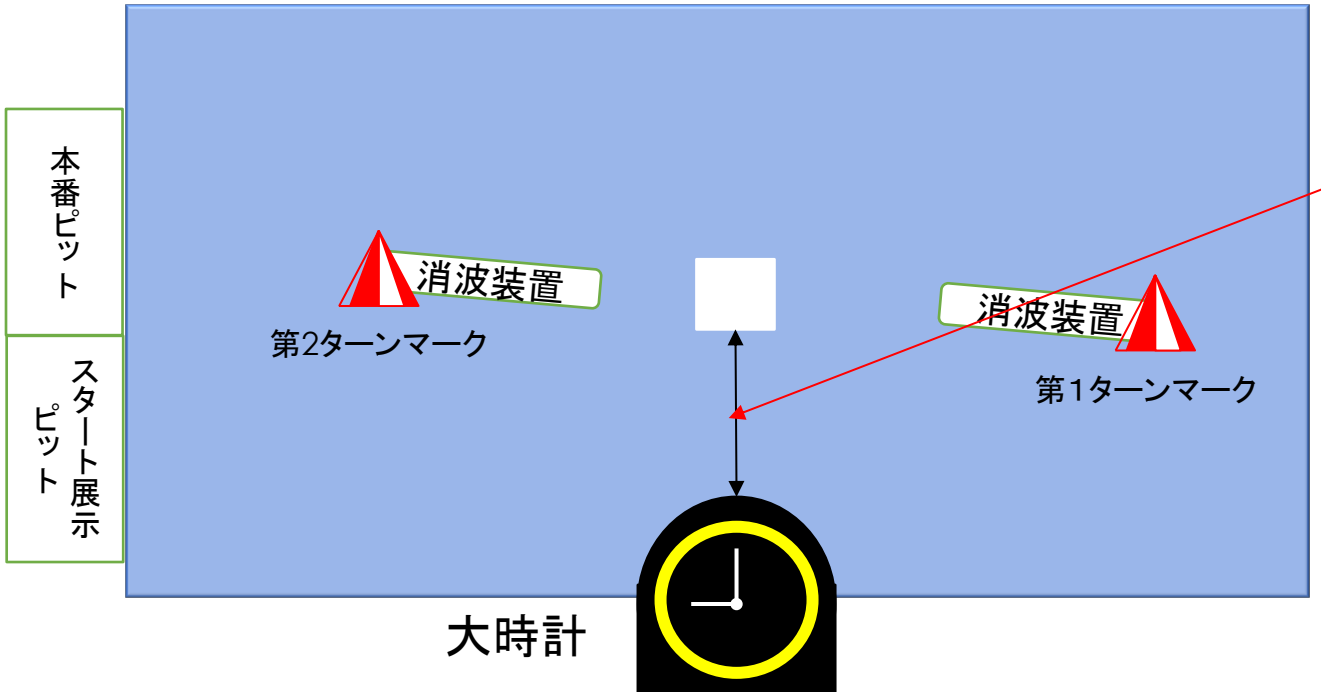
レース場の特徴量を工夫したボートレースの結果予測モデルの提案

大阪府立大学　現代システム科学域知識情報システム学類
名越　翔, 足立　匠, 玉田　拓也, 朴　太一

ボートレース(競艇)とは

- 6艇のモーターボート(1~6号艇)に選手が乗り、競争水面のコースを反時計回りに3周する競技
- レース開始直前にスタート展示ピットから選手が出発し、消波装置・第2ターンマークを回り各選手がスタート位置につく

