オンライン麻雀における実力評価基準の検討

清田　達

アブストラクト

本研究は,オンライン麻雀におけるプレイヤーの実力を公平に評価するための新たな基準を提案する. オンライン麻雀の対局結果は,配牌やツモといった確率的要素に大きく影響されるため,短期的な成績が必ずしも実力を反映するとは限らない. そこで本研究では,対局データ（牌山・ツモ履歴）を複数の統計的指標（エントロピー,ラン検定など）を用いて特徴量化し,教師なしクラスタリングによって局の展開を「良形」「悪形」「中間」に分類する手法を開発した. 自身の 10 対戦分のデータを対象としたケーススタディを通じて,特定の対局において統計的に有意な有利・不利が生じ,プレイヤーの実力評価に影響を及ぼし得ることを実証した. 本研究の貢献は,偏りの原因を問うのではなく,「その対局が実力評価の土俵として妥当か」という判定基準を提示する点にある. このアプローチは,e スポーツとしての麻雀の公平性を担保するための基盤となり,将来的には第三者認証を伴う業界標準の牌山生成アルゴリズム確立の必要性を示唆するものである.

キーワード

オンライン麻雀，実力評価，統計的検定，クラスタリング，e スポーツ，公平性

1. 序論

オンライン麻雀においては,対局成績がプレイヤーの段位や勝率に直結し,ひいてはプレイヤーの実力評価やアイデンティティに強い影響を及ぼす. しかし,各対局には配牌やツモに代表される確率的要素が存在し,短期的にはプレイヤーの実力が正しく反映されない可能性がある. 本研究では,このような偶然要因を統計的に分離し,プレイヤーの実力をより公平に評価するための検定手法を提案する. ここで重要なのは,偏りの原因が偶然であれ恣意的であれ,それが基準値から大きく逸脱している場合には,その対局結果を実力評価の根拠とすることが妥当ではない,という視点である. 本研究はこの「実力評価の土俵の妥当性」を明示する点に独自性を持つ. 研究全体の流れを図 1 に示す.

