The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.4.649
26(4), 649–659

Analysis of the Korean Baseball League using a Markov Chain Model

Hyung Woo Moon[a] · Yong Tae Woo[a*] · Yang Woo Shin[b,1]

[a]Department of Computer Engineering, Changwon National University
[b]Department of Statistics, Changwon National University

(Received May 15, 2013; Revised July 27, 2013; Accepted August 11, 2013)

Abstract

We use a Markov chain model to analyze the Korean Baseball League. We derive the distributions of the number of runs scored and the number of batters that complete their turn at bat in a baseball game using the time inhomogeneous Markov chain. The model is tested with real data produced from the 2011 Korean Baseball League.

Keywords: Baseball, distribution of the number of runs scored, distribution of the number of batters, Markov chain.

1. 서론

전통적으로 야구경기의 승패나 득점을 예측하기 위한 모형 개발은 흥미로운 연구 주제 중 하나이다. 야구 경기에서 승패나 득점을 예측하기 위한 방법은 데이타마이닝 기법이나, 회귀분석, 판별분석 등과 같은 다양한 통계기법과 마르코프 연쇄를 이용하여 타율이나 득점과 같이 승패에 주된 영향을 주는 요인들을 분석하는 연구가 진행되고 있다. 이중에서 마르코프 연쇄 모형은 타율과 진루데이타로 구성한 진루행렬을 이용하여 경기의 진행상황을 동적으로 반영할 수 있는 장점이 있다. 이에 따라 마르코프 연쇄 모형을 이용하여 야구경기의 승패나 득점을 예측하기 위한 연구가 활발하게 진행되고 있다.

Bukiet 등 (1997)은 야구경기에서 승리할 확률이나 최다 점수를 얻기 위한 최적의 타순결정을 위해 마르코프 연쇄를 이용한 득점예측 모형을 제시하였다. Hirotsu와 Wright (2003, 2005)는 Bukiet이 제시한 모형을 바탕으로 대타전략과 투수교체전략에 적용할 수 있는 개선된 모형을 제시하였다. Sokol (2003)은 마르코프 연쇄를 이용한 득점모형을 이용하여 타자의 유형을 전방위 기여자, 기회에 강한 타자, 약타자, 기회를 만드는데 강한타자 등으로 분류하여 최적의 타순 구성 전략에 반영한 방법을 제시하였다. Tesar는 팀별 진루행렬을 이용하여 각 팀의 평균 득점수와 평균 타자수를 구하였다.

최근에 국내에서도 프로야구 데이타를 이용하여 승률과 승패모형에 대한 연구가 활발히 진행되고 있다. Cho 등 (2007)은 로지스틱회귀분석, 판별분석, 의사결정나무모형을 이용하여 한국프로야구팀의 승패요

*This reach is financially supported by Changwon National University in 2011−2012.
[1]Corresponding author: Professor, Department of Statistics, Changwon National University,Changwon, Gyeongnam 641-774, Korea. E-mail: ywshin@changwon.ac.kr

Table 2.1. States of X_n

	000	100	010	001	110	101	011	111
no outs	1	2	3	4	5	6	7	8
one out	9	10	11	12	13	14	15	16
two outs	17	18	19	20	21	22	23	24
three outs				25				

인분석에 관한 연구를 하였다. 또한 Lee와 Cho (2009)는 기존의 통계모형과 신경망모형을 활용하여 야구경기에서 승패모형을 제시하였다. 그밖에 Bae 등 (2012), Choi와 Kim (2011), Kim (2011) 등과 같이 통계기법을 이용한 연구는 활발하게 진행되고 있다. 그러나 국내에서 한국프로야구 데이타를 이용한 마르코프 연쇄 모형에 대한 연구는 거의 이루어지지 않았다.

야구에서 진루규칙은 타격 결과에 대해 주자상태의 변화를 정의한 규칙이다. 이러한 진루규칙은 효율적인 타순을 결정하기 위한 모델이나 경기당 득점을 예측하기위한 마르코프 연쇄 모형 개발을 위해 중요하게 활용된다. Bukiet 등 (1997)이 제안한 마르코프 연쇄 모형에서는 D'Esopo와 Lefkowitz (1960)가 제안한 진루규칙을 사용하였다. 하지만 이 진루규칙은 야구의 기본 규칙만을 이용하여 정의한 관계로 실제 경기에 적용하기 위해서는 좀더 세밀하게 현실 상황을 반영한 진루행렬의 구성이 필요하다. 예를 들어, 주자 1루 상황에서 타자가 1루타를 쳤을 때, 주자의 상태는 1, 2루 또는 1, 3루가 될 수 있다. 야구 동호인 사이트인 아이스텟(www.istat.co.kr)의 자료에 따르면 실제 2007년 부터 2011년까지 지난 5년간 한국프로야구에서 주자가 1루 상황에서 타자가 1루타를 쳤을 때, 주자가 1, 2루가 되는 경우는 약 70%였고, 1, 3루가 되는 경우는 약 30%였다.

