決定木分析とランダムフォレストによる野球の勝因分析

2016SS005 江口慎一 2016SS094 保田理人
指導教員：小市俊悟

1　はじめに

日本プロ野球は，セ・リーグとパ・リーグそれぞれ 6 チームで構成されている．愛知県を本拠地とする中日ドラゴンズはセ・リーグに属している．これまで日本一 2 回，リーグ優勝 9 回という成績を収めている．しかし現在，中日ドラゴンズは成績不振に陥っている．本研究では 2011 年から 2018 年 (2013 年は除く) の 7 年間の様々なデータを分析対象とし，勝利に対してどのような要素が大きな影響を与えているのかを決定木やランダムフォレストといった機械学習の手法を用いて明らかにする．

2　データについて

本研究では，機械学習における学習のために，セ・リーグの 2011 年度から 2018 年度の各月ごとのデータ [1] を集めた．このデータに対して，プログラミング言語 Python においてデータ分析を支援する機能を提供するパッケージである pandas や，決定木・ランダムフォレストによる分析を提供するパッケージである scikit-learn を用いて，成績に影響する特性等を分析する．

本研究では，3・4 月，5 月，6 月，7 月，8 月，9・10 月の各期におけるセ・リーグ全チーム間の対戦成績から各期および各対戦カードについての勝ち越し，負け越し，引き分けを求め，それを成績として，勝・負・引き分けで表し，これをデータにより説明（予測）することを試みる．データは打者側と投手側のデータの 2 種類を用意し，打者側の属性 36 個と投手側の属性 31 個を表 1 に示す．

3　分析方法

3.1　決定木による分析

3.1.1　決定木

決定木とはデータを分割する条件を階層的に適用することでデータを分類するものである．その際，一番最初に適用する条件で用いられる属性が最も重要な属性だと一般には考えられる．階層を深くすることを許せば，細かい分類が可能であるが，それはデータに適合し過ぎであるとも考えられるので，本研究では階層の深さを 3 程度にした．決定木を図 1 を用いて説明する．図 1 の決定木に基づくと「負」という属性の特徴量について −0.50 を境界にまず分類する．続いて，「負」が −0.50 以上のデータについては，「UC 打数」が 134.00 を境界に分類する．「負」が −0.50 以上かつ「UC 打数」が 134.00 以上に該当する学習データは，3 つ存在しており，この決定木に基づく予測の際には，これらの条件に合致するものは負と判定される．次に，「UC 打数」が 134.00 未満のデータについては，「打席数」が 74.50 を境に分類する．「打席数」が 74.50 未満にまで当てはまる学習データは，17 個存在しており，予測の際には，これらの条件に合致するものは学習データにおける多数派である勝と判定される．

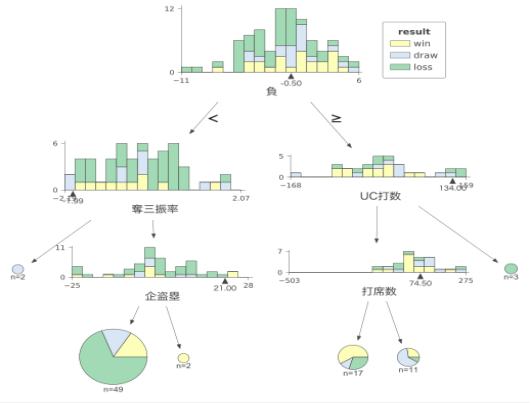

図 1　決定木分析の例

3.1.2　決定木を用いた分析方法

表 1, 2 に示した属性に関するデータを用いて，成績に影響を与えそうな属性を割り出すために，下記の 6 種類の方法をとった．いずれも勝ち負け，引き分けを判別することを目的としている．

方法 1: 2 チームのデータをそのまま用いて決定木分析を実行する方法．

方法 2: 2 チームのデータについて，対応するものの間で差を取り，それを対象に決定木分析を実行する方法．

方法 3: 投手の 31 個の属性，打者の 36 個の属性から，成績に直接影響を与えそうな属性として独断と偏見のもとで，投手 12 個の属性，打者 20 個の属性に限定して決定木分析を実行する方法．（実際に減らして用いた属性は表 2 を参照）

