Vol.2020-GI-43 No.21
2020/3/14

上海ゲームにおける面白いインスタンス生成

森長 剛志[†1]　池田 心[†1]

概要：不完全情報パズルゲームである上海ゲーム (Mahjong Solitaire) は，ゲーム開始時の初期配置によって難易度の差が激しいことや確率論的に良い手を打つと裏目に出てしまうことなどがある．ある程度は仕方ないかもしれないが，このような要素は運次第で理不尽と感じさせ不満を抱かせてしまう．そこで本研究は不完全情報ゲームにおける面白さや難しさの定量化・推定の可能性と特徴的な課題があるのかの解明を行なった．まず，大町による面白い問題の選択方法を検証し，仮説とは異なる結果を得た．被験者実験から得たデータの分析や教師あり学習による推定の結果，難しさについてはある程度推定できたが，面白さについてはあまり良い結果とはならなかった．その理由として，問題を解くまでが長いためデータを集めるのが困難であること，運要素が絡むため評価が安定しにくいことなどが挙げられる．また，各被験者の評価傾向や嗜好に偏りがあった可能性が考えられることから筆者自身が 300 問を 2 回プレイし分析を行なった．その結果，一回目と二回目の面白さ評価では平均 1.17 の違いがあり，これは被験者間の違い 1.22 と比べてもかなり大きい値であるといえる．また，2 ポイント以上離れていたのが 34.33%あったことから，上海ゲームは同じ人物であってもプレイ毎に面白さ・難しさの感じ方が変わることが確認できた．

Generation of Interesting Instances for Mahjong Solitaire

Morinaga Tsuyoshi[†1]　Ikeda Kokolo [†1]

1. はじめに

ゲームやパズルは人間の大事な文化の一つであり，コンピュータの普及によって手軽に楽しめるようになっている．また，これらはルールが簡素明快であることが多い一方で，勝利のためには知的な能力を要求するため，人工知能の良いテストベッドにもなってきた．チェッカーが解析されたり，囲碁・将棋でプロ棋士よりも強いプログラムが開発されるなど技術が大きく進歩した [1] 結果，次の段階として「人間を楽しませる」ことも注目されるようになってきている [2][3].

パズルの問題（インスタンス）生成はそのような流れの一つであり，数独，倉庫番などさまざまなパズルで問題を生成するための技術が提案されている．これらの多くは完全情報ゲームを対象としていたが，本論文では，不完全情報パズルゲームである「上海（Mahjong Solitaire)」を対象とする．Windows に標準搭載されるなど一定の人気と知名度を持つゲームであるが，ゲーム開始時の見えない部分も含めた牌の配置によってゲームクリアできるかどうかの難易度が大きく変わること，場合によっては「確率的には良いはずの手」を選択したために悪い結果を招いてしまうことがあることなど，不満がないわけではない．

以前大町らは，このような不満を解決するための方法を提案しているが [4]，その評価にまでは至っていない．本論文では，まず大町らの手法を追試し，問題の面白さを事前に推定するための教師あり学習を行なう．そして，その推定が他のゲームに比べて難しい理由を，具体的なデータとともに考察する．

2. 対象問題

2.1 上海ゲーム

上海ゲームは麻雀牌を使った一人有限確定不完全情報ゲームであり，麻雀牌を特定の型に立体的に積み上げた状態から以下のルールに従ってすべての牌を取り除くことを目標とするゲームである．

- プレイヤは 2 枚の同じ牌種の牌の組を一度に取り除

[†1] 現在，北陸先端科学技術大学院大学

くことができる．

- 何らかの牌が上に乗っている牌は取り除けない．
- 左右両方に他の牌が接している牌は取り除けない．

プレイヤはこの動作を繰り返し行い，すべての牌を取り除くことができれば解答成功（ゲームクリア）となる．しかし，着手不可能になった時点で牌が残っていた場合は解答失敗（ゲームオーバー）となる．