본 논문에서는 타자의 타격결과와 주자상태를 나타내는 확률과정을 구체적으로 정의하여 경기진행 상황을 마르코프 연쇄로 나타내고 경기별 득점 분포와 타석에 서는 타자 수의 분포를 구하는 알고리즘을 제시하였다. 또한 한국프로야구 경기에서 발생한 실제 데이타를 이용한 진루규칙을 제시하고 그 결과를 프로야구경기에 적용하여 실제 경기결과와 비교하였다. 본 논문에서는 타자별 진루행렬을 이용하므로 Tesar의 방법에 비하여 타순의 변화나 타자의 타격능력 변화를 동적으로 반영할 수 있는 장점이 있다.

이 논문의 구성은 다음과 같다. 제 2절에서는 마르코프 연쇄를 이용하여 야구경기를 나타내고 각 이닝별 점수의 분포와 타석에 서는 타자수의 분포뿐만 아니라 한 게임에서의 분포를 구하는 알고리즘을 제시하였다. 제 3절에서는 제안된 모형의 결과를 실증자료와 비교하여 모형의 정확성을 검토하고 과거 자료를 바탕으로 차년도의 득점을 예측하여 실제 경기결과와 비교한다. 제 4절에서는 이 논문의 결론 및 향후 과제에 대해서 기술한다.

2. 마르코프 연쇄 모형

주자상태를 나타내기 위하여 i번째 루(base)의 상태를 b_i $(i = 1, 2, 3)$라 하자. 이 때 i번째 루상에 주자가 있으면 $b_i = 1$ 그렇지 않으면 $b_i = 0$으로 나타내고 전체 주자상태는 $b_1b_2b_3$로 나타낸다. 예를 들어 '000'은 1루, 2루, 3루가 모두 비어있는 상태를 나타내며 '101'은 1루와 3루에는 주자가 있으나 2루는 비어있는 상태를 나타낸다. 한 이닝에서 n명의 타자가 타석을 마쳤을 때 아웃된 타자 수와 주자상태를 나타내기 위하여 확률변수 X_n을 Table 2.1과 같이 정의하자. 예를 들어 한 이닝에서 n명의 타자가 타석을 소화했을 때 한 명의 타자가 아웃을 당하고 주자상태가 '101'일 때, $X_n = 14$이다.

한 이닝에서 n번째 타자가 타석을 마쳤을 때 아웃된 타자 수와 주자 상태는 그 타자가 타석에 서기 직전의 상태와 그 타자의 타격 결과에만 의존한다고 하면 확률과정 $\boldsymbol{X} = \{X_n, n = 1, 2, \ldots\}$은 상태공간

Table 2.2. States of H_n

타격결과	1루타	2루타	3루타	홈런	사사구	아웃
H_n	1	2	3	4	5	6

이 $\mathcal{S} = \{1, 2, \ldots, 25\}$인 마르코프 연쇄가 된다. 한 이닝에서 세명의 타자가 아웃을 당하면 그 이닝은 끝나게 되므로 상태 25는 마르코프 연쇄 $\boldsymbol{X}$의 흡수상태가 된다. n번째 타자의 타격 결과를 H_n으로 두고 Table 2.2와 같이 정의하고 이 때 얻는 점수를 R_n이라 하자. $X_{n-1} = i$인 상태에서 n번째 타석에 선 타자의 타격결과가 $H_n = h$일 때, r득점하고 주자상태가 $X_n = j$가 될 확률을

$$P_H^{(r)}(h; i, j) = P\left(R_n = r, X_n = j \,|\, H_n = h, X_{n-1} = i\right)$$

이라 하고 $P_H^{(r)}(h; i, j)$를 (i, j)성분으로 갖는 24×24행렬을 $P_H^{(r)}(h) = (P_H^{(r)}(h; i, j))_{1 \le i, j \le 24}$으로 두자. 한 이닝에서 아웃된 타자 수는 감소하지 않으므로 행렬 $P_H^{(r)}(h)$는 다음과 같은 형태가 된다.

