P1):-
    comingout(G, P2, seer),
    estimate(G, P2, P1, werewolf).
```

ルール 1 は，占い師と役職表明したプレイヤ P2 が占い騙りをしている人狼プレイヤ P1 を人狼であると推測している．占い師であれば，嘘を発言している人狼プレイヤに行う当たり前の行為であるが，今回のデータセットでは，負例が 1 事例も該当しない．人狼プレイヤはあらかじめ全プレイヤの陣営を認知しているため，占い師が取るであろう一般的な行動をしなかったと考えられる．また，本実験は初心者を対象としたため，占い師と嘘をついた被験者は自らの役職をまっとうし，占い師の立場で行動を取ることが出来なかったと推察される．このことから，占い師を騙る人狼は占い師の行動を認知し，占い師のようなふるまいを行うことが求められる．

ルール 2（正例：20，負例：2）

```
s_deceive(G, P1):-
    estimate(G, P2, P1, werewolf),
    agree(G, P2, P3),
    estimate(G, P2, P4, werewolf).
```

このルールでは，プレイヤ P1 を人狼と推測したプレイヤ P2 は，プレイヤ P4 を人狼と推測したプレイヤ P3 に同意している．該当する正例数を見ると，29 事例中 20 事例と 70%弱の事例に当てはまる．占い師を騙る人狼に対し人狼と推測したプレイヤ P2 は，ただ人狼を探すだけではなく，信頼をおけるプレイヤを見つけるなどと，積極的に行動を取るプレイヤであると推測される．これより，議論の中心となる占い師と役職表明し

たプレイヤだけではなく，他のプレイヤに対し行動を取るプレイヤから人狼と推定されるといった結果が得られた．

4.2 勝利人狼

人狼プレイヤの勝敗に着目し，勝利に結びつく行動を抽出することを目的に，目標概念「ゲーム G において人狼プレイヤ P1 は勝利した (winner(G,P1))」を学習対象とした．ここで正例は勝利した人狼（61 事例），負例は敗北した人狼（102 事例）である．以下に得られたルールの例を示す．

ルール 3（正例：11，負例：2）
```
winner(G, P1):-
    team(G, P1, P2, human),
    qst_ctgr(G, P2, P3, q0),
    estimate(G, P3, P2, werewolf).
```

ゲームに勝利した人狼プレイヤ P1 は，プレイヤ P2 を人間陣営だと認知しており，その人間プレイヤ P2 はプレイヤ P3 へ質問カテゴリ q0 の質問をし，質問されたプレイヤ P3 は人間プレイヤ P2 を人狼であると推測している．人間プレイヤが他プレイヤから人狼であると推測されることで，人狼プレイヤ自身には投票の矛先が向かず，人狼陣営が勝利する事例が多かったと考えられる．このことからルール 3 は，人間プレイヤが人狼と疑われている場合，票が集中するだけではなく，人狼同士でも票が揃えやすく，人狼が勝利するゲームの特徴と言える．

ルール 4（正例：13，負例：2）
```
winner(G, P1):-
    agree(G, P1, P2),
    team(G, P2, P3, human)
    estimate(G, P3, P2, werewolf).
```

人狼プレイヤ P1 はプレイヤ P2 へ同意しており，プレイヤ P2 はプレイヤ P3 が人間であることを認知しており，人間プレイヤである P3 はプレイヤ P2 を人狼であると推測している．プレイヤ P1 および陣営情報を持つプレイヤ P2 は人狼であり，プレイヤ P3 は人間陣営のプレイヤであることがルールから分かる．陣営情報を把握した上でルールを確認すると，人間プレイヤは人狼プレイヤを見破り人狼と推測することが出来ているが，人狼陣営内で仲間へ同意することで人狼陣営が勝利していることが分かる．これより，人狼プレイヤは仲間への同意行動により，人間陣営を惑わせることが出来ていると推察される．

4.3 同調された占い師

同調（信頼）されたプレイヤは，ある側面で，他のプレイヤの説得に成功したと考えることができる．信頼の獲得や説得の成功に関する行動を抽出することを目的に，目標概念「ゲーム G においてプレイヤ P1 は信頼された (trust(G,P1))」を準備した．正例として同調された占い師（69 事例），負例として反駁された人狼（86 事例）を利用する．また正例では，占い師であると役職表明していることは自明であるため，背景知識 comingout/3 は排除した．加えて，正負例の直接的な意味を持つため，第三引数が対象となる占い師や人狼であるような agree/3 と disagree/3 も排除している．以下に得られたルールを示す．

ルール 5（正例：12，負例：0）
```
trust(G, P1):-
    inspect(G, P1, P2, villager),
    change_co(G, P3, P2, villager).
