# 마르코프 연쇄를 이용한 프로야구 승률 예측 모델 개선에 대한 연구

김도헌[1,0] 김진한[1] 박성규[1] 이진섭[1] 정재형[1] 고화정[1] 이도훈[2]
[1]부산과학고등학교 [2]부산대학교 정보컴퓨터공학부
donny7112@naver.com, awe2478223@gmail.com, psg9806@naver.com, bbben12300@gmail.com,
jaehyoung98@gmail.com, withhjko@naver.com, dohoon@pusan.ac.kr

# A Study on Improvement of Prediction Model for Baseball Winning Rate

DoHeon Kim[1,0] JinHan Kim[1] SeongGyu Park[1] JinSeob Lee[1] JaeHyeong Jeong[1] HwaJung Ko[1] DoHoon Lee[2]
[1]Busan Science High School [2]Dept. of Computer Science and Engineering, Pusan National University

**요 약**

야구 경기의 승패 결과를 예측하는 방법은 다각적으로 연구되어왔다. 특히, 마르코프 연쇄를 적용한 기존의 연구는 해당 시즌 타자의 경기기록으로 만든 전이 행렬로 기대점수분포를 계산하는 방법을 사용한다. 본 논문에서는 투수에 대한 전이 행렬을 추가로 이용하여 기존 모델을 개선하는 방법을 제안한다. 투수 반영정도, 즉 실제 승률에 가장 가까운 팀별 가중치를 추정한다. 또한 실험에서 각 팀별 가중치를 적용하여 승률 예측이 기존 모델보다 개선됨을 보이고자 한다.

## 1. 서 론

운동경기에서 상황분석이나 예측은 팀이나 감독에게 매우 중요하다. 사람에게 의존하던 분석은 수없이 생성되는 자료에 의해 정량화되는 방법으로 바뀌었다. 이러한 데이터를 기반으로 경기 결과를 예측하려는 많은 시도가 있어왔고[1] 경기의 예측 결과는 작전과 전략의 수립에 도움을 주고 선수 개인 역량을 향상시키는데 활용되었다.

승률을 계산할 때 팀 타율, 팀 평균 자책점 등 팀 자체의 데이터를 사용하는 것은 경기에 실제 출전하는 선수의 개인 기록을 반영하지 않았다는 점이 단점으로 지적된다[6]. 이에 [2]에서는 타자 데이터를 사용한 개선된 모델을 제시하였다. 그러나 승률에 영향을 미치는 요소는 타자뿐만 아니라 투수, 수비 등 다양하다. 따라서 본 논문에서는 투수 데이터를 추가하여 보다 개선된 모델을 제시하고자 한다.

## 2. 제안 모델링 연구 방법

### 2.1 관련 연구

#### (1) 야구에서 상태의 정의

야구 경기 중에 발생하는 상황을 아웃카운트와 각 베이스의 선수 유무에 따라 정의할 수 있다.

상태 S($X_1, X_2, X_3, X_4$)에서 $X_1$은 아웃카운트, $X_2$는 1루의 주자 유무, $X_3$은 2루의 주자 유무, $X_4$은 3루의 주자 유무를 의미하며, 표1과 같이 총 25가지 경우로 나타낼 수 있다. 예를 들어 $S_7$(0,0,1,1)은 노아웃, 주자 2,3루 상황을 의미한다.

표 1 : 각 번호와 상태

| 번호 | 상태 | 번호 | 상태 | 번호 | 상태 |
|---|---|---|---|---|---|
| $S_1$ | (0,0,0,0) | $S_{10}$ | (1,1,0,0) | $S_{19}$ | (2,0,1,0) |
| $S_2$ | (0,1,0,0) | $S_{11}$ | (1,0,1,0) | $S_{20}$ | (2,0,0,1) |
| $S_3$ | (0,0,1,0) | $S_{12}$ | (1,0,0,1) | $S_{21}$ | (2,1,1,0) |
| $S_4$ | (0,0,0,1) | $S_{13}$ | (1,1,1,0) | $S_{22}$ | (2,1,0,1) |
| $S_5$ | (0,1,1,0) | $S_{14}$ | (1,1,0,1) | $S_{23}$ | (2,0,1,1) |
| $S_6$ | (0,1,0,1) | $S_{15}$ | (1,0,1,1) | $S_{24}$ | (2,1,1,1) |
| $S_7$ | (0,0,1,1) | $S_{16}$ | (1,1,1,1) | $S_{25}$ | (3,0,0,0) |
| $S_8$ | (0,1,1,1) | $S_{17}$ | (1,0,1,1) | | |
| $S_9$ | (1,0,0,0) | $S_{18}$ | (1,1,1,1) | | |

#### (2) 상태의 전이 확률을 나타내는 P행렬

Buklet et.al.은 실시간으로 타자와 투수의 상황을 고려하여 현재의 상태에서 다음 상태로의 전이를 표현 하고자 특정 상태에서 특정 상태로의 전이 확률을 성분으로 나타낸 P 행렬로 제안하였다[1].