ボートレース場の水面模式図

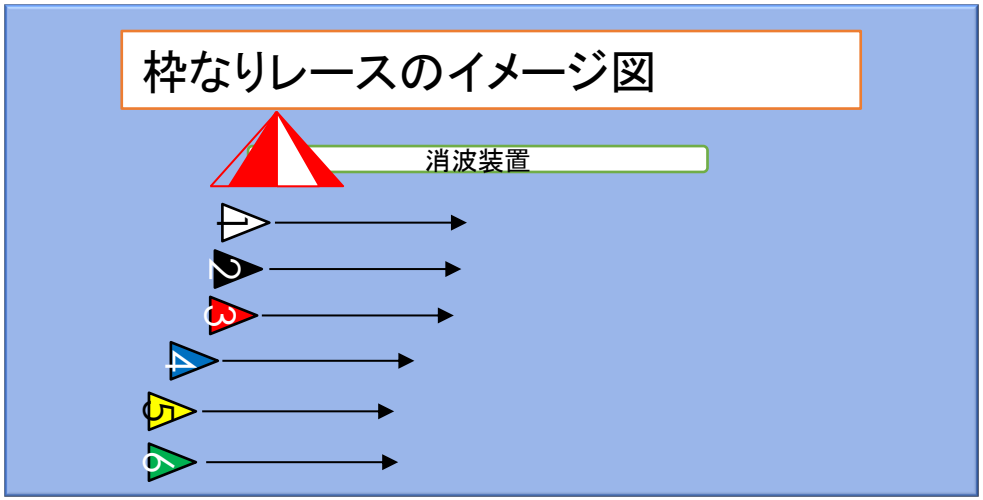


水面でボートを同じ場所にとどめることが難しいため、水面にある大時計が0秒を指してから1秒以内にスタートライン（大時計の前）を通過すればよいルールとなっている

ボートレースの競技としての特徴

- 一般的にコースの内側を走る1コースが最も有利な位置で
  逆に6コースが外側の最も不利な位置となり、1コースの勝率が最も高い

- コースの選択はスタート展示ピットからスタート位置につくまでの選手の行動によって行われ
  1号艇が1コース,2号艇が2コースと艇番号とコースが一致するレース(枠なりレース)が70%近くである
  逆に艇番号とコースが一致しないことを前付けと呼ぶ。これは、選手の中に新人選手や、他の選手と大きく年の離れたベテラン選手がいる場合が多い



ボートレースの競技としての特徴

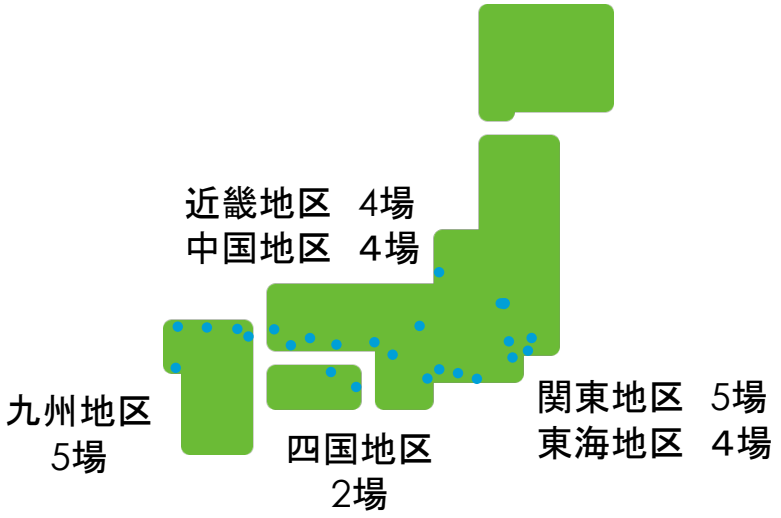
- 選手はA1,A2,B1,B2の4階級があり、半年ごとの成績によって与えられる。
  全選手（約1600人）のうち、成績の上位20%がA1,上位20%~40%がA2,上位40%~90%がB1級,それ以外がB2級となるが、主に新人選手により構成される
- 階級間で選手の実力に大きな差が生じ、上位選手で構成されるA1級と新人選手以外の下位選手で構成されるB1級とを比べると全体の勝率では2.45倍,有利な1コースでの勝率では1.8倍の差が生じる
- 多くのレースでは実力に差がある選手同士で同じレースを走る

階級	分かれ方	全体の勝率	1コースでの勝率
A1	成績上位20%	27%	72%
A2	成績上位20%~40%	20%	60%
B1	成績上位40%~90%	11%	40%
B2	A1,A2,B1以外	5%	32%