上海ゲームには同じ局面からでも牌の取り方を間違えると詰んでしまう場合がある．図 1 上部の局面を例に示す．

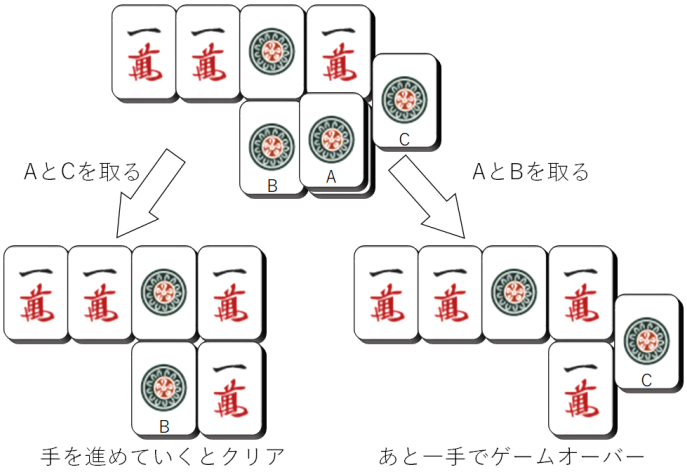

図 1: 牌の取り方を間違えると詰んでしまう局面の例

図中の局面は“一萬”と“一筒”4 枚ずつで構成されており一筒の 1 枚は 2 層目に置かれている．このときの打てる手は一筒 A,B,C の 3 枚のうち 2 枚をどのペアで取り除くかの 3 パターンが挙げられる．例えば A と B を取った場合，図 1 右下の状態となる．次の手で一萬のペアを取り除くことができるが，残りの一萬の牌と一筒の牌が相互に支配する形となり詰みとなる．もう 1 つの例として A と C を取った場合，図 1 左下の状態となる．このあとは取れる牌を取っていけばゲームクリアとなる．

2.2 上海ゲームの戦略と確定着手

本研究の目的は「強い上海プレイヤ」を作ることではないが，上海ゲームの基本的な戦略はインスタンスの面白さや難しさとも関連するため簡単に述べておく．上海ゲームでは全ての牌が見えている場合は解けるか解けないかが先読みによって確定できるが，序盤は隠れた牌が多いため「いろんな牌の隠れ方を想定し，できるだけ詰んでしまう危険性が低い取り方をする」ことが戦略の基本となる．例えば，4 枚存在するはずなのに 2 枚しか見えていない牌種を取った場合，もし残りの 2 枚が縦に重なっていればその時点でクリアは不可能になるため，これは「ある程度危険な取り方」ということになる．

一方で，ある局面において将来的に解答不能になる原因となりえない着手を「確定着手」と呼ぶことにする．これは完全情報ゲームにおける手筋や定石などの解法ロジックに近い．上海ゲームでは以下の 4 つの確定着手が存在する．

A) 残り枚数が 2 枚の牌種の牌は取る．
B) 同時に 4 枚着手可能な牌種の牌は取る．
C) 同時に 3 枚着手可能な牌種で且つ，その中に浮いている牌がある場合は，その牌以外の 2 枚の牌を取る．
D) 同時に 3 枚着手可能な牌種で且つ，その中に自身と同じ牌種の牌を直接支配している牌がある場合，その牌を含めた 2 枚を取る．

支配している牌とは，他の牌に対して着手不能にしている牌のことである．また，浮いている牌とは支配してもされてもいない牌のことである．図 2 に C と D の確定着手が行える局面の例を示す．“中”は C の確定着手，“一萬”は D の確定着手が行える．

図 2: “中”は C の確定着手，“一萬”は D の確定着手が行える局面の例

3. コンテンツ生成と面白さ評価

アルゴリズムからコンテンツの自動生成を行なう Procedual Content Generation(以下 PCG) が活発に研究されている．PCG にはコンテンツ作成におけるコストを抑制することともにプレイヤに毎度異なるプレイ経験を与えることが期待されている．

関連研究として及川らはテトリスにおける T-spin と呼ばれる重要な技術の構成力向上を目的とした「詰め T-spin 問題」の自動生成を行なった [5]．この研究は詰め問題の自動生成の手法を提案しつつ教師あり学習による推定から各プレイヤの熟練度に適した面白いまたは難しい問題の選別・提供を可能とした．

上海ゲームとテトリスは同じ不完全情報パズルゲームに分類されるが異なる点がある．それはプレイの進行とともに見えなかったものが徐々に明らかになっていく特殊な不完全情報性であること，解が一意でない，つまりある問題に対して複数の解答手順が存在しうるため，たとえ悪い手を打ったとしてもゲームクリアとなる可能性があることが挙げられる．そのため試行によって面白さの感じ方が変わることが考えられる．また，数手で解答されるテトリスの詰め問題と比べると上海ゲームは数十手を要するため被験

者の評価データが集めにくいことがいえる．このことから関連研究よりも推定が困難であることが予想される．

4. 上海ゲームにおけるインスタンス生成の先行研究

本研究は過去に行なわれた大町らの研究 [4] に沿ってインスタンスの生成と抽出を実施した．本章では先行研究におけるインスタンスの生成と抽出の方法，作成した人工プレイヤなどの提案手法について述べる．

先行研究では以下の 4 つのステップのアルゴリズムによる生成検査法に基づき，プレイヤへ提供するインスタンスの生成を行なった．

1. 解法が必ずあるようにインスタンスの内容を乱数により決定し，インスタンスを複数作成する．
2. 隠れ牌を n 回仮定し深さ d だけ読むモンテカルロプレイヤを作成し，$(d, n) = (1, 1), (3, 16)$ とした 2 つの性能の異なるモンテカルロプレイヤにインスタンスを解かせる．
3. 2 つのモンテカルロプレイヤの平均クリア率のプロット図からインスタンスを難易度や特徴ごとに分類する．
4. 分類を基に面白いインスタンスのみを抽出する．