$$P_H^{(r)}(h) = \begin{pmatrix} A_0^{(r)}(h) & B_0^{(r)}(h) & C_0^{(r)}(h) \\ O & A_1^{(r)}(h) & B_1^{(r)}(h) \\ O & O & A_2^{(r)}(h) \end{pmatrix}.$$

단, $A_k^{(r)}(h)$, $B_k^{(r)}(h)$, $C_k^{(r)}(h)$는 크기가 8×8인 블록행렬이고 O은 영행렬이다. $P_H^{(r)}(h)$의 첫번째 블록에 있는 행렬 $A_0^{(r)}(h)$, $B_0^{(r)}(h)$, $C_0^{(r)}(h)$은 각각 아웃된 타자의 수가 0인 상황에서 석에 선 타자의 타격결과가 h 라는 가정하에 r득점을 하고 아웃카운트가 $0, 1, 2$만큼 증가하면서 주자 상태의 변화를 나타낸다. 마찬가지로 두번째와 세번째 블록에 있는 행렬 들은 아웃된 타자 수가 각각 1과 2인 상황에서 타격결과에 따른 득점과 아웃된 타자수, 주자상태의 변화를 나타낸다.

$X_{n-1} = i$인 상태에서 타석에 선 타자의 타격 결과가 h일 확률을 $P_n^H(i; h) = P(H_n = h \,|\, X_{n-1} = i)$로 두면 전확률공식에 의하여 다음이 성립함을 알 수 있다.

$$\begin{aligned} P_n^{(r)}(i, j) &= P(R_n = r, X_n = j \,|\, X_{n-1} = i) \\ &= \sum_{h=1}^{6} P(R_n = r, X_n = j \,|\, H_n = h, X_{n-1} = i) P(H_n = h \,|\, X_{n-1} = i) \\ &= \sum_{h=1}^{6} P_H^{(r)}(h; i, j) P_n^H(i; h), \quad 1 \le i \le 24,\ 1 \le j \le 25,\ r = 0, 1, 2, 3, 4. \end{aligned} \tag{2.1}$$

25는 $\boldsymbol{X}$의 흡수상태이므로, 25×25행렬 $P_n^{(r)} = (P_n^{(r)}(i, j))$은 다음과 같이 나타낼 수 있다.

$$P_n^{(r)} = \begin{pmatrix} Q_n^{(r)} & \boldsymbol{q}_n^{(r)} \\ \mathbf{0} & 1 \end{pmatrix}, \quad r = 0, 1, 2, 3, 4. \tag{2.2}$$

단, $Q_n^{(r)}$은 24×24 행렬이고 $\boldsymbol{q}_n^{(r)}$은 3아웃에 대응하는 24×1행렬이다. 따라서 마르코프 연쇄 $\boldsymbol{X}$의 전이확률행렬은 다음과 같다.

$$P_n = \begin{pmatrix} Q_n & \tilde{\boldsymbol{q}}_n \\ \mathbf{0} & 1 \end{pmatrix}, \quad n = 1, 2, \ldots.$$

단, $Q_n = \sum_{r=0}^{4} Q_n^{(r)}$, $\boldsymbol{q}_n = \sum_{r=0}^{4} q_n^{(r)}$.

2.1. 한 이닝 분석

이제 k번 타자부터 시작하는 이닝에서 얻는 점수의 분포와 타자수의 분포를 구하자. P_n은 진루행렬 $P_H^{(r)}(h)$와 n번째 타석에 선 타자의 타격결과 $P_n^H(i;h)$에 의하여 정해지므로 타석에 서는 타자에 따라 서로 다른 행렬이 주어진다. 한 이닝에서 n번째 타석에 서는 타자는 그 이닝의 선두타자가 누구냐에 따라 정해지므로 표현의 명확성을 위하여 $Q_n^{(r)}$, $P_n^{(r)}$ 대신에 선두타자 k를 나타내서 $Q_{k,n}^{(r)}$, $P_{k,n}^{(r)}$과 같이 표현하자.