P행렬은 25×25행렬로 행은 이전상황, 열은 나중 상황을 나타내며, $P_{i,j}$은 $S_i$상태에서 $S_j$상태로 전이될 확률을 나타낸다. (단, 1≤i,j≤25) 그림1과 같이 $P_{1,3}$은 $S_1$(0,0,0,0)에서 $S_3$(0,0,1,0)로 전이될 확률이다.


그림 1 : $S_1$상태에서 $S_3$상태의 전이

P행렬은 그림2와 같이 부분행렬 A, B, F행렬로 나타낼 수 있다. (단, A, B는 8×8행렬, F는 8×1행렬) 여기에서 A는 아웃카운트 변화 없이 주자 상태만 변하는 상황, B는 아웃카운트가 1개 추가되는 상황, F는 아웃카운

트가 2개 추가되는 상황, 0은 불가능한 상황, 1은 경기 종료 상황을 나타낸다.

$$\mathbf{P} = \begin{pmatrix} A & B & 0 & 0 \\ 0 & A & B & 0 \\ 0 & 0 & A & F \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

그림 2. 부분 행렬로 표현된 P행렬

부분 행렬의 각 성분은 그림3과 같다.

$$A = \begin{pmatrix} P_H & P_S+P_W & P_D & P_T & 0 & 0 & 0 & 0 \\ P_H & 0 & 0 & P_T & P_S+P_W & 0 & P_D & 0 \\ P_H & P_S & P_D & P_T & P_W & 0 & 0 & 0 \\ P_H & P_S & P_D & P_T & 0 & P_W & 0 & 0 \\ P_H & 0 & 0 & P_T & P_S & 0 & P_D & P_W \\ P_H & 0 & 0 & P_T & P_S & 0 & P_D & P_W \\ P_H & P_S & P_D & P_T & 0 & 0 & 0 & P_W \\ P_H & 0 & 0 & P_T & P_S & 0 & P_D & P_W \end{pmatrix}$$

$$B = P_{\text{out}}\mathbf{I}. \qquad F = (P_{\text{out}}, \ldots, \mathbf{P}_{\text{out}})^T$$

그림 3. 행렬의 각 성분 값

그림3의 각 성분 요소는 KBO홈페이지에서 제공하는 선수들의 데이터 중 안타(H), 홈런(HR), 볼넷(BB), 사구(HBP), 타석수(PA), 1루타(H-2B-3B), 2루타(2B), 3루타(3B)를 이용하여 다음과 같이 구할 수 있다.

$$P_W = \frac{BB+HBP}{PA} \text{ (볼넷, 사구로 인한 1루 출루율)}$$

$$P_S = \frac{H-(2B+3B+HR)}{PA} \text{ (1루타 칠 확률)}$$

$$P_D = \frac{2B}{PA} \text{ (2루타 칠 확률)}$$

$$P_T = \frac{3B}{PA} \text{ (3루타 칠 확률)}$$

$$P_H = \frac{HR}{PA} \text{ (홈런칠 확률)}$$

$$P_{out} : 1-(P_W+P_S+P_D+P_T+P_{HR}) \text{ (아웃될 확률)}$$

식 1 : P 행렬의 성분 제작 방법

### (3) 점수를 낼 확률을 나타내는 U행렬

U 행렬은 21×25 행렬로 행은 점수를 열은 상황을 나타낸다. $U_{i,j}$는 i번째 이닝에서 j번째 타자가 들어설 때 점수 및 주자 상태별 확률을 행렬로 표현한 것으로 각 상황의 확률과 시합의 승률들을 예측할 수 있다[2].

$$U_{i,j} = \begin{bmatrix} p_{0,1} & p_{1,2} & p_{1,3} & \cdots & p_{0,24} & p_{0,25} \\ p_{1,1} & p_{1,2} & p_{1,3} & \cdots & p_{1,24} & p_{1,25} \\ p_{2,1} & & & \cdots & & p_{2,25} \\ \vdots & & & \ddots & & \vdots \\ p_{20,1} & p_{20,2} & p_{20,3} & \cdots & p_{20,24} & p_{20,25} \end{bmatrix}$$

$p_{k,s}$ : 점수가 k이고 $S_S$에 있을 확률

식 2 : U Matrix

U 행렬의 전이는 크게 한 이닝에서 다음 타자로 넘어가는 경우와 이닝의 전환으로 나누어진다[2]. 그림4는 한 이닝에서 타자의 변화와 다음이닝으로 옮겨가는 예를 보여준다.