先行研究では二つの性能の異なるモンテカルロプレイヤを用いて平均クリア率の二次元プロットを作成し，その分布から特徴的なインスタンスを抽出するという手法を提案した．弱い人工プレイヤを x 軸，強い人工プレイヤを y 軸としたときの平均クリア率の二次元プロットを図 3 に示す．

【クリア率の分布と領域分け】

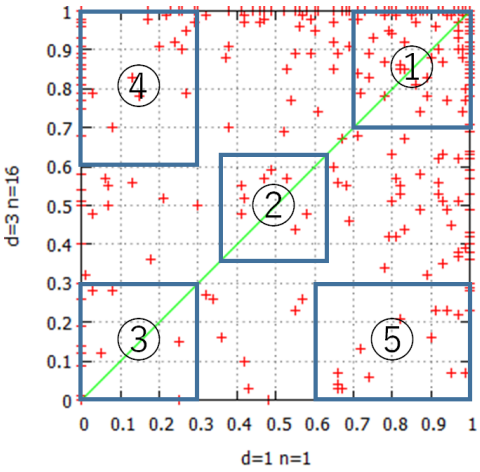

図 3: 二つの性能の異なる人工プレイヤの二次元プロット

図 3 を見ると，強い人工プレイヤは解けるが弱い人工プレイヤには解けないインスタンス，強い人工プレイヤは解けないが弱い人工プレイヤには解けるインスタンスなどが見受けられる．一見，弱い人工プレイヤが解けるインスタンスなら強い人工プレイヤも同様に解けるのではないかと思われるが，上海ゲームには運が絡むことや確率的に悪い手がクリアに繋がることも考えられるため，このようなインスタンスが存在する．

先行研究はこの結果を基に平均クリア率を大まかに高・中・低の三つに分類し，以下のように各インスタンスに対する仮説を示した．

- 領域 1 : 強い人工プレイヤ : 高，弱い人工プレイヤ : 高
  読みを必要とする場面が少ない．簡単すぎるだろう．
- 領域 2 : 強い人工プレイヤ : 中，弱い人工プレイヤ : 中
  運に左右される選択肢が含まれる場合があるだろう．
- 領域 3 : 強い人工プレイヤ : 低，弱い人工プレイヤ : 低
  必要な読みが難しすぎる．運に左右される選択肢が多く含まれるだろう．
- 領域 4 : 強い人工プレイヤ : 高，弱い人工プレイヤ : 低
  読みが必要な局面が多く，理不尽さや運要素は少ないだろう．
- 領域 5 : 強い人工プレイヤ : 低，弱い人工プレイヤ : 高
  本来良い手が裏目に出る，理不尽さを感じる選択肢が多いだろう．

強い人工プレイヤが高い平均クリア率，弱い人工プレイヤが低い平均クリア率（領域 4）のインスタンスは，理不尽さや運要素の少ない解き応えのある傾向にあるため，プレイヤに提供するにあたって有望であることを大町らは主張した．

【運要素と理不尽さ】

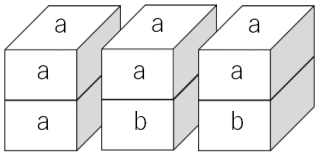

図 4: 運要素がある局面例

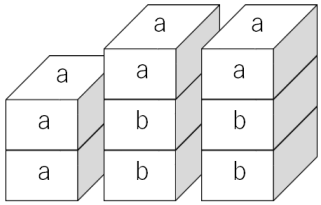

図 5: 理不尽さを感じる場合例

図 4 と図 5 は運が絡むまたは理不尽さを感じる局面の例である．図 4 では 2 層目に “a” が 3 つあり，その中の 2 枚を選択する局面である．もう 1 つの “a” を見つけるとクリアとなるが，見えている牌だけではどこにあるのか分からない．ランダムな場所にあると仮定すると 2/3 の確率でクリアとなる．これは，上級者プレイヤであってもそれ以上の確率にすることはできない．

図 5 では，3 層目に “a” が 2 枚，2 層目に “a” が 1 枚あり，その 3 枚の中から 2 枚を選択する局面である．見えている牌だけから判断すると 3 層目の “a” を 2 枚取る場合は 4/5 の確率，2 層目の “a” を取る場合は 3/5 の確率でクリアできる．当然高い確率の手を選択するが，それが裏目に出てしまうこともある．つまり，良いはずの手が分かる上級者プレイヤほど結果的に損をしてしまう理不尽なイ

ンスタンスであるといえる．

5. 被験者実験

本研究では，4 章で述べた「読みが必要な局面が多く，理不尽さや運要素が少ない」インスタンスは人間にとって本当に面白いのか，面白さや難しさを感じさせる要素は何か調査するため被験者実験を行なった．本章では，被験者実験における実験設定や結果を説明し被験者の傾向を述べる．

5.1 実験設定

本研究では「上海ゲームの経験者」や「他パズルゲームの経験者」など何かしらのパズルゲームに触れたことがある男性 14 人を対象に，インスタンスごとの面白さ・難しさを評価してもらう被験者実験を行なった．

