

潜在クラスマルコフモデルによるプロ野球先発投手の失点予測に関する一考察

1X16C020-1 上原諒介
指導教員 後藤正幸

1. 研究背景と目的

プロ野球において、投手の運用はシーズンで優勝するための重大な要因の一つである。その中でも、試合中の先発投手の交代タイミングは、勝敗に直結する重要な意思決定事項であると考えられる。一方、最近では1球毎の状況や結果に関するデータが全試合に渡って記録されており、これらを活用することで投手交代の意思決定支援に繋がる可能性がある。そこで本研究では、過去のデータに基づき投手のインニングごとの期待失点を推定することにより、投手交代の意思決定を支援する情報の提供を考える。インニング毎の期待失点(得点)を算出している従来研究 [1] では、アウトカウントと走者状況を状態とし、打者の能力のみを考慮して、現在の状態から次の状態へ遷移する確率(以後、遷移確率)を推定して期待得点を算出していた。しかし、チームの得点には打者の能力のみでなく、投手の能力や当日の調子も大きく影響すると考えられる。そこで、打者情報に加えて投手の能力と当日の調子を反映し、期待失点を算出するモデルを構築するとともに、それを用いた分析を行うことを本研究の目的とする。

具体的には、投手と打者の組み合わせごとに遷移確率が異なることを考慮し、インニング毎の期待失点を算出するモデルを提案する。しかし、全ての投手と打者の組み合わせごとに遷移確率行列を推定すると、データ数が足りず遷移確率の推定精度が悪くなってしまふ。そこで、潜在クラスモデルを用いて投手と打者の組み合わせを少数の潜在変数でグルーピングしながら、遷移確率を推定することを考える。これにより、投手と打者の組み合わせの潜在クラスごとに、遷移確率を推定して精度を高めたモデルを構築する。提案モデルを実際の日本プロ野球のデータに適用し、期待失点の算出と潜在クラスの解釈を行うことで、提案モデルの有用性を示す。

2. 遷移行列を用いた期待得点の算出

インニングごとの期待得点に着目した研究として大澤らの研究 [1] がある。大澤らは打者ごとに算出された遷移確率をもとにインニング毎の期待得点を算出する手法を提案している。1インニングで挙げる最大得点を R_{max} と定義し、状態として、アウトカウント(0アウト, 1アウト, 2アウト)と走者状況(無走者, 一塁, 二塁, 三塁, 一二塁, 一三塁, 二三塁, 満塁)の組み合わせ、さらに3アウトの攻撃終了を加えた25個のアウトカウント・走者状況を定義する。また、1インニング中に挙げる得点を行、状態を列とし、得点と状態の確率分布を表す行列を $U \in \mathbf{R}^{R_{max}+1 \times 25}$ 、ある打者が打席に入った時点の攻撃状態を行、ある打者が打撃を行った結果の攻撃状態を列とする遷移行列を $P \in \mathbf{R}^{25 \times 25}$ と定義すると、これらを掛け合わせることで攻撃状態の遷移をシミュレートすることが可能である。ただし、インニング開始から n 人の打席終了時における U を U_n 、さらに U_n の i 行目を $U_{n|i}$ と表す。期待得点の算出アルゴリズムを以下に示す。

Step 1) 対象期間中の打撃結果から、打者ごとに遷移行列 P を算出する。

Step 2) U_{n-1} と P より、式(1)を用いて U_n を算出する。

$$U_{n|i} = \sum_{r=0}^4 U_{n-1|i-r} P^{(r)} \quad (1)$$

$(i = 1, 2, \dots, R_{max}+1, i > r)$

ただし、 $P^{(r)}$ は遷移行列 P のうち、1打席で得られる点数が r 点となる遷移のみを抜き出した遷移行列である ($r \in \{0, 1, 2, 3, 4\}$)。