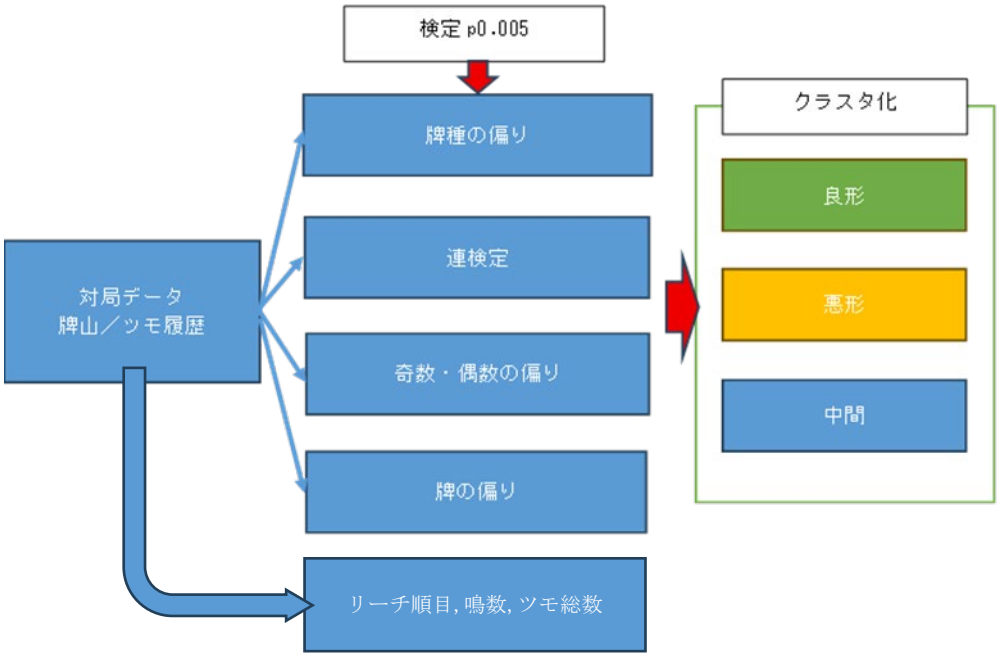


図 1 分析フロー（p 値による偏り検定と教師なしクラスタ化）

2. 研究の背景

オンライン麻雀における実力評価は古くから議論されてきた. 既存の研究では,得点効率,和了率,放銃率などの指標が提案されてきたが,それらは必ずしも局面ごとの有利・不利を考慮していない. 実力を公平に測定するためには,局面の偏りを統計的に検出し,その影響を排除する手法が必要である. さらに従来の枠組みでは「偶然か恣意か」という原因論に焦点が当てられてきたが,本研究は「その対局結果を実力評価に用いてよいか」という妥当性の判断に着目している.

3. 手法

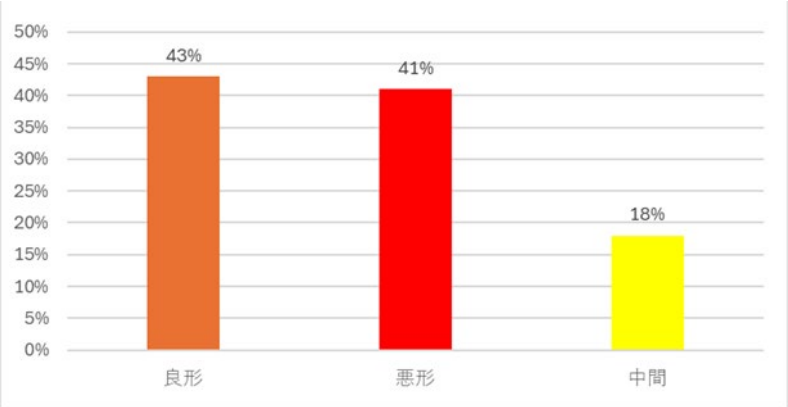
本研究では,対局（対戦）に使用する牌山データを局ごとにそれぞれのプレイヤーの配牌 13 枚とツモ列 18 枚又は 17 枚を足した牌列 31 枚又は 30 枚を以下の特徴量を用いて数値化した.この時,配牌は理牌を実施している.（大規模シミュレーション 4 万局における各クラスターの平均的特徴は表 1 参照）. 牌種の偏り,連検定,偶数・奇数の偏り,牌の偏りをその牌列が持つ特徴量とした. 各指標（entropy_bits, p_runs, p_empirical, markov_p, cramers_v）についての詳細は付録 1 に示す. これらを入力として教師なしクラスタリングを行い,局毎の牌列を「良形」「悪形」「中間」に分類し,Isolation Forest を用いて異常局の検出も行った. ランダムシミュレーションとして約 4 万局を生成し,公

平な場合のクラスターの平均的な特徴に対する遭遇率をグラフに示した（良形 ≈ 43%, 悪形 ≈ 41%,中間 ≈ 16%, 図 2 参照）. 分析手法の詳細は付録 1 に示す.

表 1 大規模シミュレーションにおける各クラスターの平均的特徴

指標	中間	良形	悪形
p_runs	0.4	0.08	0.08
p_empirical	0.55	0.25	0.78
entropy_bits	0.57	0.48	0.67
markov_p	0.48	0.17	0.16
cramers_v	0.16	0.17	0.14

図 2 40,000 局における各クラスター遭遇率



4. ケーススタディ

4.1. サンプリング設計

本ケーススタディでは,提案手法の有効性を検証するため,筆者自身の 10 対戦のデータに加え,国内で広く利用されている主要なオンライン麻雀プラットフォーム（以下、プラットフォーム A）の中～上級者卓から 90 対戦分のデータを収集した.

対象母集団: 本研究の目的が「実力評価」の妥当性検証にあることから,プレイヤーの技術介入が勝敗に大きく影響すると考えられる中～上級者レベルの対局を分析対象とした.

データソース: 上記母集団を代表するものとして,プラットフォーム A における「中級者卓」「上級者卓」「最上級者卓」に相当する 3 つのレベル帯の公開牌譜をデータソースとした.