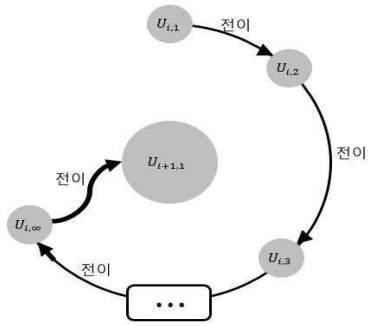


그림 4 : U Matrix의 전이

$U_{i,j+1}$의 열을 한 줄씩 바꿔 넣어 $U_{i,j}$를 $U_{i,j+1}$로 만든다. $U_{i,j+1}$는 j번째 타자의 상태에서 점수를 낸 상황을 반영하므로 식은 다음과 같다[1].

$$\begin{aligned} U_{i,j+1}(row\,k) = {} & U_{i,j}(row\,k)P_0 + U_{i,j}(row\,k-1)P_1 \\ & + U_{i,j}(row\,k-2)P_2 + U_{i,j}(row\,k-3)P_2 \\ & + U_{i,j}(row\,k-4)P_4 \end{aligned}$$

식 3 : U 행렬의 전이 방법

이때 사용되는 $P0 \sim P4$ 행렬은 기존 P 행렬에서 득점에 관련된 0득점 ~ 4득점으로 각각 분할한 행렬이다.

실제 야구에서는 3아웃이 될 경우 다음이닝으로 넘어가게 된다. 이를 연산에 적용하기 위해서는 위의 연산을 통해 $U_{i,\infty}$에서 행렬 값들이 확률임을 고려해 3아웃 이외의 상태 값들의 합이 0.0000001보다 작으면 0이라고 가정하고 연산을 중단한다. 그 후, $U_{i,\infty}$의 25열(3아웃으로 끝날 확률)을 $U_{i+1,1}$의 1열로 사용한다. 이 때 타자는 $U_{i,\infty}$의 마지막 타자 이후의 타자를 $U_{i+1,1}$의 타자로 사용한다.

## 2.2 투수력을 반영한 모델

투수력은 경기 전반에 걸쳐 영향을 미치는 중요한 요소이다. 이 투수력을 경기력에 포함하는 방법을 다음과 같이 제안한다. 각 팀별 타자수 상위 20명의 투수 기록을 토대로 확률 행렬P′ 을 생성한다. P′ 행렬은 P 행렬과 같이 상태가 전이 될 확률을 행렬로 표현한 것이다.

기존 모델에 P′ 행렬을 도입하기 위해서 가중치를 사용하여 확률 연산을 한다. 팀마다 투수가 경기에 미치는 영향이 서로 다르기 때문에 가중치는 팀별로 다른 값을 가진다. 가중치 $\alpha\,(0 \le \alpha \le 1)$에 대하여 P행렬과 P′ 행

렬의 가중치 연산을 다음과 같이 정의한다.

$$\alpha P+(1-\alpha)P'$$

식 4 : 가중치 연산

위의 식 4를 기존 연구에서 식 3의 P에 대입하여 각 팀의 투수력을 반영한 승률을 구할 수 있다.

## 3. 실험

P 행렬과 P′ 행렬을 적절히 반영하기 위해서는, 각 팀별로 최적의 $\alpha$값을 찾아야하기 때문에 팀들 간의 경기 결과에 투수데이터의 가중치$\alpha$가 얼마만큼 영향을 주는지 알아야한다. $\alpha$값을 0부터 1까지 0.1 단위로 증가시켰을 때 실제 승률과의 차이가 가장 작은 경우가 최적의 $\alpha$값이라고 할 수 있을 것이다. 이 실험에서는 두산, SK, KIA의 $\alpha$를 구하기 위해 임의의 5팀과의 $\alpha$별 예상 승률과 실제 2015시즌의 승률을 비교하였다. 그 결과 가장 작은 오차를 갖는 $\alpha$값을 팀의 $\alpha$값으로 하여 $\alpha$ = 1인 기존 모델의 경우와 비교하였다. 오차는 실제승률 - 모의실험 승률의 절댓값을 의미한다.

