卒業論文

ランダムフォレストによるサッカー W 杯の優勝国予想

2014011　石田亜斗武

指導教員　六井 淳 教授
副査　渡邊貴之 教授

2024 年 1 月

静岡県立大学経営情報学部

概要

スポーツの勝敗予測は、選手のコンディション、戦術の変化、チームの相互作用など、多数の変数に影響される。本研究では、私自身の長年のサッカーの経験と、人工知能や機械学習の知識を使い、FIFA ワールドカップ優勝国の予測を行う。

本研究ではスポーツの勝敗予測における変数の複雑性に対応するため、ランダムフォレストという機械学習手法を用いる。ランダムフォレストはモデル構築とチューニングの容易さから、活用範囲の広い手法であることが知られている。FIFA ワールドカップの優勝国を予測するために過去の大会の結果からデータセットを作成し、ランダムフォレストで予測モデルを構築し予測を行った。

Abstract

Predicting winners and losers in sports is extremely difficult because it is influenced by numerous variables, such as player conditions, tactical changes, and team interactions. In recent years, statistical methods and machine learning have been used to improve the accuracy of predicting the outcome of difficult-to-predict sports matches. Therefore, in this study, I will use my own many years of soccer experience and knowledge of artificial intelligence and machine learning to predict the FIFA World Cup winning country.

To cope with the complexity of variables in sports win/loss prediction, this study uses a machine learning method called random forests. Compared to other methods, Random Forest captures the diversity and complexity of the data by utilizing a large number of decision trees, reducing the risk of overlearning and providing reliable prediction results. In addition, it is easy to build and tune models, enabling results to be achieved within a limited research timeframe. The results of this research are expected to provide useful information for those involved in the FIFA World Cup in response to the growing use of data in the field of sports analysis.

目次

第 1 章

はじめに

1.1 研究背景

スポーツの勝敗予測は、選手のコンディション、戦術の変化、チームの相互作用など、多数の変数に影響されるため、予測が非常に困難である [1]。近年では、予測が困難なスポーツの試合結果を科学的に分析するアナリティクスの分野が注目され、予測の精度を高めるために統計的手法や機械学習が利用されるようになっている [2]。本研究では、私自身の長年のサッカーの経験と、人工知能や機械学習の知識を使い、FIFA ワールドカップ優勝国の予測を行う。

1.2 研究目的

本研究の目的は、過去の FIFA ワールドカップのデータをもとに高精度な予測モデルを構築し、優勝国を予測することである。サッカーの試合結果予測においては、様々な機械学習手法が採用されている。イングランドのプロサッカーリーグの試合結果予測を行った研究 [3] では、決定木 [4] によってデータの特徴間の関係をモデル化しやすくし、ナイーブベイズ [5] によって特徴間の独立性を仮定することで、サッカーのような多くの変数が影響するデータに対応し予測を行っている。また同じくイングランドのプロサッカーリーグに対して試合結果予測を行った別の研究 [6] では、サポートベクタマシン [7] の過学習 [8] を防ぎながら限られたデータからでも高い予測精度を達成するという特性を活かして予測を行っている。

本研究ではこれらの手法の中からランダムフォレスト [9] を採用する。ランダムフォレストは、複数の決定木を組み合わせることでデータの多様性と複雑な関連性を捉え、過学習のリスクを減らすことができる [10][11] ため、ナイーブベイズの変数同士の複雑な関係やデータパターンを捉えるのには限界があるという欠点 [12] を解決することでき、決定木の訓練データに過剰に適合する傾向がある [13] という問題を解決することができる。また、ランダムフォレストにはモデル構築とチューニングの容易であるという特徴があり [14]、サポートベクタマシンのモデル設計やチューニングが困難であるという欠点 [15] を解決することができる。

さらに、本研究ではランダムフォレストを用いた先行研究 [16][17] では取り扱わないジニ係数 [18] を特徴量としてを加える。ジニ係数とは国の貧富格差を示す指標であり、貧困がサッ