한 이닝에서 얻는 점수의 분포. k번 타자부터 시작하는 이닝에서 n명의 타자가 타석을 마쳤을 때, r 득점하고 주자 상태(아웃카운트 포함)가 i에 있을 확률을 $U_{k,n}(r,i)$라 하고

$$U_{k,n}(r) = (U_{k,n}(r,1), U_{k,n}(r,2), \ldots, U_{k,n}(r,25)), \quad n = 0,1,2,\ldots$$

이라 하자. 단, 이닝이 시작될 때는 루상에 주자가 없고 아웃당한 선수가 없을뿐만 아니라 그 이닝에서의 득점도 없으므로 $U_{k,0}(0) = (1,0,\ldots,0)$이다. k번 타자부터 시작하는 이닝에서 n명의 타자가 타석을 마쳤을 때, r 득점할 확률은 $n-1$명의 타자가 타석을 마쳤을 때까지 $r-j$ 득점하고 n번째 타자가 타석을 마친 다음 j $(0 \le j \le \min(r,4))$득점하면 되므로 전확률공식에 의하여 다음이 성립함을 알 수 있다.

$$U_{k,n}(r) = \sum_{j=0}^{\min(r,4)} U_{k,n-1}(r-j)P_{k,n}^{(j)}, \quad r = 0,1,\ldots,n. \tag{2.3}$$

따라서 k번 타자부터 시작한 이닝에서 그 이닝이 끝날 때까지 r득점할 확률 $v_k(r)$은 다음과 같다.

$$v_k(r) = \lim_{n\to\infty} U_{k,n}(r,25), \quad r = 0,1,2,\ldots,\ 1 \le k \le 9. \tag{2.4}$$

참고. $v_k(r)$을 계산하기 위해서는 주어진 $\epsilon > 0$(예 $\epsilon = 10^{-4}$)에 대하여 $v_k(0) + v_k(1) + \cdots > 1 - \epsilon$을 만족할 때까지 n을 증가시킨다.

한 이닝에서 타석에 서는 타자수의 분포. k번 타자부터 시작하는 이닝에서 세 명의 타자가 아웃 될 때까지 타석에 서는 타자수 T_k, $k = 1,2,\ldots,9$의 분포는 흡수상태를 갖는 마르코프 연쇄에서 흡수상태에 도달할 때까지 전이의 수를 나타내는 이산 상형분포(phase type distribution)가 되어 다음과 같다 (Shin, 2011).

$$P(T_k = n) = \begin{cases} U_{k,0}(0)Q_{k,1}Q_{k,2}\cdots Q_{k,n-1}\boldsymbol{q}_{k,n}, & n \ge 3, \\ 0, & n \le 2. \end{cases} \tag{2.5}$$

2.2. 한 게임 분석

첫 번째 이닝의 선두타자는 1번 타자이지만 그 이후 이닝의 선두타자는 직전 이닝의 마지막 타자에 따라 달라지게 된다. 따라서 한 게임을 분석하기 위해서는 i이닝 동안 타석에 서는 타자 수의 분포를 구할 필요가 있다. 이를 위하여 직전이닝까지 j명의 타자가 타석에 섰을 때, 다음 이닝의 선두타자는 다음과 같음을 알 수 있다.

$$\mathrm{LH}(j) = \begin{cases} (j+1) \bmod 9, & (j+1) \bmod 9 \not\equiv 0, \\ 9, & (j+1) \bmod 9 \equiv 0. \end{cases}$$

단, $n \bmod 9$는 n을 9로 나누었을 때 나머지이다. 따라서 세번째 이닝까지 13명의 타자가 타석에 섰다면 네번째 이닝의 선두타자는 $\mathrm{LH}(13) = 5$번 타자이다.

한 게임에서 타석에 서는 타자수의 분포. i이닝 동안 타석에 서는 타자 수를 N_i라 하자. 첫 번째 이닝은 1번 타자부터 시작하므로 $N_1 = T_1$이 된다. 또한 $i-1$ 이닝 동안 j명의 타자가 타석에 섰다면 i번째 이닝의 선두 타자는 LH(j)이므로 $N_i = T_{\mathrm{LH}(j)}$이다. 따라서 전확률공식에 의하여 다음을 얻는다.

$$P(N_i = n) = \begin{cases} P(T_1 = n), & i = 1, \\ \sum_{j=1}^{n} P(T_{\mathrm{LH}(j)} = n - j) P(N_{i-1} = j), & i = 2, 3, \ldots. \end{cases} \tag{2.6}$$

한 게임에서 얻는 점수의 분포. n번째 이닝에서 r득점할 확률을 $w_n(r)$이라 하면 $\boldsymbol{w}_n = (w_n(r), r = 0, 1, \ldots)$은 다음과 같다.

$$\boldsymbol{w}_n = \begin{cases} \boldsymbol{v}_1, & n = 1, \\ \sum_{k=1}^{\infty} P(N_{n-1} = k) \boldsymbol{v}_{\mathrm{LH}(k)}, & n = 2, 3, \ldots. \end{cases} \tag{2.7}$$