今後，これらの評価を教師あり学習に用いることを想定すると，できるだけ多くの面白さ・難しさを収集したい．一方で，こういった「一人の意見」,「一回の意見」などの感性評価はブレやノイズといった誤差が生じやすいことが知られており，できるだけ多くの人に評価してもらい，その平均値をインスタンスの面白さ・難しさとして使いたい．しかしながら，被験者を雇用するには謝金が必要であること，上海ゲームは 1 つのインスタンスをプレイするのにも数分〜十分程度要することを考えると，どちらも満足いくほどの数を集めることはできないと思われる．

今回は，問題数と 1 問当たりの被験者数のバランスを考慮し，被験者を 2 つのグループに分け，それぞれ 50 問ずつ評価してもらうことにした．50 問の評価に掛かった時間は 4 時間ほどであり，被験者の疲労を考慮して実験は 2 時間ずつ（25 問ずつ）に分けて行なった．

設定した問題として 9 種 36 枚の麻雀牌で構成されている図 6 のレイアウトを用いた．また，被験者には 4 章で提示した領域 1〜領域 4 までの上海問題をプレイしてもらった．表 1 に問題数と配分を示す．

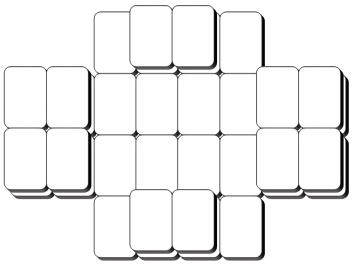

図 6: 使用したレイアウト

表 1: 各グループにどの領域から何問出題したか

	領域 1	領域 2	領域 3	領域 4	合計
被験者 A グループ	10	9	18	13	50
被験者 B グループ	13	9	13	15	50

出題形式として，最初の 2 問をチュートリアルとして用意し動作やルールの確認をしてもらった．また，被験者にプレイさせる問題はランダムに出題するようにした．領域 5 の問題は生成されることが稀であり数を揃えることは困難なことから今回の被験者実験では用いないこととした．

被験者実験から取得したデータは，「面白い，難しい」の 5 段階評価，プレイしてどのような点が面白かった等の感想，プレイヤが打った手順と 1 手ずつの時間，1 問に掛かった時間，問題の正解／不正解である．

5.2 実験結果

各領域の問題に対する被験者全体の平均評価を表 2 に示す．全体的な面白さ・難しさの標準偏差（問題ごとに被験者全員の平均値を取り，問題間の評価の標準偏差）はそれぞれ 1.24，1.18 であった．

領域 1〜3 は，両方の人工プレイヤにとって簡単・中程度・難しい問題である．領域 4 は，強い人工プレイヤのクリア率のみ高い問題であり，先行研究ではこれを面白いはずとしていた．以降，いくつかの視点で実験結果を考察していく．

表 2: 被験者全体における各領域の評価

	AI プレイヤ正解率		面白さ	難しさ	
	弱い側	強い側	平均値	平均値	平均クリア率
領域 1	高	高	0.58	0.16	0.78
領域 2	中	中	0.13	0.60	0.52
領域 3	低	低	0.05	1.06	0.36
領域 4	低	高	0.30	0.54	0.56
全体			0.25	0.60	0.54

5.2.1 領域ごとの面白さと難しさの傾向

表中の各領域の平均面白さは $1 > 4 > 2 > 3$ となっており，簡単すぎるはずとされた領域 1 のインスタンスが最も面白く，反対に最も難しい領域 3 のインスタンスが最もつまらないという結果になった．大町らの研究では領域 4 が最も面白いと評価されるであろうと期待していたが，今回の実験結果からは（統計的検定はおこなっていないが）それは否定された．この原因として，被験者は初級者が多いため領域 1 のインスタンスでも簡単すぎるほどではなかったことが考えられる．被験者の中には領域 4 のインスタンスが最も面白いと評価している人もいるが，合計でこのような形となったのではないかと思われる．

平均難しさについては $1 < 4 < 2 < 3$ となっており，領域 1 のインスタンスは最も簡単，反対に領域 3 の問題は最も難しいという結果になった．このことは大町らが主張したインスタンスの分類と人間プレイヤにとっての難しさの感じ方がおおよそ一致しているということがいえる．

5.2.2 平均クリア率と面白さ・難しさの関係

表中の各領域の平均クリア率は 1 > 4 > 2 > 3 となっており，簡単なインスタンスや読みを必要とするインスタンスはクリア率が高く，反対に難しいインスタンスはクリア率が低い．また，運要素が強いインスタンスは 50% 程度のクリア率という想定通りの結果となった．

それぞれ平均面白さ 1 > 4 > 2 > 3，平均難しさ 1 < 4 < 2 < 3 と各領域順が対応していることから，正解できなかったインスタンスは難しくてつまらない，正解できたインスタンスは簡単で面白いという傾向であることが考えられる．これを踏まえると「なぜ最も面白いと期待された領域 4 のインスタンスが 2 番目に面白いという評価なのか」について，人間プレイヤは少なからず誤った読みをしてしまうことで不正解となり，面白くないと感じたからだと考える．その意味では，初級者相手であれば，簡単めの問題を出すなり，「(相対的に) 強いプレイヤ」としてもっと弱い $(d, n) = (2, 8)$ などを利用することも考えられるかもしれない．