カーにおけるスキルの習得と知覚運動スキルの発達にプラスの影響を与えるといった研究 [19] もある。これは、従来のサッカーの勝敗予測ではあまり注目されてこなかった選手の心理的側面を考慮することにより、試合結果に影響を与える可能性のある新たな要因をモデルに組み込む試みである。このアプローチにより、単に試合結果の統計的な情報だけでなく、選手の心理状態も試合結果予測の重要な要素として取り入れることができると考える。ランダムフォレストを用いたクラス分類によって FIFA ワールドカップの優勝国予測を目的とした検証実験を行う。

1.3　論文の構成

第 2 章では、本研究の関連技術と用語について述べる。
第 3 章では、検証実験について述べる。
第 4 章では、まとめと今後について述べる。

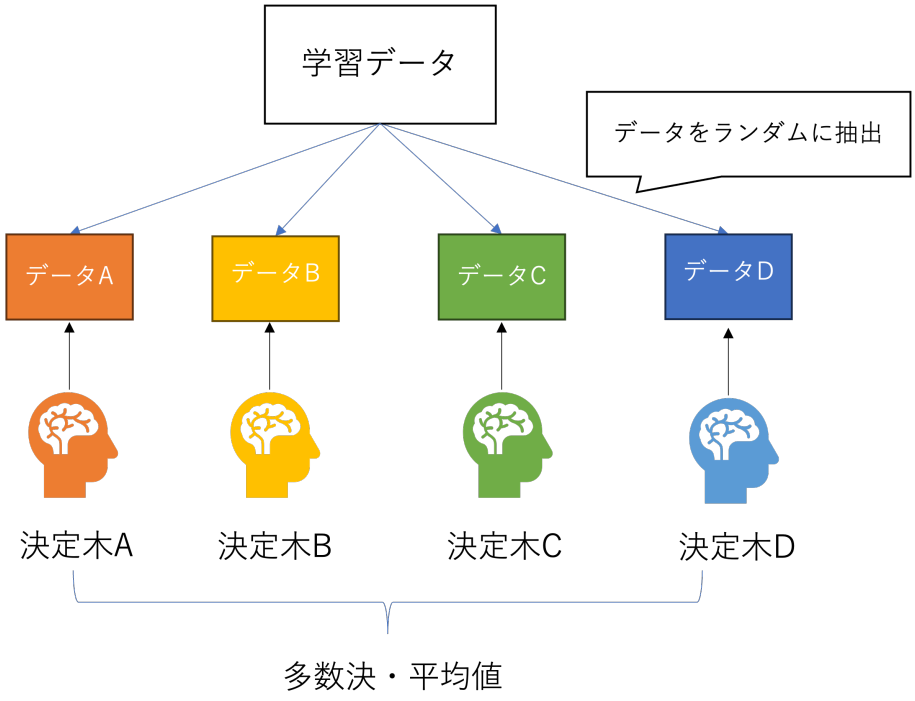
第 2 章

関連技術

本章では、本研究に関連する技術について述べる。

2.1 ランダムフォレスト

ランダムフォレストは、決定木を基にした集団学習アルゴリズムである。この手法では、多数の決定木がデータの異なるランダムなサブセットに基づいて訓練され、それぞれの出力に基づいて最終的な予測結果が多数決または平均値によって決定される (図 2.1)。ランダムフォレストの大きな特徴は、個々の決定木の学習にランダム性を導入している点にある。これによって、各決定木が相補的に機能し、過学習を防ぎつつ高い汎化性能を得ることが可能となる。



ランダムフォレストの大きな特徴は、個々の決定木の学習にランダム性を導入している点にある。これによって、各決定木が相補的に機能し、過学習を防ぎつつ高い汎化性能を得ること

が可能となる。いかにその流れを説明する [20]。

ランダムフォレストの学習メカニズムは、ランダム性を 2 つの主要な側面で取り入れる。まず、バギング（Bootstrap Aggregating）を用いて、各決定木の訓練データを選定する。これは、元のデータセット S から重複を許してランダムにサンプル S_{0t} を抽出するプロセスである。次に、決定木の各ノードでの分割関数の学習にランダム性を導入する。分割関数のパラメータ θ_j に対する候補集合 τ は、一般に広範囲にわたるが、そのサブセット $\tau_j \subseteq \tau$ のみを使用し、その選択をランダムに行う。