抽出方法: データの恣意的な選択を避けるため,（2025/9/04）以降に記録された公開

牌譜の中から,各レベル帯ごとに,時系列で最も早いものから順に 30 戦分,合計 90 対戦を系統的に抽出した.これにより,サンプルサイズは計 100 対戦となった.

表 2 特定対局における各プレイヤーの局パターン遭遇率とベースラインの比較

プレイヤー名	局数	良型ポテンシャル	悪形ポテンシャル	中間
A さん	8	12.5%	62.5%	25.0%
B さん	8	12.5%	12.5%	75.0%
C さん	8	37.5%	0.0%	62.5%
D さん	8	25.0%	37.5%	37.5%
ベースライン	40,000	43.2%	41.1%	15.7%

5. 関連研究との比較

近年,将棋や囲碁における人工知能研究では,局面評価関数の学習や良形・悪形に関する統計的分析が進展している(例:AlphaGo,AlphaZero など). これらの研究は主に「最適手の探索」を目的としており,プレイヤーが直面する状況的な不利や偏りの検出を目的とするものではない. 本研究は,従来の「最適性」に関する研究とは異なり,対局が実力評価として公平であったかを統計的に検証する枠組みを提示する点に新規性がある.

6. 再現性と手法選択の根拠

本研究で用いたクラスタリングおよび異常検知手法については,以下のように設定した. まず,クラスタ数 k=3 はシルエット係数の最大化に基づいて選定した. 特徴量として指標(entropy_bits, p_runs, p_empirical, markov_p, cramers_v,リーチ順目,ツモ総数,鳴数)を採用したのは,局面の複雑性や配牌・ツモの独立性を定量的に評価するためである. 再現性については,著者が分析に用いたプログラムとデータを GitHub リポジトリに公開している[1]. これにより,他の研究者は手順に従って同一の分析を追試することが可能である. クラスタリングの例示結果を図 3 に示した.※ランダム 4 万局のシミュレーション結果と分布を比較しやすいようにリーチ順目,鳴数,ツモ総数は除外した.

良形(オレンジ色)

- 特徴: 全体的に多角形は小さいが,entropy_bits(複雑性・柔軟性)が比較的

高くなる傾向がある.

- 解釈: これは「完成形に近いが故に選択肢は少ないものの,多くの可能性を秘めた質の高い手」と解釈できる.例えば,高得点が期待できる役の聴牌(テンパイ)に近い状態などがこれに該当すると考えらる.

悪形(緑色)

- 特徴: p_empirical(牌の平凡さ)が突出して高くなる傾向がある.一方で p_runs(連続性)や markov_p(接続性)などは低くなる傾向を示す.
- 解釈: これは「ありふれた中張牌ばかりで,順子や刻子の元となる搭子(ターツ)が少なく,方向性が全く見えない手」を典型的な悪形として捉えていることを示している.いわゆる「バラバラな手」の平均像である.

中間(青色)

- 特徴: 多角形が最も大きく,p_empirical に加えて p_runs や cramers_v(グループ性)なども比較的高く,全体を覆うような分布になっている.
- 解釈: これは「平凡な牌が多いものの,ある程度の連続性やグループ性も持ち合わせている,最も一般的で,良くも悪くもないスタート地点の手」を示している.こからプレイヤーの選択によって良形にも悪形にもなりうる、ポテンシャルを秘めた状態と言える.

アルゴリズムの詳細設定は付録 1 にまとめてある.