단, $\boldsymbol{v}_k = (v_k(r), r = 0, 1, \ldots)$. 따라서 m이닝 만에 경기를 마치는 게임(연장전으로 갈 경우는 9이닝 이상 경기할 수 있다)에서 한 팀이 얻는 점수의 분포 $\boldsymbol{w}^{(m)}$은 다음과 같다.

$$\boldsymbol{w}^{(m)} = \boldsymbol{w}_1 * \boldsymbol{w}_2 * \cdots * \boldsymbol{w}_m. \tag{2.8}$$

단, $\boldsymbol{c} = \boldsymbol{a} * \boldsymbol{b}$는 $\boldsymbol{a}$와 $\boldsymbol{b}$의 합성곱을 나타낸다. 즉, $\boldsymbol{c} = (c_0, c_1, \ldots)$의 n번째 성분은 $c_n = \sum_{i=0}^{n} a_i b_{n-i}$, $n = 0, 1, 2, \ldots$.

위의 결과는 다음과 같은 알고리즘으로 요약된다.

알고리즘

단계 1. 진루행렬의 구성

(1) 자료로부터 $P_H^{(r)}(h; i, j)$와 n번타자에 대한 $P_n^H(i, h)$을 계산한다.

(2) 위의 (1)에서 구한 자료와 식 (2.2)을 이용하여 각 선수별 진루행렬 $P_n^{(r)}$을 구한다.

단계 2. 한 이닝에서의 득점분포와 타자수 분포 계산

각각의 $k = 1, 2, \ldots, 9$에 대하여, 식 (2.4)와 (2.5)를 이용해 $v_k(r)$와 T_k의 분포를 각각 구한다.

단계 3. 한 경기에서의 득점분포와 타자수 분포 계산

식 (2.6)과 (2.8)을 이용하여 한 경기에서 타석에 서는 타자수 분포와 한 경기에서 얻는 점수의 분포를 각각 구한다.

3. 적용결과

3.1. 자료수집과 모형의 적용

본 논문에서는 실증자료분석을 위하여 아이스텟(www.istat.co.kr)에서 제공하는 2007년부터 2011년까지 한국프로야구 경기에서 발생한 타자와 투수에 대한 자료를 사용하였다. 이 사이트에서 제공하는 대량의 자료를 효과적으로 관리하기 위한 데이터 아키텍처를 설계하여 관계형 데이터베이스에 저장하고 자료 분석에 필요한 타율이나 진루확률에 필요한 자료는 SQL 언어를 사용하여 검색하였다. 모형의 적용대상으로는 투수진이 안정된 삼성라이온즈, 타순이 안정적인 롯데자이언츠 그리고 상대 투수에 따라 타순이 자주 바뀌는 SK와이번스팀의 2011년 경기를 선택하였다.

Table 3.1. Virtual lineup of each team in 2011

팀명	타순 (1–9번 타자까지 순서대로)
삼성 라이온즈	배영섭, 박한이, 박석민, 최형우, 채태인, 조영훈, 신명철, 진갑용, 김상수
롯데 자이언츠	전준우, 김주찬, 손아섭, 이대호, 홍성흔, 강민호, 조성환, 황재균, 문규현
SK 와이번스	정근우, 박재상, 최 정, 이호준, 박정권, 김강민, 박진만, 정상호, 조동화

Table 3.2. Expected number of runs scored and number of batters at bat in a game in 2011

팀명	경기당 평균득점 수			경기당 평균타자 수		
	MC모형	실제경기	오차	MC모형	실제경기	오차
삼성 라이온즈	4.69	4.70	−0.01	39.8	38.7	1.1
롯데 자이언츠	5.37	5.36	0.01	40.4	39.4	1.0
SK 와이번스	4.44	4.39	0.05	38.9	38.5	0.4

Table 3.3. Prediction of the expected number of runs scored and number of batters at bat in a game

연도	팀명	경기당 평균득점 수			경기당 평균타자 수		
		MC모형	실제경기	오차	MC모형	실제경기	오차
2009	삼성 라이온즈	4.21	5.15	−0.94	39.0	39.2	−0.2
	롯데 자이언츠	4.99	4.79	0.20	39.9	38.0	1.9
	SK 와이번스	4.56	5.50	−0.94	39.1	40.4	−1.3
2010	삼성 라이온즈	5.07	5.12	−0.05	40.1	39.8	0.3
	롯데 자이언츠	5.29	5.81	−0.52	40.2	39.6	0.6
	SK 와이번스	5.20	5.29	−0.09	40.3	39.2	1.1
2011	삼성 라이온즈	5.27	4.70	0.57	40.5	38.7	1.8
	롯데 자이언츠	5.57	5.36	0.21	40.2	39.4	0.8
	SK 와이번스	5.20	4.39	0.81	40.0	38.5	1.5