ランダム性の度合いは、比率 $|\tau_j|/|\tau|$ によって調整可能である。ここで、$\rho = |\tau_j|$ はランダム性の指標として機能し、ρ の値に応じて、木間の相関が変化する。例えば、$\rho = |\tau|$ の場合、すべての決定木で同一の学習データが使用されるため、ランダム性はない。逆に、$\rho = 1$ の場合、完全にランダムな選択が行われ、互いに相関のない決定木が得られる。

各決定木の構築では、根ノードから始めて、最適な分割関数を選択し、データを左右の子ノードに割り当てる。この過程は、すべてのノードにおいて再帰的に繰り返される。具体的には、ノード番号 i のデータセット S_i が分割関数によって分割され、左右の子ノードに対応するデータセット S_{iL} と S_{iR} に割り当てられる。このプロセスは、決定木のすべてのノードで順に実行され、各ノードで $S_j = S_{jL} \cup S_{jR}$ および $S_{jL} \cap S_{jR} = \emptyset$、$S_{jL} = S_{2j+1}$、$S_{jR} = S_{2j+2}$ が成立する。

分割関数 $h(v, \theta_j)$ の学習は、ノード j における最適なパラメータ θ_j^* を選択するプロセスであり、以下のように定義される：

$$\theta_j^* = \arg\max_{\theta_j \in \tau_j} I_j(S_j, S_{jL}, S_{jR}, \theta_j) \tag{2.1}$$

ここで、I_j は情報利得を表す目的関数である。情報利得は以下のように計算される：

$$I_j = H(S_j) - \sum_{i \in \{L,R\}} \frac{|S_{ji}|}{|S_j|} H(S_{ji}) \tag{2.2}$$

ここで、$H(S)$ はデータセット S のエントロピーであり、S_{ji} はノード j の左右の子ノードに割り当てられるデータセットを示す。

以上の学習アルゴリズムは、決定木のノードに割り当てられる学習データ集合の数が 1 になるまで継続される。ただし、過学習を防ぐために適切な停止条件を設定する必要がある。そのための方法としては、あらかじめ設定した最大高さ D に達した場合、ノードに割り当てられた学習データの個数が一定値以下になった場合、または分割による情報利得が一定値以下になった場合などがある。

2.1.1　特徴選択

ランダムフォレストは特徴量選択をすることで、各決定木でランダムに選択された特徴量のサブセットを使用することで、データセットの異なる側面を捉え、全体としてバランスの取れた分析を行うことができる。本研究で用いる scikit-learn での分類問題における計算を例に以下に説明する [21]

特徴量の重要度の計算においてジニ不純度が重要である。ジニ不純度 $G(k)$ は、特定のノードにおけるサンプルの分布の均一性を測定する。式

$$G(k) = \sum_{i=1}^{n} p(i) \times (1 - p(i)) \tag{2.3}$$

は、ノード k におけるターゲットラベル i の出現確率 $p(i)$ を用いて、そのノードの不純度を算出する。完全に均一なサンプルの集合（すべてが同じラベル）の場合、ジニ不純度は 0 となり、サンプルが多様なラベルを持つほどジニ不純度は高くなる。

特徴量の重要度 $I(j)$ は、特徴量がデータの分割にどれだけ効果的に寄与しているかを示す。特徴量 j を用いてノードを分割することで得られるジニ不純度の減少量を、すべてのノードにわたって集計する。この計算は、式

$$I(j) = \sum_{i=1}^{n \in F(j)} (N_{parent}(i) \times G_{parent}(i)) - (N_{leftchild}(i) \times G{*}leftchild(i) + N{*}rightchild(i) \times G{*}rightchild(i)) \tag{2.4}$$

によって表される。ここで、$F(j)$ は特徴量 j が分割に使用されるノードの集合を示し、$N * parent(i)$、$N_{leftchild}(i)$、$N * rightchild(i)$ はそれぞれ親ノードとその子ノードのサンプル数を表す。$G * parent(i)$、$G_{leftchild}(i)$、$G * right\_child(i)$ はそれぞれのノードにおけるジニ不純度を示す。