方法 4: 投手だけのデータを用いて決定木分析を実行する方法．

方法 5: 打者だけのデータを用いて決定木分析を実行する方法．

方法 6: データの各要素を試合数で割り，1 試合の平均を出して決定木を実行する方法．

表 1　打者 36 個と投手 31 個のデータ

打者	投手
打率	防御率
打点	勝ち数
本塁打	負け数
安打数	セーブ数
単打	奪三振数
2 塁打	試合数
3 塁打	投球回
出塁率	奪三振率
長打率	投球数
OPS	打者数
得点圏打数	被安打
得点圏安打	被本塁打
得点圏打率	与四球
UC 打数	与死球
UC 安打	敬遠
UC 率	失点
UC 本塁打	自責点
試合数	完投
打席数	完封
打数	無四球
得点	被打率
四球	QS 率
死球	援護点
企盗塁	援護率
盗塁	WHIP
盗塁成功率	UC 打数
企犠打	UC 被安打
犠打	UC 被本塁打
犠打成功率	最高球速
犠飛	最低球速
代打数	球速差
代打安打	-
代打率	-
併殺	-
失策	-
三振	-

表 2　打者 20 個と投手 12 個のデータ

打者	投手
打率	防御率
本塁打	S
安打数	奪三振数
単打	奪三振率
2 塁打	投球数
3 塁打	被安打
出塁率	被本塁打
長打率	与四球
得点圏打率	与死球
四球	敬遠
死球	完投
企盗塁	被打率
盗塁	-
企犠打	-
犠打	-
代打数	-
代打率	-
併殺	-
失策	-
三振	-

3.2　ランダムフォレストによる分析

3.2.1　ランダムフォレストについて

ランダムフォレストとは，決定木の欠点を補いつつ，その高い分類能力を維持，さらには高精度にしようとして考案された手法である．ランダムフォレストでは，学習データを全て使うのではなく，一部をランダムに取り出す．適当な数だけそのような学習データを用意し，それらそれぞれに対して，決定木を作成することを行う．また，学習データの部分集合を作成する際に，ランダム性のあるサンプリングを行うが，決定木を作成する際にもランダム性が導入されている．決定木を構成している条件は，それぞれ属性に対して定義されているが，決定木を作成する際に，使用できる属性をランダムに制限する．決定木を複数作成することの副産物として，作成された決定木において，よく採用されている属性というものが存在する．そのような属性，さらには条件は，「本質的に」重要であると認識すべきものと考えられる．多くの決定木の多くの条件に現れる属性はデータを分類するのに重要な属性であると考え，それを定量的に評価したものが，ランダムフォレストによって算出される属性（特徴量）の重要度である．その値が大きい属性ほど重要であるとされる．

4　決定木分析の結果

4.1　6 つの方法の分析結果のまとめ

6 つの方法で分析を行った結果から，どの属性が何回結果として出てきたかをまとめた．打者属性を見てみると，「出塁率」が 10 回，「得点圏打率」が 18 回と打者にとってプラスな属性も見られたが，「併殺」が 10 回，「三振」が 23 回と打者にとってあまり良くない属性も多く見られた．また，「安打数」「試合数」「打数」はあまり出ず，勝敗にあまり影響を与えないと感じた．次に，投手属性を見てみると，「負け数」が 14 回，「セーブ数」が 10 回と勝敗に直結しそうな要素が多く見られた．また，「奪三振率」が 10 回，「与死球」が 14 回，「被打率」が 15 回，「援護点」が 12 回とこちらの要素も多く見られた．