ボートレース場について

- 日本には24場の競艇場があり、海水や淡水など水質の違いや、場所による気候の違いが存在する。これにより、レース結果においても異なる特徴が存在する
- ボートレースは1つの大会に4~7日ほど期間があり、初日に選手はくじ引きで自身が使うモーターを決める
  そのモーターを自身の好みやレース場・天候等に合わせて日々整備し、レースに臨む。
  また、モーターの性能はすべて等しいわけではなく、強弱が存在する

全国24場のボートレース場の分布



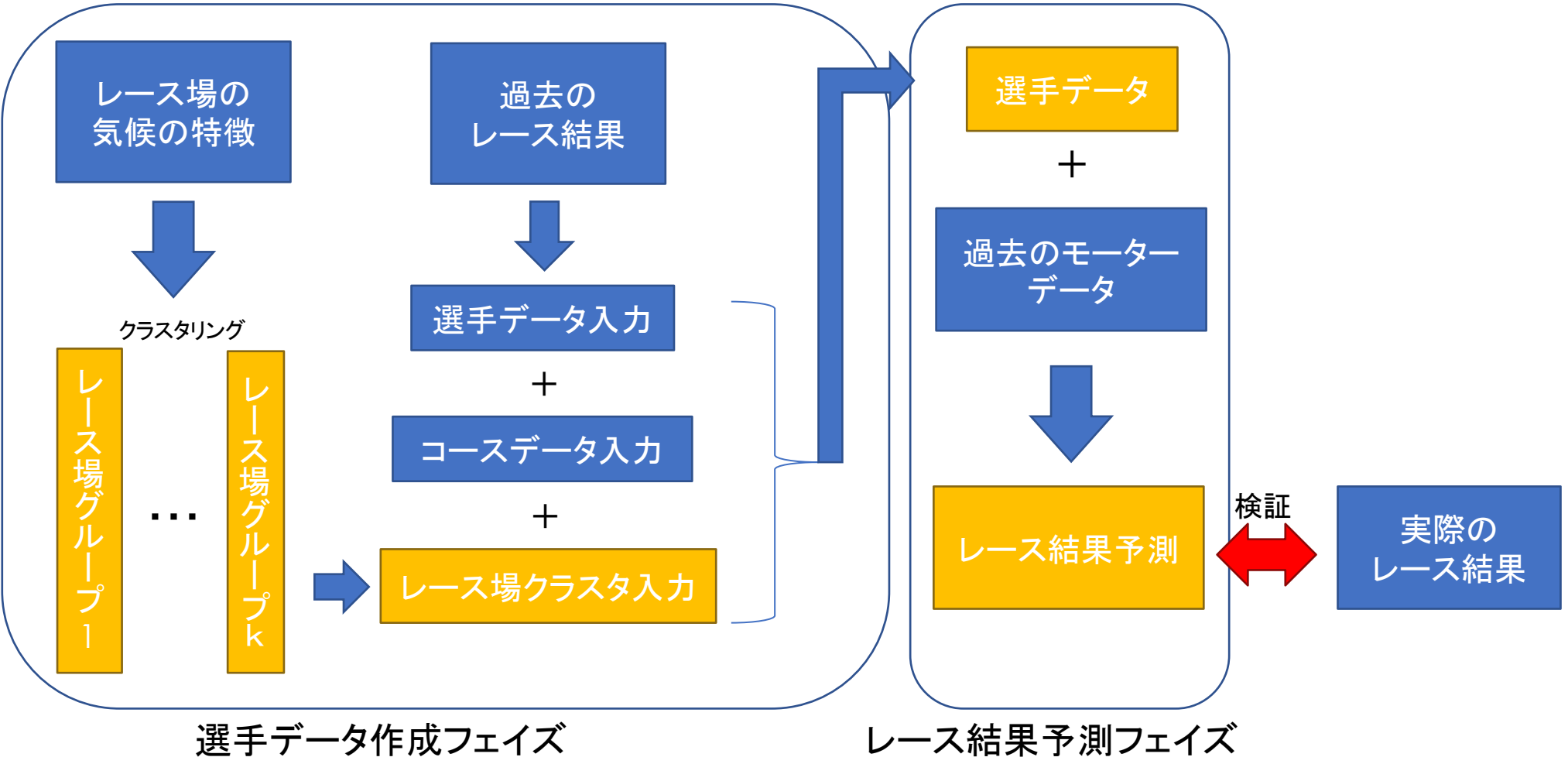
結果予測モデルの目的

- 目的
  前述したように実力に差がある選手同士で同じレースを走ることが多く、実際に上位選手の勝率は高くなっているが下位の選手でも1コースを走る場合は比較的高い勝率を見込むことができるため予想は難しい
  そこで1コースの選手が勝利することができるかを予想することを本予測モデルの目的とする