한국프로야구는 경기별로 선수와 타순의 변화가 많고, 대타 기용이나 투수 교체가 빈번하게 이루어지는 경향이 있다. 따라서 한국프로야구 데이터로부터 마르코프 연쇄 모형을 각 경기에 그대로 적용하기에는 충분한 자료를 구하기 어려웠다. 따라서 2011년에 출장한 타자 중에서 각 팀별로 포지션과 타순별로 출장 횟수가 가장 많은 선수들을 추출하여 Table 3.1과 같이 가상의 타순을 구성하였다. 선수별 타격확률은 2011년도 결과를 이용하였다. Table 3.2는 가상의 타순으로 구성된 팀에 마르코프 연쇄 모형을 적용한 결과(MC모형)와 실제 경기결과와 비교한 표이다. Table 3.2을 구하는 과정 중 일부는 부록에 제시하였다. 마르코프 연쇄 모형과 실제 경기간의 평균득점 차이는 0.01–0.05점이고, 평균타자 수의 차이는 약 1명 정도로 나타났다.

3.2. 과거의 자료로부터 차년도 경기예측

Table 3.3은 2009년 부터 2011년까지 3년동안 각각 과거 2년간의 자료를 바탕으로하여 마르코프 연쇄 모형을 이용한 팀별 게임당 평균득점 및 타자수를 예측한 결과이다. 예를 들어 2009년도 경기결과는 2009년도에 출현빈도가 가장 높았던 선수들을 추출하여 가상의 팀을 구성하고 각 선수의 2007년과 2008년 2년간의 타격결과를 바탕으로 타격확률을 구하여 경기당 평균 득점과 타자수를 계산하였다. 이때 타석 수가 200타석 이하로 상대적으로 적은 선수는 과거 5년간의 통산 성적에 의해 타격확률을 계산하였다. 2010년과 2011년의 결과도 같은 방법으로 구하여 예측하였다. 예측결과와 실제경기결과의 차이가 가장 적게 나타난 경우는 2010년 삼성 라이온즈의 경우로써 마르코프 연쇄 모형을 이용한 평균득

점은 5.07점이었고 실제 평균득점은 5.12점으로 차이가 0.05점이었다. 그리고 차이가 가장 크게 나타난 경우는 2009년 삼성과 SK의 경우로써 예측점수와 실제 점수간의 차이가 0.94점이었다. 그 이유는 본 모형에서는 포지션별로 출전 빈도수가 가장 높은 선수와 고정된 타순을 적용하였지만, 실제 경기에서는 경기마다 타자와 타순이 바뀌고 경기 중에도 선수가 교체되기 때문으로 여겨진다. 그리고 각 선수들의 기량이 해마다 변화하는 것도 예측 결과와 실제 결과간의 차이를 보이는 요인이 될 수 있다고 생각된다. 선수별 기량 변화에따른 각 타자의 타격확률 $P_n^H(i;h)$의 변화는 전년도 경기 결과를 이용하여 2011년도 경기를 예측한 결과 Table 3.3가 당해년도 경기결과를 이용하여 적용한 2011년도의 경기결과 Table 3.2와 차이를 보이는 요인이 됨을 알 수 있다.

4. 결론

마르코프 연쇄로 모형을 프로야구 경기에 적용하기 위해서는 타격 결과에 대해 주자상태의 변화를 나타내는 진루행렬과 타자의 타격확률의 계산에 실제 경기 상황이 정확하게 반영되어야 한다. 본 논문에서는 마르코프 연쇄를 이용하여 한국프로야구의 경기결과를 예측하고 분석하였다. 타자의 타격결과와 주자상태를 나타내는 확률과정을 구체적으로 정의하여 경기진행 상황을 동적으로 반영한 마르코프 연쇄를 구성하여 경기당 득점 분포와 타석에 서는 타자 수의 분포를 구하였다. 실제 데이터를 바탕으로 주자 상태를 고려한 진루행렬과 각 선수별 타격 확률을 구하여 한국프로야구 경기에서 평균득점과 타석에 서는 타자 수의 평균을 구하고 실제 모형과 비교 하였다.