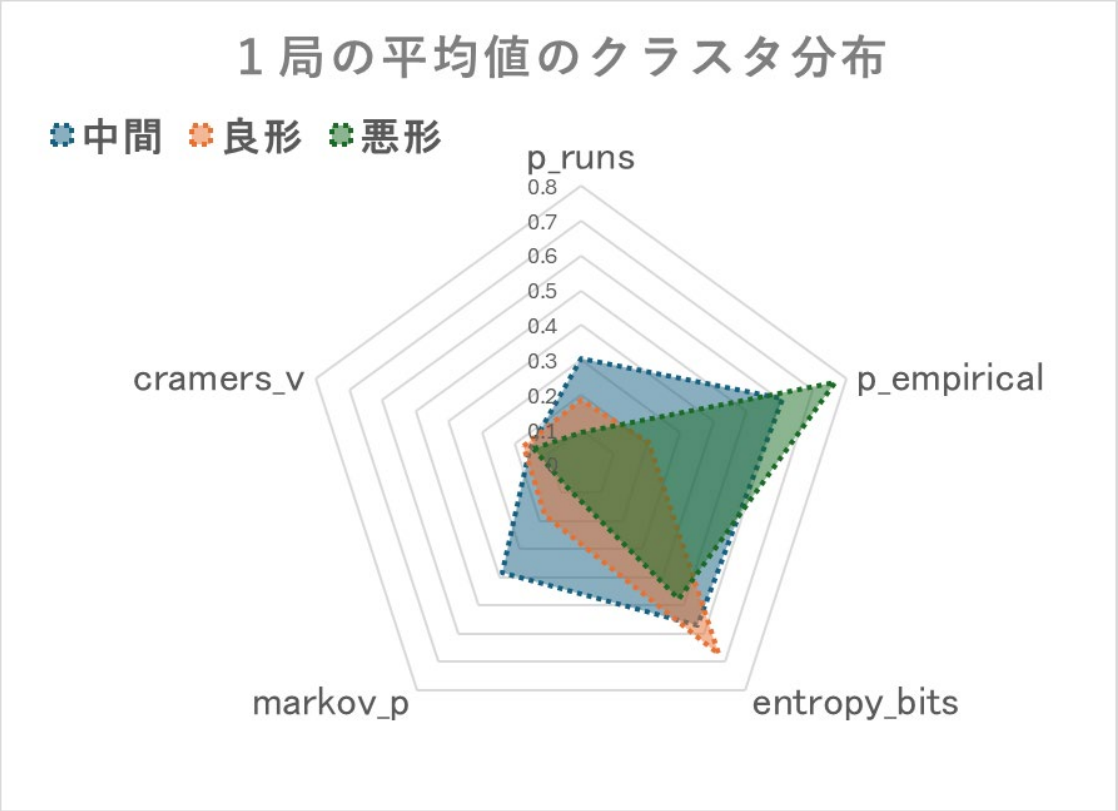


図 3 1 対局データにおけるクラスターの平均的特徴

7. 社会的意義

本研究の成果は,オンライン麻雀を単なる娯楽としてではなく,e スポーツ競技として捉えた際の公平性評価に新たな視点を提供するものである. 具体的には,牌山の極端な偏りをスポーツ競技における「不正な外部要因」に例えることができる. 例えば,陸上競技においてドーピングによって一部の選手が不当な優位を得た場合や,競技コース上に障害物や異常な環境条件が存在した場合,記録や順位は「実力の正当な反映」とは認められない. 同様に,麻雀においても牌山の状態が統計的に大きく逸脱している場合,その対局結果を基にプレイヤーの実力を評価することは妥当ではない. 本研究は「不正を検出する」ことを目的とするものではなく,むしろ「実力を正しく測定できる条件を定義する」ことを目指している. その意味で,スポーツ倫理における「公正な競技環境」の概念を,e スポーツ麻雀に適用する枠組みを提示したと言える. この視点を踏まえれば,業界においては牌山生成アルゴリズムの第三者認証が不可欠となる. これはスポーツにおけるアンチドーピング規制や競技環境の公平性チェックに相当する役割を果たすものであり,オンライン麻雀を持続可能な競技文化として発展させるための基盤となる.

8. 限界と展望

限界 本研究で対象とした実データは著者自身の 10 対局とプラットフォーム A における「中級者卓」「上級者卓」「最上級者卓」に相当する 3 つのレベル帯から恣意性を排除した系統データ 30 戦分の計 100 戦に限定されており,サンプルサイズが小さい. さらに,検出された偏りは特定のプレイヤーや時期に依存している可能性があるため,一般化には慎重さを要する.