- 予測方法
  前述の気候や水質で競艇場のクラスタリングを行う
  そのうえで、選手データとコース、競艇場クラスタ等の組み合わせで
  そのシチュエーションでの選手の能力データの作成を行う

これを各レース6人全員に行い、結果を決定木モデルで予測する

結果予測モデル


レース場の
気候の特徴
過去の
レース結果
クラスタリング
レース場グループ1
…
レース場グループk
選手データ入力
＋
コースデータ入力
＋
レース場クラスタ入力
選手データ
＋
過去のモーター
データ
レース結果予測
検証
実際の
レース結果
選手データ作成フェイズ
レース結果予測フェイズ

レース場クラスタリング：目的

- 前述の理由からレース場による選手の得意不得意が生じるため結果を予測する場合は当該レース場での成績を使用する必要がある
- しかし各選手のレースへの出場は年間約10～30回となっており、また全レース場へ等しく出場するわけではないためレース場ごとの出走数に隔たりが生じる
- そのため気象条件によりレース場をクラスタリングし各クラスタの場での選手のレース結果を用いることでデータ量の不足を補う
- データが増えることでの安定した結果予測とそのレース場で初出場、もしくは久しぶりの出場となった選手の結果予測に対しても高い精度が期待される

レース場クラスタリング：手法・変数

- 手法
  Visual Mining Studioを使用し
  k-means法・階層型クラスタリング(ward法)で検討

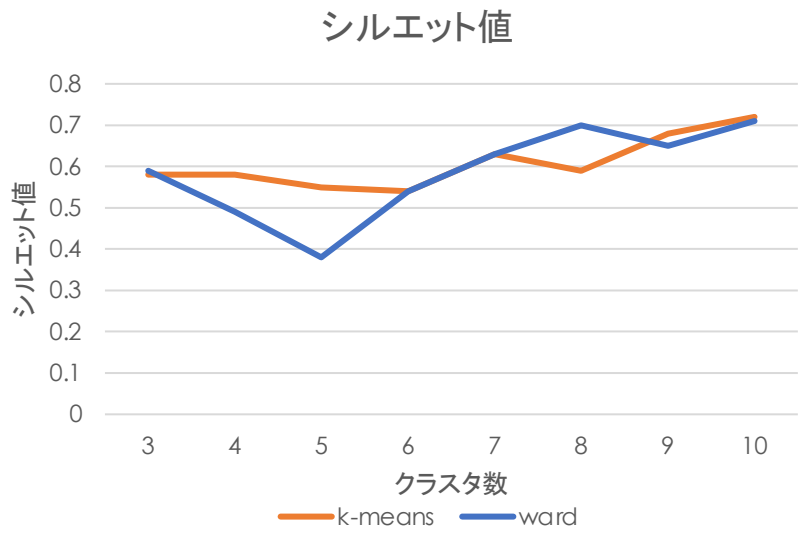
- 使用した変数
  - 天候(気温・波高・風速)
  - レース間の気温差：大きいほどモーターの調整が難しい
  - 水質：海水・淡水・汽水(海水と淡水が混ざっている)の3種類を3段階の数値で表現
  - スタートタイム分散：大きいほどスタートが難しい
  - レースタイム平均：短いほど水面が穏やか,走りやすいレース場

レース場クラスタリング：結果

k-means法,階層型クラスタリング(ward法)のそれぞれのクラスタ数でシルエット値を計算、右下の図のようになった

k-means法は、クラスタ数10で最高値 0.72 をとり
ward法は、クラスタ数10で最高値 0.71 をとった

しかし、各クラスタ内でのレース場の数が1のクラスタが多く発生したため
k-means 法でのクラスタ数7を採用した



各クラスタの特徴

1:水質が0(淡水)で風強め、波高め
2:風が非常に強く、波も非常に高い
3:水質が2(海水)
4:水質が0(淡水)で風弱め、波低め
5:気温が高く、風が吹かず、波も低い
6:気温差大きい
7:水質が1(汽水)