一方で，領域 4 のような選択に価値がないわけではない．同程度の平均クリア率である領域 4 と領域 2 であるが，被験者は少なくとも運要素が強い領域 2 よりも読みを必要とされる領域 4 の方が面白いと感じていることが確認できる．

5.2.3 人間プレイヤと人工プレイヤの平均クリア率における傾向

人間プレイヤとモンテカルロプレイヤの平均クリア率を比較してみると傾向が似ていることが分かる．どちらも領域 1 の平均クリア率が一番高く，領域 3 の平均クリア率が一番低い．また，運要素が強い領域 2 のインスタンスについては両者とも 50 %ほどの平均クリア率であることから，人間プレイヤと人工プレイヤの差はあまりないことが分かる．すなわち，人工プレイヤは，ある程度は人間プレイヤにとっての難しさを推定できていることがいえる．

一方で，両方の人工プレイヤが 20% 以下しかクリアできなかった領域 3 でも人間プレイヤは 36%クリアできていること，逆に両方の人工プレイヤが 90% 以上クリアできている領域 1 でも人間プレイヤが 78% しかクリアできていないことは面白い結果である．人間には，1 手または 3 手の深さしか読んでいない人工プレイヤとは異なり，もっと奥深くまで読める能力がある．しかし，人間には，完全な記憶を持つ人工プレイヤと異なり，すでに取った牌種を忘れてしまうという欠点もある．このことから，より人間に近いテストプレイヤを作ろうとするならば，これらの要素も考慮に入れるべきかもしれない．

5.2.4 面白さと難しさの関係

5.2.2 節では平均値を見て“簡単な問題が面白い”という傾向を得たが，個別の問題ごとの分析はしていなかった．面白さと難しさに相関関係はあるのか調べるため，図 7 のように縦軸を平均面白さ，横軸を平均難しさとしてプロットし，特徴を割り出した．

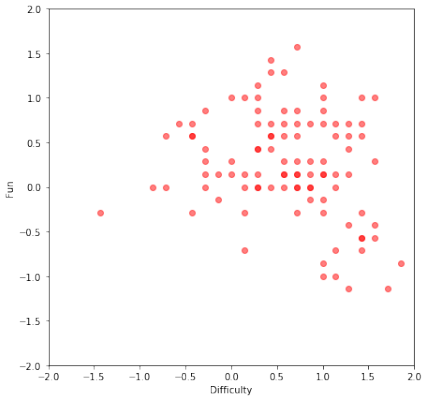

図 7: 平均面白さと平均難しさのプロット

図中のグラフは典型的な山形となっており，難しさが上昇するに従って面白さも上昇しているが難しくなりすぎると途端につまらないという傾向になっている．5.2.2 節では各項目の平均値から簡単な問題ほど面白いという見方であったが，個別のインスタンスごとにみると，難しすぎず簡単すぎないインスタンスの方が望ましいことが分かった．

5.2.5 被験者同士の意見の相違

本研究のような被験者を用いた感性評価実験では，人によって面白さ難しさの基準が異なったり，5 段階評価のスケーリングが異なる可能性が高い．そこで，被験者同士の評価を度数分布化し，評価傾向やその度合いが人によってどれほど異なるのか調査した．図 8 と図 9 は面白さ評価についての度数分布であり，縦軸，横軸は各被験者の評価である．

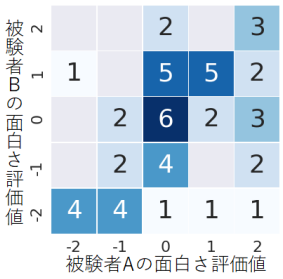

図 8: 験者群の中で最大相関となったペアの面白さ評価 50 問分の度数分布

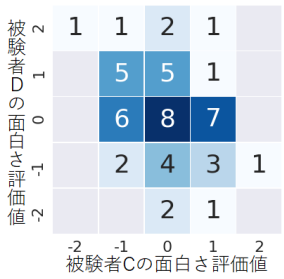

図 9: 験者群の中で最低相関となったペアの面白さ評価 50 問分の度数分布

図 8 と図 9 の相関係数はそれぞれ 0.46，-0.30 であり最高値と最低値を示している．これらの被験者について分析すると，傾向が近い被験者同士の平均クリア率は同程度であること，反対に傾向が異なる被験者同士の平均クリア率には大きな差があることが分かった．このことから，上海ゲームはプレイヤの実力などの違いによって面白さの感

じ方が異なり，全員向けの面白いインスタンスの実現は難しいと考えられる．しかし，同様の傾向を持つ人同士でグルーピングを行なうことで精度の良い面白さ推定が期待できると思われる．