표 3 : KIA의 실제 승률에 따른 $\alpha$값 오차 비교

| | 두산 | 한화 | NC | 넥센 | SK | 평균 |
|---|---|---|---|---|---|---|
| 실제 승률 | 0.5000 | 0.5625 | 0.3125 | 0.2500 | 0.6250 | |
| $\alpha$ = 1 | 0.3167 | 0.4510 | 0.2452 | 0.2961 | 0.2245 | |
| 오차 | 0.1833 | 0.1115 | 0.0673 | 0.0461 | 0.4005 | 0.1617 |
| $\alpha$ = 0.4 | 0.4140 | 0.4280 | 0.5554 | 0.2846 | 0.4500 | |
| 오차 | 0.0860 | 0.1345 | 0.2429 | 0.0346 | 0.1750 | 0.1346 |

표 4 : 두산의 실제 승률에 따른 $\alpha$값 오차 비교

| | 한화 | KIA | SK | NC | 넥센 | 평균 |
|---|---|---|---|---|---|---|
| 실제 승률 | 0.5625 | 0.5 | 0.6875 | 0.5 | 0.5 | |
| $\alpha$ = 1 | 0.5331 | 0.5290 | 0.4009 | 0.5055 | 0.4116 | |
| 오차 | 0.0291 | 0.0290 | 0.2866 | 0.0055 | 0.0884 | 0.0877 |
| $\alpha$ = 0.7 | 0.4903 | 0.5029 | 0.4341 | 0.4367 | 0.4721 | |
| 오차 | 0.0722 | 0.0029 | 0.2534 | 0.0633 | 0.0279 | 0.0839 |

표 5 : SK의 실제 승률에 따른 $\alpha$값 오차 비교

| | 두산 | 한화 | KIA | NC | 넥센 | 평균 |
|---|---|---|---|---|---|---|
| 실제 승률 | 0.3125 | 0.5625 | 0.375 | 0.3125 | 0.4375 | |
| $\alpha$ = 1 | 0.5596 | 0.7803 | 0.7310 | 0.5666 | 0.6820 | |
| 오차 | 0.2471 | 0.2178 | 0.3560 | 0.2540 | 0.2445 | 0.2639 |
| $\alpha$ = 0.4 | 0.3168 | 0.3479 | 0.3546 | 0.4489 | 0.2691 | |
| 오차 | 0.0043 | 0.2145 | 0.0203 | 0.1364 | 0.1683 | 0.1088 |

두산의 $\alpha$ = 0.7, KIA와 SK의 $\alpha$ = 0.4의 경우가 $\alpha$ = 1인 기존 모델에 비해 실제와의 오차가 줄어든 것을 확인할 수 있다. 따라서 두산의 경우 투수를 0.4, KIA와 SK의 경우 투수를 0.7 정도 반영할 때 가장 정확한 예측을 할 수 있다는 것을 실험결과가 보여 주고 있다.

## 4. 결론 및 추후연구

본 연구에서는 야구 경기를 예측하는데 있어서 타자뿐만 아니라 투수에 관한 정보까지 고려하여 보다 정확한 예측 결과를 얻고자 하였다. 또한 각 팀별로 투수진과 타자진의 편차가 존재한다는 점에서 기인하여 기존의 타자만 고려한 행렬인 P 행렬보다 투수를 함께 고려한 P' 행렬에 최적화된 가중치를 가했다. 그 결과 두산은 0.189, KIA는 0.1357, SK는 0.7757 만큼의 오차가 감소한 것을 확인할 수 있었다. 뿐만 아니라 실제 투수력을 반영하였을 때 반영된 팀의 타자진이 약할수록 오차가 줄어드는 것을 관찰할 수 있었다. 이를 통해 투수, 구장 정보와 같은 다른 정보들을 이러한 방식으로 추가할 경우 보다 더 정확한 예측 결과를 얻을 수 있을 것으로 기대한다.

## 5. 참고문헌


[1] Buklet, et.al. “A MARKOV CHAIN APPROACH TO BASEBALL” , Operations Research, Vol.45, No.1, pp. 14-23, INFORMS, 1997

[2] 여재룡, “Markov Chain을 이용한 한국프로야구 모델링 및 실시간 승률 예측 모델 구현”, 대한산업공학회 추계학술대회 논문집, pp. 1062-1076, 2012

[3] Shane T. Jensen, “Hierarchical Bayesian Modeling of Hitting Performance in Baseball”, Bayesian Analysis, pp.631-652, 2009

[4] Nobuyoshi Hirotsu, “A MARKOV CHAIN APPROACH TO OPTIMAL PINCH HITTINGSTRATEGIES IN A DESIGNATED HITTER RULE BASEBALL GAME”, The Operations Research Society of Japan, pp. 353-371, 2003

[5] Ursin, Daniel Joseph, “A Markov Model for Baseball with Applications”, University of Wisconsin Milwaukee UWM Digital Commons, pp. 3-29, 2014

[6] 오윤학, 김한, 윤재섭, 이종석, “데이터마이닝을 활용한 한국프로야구 승패예측모형 수립에 관한 연구”, 대한산업공학회지 제40권 제1호, pp. 8-17, 2014