Step 3) 3アウトの攻撃終了状態、すなわち U_n の第25列の総和が1に収束したら終了、さもなければStep1へ。

Step 4) Step 3より得られた収束状態の U の第25列を U_∞ とし、1インニングあたりの期待得点を式(2)から算出する。

$$\text{期待得点} = \sum_{i=1}^{R_{max}+1} U_{\infty|i} \times (i-1) \quad (2)$$

3. 提案手法

3.1. 概要

本研究では、インニング毎に投手の期待失点を算出し、先発投手を交代するべきか否かの意思決定に有益な情報を提供することを考える。しかし従来研究と同様の方法で、投手と打者の全組み合わせに対し、個々に遷移確率を推定すると、データ数が足りず遷移確率の推定精度が悪くなってしまふ可能性がある。そこで、投手と打者の組み合わせと状態遷移の背後に潜在クラスを仮定する。すなわち、投手指標と打者打順が潜在クラスから共起するような構造にすることで、潜在クラス毎に異なる遷移行列を仮定したモデルを構築する。これにより、潜在クラスへの所属確率を重みとした遷移行列を投手と打者の組み合わせごとに推測することが可能となる。ここから、投手と打者の組み合わせが決まると遷移行列も決定され、その遷移行列をもとに期待失点を算出することができる。

3.2. 提案モデル

プロ野球の投手が失点する要因として、主に投手の能力、対戦打者の能力、投手の当日の調子が考えられる。本研究では、投手の能力としてその投手の試合前までのシーズン成績を、対戦打者の能力として打者の打順を、投手の当日の調子として、その投手の当日の前インニングまでの投手成績を用いる。このとき投手の能力や調子を表現する指標として、投手の特徴を表し、かつ1インニングの失点数と相関が高いと考えられる以下の4指標を用いる。

- WHIP : 1インニングに出すランナーの数
- 失点率 : 1インニングあたりの失点数

- 奪三振率：1 イニングあたりに奪う三振数
- ゴロアウト/フライアウト比率：ゴロアウトとフライアウトの比率

上記の指標をそれぞれ、クラスごとの特徴が明確に表れるように3段階に分割し、それと打者の打順を用いて投手と打者の組み合わせを表現する。いま、 l 番目の投手指標・打者打順 ($1 \leq l \leq L$) の要素数を M_l 、要素集合を $\mathcal{X}^l = \{x_{m_l}^l : 1 \leq m_l \leq M_l\}$ 、 l 番目の投手指標・打者打順を表す変数ベクトルを $\mathbf{x} = (x_{m_1}^1, x_{m_2}^2, \dots, x_{m_l}^l)^T$ 、打撃前のアウトカウント・走者状況集合を $\mathcal{S} = \{s_i : 1 \leq i \leq 25\}$ 、打撃後のアウトカウント・走者状況集合を $\mathcal{Y} = \{y_j : 1 \leq j \leq 25\}$ 、 K 個の潜在クラス集合を $\mathcal{Z} = \{z_k : 1 \leq k \leq K\}$ と定義すると、同時確率は以下の式 (3) で定義される。

$$p(\mathbf{x}, y_j, z_k | s_i) = p(z_k) p(y_j | s_i, z_k) \prod_{l=1}^L p(x_{m_l}^l | z_k) \quad (3)$$

ただし、投手指標・打者打順と打撃後の状態の確率分布 $p(x_{m_l}^l | z_k)$ 、 $p(y_j | s_i, z_k)$ には多項分布を仮定し、パラメータは EM アルゴリズムを用いて推定する。提案モデルのグラフィカルモデルを図 1 に示す。