ID	クラスタ	レース場の数	レースタイム平均	気温	波高	平均風速	気温差	スタート分散	水質
1	淡水・荒れ水面クラスタ	2	109.621	16.186	2.307	2.806	3.097	0.250	0
2	大荒れ水面クラスタ	1	112.866	19.009	8.777	4.916	2.368	0.329	1
3	海水クラスタ	9	109.913	18.356	2.654	2.695	2.729	0.241	2
4	淡水・穏やか水面クラスタ	6	109.410	18.597	1.752	2.270	2.942	0.234	0
5	温暖静穏クラスタ	1	109.426	21.099	1.543	2.064	3.944	0.226	2
6	調整難クラスタ	2	109.316	18.724	2.346	3.076	4.015	0.268	0.5
7	汽水クラスタ	3	110.080	18.493	2.269	3.053	1.999	0.276	1

表:k-means法のクラスタ数7における各項目の平均

選手データ:目的

- ボートレースは水面を走るスポーツであり、選手が走るルートによって波が発生し、他の選手に影響を与えるため
  選手の強さだけでなく、走り方の特徴を指標に表す変数を計算した

- 走り方の特徴を表すために、決まり手を使用した
  決まり手とは、そのレースの1着がどのような勝ち方をしたのかを表すもので、主なものは逃げ、差し、まくり、まくり差しの4種類が存在する

- ある選手がレースを走ったときのこれらの決まり手の発生率を計算することで、その選手の走り方を数値で表す

選手データ：用いたデータ・変数

- 用いたデータ
  全国24場のボートレース場における2019年10月1日~2021年10月31日までのレース出走表・結果
  過去１年間を変数作成に使用

- 説明変数
  - 勝率・２連対率・３連対率
  - スタートタイミングの平均
  - 決まり手の確率(逃げ・差し・まくり・まくり差し)
  - 使用しているモーターの勝率・２連対率・３連対率