야구경기의 승패를 결정하는데 중요한 요인의 하나는 상대팀의 투수능력이다. 또한 도루 등과 같이 타자의 타격 결과 이외에 주자의 주루플레이 등도 득점에 영향을 미칠 수 있는 요인이 된다. 본 연구에서 적용한 모형은 타자의 타격능력을 중심으로 한 모형으로 이와 같은 형태의 요소는 고려하지 않았다. 보다 정확한 득점예측 등과 같은 경기 결과 예측을 이와 같은 요소들을 반영한 마르코프연쇄를 구성하는 것이 필요할 것으로 여겨진다.

부록

여기서는 실제 경기결과를 이용하여 득점수 및 타석에 선 타자수 분포를 구하는 과정을 간략하게 설명한다.

(1) 프로야구경기 자료를 바탕으로 $P_H^{(r)}(h)$을 구한다. 다음은 2007년부터 2011년까지 한국프로야구 경기에서 나타난 결과를 이용하여 구한 진루행렬 $P_n^H(i;h)$에서 영이 아닌 블록의 행렬이다. 예를 들어 행렬 $A_0^{(1)}(1)$의 $(3,2)$-성분 0.307은 0아웃에 한 명의 주자가 2루에 있는 상황 ('010')에서 타자가 1루타를 쳤을 때 즉, $H_n = 1$일 때, 1득점하고 타자는 1루에 나갈 확률을 나타낸다.

$$A_0^{(1)}(1) = \begin{pmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.307 & 0.051 & 0.007 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.812 & 0.071 & 0.024 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.323 & 0.154 & 0.018 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.640 & 0.223 & 0.052 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.031 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.444 \end{pmatrix}$$

$$B_0^{(1)}(1) = \begin{pmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.016 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.035 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.015 & 0.013 & 0.008 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.024 & 0.033 & 0.009 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.015 & 0.031 & 0.015 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.016 & 0.008 & 0.008 & 0.000 \end{pmatrix}$$

$$A_1^{(1)}(1) = \begin{pmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.409 & 0.084 & 0.007 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.782 & 0.130 & 0.010 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.332 & 0.180 & 0.042 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.687 & 0.201 & 0.034 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.004 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{pmatrix}$$

$$B_1^{(1)}(1) = \begin{pmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.040 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.068 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.011 & 0.034 & 0.012 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.011 & 0.029 & 0.020 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.039 & 0.018 & 0.039 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.024 & 0.005 & 0.012 & 0.000 \end{pmatrix}$$

$$A_2^{(1)}(1) = \begin{pmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.003 & 0.002 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.502 & 0.173 & 0.012 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.785 & 0.143 & 0.010 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.363 & 0.252 & 0.064 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.663 & 0.253 & 0.037 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.145 \end{pmatrix}.$$

또한 아웃된 타자 수가 0 혹은 1인 상황에서 1루타를 쳤을 때 1득점하고 3아웃이 발생할 확률은 0이었고, 두명의 타자가 아웃인 상황에서 1루타를 쳤을 1득점하고 3아웃이 발행하여 이닝이 끝날 확률은 다음과 같았다.

$$q_H^{(1)}(1) = (0.000, 0.000, 0.065, 0.062, 0.083, 0.041, 0.096, 0.099)$$

예를 들어 $q_H^{(1)}(1)$의 세번째 성분 0.065는 2아웃이고 주자상태가 '010'인 상황에서 타자가 1루타를 쳤을 때 1득점한 후 타자가 아웃이 되어 3아웃이 발생할 확률을 나타낸다.

(2) 프로야구경기 자료를 바탕으로 각 타자의 타격확률 $P_n^H(i;h)$을 구한다. 여기서는 주자 상태를 고려하지 않고 2011년 1년 통산 타격확률을 사용하였다. 예를 들어 롯데 자이언츠 선수의 1번, 2번,

Table A.1. Hitting rates of batters

선수명	1루타	2루타	3루타	홈런	사사구	아웃
전준우	0.178	0.065	0.008	0.018	0.095	0.636
김주찬	0.213	0.032	0.008	0.016	0.092	0.639
손아섭	0.201	0.051	0.010	0.031	0.091	0.616

3번타자의 2011년 1년 통산 타격확률 $P_n^H(i;h)$은 Table A.1과 같다.