このようにして計算された特徴量の重要度は、モデルにおける各特徴量の寄与度を示し、モデルの解釈に役立つ。特に、予測結果に大きな影響を与える特徴量を特定することで、より効果的な特徴量エンジニアリングやモデルの改善に繋がる。また、データセット内の特徴量間の関係性を理解する上で重要な手がかりを提供する。

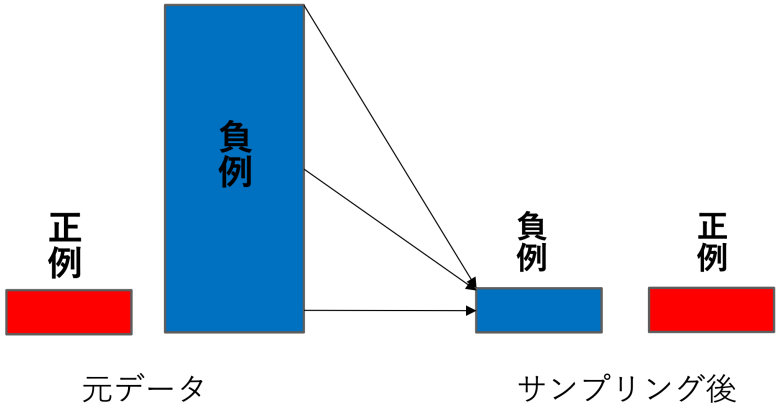
2.2 データの前処理

本章では、データ不均衡問題への対処法として本研究で用いるアンダーサンプリングと SMOTE（Synthetic Minority Over-sampling Technique）の適用について説明する [22]。これらの技術は、特に分類問題におけるクラス間の不均衡を解消し、モデルの予測性能を高めるために重要である。データセットにおけるクラス間の不均衡は、分類モデルにおいて重大な課題である [23]。正例（少数クラス）と負例（多数クラス）のデータが不均等である場合、分類モデルは多数クラスの特徴を過剰に学習する傾向があり、少数クラスの重要な特徴を見落とす可能性が高まる。これは、少数クラスの予測精度が著しく低下する結果を招き、誤った結果を導いてしまう可能性がある。

2.2.1 アンダーサンプリング

アンダーサンプリングは、多数クラスのデータをランダムまたは戦略的に減少させる手法であり、データセット内のクラス間のバランスを改善する (図 2.2)。この手法は、分類モデルが多数クラスに偏らず、少数クラスの特徴を適切に学習する機会を提供する。アンダーサンプリ

ングは、データセットから重要な情報を保持しつつ、クラス間の不均衡を軽減することで、モデルの全体的な分類精度を向上させる効果が期待される。



2.2.2　SMOTE

SMOTE（Synthetic Minority Over-sampling Technique）は、少数クラスの既存サンプルから新たな合成サンプルを生成する手法であり、少数クラスのデータ量を増やすことでデータセットのバランスを改善する。具体的には、SMOTE では以下のステップでサンプルを生成する。

1. マイノリティクラスのサンプルからランダムに一つ選択する
2. 選択したサンプルの最も近い隣人（通常は k-最近傍法により選定）を見つける。
3. 選択したサンプルとその隣人との間でランダムな点を生成する

このプロセスを繰り返すことで、マイノリティクラスのサンプル数を増やし、データセットのバランスを改善し、モデルが少数クラスの特徴をより深く学習することで偏りのない予測を行えるようにする。SMOTE によって増やされたデータは、モデルのトレーニング中に多様性を提供し、特に少数クラスの識別能力を強化する効果がある。

2.2.3　アンダーサンプリングと SMOTE の統合

本研究では、アンダーサンプリングと SMOTE を組み合わせることで、多数クラスのデータ削減と少数クラスのデータ増加を効果的に行い、クラス間のバランスを適切に調整する。この統合されたアプローチは、分類モデルが全クラスの特徴を均等に学習し、各クラスに対して公平な予測を行うための基盤を提供する。特に、クラス間の不均衡が顕著なデータセットにおいて、この組み合わせはモデルの予測性能の向上に貢献することが期待される。