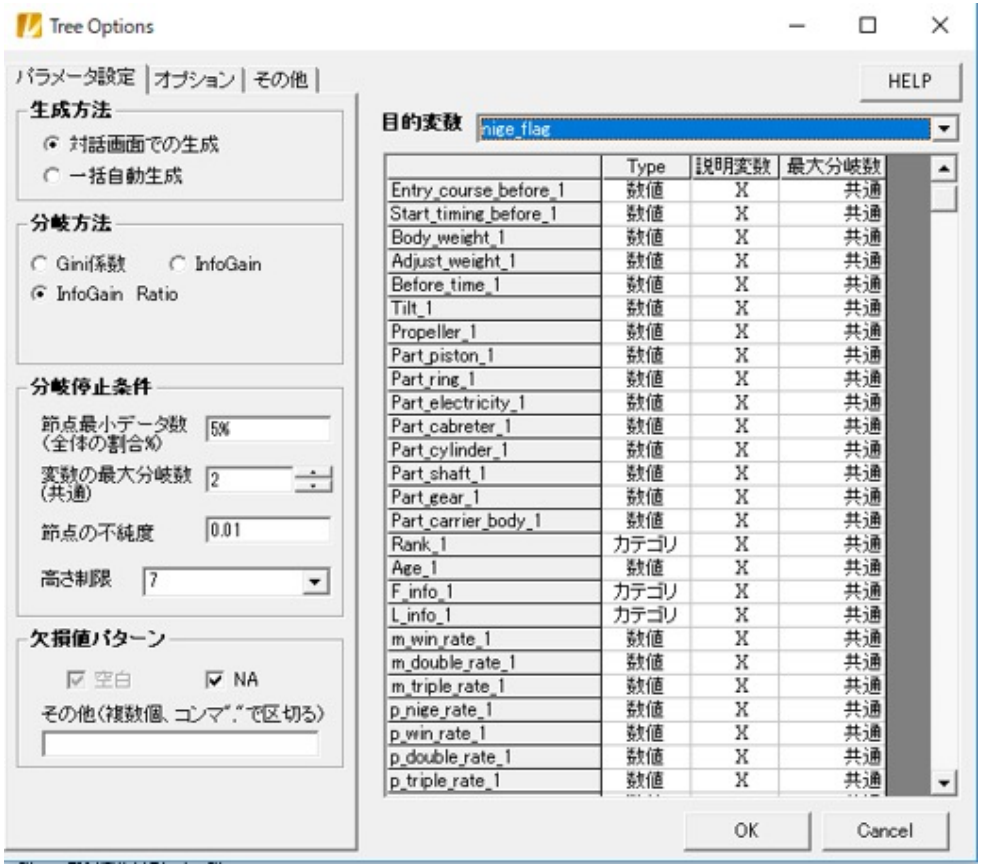
  を各クラスタで選手ごとに作成

- 目的変数
  - 1コースを走った選手(以降1コース選手と呼ぶ)が勝つかどうか

使用モデル・実験環境

- 使用モデル：可読性を確保するため決定木を使用

- 使用した設定
  ＜分岐方法＞
  InfoGain Ratio
  ＜分岐停止条件＞
  節点最小データ数：5%
  変数の最大分岐数：2
  高さ制限：7



使用モデル・実験環境

- 比較手法
  - レース場クラスタリングあり
  - レース場クラスタリングなし（全レース場をまとめる）
  - レース場クラスタリングなし（各レース場を1クラスタとする）
- テスト方法：ホールドアウト法(train:test=7:3)でランダムに分割
- 評価指標
  - accuracy
  - recall
  - precision
  - F値

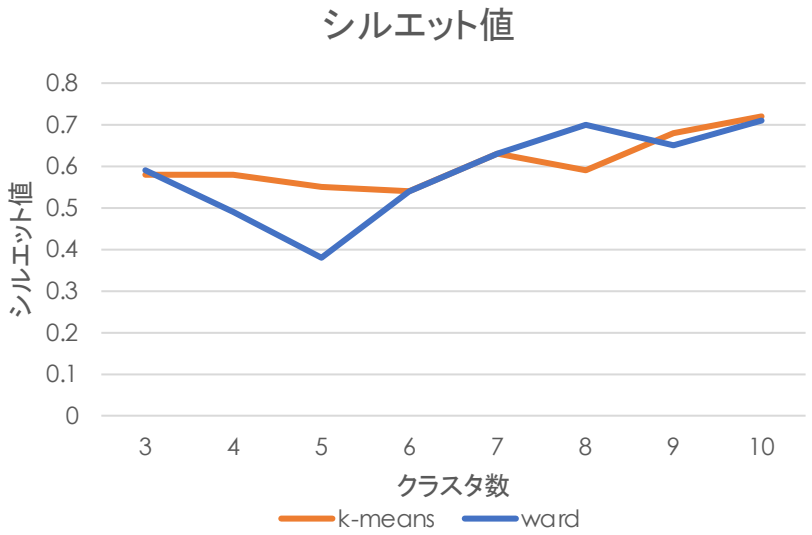
決定木分析・実験

- レース場クラスタリングをした場合、
  全てのレース場を1つのクラスタとした場合と比べて
  特にrecallにおいて、有意に精度が向上した
- また各レース場をそれぞれ1クラスタとした場合と比べてもデータ量不足を補いつつも同等の精度を確保することができた
- レース場のクラスタリングは、1コースが勝つレースを多く選択する上で有効であることが示された

クラスタリング条件	accuracy	recall	precision	F値
クラスタ数：７	0.63	0.79	0.62	0.69
全レース場をまとめる(クラスタ数：１)	0.64	0.67	0.66	0.66
レース場を個別に見る（クラスタ数：２４）	0.62	0.81	0.63	0.71

クラスタ数の調整①

- 適切なクラスタ数を調査するために、各クラスタに属するレース場が多くなるようにクラスタ数を減らして比較を行う
- クラスタ数は下図のシルエット値と各クラスタのレース場の数からクラスタ数が4で再計算した