図 8 と図 9 をみると被験者の評価のスケーリングが異なることが確認できる．図 9 の横軸の被験者 C は面白い，つまらない (＋ 2,-2) といった最大値・最小値について 1 回ずつしか付けていないのに対し，図 8 の被験者 A や被験者 B はそれぞれ 5 回〜10 回程度付けている．各被験者の評価のスケーリングが異なりすぎると，あるインスタンスに過大評価・過小評価をしてしまう問題点がありうるため，場合によっては正規化をしなければならない可能性がある．

今回の被験者実験では，全体的な被験者の評価として簡単で正解できるインスタンスが面白いこと，しかしある程度難しさが欲しいこと，各被験者で嗜好などの差異があることといった，データや問題点を得られた．教師あり学習による面白さ・難しさ推定では，これらを考慮しつつデータの加工やパラメータ調整などを行なっていく．

6. 面白さ・難しさ推定

本研究では，前章で述べた被験者実験での被験者の感想を参考にして，インスタンスの面白さや難しさに関係ありそうな特徴量を考案した．それらの特徴量の値を説明変数，被験者の主観評価を目的変数とし，教師あり学習を用いて推定を行なった．本章は，考案した特徴量と比較指標について述べる．

6.1 計測した指標

本章では，教師あり学習によって各インスタンスの面白さや難しさを推定する．このような推定はよく行われており [5]，その際には平均二乗誤差の平方根 (RMSE) が精度の指標として使われることが多い．

6.2 考案した特徴量

本研究では，2 か月前後の試行錯誤の結果，72 個の計算可能な特徴量をインスタンスの面白さ・難しさの推定に用いることにした．72 個のうち 24 個は盤面から静的に得られる特徴量で，6.2.1 節で詳述する．残り 48 個は人工プレイヤを用いて得る特徴量で，6.2.2 節で詳述する．

6.2.1 盤面から静的に得られる特徴量

被験者実験を行なった際，被験者にはインスタンスについてどのような特徴が面白さや難しさを感じるのか感想を得た．感想は全部で 500 個近くあったが，その中で重複して何度も言及されているような要素もかなり多かった．

これらの感想に出てくる計算可能な数値は，インスタンスの面白さや難しさを推測するのに役立つと考えた．そこで，その感想を基にインスタンスに関する特徴量 24 個を考案し算出を行なった．以下に一覧を示す．

- 同種牌の横並びの総数
- 同種牌が 2,3,4 枚横並びになっている数
- 同種牌の重なりの数
- 1,2 層目の牌種数
- 初期盤面時 0,1,2,3,4 枚見えている牌種数
- 1 層目に 0,1,2,3,4 枚ある牌種数
- ある列に同種牌が 2,3,4 枚ある数
- ある長列に同種牌が 2,3,4 枚ある数
- 初期盤面時の合法手数

6.2.2 人工プレイヤから得られる特徴量

5.2.3 節で述べたように，人工プレイヤの“勝率”は人間プレイヤにとっての難しさとある程度相関があることが分かっている．そこで，2 つの人工プレイヤだけでなく $(d, n) = (1, 1), (1, 16), (3, 1), (3, 16)$ の 4 つの人工プレイヤを用いる，さらに勝率以外のプレイ時の計 12 個の統計量を用いる，という追加の工夫を行うことでより精度の高い推定を行うことを目指した．以下に一覧を示す．

- 平均クリア率
- 平均と最大の総合法手数
- 平均残り牌数
- 確定着手 A,B,C,D の平均回数
- 確定着手総数の平均回数
- 同種牌の重なりや 2,3 種類の牌による詰み数

これらの特徴量を入れている理由として，合法手数や確定着手はクリアする上で重要な要素の 1 つであることから，プレイヤの実力に関わるものといえる．また，クリア率や詰み方はプレイヤが選択した着手の結果であるため，面白さ・難しさを感じる可能性が考えられる．

7. 教師あり学習の推定結果

本節では，5 章で述べた被験者実験を通して 14 人分合計 100 問のデータを収集し，6.2 節で挙げた特徴量を基に面白さ・難しさ推定を行なった．

学習には LightGBM と呼ばれる，決定木アルゴリズムに基づいた勾配ブースティングの機械学習フレームワークを用いた．パズルインスタンスの面白さ推定を行なった同様の研究では LightGBM を使用し高精度な推定結果を示していることから [5]，本研究もこれに倣うこととした．

このときの実験環境は Google Colab の Python3.6，LihgtGBM は標準搭載であり，バージョンは 2.2.3 であった．LightGBM のパラメータは learning_rate:0.01，max_depth:7 とし，他はデフォルトとした [6]．また，Boruta と呼ばれる，ランダムフォレストと検定を用いた特徴量選択ライブラリを使用した [7]．さらに，評価精度の過大評価・過小評価を抑制するため 4 分割交差検証を行なった．

7.1 ベースラインの算出

本研究において，推定結果を正確に比較できるような既存研究やデータは無いため，被験者の評価を用いた独自のベースラインを設定した．2 つのグループの中で被験者 1 人の評価を予測値，他被験者全員の評価の平均値を実値として RMSE など推定精度を算出することで，一人の評価がその他被験者の平均評価を推定できるのか，傾向は似ているのかを調べることができる．