2.3 ハイパーパラメータチューニング

ハイパーパラメータとは、モデルの学習プロセス中に事前に設定されるパラメータであり、モデルの性能に直接影響を与える。特にランダムフォレストモデルの場合、ハイパーパラメータはモデルの複雑性や学習能力を決定するため、その設定はモデルの成功に不可欠である。ハイパーパラメータの適切な設定は、モデルがデータから複雑なパターンを把握し、予測精度を最大化する上で重要な役割を果たす。これは、モデルがデータをどのように処理し、学習するかを決定するためである。例えば、決定木の数や各木の最大深度は、モデルの複雑さと一般化能力のバランスに影響を与える。過度な複雑さは過学習を引き起こす可能性があり、一方で単純すぎるモデルは未知のデータに対して十分な予測能力を発揮できない。ハイパーパラメータの適切な設定は、モデルの性能を向上させるだけでなく、データの特性や機械学習アルゴリズムの挙動に対する理解を深める効果もある。したがって、ハイパーパラメータの選択とチューニングは、データサイエンスの分野において、効果的なモデル構築に不可欠なスキルであると言える [24]。

ランダムフォレストモデルにおけるハイパーパラメータの適切な設定は、モデルの性能を決定する上で重要な役割を果たす [25]。本研究で用いる scikit-learn の RandomForestClassifier で設定するハイパーパラメータには、決定木の数（n_estimators）、最大深度（max_depth）、最小サンプル分割（min_samples_split）、最小サンプル葉（min_samples_leaf）等がある [26]。決定木の数は、モデルによって生成される決定木の総数を表し、モデルの堅牢性と直接的に関連する。木の数が多ければ多いほど、モデルはより詳細なデータ構造を学習する可能性が高まるが、同時に計算コストも増加する。したがって、モデルの性能と計算効率を考慮して、最適な木の数を決定する必要がある。最大深度は、モデルが学習する際に各決定木が到達することのできる最大の深さを定義する。深い木はより複雑なデータ構造を捉えることができるが、過学習を引き起こすリスクも伴う。適切な最大深度の設定は、モデルが過学習せずにデータを効果的に学習するために重要である。最小サンプル分割は、ノードを分割するために必要な最小サンプル数を指定し、モデルの学習プロセスにおける分割の決定に影響を与える。このパラメータが高い値に設定されると、モデルはデータの小さな変動に対して過敏に反応しなくなり、過学習のリスクが低減される。最小サンプル葉は、葉ノードに存在すべき最小サンプル数を定義する。このパラメータはモデルがデータのノイズに対してどれだけ敏感に反応するかを制御し、特にノイズが多いデータセットにおいて重要である。これらのハイパーパラメータは、モデルの予測精度、訓練時間、複雑性に大きく影響を与える。適切なパラメータの選択により、特定のデータセットに適応した最適なモデルの構築が可能となる。

ランダムフォレストモデルの最適なハイパーパラメータを見つけるためには、効果的なチューニング方法が必要である。主要なチューニング手法にはグリッドサーチとランダムサーチがあり、それぞれが特定の利点を持つ [27]。グリッドサーチは、指定されたハイパーパラメータの全ての組み合わせを網羅的に試す手法である。ハイパーパラメータの候補値をグリッド状に配置し、各組み合わせに対してモデルのパフォーマンスを評価する。この方法の利点

は、与えられた値の範囲内で最適な組み合わせを見つける可能性が高いことである。しかし、大きなパラメータ空間を持つモデルでは計算コストが非常に高くなる可能性がある。一方、ランダムサーチは、パラメータ空間内のランダムな点を選択してモデルを評価する手法である。グリッドサーチと比較して、ランダムサーチは計算コストが低く、広範なパラメータ空間の探索が可能である。しかし、ランダム性により、最適なパラメータを見逃す可能性もある。

ハイパーパラメータチューニングは反復的なプロセスであり、異なるパラメータのセットでモデルを評価し、得られた結果を比較して最適な設定を見つける。このプロセスは、モデルの予測性能を最大化し、特定のデータセットに対する最適な適応を実現するために不可欠である。

第 3 章

検証実験

3.1 データ詳細

3.2 評価指標

3.2.1 混合行列

3.2.2 正解率

3.2.3 適合率

3.2.4 再現率

3.2.5 F 値

3.3 検証内容

3.4 検証結果

3.5 考察

第 4 章

まとめと今後の課題

4.1　まとめ

4.2　今後の課題

謝辞

参考文献

[1] Milad Keshtkar Langaroudi, Mohammad Reza Yamaghani,“Sports Result Prediction Based on Machine Learning and Computational Intelligence Approaches A Survey”, P1, 2019