```

占い師プレイヤ P1 がプレイヤ P2 の占い結果が村人であると報告し，プレイヤ P3 がプレイヤ P2 の怪盗結果が村人であると報告しているルールである．プレイヤ P2 は，占い師及び怪盗と役職表明したプレイヤの両方から村人であったと報告を受けており，2 人の能力者による能力結果が等しいため，占い師および怪盗の結果の信憑性が上がっていると言える．これより，占い師の能力結果だけではなく，同陣営の能力者である怪盗結果を得ることで，占い師が信頼されるということが分かった．

ルール 6（正例：16，負例：1）
```
trust(G, P1):-
    inspect(G, P1, P2, werewolf),
    act_cnt(G, P2, c1).
```

占い師プレイヤ P1 がプレイヤ P2 の占い結果が人狼であると報告し，そのプレイヤ P2 の行動回数は c1（1–2 回）であるというルールである．通常，人狼と占い結果を言われた人間は，自らが人狼ではなく占い師と役職表明したプレイヤこそが人狼であるという説得行動をすると考えられる．したがって，発言回数が増えることで自然と行動回数が多くなる．しかし，抽出されたルールでは，人狼という占い結果を得たプレイヤの行動回数が少ない．このことから，占い師の行動ではなく人狼の行動が影響し，占い師の信頼性が向上していると推察される．

4.4 役職推定

他プレイヤの役職推定は，人狼ゲームにおいて必要不可欠な要素である．目標概念「ゲーム G においてプレイヤ P1 の役職は R である (p_role(G,P1,R))」を準備し，各役職に特有の行動を抽出するとともに，役職推定モデルを構築する．正例を各プレイヤとその実役職（村人 171 事例，占い師 79 事例，怪盗 87 事例，人狼 163 事例）とし，負例を実役職とは異なる役職（1500 事例）とする．また，人狼以外の対象プレイヤが自身の役職を表明していることは自明であるため，背景知識 comingout/3 は排除している．以下に得られたルールを示す．

ルール 7（正例：93，負例：15）
```
p_role(G, P1, villager):-
    estimate(G, P1, P2, werewolf),
    estimate(G, P3, P1, villager).
```

村人プレイヤ P1 がプレイヤ P2 を人狼と推測しており，プレイヤ P3 が村人プレイヤ P1 を村人と推測しているルールである．負例が 15 事例と決して少なくないが，多くの村人に当てはまるルールであることが分かる．人狼を探すための推測行動は，情報を持たない村人にとって自然な動きであることが分かる．それに対して，能力者を騙らない人狼は村人の動きをしなければならない．しかしながら，負例の数は 15 であるため該当する事例が少なかったと言える．人狼を探すための推測行動をするプレイヤは村人と推測されることから，村人を騙る人狼は，人狼を探す村人のようなふるまいを行うことを心掛ける必要がある．

ルール 8（正例：35，負例：2）
```
p_role(G, P1, seer):-
    estimate(G, P2, P1, seer),
    inspect(G, P1, P3, werewolf).
```

プレイヤ P2 が占い師プレイヤ P1 を占い師と推測しており，占い師プレイヤ P1 がプレイヤ P3 を人狼と推測しているルールである．50%弱の占い師に当てはまるルールであることから，占い先で人狼を当てられたとき，他プレイヤから占い師と推測される占い師が多いことが分かる．

ルール 9（正例：17，負例：6）
```
p_role(G, P1, thief):-
    lteq(G, P1, 2),
    agree(G, P2, P1),
    estimate(G, P3, P1, seer).
```

プレイヤ P3 に占い師であると推測されているプレイヤ P2 に，怪盗プレイヤ P1 が同意されており，その怪盗プレイヤ P1 は 2 番目以内に役職表明している．また，ゲームにおいて重要視される存在である占い師と表明したプレイヤが，怪盗を表明したプレイヤに対して同意していることから，2 種類しかいない重要な能力者の存在を，能力者自身も気にかけていることが分かった．

ルール 10（正例：100，負例：38）
```
p_role(G, P1, werewolf):-
    estimate(G, P2, P1, werewolf),
    agree(G, P3, P2).