展望 今後は,数千対局規模の実データを収集することで統計的な有意性をより強固に検証する予定である. しかし,個人で数千対局のサンプルを収集するのは限界があるため,プラットフォームの協力が不可欠である.また,他のプレイヤー集団への適用によって,個人差や戦略スタイルの影響を分離することも可能になると考えられる. さらに,麻雀以外の e スポーツやカードゲームにも手法を拡張することで,広く「不利な状況下での実力評価」を定量化できるフレームワークへと発展させられる余地がある. その際には,オンラインゲーム業界全体で共有可能な「牌山生成の業界標準化」と「第三者認証」の導入が,次のステップとして求められるだろう.

9. 結論

ここでオンライン麻雀と全自動麻雀卓による麻雀の根本的な違いを比較することで本研究の意義を理解してほしい.全自動麻雀卓ではプレーヤーの目の前で牌がかき混ぜられ,親が振ったサイコロの目によって配牌の開始場所が決まる二段構えの構造で公平性を担保している.結果的に全自動麻雀卓の混ぜ方に意義唱えるプレイヤーは存在しない,対してオンライン麻雀のほとんどのプラットホームでは牌山生成の透明性が担保されていない.この公平性が不透明なオンライン麻雀の実力評価について本研究は,牌譜データを基に実力評価の公平性を統計的に検定する枠組みを提示した. シミュレーションを参照基準とした分析により,特定の対局において統計的に有意な不利が生じ得ることを示し,これがプレイヤーの実力評価に影響を及ぼし得ることを明らかにした. 重要なのは,偏りの原因が偶然か恣意的操作かに関わらず,数値的な基準から大きく逸脱している場合には,その対局は実力を評価する土俵たり得ないという点である. 例えば,10 局すべてにおいて他プレイヤーが有利な配牌やツモを得た場合,それは「運が悪い」で済む話ではなく,その局面ではそもそも実力の比較が成立していない. 本研究の貢献は,原因の追及ではなく,実力評価に用いることが妥当か否かという新しい判定基準を提示する点にある. さらに本研究は,オンライン麻雀の健全化に向けた業界標準化の必要性を示唆しており,第三者認証を伴う業界標準的な牌山生成アルゴリズムの確立が,今後の社会的課題となる. 今後の拡張研究を通じて,オンライ

ンゲームにおける技能評価の客観性と信頼性を高める基盤の構築に寄与できると考える.

10.今後の研究課題

新たな評価指標の提言

新たな**指標値:【特殊手ポテンシャル指標(仮)による数値化】**
今回の検定手法,実力評価が可能な土俵であるかを数値的に可視化する試みの中では採用していない指標だが,例えば,表 3 に示した特定の局の指標値が示す値では「牌の連続性(p_runs, markov_p)が低いにもかかわらず,牌の関連性(cramers_v)は比較的高く,情報量(entropy_bits)も中程度」という相反する相関の組み合わせが確認されている.この相関の組み合わせそのものが,統計的にほとんど現れないパターンだった可能性がある.
これは,七対子(チートイツ)や対々和(トイトイホー)のような,順子系とは異なる特殊な手役に向かいやすい,特異な構成の手牌であった可能性を示唆している.「順子系ではない特殊な手の可能性」を評価する指標として計算式を考案できる. 式は付録 1 参照されたい.大規模サンプルでこの指標の有効性が評価できれば土俵の実力評価する際の指標となる可能性を秘めている.

指標	値
p_value	0.007940835
p_runs	0.048207206
p_empirical	0.1845
entropy_bits	0.583659167
markov_p	0.0461957
cramers_v	0.175520811

表 3 特定局の各指標値

参考文献

[1] Toru1968, "mahjong-performance-analysis", (2024). https://github.com/Toru1968/mahjong-performance-analysis (2025 年 9 月 2 日アクセス).

付録 1. 分析手法の詳細

付録 1.1. 特徴量（指標）の定義

付録 1.1.1. entropy_bits（情報エントロピー）

意味: 牌山や手牌の不確実性の度合いをビット単位で表す.