ID	クラスタ	レース場の数	レースタイム平均	気温	波高	平均風速	気温差	スタート分散	水質
1	穏やか水面クラスタ	10	109.43	18.14	1.98	2.54	3.19	0.24	0.1
2	大荒れ水面クラスタ	1	112.87	19.01	8.78	4.92	2.37	0.33	1
3	汽水クラスタ	3	110.08	18.49	2.27	3.05	2	0.28	1
4	海水・温暖クラスタ	10	109.86	18.63	2.54	2.63	2.85	0.24	2

クラスタ数の調整②

- クラスタリング数7と4を比較すると
  評価指標のaccuracy, recall, F値において

  大きな変化が見られなかったが
  recallにおいてはクラスタ数7で優位となることが確認された

クラスタリング条件	accuracy	recall	precision	F値
クラスタ数： 7	0.63	0.79	0.62	0.69
クラスタ数： 4	0.63	0.74	0.63	0.68

モデルの検証

- 実際のレースを用いて、今回のクラスタ数を7に設定したモデルと全レースの1コース平均勝率（計算すると約56%）の確率での予測の精度を検証する
- 検証方法
  対象：テストデータ約13000レース
  決定木モデルの予測と平均勝率約56%での予測を的中率を比較する

- 結果
  単純な平均勝率を用いたモデルより的中率は8.2ポイント高いため、分類予測モデルの有意性を示す結果となった

モデル	的中率
クラスタ数：7	62.3%
平均勝率	54.1%

※平均勝率での的中率は100回繰り返した結果の平均

決定木分析：木の詳細

可読性の高いモデルである決定木を採用したが
主な分岐の要因になっている変数を見ると
1コース選手の勝率・階級に大きく依存しているため

1コース選手の勝率やそれに準ずる項目を使用しない場合の各要素の影響度も考察することとする

分岐している変数	変数の説明
Rank_1	1コース選手の階級
p_win_rate_1	1コース選手の勝率(1着確率)
Rank_3	3コース選手の階級

決定木の一部

決定木分析：変数比較

①説明変数：すべて

②説明変数：1コース選手の勝率やそれに準ずる項目以外

の2つの場合で、最初のノードから強制分岐メニューを開き

上位に現れる変数の変化を見る

①で現れた変数（説明変数重要度）	②で現れた変数（説明変数重要度）
1コース選手の階級(0.047)	1コース選手がまくりで負ける確率(0.014)
1コース選手の1着確率(0.033)	1コース選手の平均スタートタイミング(0.013)
1コース選手の2着以内確率(0.031)	1コース選手が差しで負ける確率(0.012)
1コース選手の3着以内確率(0.026)	1コース選手がまくり差しで負ける確率(0.012)
1コース選手がまくりで負ける確率(0.014)	1コース選手の年齢(0.008)

決定木分析：変数比較

下表を見ると①でも②でも1コース選手が勝つかどうかにおける
重要な説明変数としては、1コース選手に関係するものしかなかった
つまり、実力の異なる選手同士で行うことが多いボートレースだが
有利な1コースの選手が勝つかどうかには
周りの選手が大きな影響を及ぼしているわけではないことが考えられる

①で現れた変数（説明変数重要度）	②で現れた変数（説明変数重要度）
1コース選手の階級(0.047)	1コース選手がまくりで負ける確率(0.014)
1コース選手の1着確率(0.033)	1コース選手の平均スタートタイミング(0.013)
1コース選手の2着以内確率(0.031)	1コース選手が差しで負ける確率(0.012)
1コース選手の3着以内確率(0.026)	1コース選手がまくり差しで負ける確率(0.012)
1コース選手がまくりで負ける確率(0.014)	1コース選手の年齢(0.008)

※前スライドと同様の表

結論→課題

結論

- 我々が提案したボートレースにおけるレース場のクラスタリングを用いた決定木モデルはデータ量を数倍に増やした上で、あまり精度を下げないことが示された
- モデルに可読性のある決定木を用いたことで重要な変数が1コース選手の能力であることが示された

課題

- recallに関しては他評価指標に比べ高くなったが、本来の目的でもあるprecisionの精度に課題の残る結果となった
- データ量が増えたことを活用した成果を示すことができなかった