```

プレイヤ P2 がプレイヤ P1 を人狼と推測しており，推測したプレイヤ P2 はプレイヤ P3 に同意されている．正例が 100 事例と半数以上が該当すると同時に，38 事例と多くの負例にも該当するルールである．すなわち，信頼されているプレイヤに人狼と推測されることで，人狼であると誤認される可能性が高まると推測できる．

以下に，一連の実験結果をまとめる．まず，本データセットにおいては，人狼プレイヤは騙る役職（占い師）が取るべき行動を取っていないことが分かった．また，勝利した人狼のゲームでは，投票が集まりやすい人間プレイヤがいることや人狼同士の連携が取れていることが，人狼が勝利した場合の特徴であることが分かった．人狼は，占い師や怪盗といった能力者だけではなく，情報を持たない村人の取る行動も十分に認知する必要がある．人狼の時にどのような行動を取るのか考えるよりも，人間であればどのような行動を取るのかを意識することで，村人や占い師らしい行動をすることが出来るため，本実験で得られた差異が見られなくなると推察される．

5 まとめと今後の課題

本論文では，ワンナイト人狼 100 ゲームのログデータに対して帰納論理プログラミングを適用し，占い師を騙る人狼や信頼されやすい占い師に見られる特徴的な行動の抽出を行うと共に，役職推定のためのモデルを構築した．

今後の課題としては，経験の異なる被験者によるゲームログを収集・分析することや，今回得られた分析結果と他の人狼ゲームに対する分析結果とを比較することがあげられる．特にこれまで，掲示板型の人狼である人狼 BBS[1]のログデータを対象に，帰納論理プログラミングを用いた分析が行われており [9, 10]，これら

[1] http://www.wolfg.x0.com/

の結果との比較は急務であると言える．一方，BDI モデル [11] や確率論理 [12, 13] などを用いた分析にも取り組む予定である．

参考文献

[1] 鳥海 不二夫，片上 大輔，大澤 博隆，稲葉 通将，篠田 孝祐，狩野 芳伸：『人狼知能』，森北出版，2016.

[2] 狩野 芳伸，大槻 恭士，園田 亜斗夢，中田 洋平，箕輪 峻，鳥海 不二夫（著），人狼知能プロジェクト（監修）:『人狼知能で学ぶ AI プログラミング』，マイナビ出版，2017.

[3] G. Chittaranjan and H. Hung : Are you a werewolf? Detecting deceptive roles and outcomes in a conversational role-playing game, *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.5334–5337, 2010.

[4] 西崎 絵麻，尾崎 知伸 : ワンナイト人狼を対象とした投票行動の特徴分析，人工知能学会 第 112 回知識ベースシステム研究会，SIG-KBS-B508, pp.52–59, 2017.

[5] 西崎 絵麻，坂口 早紀，尾崎 知伸：ワンナイト人狼における投票行動の分析，2017 年度人工知能学会全国大会（第 31 回），2H1-5，2017.

[6] 古川 康一，尾崎 知伸，植野 研 :『帰納論理プログラミング』，共立出版，2001.

[7] S. Muggleton : Inverse Entailement and Progol, *New Generation Computing*, Vol.13, Issue 3-4, pp.245–286, 1995.

[8] A. Srinivasan : The Aleph Manual, `http://www.cs.ox.ac.uk/activities/-machinelearning/Aleph/aleph`.

[9] E. Nishizaki and T. Ozaki : Behavior Analysis of Executed and Attacked Players in Werewolf Game by ILP, *Proc.* of the 26th International Conference on Inductive Logic Programming (Short papers), pp.48–53, 2016.

[10] S. Sakaguchi and T. Ozaki: An Experimental Analysis of Whispers' Effect in Werewolf BBS by Relational Association Rules, *Proc.* of the 26th International Conference on Inductive Logic Programming (Short papers), pp.60–65, 2016.

[11] N. Nide and S. Takata : Tracing Werewolf Game by Using Extended BDI Model, *IEICE Transactions on Information and Systems*, Vol.E100.D, No.12, pp.2888–2896, 2017.

[12] L. De Raedt, P. Frasconi, K. Kersting and S. Muggleton (Eds.) : *Probabilistic Inductive Logic Programming*, Springer, Berlin, Heidelberg, 2008.

[13] T. Sato and Y. Kameya : Parameter Learning of Logic Programs for Symbolic-statistical Modeling, *Journal of Artificial Intelligence Research*, Vol.15, Issue.1, pp.391–454, 2001.