被験者実験は，50 問を 7 人のグループ，別の 50 問を別の 7 人のグループで行った．従って，このような「一人分を抜いて残り 6 人の平均値と比較」は，14 通り行えることになる．この 14 通りのうち，最も推定が上手くいった場合，つまり「他人の意見をよく当てられた人」の場合と，14 通りの平均値を表 3 に示し，これをベースラインとする．

7.2 被験者全体の学習結果

5 章の被験者実験では，被験者ごとの嗜好やスケーリングの問題が挙げられたが，実際にこれらの問題点が推定にどれくらい悪影響を及ぼすかみるために，本節では被験者全員の生データを用いて学習を行なった．Boruta のパラメータの 1 つであり特徴量選別の緩急を担う perc 値を 50（デフォルト値は 100）としたとき，面白さ・難しさの特徴量数はそれぞれ平均 12 個，27.5 個であった．表 3 に被験者全員の評価を用いた学習の結果を示す．

表 3: 被験者全体の学習結果

	面白さ RMSE	難しさ RMSE
被験者全体の学習結果	0.62	0.45
ベースライン（最良値）	1.00	0.96
ベースライン（平均値）	1.22	1.16

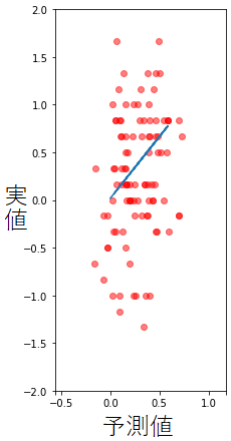

図 10: 面白さ推定 (perc:50)

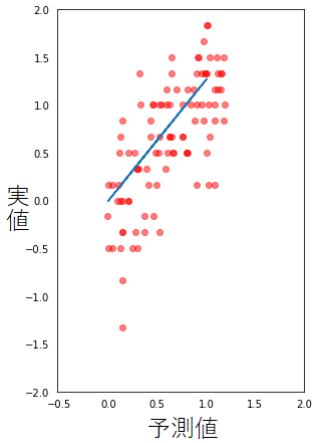

図 11: 難しさ推定 (perc:50)

表 3 をみると，面白さよりも難しさの方が高い推定精度であることが確認できる．被験者が感じる難しさの要素と選別された特徴量の要素が合致していることが考えられる．一方で面白さについては特徴量として出せていないことも挙げられるが，サンプル数の少なさや各被験者の好みがさまざまであることなど，複数の原因が考えられる．

図 10 と図 11 は縦軸を被験者全体の評価の平均値（実値），横軸を推定モデルによる予測値としたときの面白さと難しさのプロット (perc:50) である．図 10 の面白さ推定をみると予測値が 0.0〜0.5 付近に集まっており，無難な推定しかできていないことが見受けられる．図 11 の難しさ推定では，推定値と実値に相関があるようにみえることから，ある程度の推定ができていることが考えられる．

8. 推定を阻害している要因

今回用いた方法と設定では推定精度が不十分だったため，推定を阻害している要因を明らかにすることは重要である．そこで，本研究では以下の 5 つを要因として仮説を立てた．

(1) サンプル数が少ない
(2) 人によって好みが異なる
(3) スケーリングの問題
(4) 試行ごとに評価が異なる
(5) 特徴量，ハイパーパラメータが悪い

これらの仮説のうちどれか，または複合的な影響が考えられる．このうち，(3) のスケーリングの問題については被験者全体の評価を平均値と標準偏差について正規化し，再度推定を試したが限定的な改善のみ得られた結果だった．また，(5) の特徴量やハイパーパラメータについては，試行錯誤を行ない調整した結果が前述したものである．そのため，(3) と (5) についてはこれ以上の対処は難しいだろうと考える．

残りの (1)，(2)，(4) の要因について，それぞれどの程度影響するのか，筆者が 300 問とサンプル数を増やし，さらに 2 週間ほど時間を空けて再度解くことで，同一人物でどれほど意見が異なるのか検証を行なった．

8.1 推定を阻害している要因の検証

筆者が 300 問を 2 回行ない，各評価を平均したもので学習してみたが，あまり良い推定にはならなかった．そこで，各仮説に対し検証を行なった．

一回目と二回目の平均評価，一回目のみの評価，二回目のみの評価を用いてそれぞれ学習し，各インスタンスに対する推定値を得た．平均評価をテストデータとして 3 つの推定値との面白さ RMSE を算出すると，それぞれ 1.21（平均時），1.25（一回目のみ），1.26（2 回目のみ）であったことから，平均化による推定精度向上がみられた．

また，サンプル数が 100 問のときと 300 問のときでどれほど推定精度が変わるのか確認を行なった．100 問の選出は，300 問から各領域の割合に従いながらランダムで抜