[2] 谷岡広樹, ”スポーツアナリティクスにおけるデータと AI 活用”,P1, 2020

[3] Saurabh Vaidya ,”Football Match Winner Prediction”, P1,P2, 2016

[4] Lucidspark, ”決定木分析とは？メリットとやり方”, https://lucidspark.com/ja/blog/how-to-make-a-decision-tree

[5] AVINTON ,”機械学習入門者向け Naive Bayes(単純ベイズ) アルゴリズムに触れてみる”,https://avinton.com/academy/naive-bayes/

[6] Chinwe Peace Igiri,”Support Vector Machine-Based Prediction System for a Football Match Result”, P1, 2015

[7] AIsmiley, ”サポートベクターマシン（SVM）とは？特徴やメリットと活用事例”, https://aismiley.co.jp/ai_news/svm/

[8] TRYETING”機械学習における過学習とは何か?原因・回避方法をくわしく解説”, https://www.tryeting.jp/column/6846/

[9] IBM”ランダムフォレストとは”,https://www.ibm.com/jp-ja/topics/random-forest

[10] Matthias Schonlau, Rosie Yuyan Zou, ”The random forest algorithm for statistical learning”, P4, 2020

[11] Kai Liang,”Analysis and Evaluation of Sports Effect Based on Random Forest Algorithm under Big Data”,P2, 2022

[12] OpenGenus”9Advantages and 10disadvantages of Naive Bayes Algorithm”, https://iq.opengenus.org/advantages-and-disadvantages-of-naive-bayes-algorithm/

[13] InsideLearningMachines”8 Key Advantages And Disadvantages Of Decision Trees”, https://insidelearningmachines.com/advantages_and_disadvantages_of_decision_trees/

[14] ” Hyperparameters and Tuning Strategies for Random Forest”, https://ar5iv.labs.arxiv.org/html/1804.03515, 2019

[15] Rosita Guido, ” A hyper-parameter tuning approach for cost-sensitive support vector

machine classifiers”,P3,2022

[16] Ayush Majumdar,https://ieeexplore.ieee.org/author/37089837731,” Football Match Prediction using Exploratory Data Analysis and Multi-Output Regression”,P2, 2022

[17] Pakawan Pugsee, ”Football Match Result Prediction Using the Random Forest Classifier”,P2, 2019

[18] 野村證券”ジニ係数｜証券用語解説集”, https://www.nomura.co.jp/terms/japan/si/A02571.html

[19] Luiz Uehara ほか.” The Poor “Wealth” of Brazilian Football: How Poverty May Shape Skill and Expertise of Players”. https://www.frontiersin.org/articles/10.3389/fspor.2021.635241/

[20] 波部斉”ランダムフォレスト”研究報告コンピュータビジョンとイメージメディア(CVIM),Vol.2012-CVIM-182 No.31,pp1-8, 2012

[21] Oracle AI Data Science Blog, ”Random Forests, Decision Trees, and Ensemble Methods Explained”, https://blogs.oracle.com/ai-and-datascience/post/random-forests-decision-trees-and-ensemble-methods-explained

[22] スタビジ”不均衡データの扱い方と評価指標!Smote を Python で実装して検証していく!”, https://toukei-lab.com/imbalance-data-smote#i-4

[23] Qiita”不均衡データ”, https://qiita.com/tk-tatsuro/items/10e9dbb3f2cf030e2119

[24] aws”ハイパーパラメータチューニングとは何ですか?”, https://aws.amazon.com/jp/what-is/hyperparameter-tuning/

[25] Chang-Yun Lin, ”Multiresponse surface methodology for hyperparameter tuning to optimize multiple performance measures of statistical and machine learning algorithms”, P2, 2023

[26] scikit-learn”sklearn.ensemble.RandomForestClassifier”,https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[27] キカガク”機械学習の基礎”,https://free.kikagaku.ai/tutorial/basic_of_machine_learning/learn/machine_learning_hyperparameters

14

付録 A