使い方: 各牌が出る確率 pi を計算し,シャノンエントロピーを求める.

例: 同じ牌ばかり出る偏った牌山は低エントロピー,ランダム性が高い牌山は高エントロピー.

検定内容: 牌山の公平性評価,プレイヤー間のランダム性の比較（不正牌山の検出）.

付録 1.1.2. p_runs（ラン検定の p 値）

意味: 牌の連続パターンがランダムかどうかの統計的指標.

使い方: 同じ色の牌（萬子・筒子・索子）が連続して出る回数を観測し,ランダムな順列から期待される連続回数と比較する. p 値が小さいと「連続が偶然ではない」＝偏りありと判断する.

検定内容: 牌の偏り検出,偏った順番でゲーム進行していないかの確認.

付録 1.1.3. p_empirical（経験的 p 値）

意味: シミュレーションや実データから直接計算した p 値.

使い方: 実際の牌山データを n 回シミュレーションし,観測された統計量（例：特定牌が特定位置に出る頻度）がどのくらい珍しいか評価する.

検定内容: 理論的な完全ランダム分布ではなく,実ゲームデータに基づく公平性評価（例：「このプレイヤーが連続してツモに恵まれているのは偶然か？」）.

付録 1.1.4. markov_p（マルコフ確率／p 値）

意味: 牌の順序性を考慮した確率.

使い方: 牌の出現が直前の牌に依存するかをマルコフモデルで評価する.

検定内容: 牌山生成の偏りを検出,偏った順序が連続するとプレイヤー有利になるかの分析.

付録 1.1.5. cramers_v（クラメールの V）

意味: 2 つのカテゴリ変数の関連度（0〜1）.

使い方: カイ二乗検定から計算する.

検定内容: 特定の条件（役満成立・ドラ牌）と牌の出現の依存度を定量化する. 0.5 以上なら強い依存関係,0.1 未満ならほぼ独立と見なす.

付録 1.1.6. 特殊手ポテンシャル評価指数

特殊手ポテンシャル = (cramers_v × entropy_bits) / (p_runs + markov_p)

牌の関連性(cramers_v)と選択肢の多さ(entropy_bits)が高いほど,そして牌の連続性(p_runs, markov_p)が低いほど高い値を示す.

「順子になりやすい手」は極めて低い値を示すが特殊手は高い値を示す。

※この指標が対局に及ぼす影響は,今後の大規模検証で判断可能となる.

付録 1.2. 分析手法の設定

クラスタリング: K-means を用い,クラスタ数 k=3 とした. これはシルエット係数の最大化に基づいて選定した.

異常検知: Isolation Forest を利用した.

付録 1.3. 再現性

著者が用いた分析データおよびプログラムは GitHub 上で公開されている[1].

公開された手順に従うことで,同一の分析を再現可能であり,研究者が追試できる水準の透明性を確保している.

A Study on Performance Evaluation Criteria in Online Mahjong

Toru Kiyota.

Abstract

This study proposes a new framework for fairly evaluating player skill in online mahjong. Game outcomes in online mahjong are heavily influenced by probabilistic factors such as initial hands and tile draws, meaning short-term results do not always reflect true skill. To address this, we developed a method that uses several statistical indicators (e.g., entropy, runs test) to featurize game data (tile walls and draws) and classifies game situations into "Favorable," "Unfavorable," and "Neutral" forms using unsupervised clustering. A case study of our own 10 matches demonstrates that statistically significant advantages and disadvantages can occur in specific games, potentially affecting player evaluation. The contribution of this research lies not in questioning the cause of bias, but in presenting a new criterion for judging whether a game is a "valid basis for skill assessment." This approach provides a foundation for ensuring the fairness of mahjong as an e-sport and suggests the future necessity of establishing an industry-standard tile-shuffling algorithm with third-party certification.

Keywords
Online Mahjong, Skill Evaluation, Statistical Test, Clustering, e-Sports, Fairness