(3) 각 선수의 진루확률 $P_n^{(r)}$은 다음과 같이 구한다.

$$P_n^{(r)} = \sum_{h=1}^{6} \Delta_n(h) P_H^{(r)}(h), \quad r = 0, 1, 2, 3, 4.$$

단, $\Delta_n(h) = \mathrm{Dia}[P_n^H(1,h), \ldots, P_n^H(24,h), 0]$은 대각성분이 $(P_n^H(1,h), \ldots, P_n^H(24,h), 0)$인 대각행렬이다. 롯데 자이언츠의 1번타자 전준우 선수의 진루행렬은 1루타를 칠 확률

$$\Delta_{\text{전준우}}(1) = \mathrm{Dia}[0.178, 0.178, \ldots, 0.178, 0]$$

과 2루타를 칠 확률

$$\Delta_{\text{전준우}}(2) = \mathrm{Dia}[0.065, 0.065, \ldots, 0.065, 0]$$

등과 같이 타격확률을 이용하여 다음과 같이 구한다.

$$P_{\text{전준우}}^{(r)} = \sum_{h=1}^{6} \Delta_{\text{전준우}}(h) P_H^{(r)}(h), \quad r = 0, 1, 2, 3, 4.$$

같은 방법으로 2번 타자, 3번 타자 등 9명의 선수에 대한 진루확률을 구한다.

득점 분포 및 타자수 분포를 구하는 다음 단계는 알고리즘의 단계를 따라서 쉽게 구현할 수 있으므로 생략한다.

References

Bae, J. Y., Lee, J. M. and Lee, J. Y. (2012). Predicting Korea Pro-baseball rankings by principal component regression analysis, *Communications of the Korean Statistical Society*, **19**, 367–379.

Bukiet, B., Harold, E. R. and Palacios, J. L. (1997). A Markov chain approach to baseball, *Operations Research*, **45**, 14–23.

Cho, Y. S., Cho, Y. J. and Shin, S. G. (2007). A study on winning and losing in Korean professional baseball league, *Journal of the Korean Data Analysis Society*, **9**, 501–510.

Choi, Y. G. and Kim, H. M. (2011). A statistical study on Korean baseball league games, *The Korean Journal of Applied Statistics*, **24**, 915–930.

D'Esopo, D. A. and Lefkowitz, B. (1960). The distribution of runs in the game of baseball, SRI Internal report.

Hirotsu, N. and Wright, M. (2003). A Markov chain approach to optimal pinch hitting strategies in a designated hitter rule baseball game, *Journal of the Operations Research Society of Japan*, **46**, 353–371.

Hirotsu, N. and Wright, M. (2005). Modelling a baseball game to optimise pitcher substitution strategies incorporating handedness of players, *IMA Journal of Management Mathematics*, **16**, 179–194.

Kim, H. (2011). Suggestion of a new method of computing percentage of victories for the Korean professional baseball, *The Korean Journal of Applied Statistics*, **6**, 1139–1148.

Lee, J. T. and Cho, H. S. (2009). Win-loss models when two teams meet using data mining in the Korean pro-baseball, *Journal of the Korean Data Analysis Society*, **11**, 3417–3426.

Shin, Y. W. (2011). *Introduction to Stochastic Processes*, Kyungmoon Publishers, Seoul.

Sokol, J. S. (2003). A robust heuristic for batting order optimization under uncertainty, *Journal of Heuristics*, **9**, 353–370.

Tesar, N. Estimating expected runs using a Markov model for baseball, https://www.edsolio.com/media/2/265/files/TesarFinal_Draft.pdf.

마르코프 연쇄를 이용한 한국 프로야구 경기 분석

문형우[a] · 우용태[a*] · 신양우[b,1]

[a]창원대학교 컴퓨터공학과, [b]창원대학교 통계학과

(2013년 5월 15일 접수, 2013년 7월 27일 수정, 2013년 8월 11일 채택)

요 약

본 논문에서는 마르코프 연쇄로 모형을 이용하여 한국프로야구의 경기결과를 예측하고 분석하였다. 타자의 타격결과와 주자상태를 나타내는 확률과정을 구체적으로 정의하여 경기진행 상황을 동적으로 반영한 프로야구 경기를 마르코프 연쇄를 구성하여 실제 데이터를 바탕으로 주자 상태를 고려한 진루행렬과 각 선수별 타격 확률을 구하여 경기당 득점 분포와 타석에 서는 타자 수의 분포를 구하였다.

주요용어: 득점 분포, 마르코프 연쇄, 야구, 타자수 분포.

*이 논문은 2011−2012년도 창원대학교 연구비지원을 받았음.

[1]교신저자: (641-773) 경남 창원시 의창구 창원대학로 20, 창원대학교 통계학과, 교수.
E-mail: ywshin@changwon.ac.kr