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

2.1 ランダムフォレスト

2.1.1 特徴選択

2.2 データの前処理

2.3 ハイパーパラメータチューニング

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/

付録 A