4.2　正答率と重要度

次に，各年度のデータをもとに作成した決定木に他年度のデータを入力し，正答率を出した．これを各年度，各方法で行った．全体的に見てみると，正答率は 50 %前後とあまり高い数字は得られなかった．しかし，方法 2 では，2018 年度，2017 年度，2014 年度のデータで作成した決定木は比較的正答率が高く，70.34 %，66.00 %，64.85 %となった．今後は比較的高い正答率となった方法 2 を用いて分析を行う．また高い正答率が得られなかった原因として，引き分けが影響していると予想し，引き分けを年間の対戦成績に基づいて，勝と負に変換することにした．結果として，2018 年では 64.28 %，2017 年では 65.52 %，2016 年では 62.36 %，2015 年では 60.39 %，2014 年では 67.96 %，2012 年では 65.29 %，2011 年では 69.13 %となった．さらに，この条件のもと決定木を作成し，決定木に用いられた属性の重要度を次のように計算した．決定木の階層を 3 までとするとき，最終的な分類までに適用される条件は 3 つである．そこで，一番最初に適用する条件に用いられる属性の重要度を 3，2 番目を 2，3 番目を 1 というように定める．各決定木と各属性についてこの重要度を求めた．打者属性を見てみると，重要度 8 の「打率」を除き他の属性の重要度にあまり差は出なかった．次に，投手属性を見てみると「勝」「負」といった属性の重要度が高くなった．次に，表 2 のデータを用いて決定木分析と重要度の計算を行った．打者属性では「打率」「3 塁打」「出塁率」「長打率」「得点圏打率」「四球」「死球」「企犠打」「代打率」「三振」といった属性の重要度が高くなった．投手属性では「セーブ数」の重要度が高くなった．

4.3　実際の順位と決定木分析による予想順位の違い

まずはじめに，作成した決定木を利用して，どのように順位を予想したのかを説明する．作成した決定木に，各期ごとの対戦 2 チームのデータを入れ，その期での勝ち負けを予測する．各期ごとに予測された勝敗を集計し，各チームの勝率を計算し，その勝率を元に順位を予想した．この

ように予想した順位と実際の順位を表 3 から表 9 に示す．2016 年と 2012 年の予想順位は実際の順位と完全に一致したが，2015 年は精度が悪かった．しかし，全体的に良い精度で順位予想ができており，勝敗を予想できる属性の抽出が妥当なものであると考えられる．

表 3　2018 年

球団	順位	予想
広島	1 位	1 位
ヤクルト	2 位	2 位
巨人	3 位	4 位
DeNA	4 位	5 位
中日	5 位	6 位
阪神	6 位	3 位

表 4　2017 年

球団	順位	予想
広島	1 位	1 位
阪神	2 位	3 位
DeNA	3 位	4 位
巨人	4 位	2 位
中日	5 位	5 位
ヤクルト	6 位	6 位

表 5　2016 年

球団	順位	予想
広島	1 位	1 位
巨人	2 位	2 位
DeNA	3 位	3 位
阪神	4 位	4 位
ヤクルト	5 位	5 位
中日	6 位	6 位

表 6　2015 年

球団	順位	予想
ヤクルト	1 位	4 位
巨人	2 位	1 位
阪神	3 位	6 位
広島	4 位	3 位
中日	5 位	5 位
DeNA	6 位	2 位

表 7　2014 年

球団	順位	予想
巨人	1 位	3 位
阪神	2 位	1 位
広島	3 位	2 位
中日	4 位	4 位
DeNA	5 位	5 位
ヤクルト	6 位	6 位

表 8　2012 年

球団	順位	予想
巨人	1 位	1 位
中日	2 位	2 位
ヤクルト	3 位	3 位
広島	4 位	4 位
阪神	5 位	5 位
DeNA	6 位	6 位

表 9　2011 年

球団	順位	予想
中日	1 位	1 位
ヤクルト	2 位	2 位
巨人	3 位	3 位
阪神	4 位	5 位
広島	5 位	4 位
DeNA	6 位	6 位

5　ランダムフォレストを用いた分析の結果

5.1　重要な属性の抽出

投手と打者の全データの中で，最も重要だと考えられる属性を出すため，表 2 のデータを用いて，各年度・各期における属性の重要度をランダムフォレストにより求めた．また，各属性について 6 年分の重要度を足し合わせ，それを表 10，11 に示す．この結果から 6 年分の重要度を総合的に判断できると考えた．その値が高かった 10 個の属性は長打率，打率，出塁率，犠打，防御率，奪三振率，代打率，2 塁打，盗塁，セーブ数である．

5.2　正答率の確認

前節で出した 10 個の属性が，ランダムフォレストによって抽出された勝敗を決めるのに重要な属性である．抽出された属性の重要性を確かめるため，以下の 2 つの方法を用いた．