き出した．選出した 100 問で学習し，これを 5 回行ない RMSE の平均を算出した．このときの面白さ RMSE はそれぞれ 1.12（100 問），1.21（300 問）であったことから，サンプル数を増やしても効果は薄いことが分かった．

このことから，(4) の試行ごとに評価が異なることによる影響は他の要因よりも甚大である可能性が考えられる．次に (4) の影響がどの程度なのか確認するため，筆者の一回目と二回目でどの程度評価が違うのか調査を行なった．

8.2 一回目と二回目の評価の違い

図 12 は筆者が 300 問を解いたときの一回目と二回目の面白さ評価の度数分布である．これをみると，一人の評価にもかかわらず 3 ポイント以上離れている箇所が見受けられ，人一人による評価でも大きく異なることが確認できる．

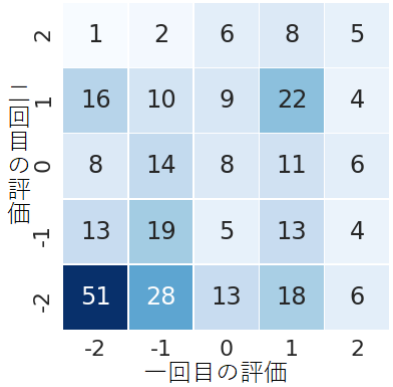

図 12: 筆者の一回目と二回目の面白さ評価の度数分布

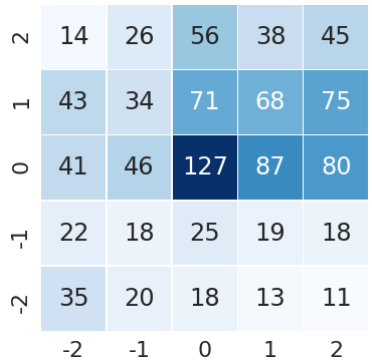

図 13: A グループ全体の面白さ評価の度数分布

表 4: 筆者一人と被験者 7 人の評価の違い

	筆者一人	A グループ被験者全体
相関係数	0.29	0.18
評価の違いの平均値	1.17	1.22
2 ポイント以上異なる割合	34.33%	35.52%

図 13 は A グループ 7 人の被験者全員の面白さ評価の度数分布である．異なる人同士であるため当然評価はばらつくが，表 4 の各項目をみると，人一人による評価のばらつきとの違いがあまり変わらないことが分かる．

(4) の試行によって評価が大きく異なることによる影響は (2) の人の好みのばらつきによる影響よりも大きく，不完全情報ゲームにおける面白さ推定を行なうにあたっての大きな課題であることが解明できた．

9. おわりに

本研究では，不完全情報パズルゲームである「上海ゲーム」においてインスタンスの面白さ推定を行なった．先行研究で主張した仮説を検証するため，被験者実験を行ない，各インスタンスに対する評価と感想を得た．その結果，先行研究の主張は異なっていたこと，難しさと平均クリア率において人と人工プレイヤとの類似性などが得られた．

次に，被験者の感想を基に特徴量となりそうな要素を考案し，各インスタンスについて面白さ・難しさを推定するモデルを生成した．特徴量の選別などさまざまな工夫を施した結果，難しさについてはある程度正確な推定が可能であることを解明した．一方で面白さについての推定は不十分であり満足いくものではなかった．

推定がうまくいかなかった理由としてサンプル数の少なさ，被験者の好みのばらつき，試行ごとの評価のばらつきなどの複合的影響が考えられる．それぞれがどの程度影響を及ぼすのかを確かめるため，筆者が 300 問を時間をあけて 2 回解き，面白さ・難しさを評価する実験を行なった．

結果として，試行毎で評価が異なることによる影響は人の好みのばらつきなどによる影響よりも大きく，推定精度の著しい低下に繋がっていたと考えられる．

今後の展望として，さらに被験者実験を行なうことで試行ごとの評価の安定を図る．

参考文献

[1] Cambell Murray A. Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. Artificial intelligence, 134.1 pp.57-83, 2002.
[2] 村瀬 芳生，松原 仁，平賀 譲，“「倉庫番」の問題の自動生成”, 情報処理学会論文誌 Vol.39 No.3 p.567-574,1998.
[3] 土出 智也，真貝 寿明，“数独パズルの難易度判定—解法ロジックを用いた数値化の提案—”, 大阪工業大学紀要. 理工篇 56(1), 1-18, 2011.
[4] 大町 洋, 池田心. “強さの異なる人工プレイヤを用いた，不完全情報パズルの面白いインスタンス生成”. 北陸先端科学技術大学院大学修士論文, Mar-2014.
[5] 及川 大志，池田心. Improving Human Players’ T-spin Skill in Tetris with Procedual Problem Generation. The16th International Conference on Advances in Computer Games(ACG 2019).
[6] LightGBM. [https://lightgbm.readthedocs.io/en/latest/genindex.html]. （アクセス：2020/02/04）
[7] boruta_py. [https://github.com/scikit-learn-contrib/boruta_py]. （アクセス：2020/02/04）