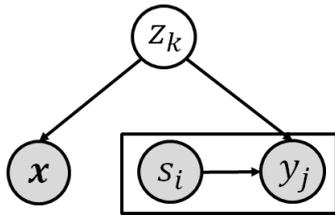


図 1: 提案モデルのグラフィカルモデル

4. 実データを用いた分析

4.1. 分析条件

本研究では、データスタジアム株式会社より提供された 2016 年から 2018 年までの日本プロ野球シーズンデータ 3 年分 (打席ごと) を分析対象とする。そのうち、2016, 2017 年のデータを学習に、2018 年のデータをテストに用いた。分析対象は先発投手のみとし、学習データ数は 84,510 打席、テストデータ数は 35,825 打席である。潜在クラス数は、予備実験により探索的に求め、 $K=6$ と設定した。また野球の特性上、失点が 0 となるイニングが非常に多いため、全てのイニングで予測値を 0 としても高い精度が出てしまうという問題がある。そこで、テストデータにおける全イニングのうち、失点しているイニングのみで失点数の予測誤差を比較し、予測精度を検証する。またベースラインとなる比較手法としては、投手ごとの 2016, 2017 年の総失点数を総イニング数で割って得られる、平均失点数を用いた。分類の評価指標には平均絶対誤差 (MAE)、平均平方二乗誤差 (RMSE) を用いる。

4.2. 分析結果

4.2.1. 期待失点の精度検証

テストデータにおける、実際の失点と各手法の期待失点との誤差を表 1 に示す。

表 1: 提案手法と比較手法の MAE と RMSE

	提案手法	比較手法
MAE	1.3519	1.8055
RMSE	1.4019	1.8292

表 1 より、提案手法は比較手法よりも MAE, RMSE ともに予測精度が高いことがわかる。このことから、提案手法は期待失点の算出に有効であると考えられる。

4.2.2. 潜在クラスに関する分析

潜在クラスごとの解釈の一例を表 2 に示す。

表 2: 潜在クラスへの所属確率と解釈 (一部抜粋)

クラス名	クラス 1	クラス 2
所属確率	0.32514	0.06637
クラスの解釈 (シーズンの指標)	WHIP, 失点率が非常に優れるが、奪三振率が悪い	WHIP, 失点率、奪三振率が良い
クラスの解釈 (当日の指標、打順)	WHIP, 失点率があまり優れなく、1~4 番打者の所属確率が高い	WHIP, 失点率が優れ、フライアウト率が高い、5~8 番打者の所属確率が高い
所属例	岸孝之投手と西川 遥輝選手 (1 番打者)	菅野智之投手と大野 奨太選手 (8 番打者)

表 2 において、クラス 1 はシーズン全体としては打者を抑える確率がかなり高いが、投手の当日の調子があまり良くないクラスとなっている。すなわち、投手の基本的な能力を考えると抑える確率の高い対戦であるが、当日の調子や相性などによって抑える確率が低くなっている状況である。一方、クラス 2 は投手がシーズン、当日ともに良い成績を残しており、打順も下位であることから、投手が抑える確率の高い対戦であると言える。このように、提案手法を用いることで、期待失点を計算するだけでなく、潜在クラスの特徴から対戦の分析を行うこともできる。

5. まとめと今後の課題

本研究では、投手指標と打者打順の組み合わせと状態遷移の背後に潜在クラスを仮定し、潜在クラスごとに遷移確率を推定した。そして、潜在クラスへの所属確率を重みとして、投手と打者の組み合わせごとに異なる遷移行列を構築し 1 イニングごとの期待失点を算出した。そして、日本プロ野球の実データを用いて、期待失点の算出や潜在クラスの解釈を行うことで、本提案モデルの有用性を示した。

今後の課題として、投手交代タイミングへの言及や分析対象を先発投手以外の投手に広げること、打順以外の打者指標の導入などが挙げられる。

謝辞：本研究は、情報・システム研究機構の新領域融合研究プロジェクト『社会コミュニケーション』データ中心科学研究リサーチコンズ事業『人間・社会データ』の支援により、データスタジアム株式会社からデータを提供して頂いております。

参考文献

- [1] 大澤清, 合田憲人, “野球における走者の進塁状況を考慮した勝率計算方法,” 日本応用数理学会論文誌, Vol.18, No.3, pp.321-346, 2008.
- [2] T.Hoffman, “Probabilistic Latent Semantic Analysis,” Proc. of UAI99, pp.289-296, 1999.