5.2.1　方法 1

方法 1 では，決定木同様に順位を予想し，実際の順位と比較した．結果として，2017 年と 2016 年の予想順位は実際の順位と完全に一致したが，2015 年と 2011 年の予測は悪かった．しかし，全体的に良い精度で順位予想ができており，勝敗を予想できる属性の抽出は妥当なものであると考えられる．

表 10　ランダムフォレスト結果 (打者)

属性	重要度の和
打率	0.232
本塁打	0.084
安打数	0.111
単打	0.063
2 塁打	0.118
3 塁打	0.070
出塁率	0.208
長打率	0.234
得点圏打率	0.096
四球	0.104
死球	0.065
企盗塁	0.083
盗塁	0.116
企犠打	0.096
犠打	0.143
代打数	0.097
代打率	0.122
併殺	0.066
失策	0.066
三振	0.063

表 11　ランダムフォレスト結果 (投手)

属性	重要度の和
防御率	0.143
S	0.113
奪三振率	0.075
奪三振数	0.124
投球数	0.095
被安打	0.090
被本塁打	0.069
与四球	0.083
与死球	0.090
敬遠	0.075
完投	0.021
被打率	0.101

5.2.2　方法 2

方法 2 では，各年度について，10 個の属性だけを使用して決定木を作成し，その決定木を各年度のデータに適用することで，作成した決定木の勝敗予測の正答率を確認した．結果として，2013 年と 2014 年のデータで作成した決定木の精度が高く，反対に 2015 年の精度が低かった．2016 年と 2014 年の決定木は，どの年度のデータを使用しても，勝敗を一定程度に予想することが可能である．2015 年のデータを用いて作成された決定木の正答率が低かった理由として 2015 年がセ・リーグの全球団が借金を抱えた野球史に残る年であったため，勝敗を予想することが困難であったことが考えられる．方法 2 において一番正答率の良かった 2016 年の決定木を 2016 年のデータに適用した結果を図 2 に示す．

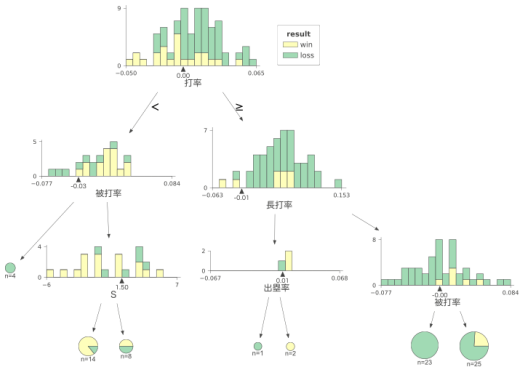

図 2　2016 年-2016 年の決定木

5.2.3　決定木分析の結果との関係性

決定木分析とランダムフォレストを用いて，重要な属性を抽出したが，ここで一つ問題となるのは，決定木分析に

より重要であるとした属性とランダムフォレストによって重要であるとした属性が一致していないということである．これを理解するために，決定木分析で得た属性「打率」「出塁率」「得点圏打率」とランダムフォレストから得た属性「打率」「2 塁打」「出塁率」「長打率」の間で，相関関係を求めることにした．例として全球団のデータを用いた相関係数を表 12 に示す．各球団・全球団においても相関関係があると言える相関係数を得た．よって，決定木分析で頻出した「打率」「出塁率」「得点圏打率」はランダムフォレストで頻出した「打率」「2 塁打」「出塁率」「長打率」と互換性があると考えられる．つまり，決定木分析でも，ランダムフォレストを用いた分析でも，重要な属性を見付けることができたのではないかと考える．

表 12　全球団のデータを用いて

全球団	打率	2 塁打	出塁率	長打率	得点圏打率
打率	—	0.639	0.859	0.817	0.739
2 塁打	—	—	0.621	0.767	0.587
出塁率	—	—	—	0.749	0.571
長打率	—	—	—	—	0.697

6　決定木とランダムフォレストを用いた分析の結果の考察

決定木とランダムフォレストを用いて勝敗に影響を与える属性とは何かを分析し，得点圏打率，長打率，出塁率，2 塁打，盗塁，犠打の 6 個の重要属性を抽出した．分析結果によれば重要属性値が高い方が勝ちに繋がり，低い方が負けに繋がる．よって，この属性値さえリーグ内トップであれば優勝できるのではないかと予想し，実際の優勝球団の各属性のリーグ内順位をまとめた．結果として，2014 年と 2011 年以外の年で，優勝球団は重要属性値が比較的高い結果となった．特に 2018 年と 2017 年では，ほとんどの属性値が 1 位となっており，近年にかけて，よりこの重要属性が勝敗への影響を与えていると考えられる．反対に 2014 年と 2011 年では，属性値が高くなくても優勝することができており，特に 2011 年に優勝した中日ドラゴンズは，重要属性値が著しく低いにもかかわらず優勝している．私たちが抽出した重要属性は，近年では高い程優勝することができたが，2014 年以前では低くても優勝することができていた．そこで，2014 年以前と以降で何が違うのかを考える．

6.1　投高打低から打高投低の時代へ

分析結果により，投手よりも打者の方に勝敗を分ける要因があることが示唆される．しかし，2014 年以前については，重要属性と優勝球団との間にずれが生じたので，2014 年以前では打者ではなく投手が重要であったのではないかと考えた．そこで，2014 年以前と以後で NPB が使用していた統一球に着目した．実際に 2011 年と 2012 年で使用された統一球は，ミズノ製の低反発ゴム材を用いた統一球を採用しており，セ・リーグ，パ・リーグ合わせた本塁打数が 2010 年の 1605 本から 2011 年には 939 本に激減した．また，2012 年では，開幕から 4 月 25 日まで全球団で 123 試合中 38 試合が完封試合という異常な投高打低を記録した．次に，2011 年から 2018 年の優勝球団の主な投手成績をまとめてみたところ，2014 年を境に投手成績が大きく変わっていた．防御率では，2 点台の 2013 年・2011 年に対し，2018 年では 4 点台でも優勝することができている．被本塁打では，2013 年の 64 本に対して 2014 年では 122 本と，ほぼ倍増しており，飛ぶボールへの変化は，投手に多大な影響を与えていたことが分かる．よって，投高打低から打高投低の時代へと変わってきており，点を取られてもそれ以上の点を取れる打力のある球団が優勝していると考えられる．

6.2　中日ドラゴンズが優勝するためには

本研究では，セ・リーグ での勝敗判別について考えてきた．抽出した重要属性から投手よりも打者に勝敗を決する要因があり，近年においてよりその傾向があることがわかった．そこで，2014 年から B クラスを抜け出せない中日ドラゴンズが優勝するためにどうすべきかを，重要属性に基づいて考える．まず，中日ドラゴンズの重要属性のリーグ内順位をまとめてみたところ，ほとんどの順位が B クラス (4 位以下) であり，盗塁や犠打などを多く用いている年も見られるが，長打率や出塁率や 2 塁打などの属性値が低く，打力不足が露呈している．よって，中日は投手依存のチームから脱却し，打者の強化を図ることが重要であると感じた．戦略的には打つだけでなく，盗塁や犠打も使いながらランナーを得点圏へ送ることが重要であると言えよう．将来的には，根尾選手のような安打量産が期待される選手や，日本を代表するスラッガーになれると期待される石川選手のような若手が台頭してくると中日ドラゴンズはセ・リーグ A クラスに戻ってこられるのではないかと考える．

7　おわりに

本研究では，中日ドラゴンズの弱体化の要因を決定木やランダムフォレストといった機械学習を用いて分析した．その結果，打者の成績が近年は重要であることがわかり，飛ばないボールから飛ぶボールに変化したこともあり，投手型のチームであった中日ドラゴンズは結果として成績を落とすことになったと考える．これを踏まえて，中日ドラゴンズをいかに A クラスに戻すのかについて考えると，打者の強化を図ることが重要であり，中でも出塁率が高い選手，長打率の高い選手の両者を補強することが大事なのではないかという結論に至った．戦略的には打つだけでなく，盗塁や犠打も使いながらランナーを得点圏へ送ることも重要であると感じた．

参考文献

[1] データで楽しむプロ野球 (2019/01/08 アクセス) : 『https://baseballdata.jp/2018/4/ctop.html』