Journal of The Korea Society of Computer and Information
Vol. 24 No. 1, pp. 41-47, January 2019
www.ksci.re.kr
https://doi.org/10.9708/jksci.2019.24.01.041

A Baseball Batter Evaluation Model using Genetic Algorithm

Su-Hyun Lee*, Yerin Jung**, Hyung-Woo Moon***, Yong-Tae Woo*

Abstract

In this paper, we propose a new batter evaluation model that reflects the skill of the opponent pitcher in Korean professional baseball. The model consists of evaluation factors such as Run Value, Contribution Score and Ball Consumption considering the pitcher grade. These evaluation factors are calculated as different data. In order to include the evaluation factors having different characteristics into one model, each evaluation factor is weighted and added. The genetic algorithms were used to calculate the weights, and the data were based on the 2016 records of Korea Professional Baseball and the salary data of the players of 2017. As a result of calculation of the weight, the weight of the Run Value was high and the weight of the Contribution Score was very low. This means that when calculating the annual salary, it reflects much of the expected score according to the batting result of the batter. On the other hand, the contribution score indicating the degree to which the batting result contributed to the victory of the team according to the state of the economy is not reflected in the salary or point system.

▸Keyword: Professional baseball, Batter evaluation, Genetic algorithm, Baseball database, Pitcher weight

I. Introduction

야구는 다른 스포츠보다 훨씬 많은 데이터를 양산하며, 타율, 방어율 등과 같이 확률을 자연스럽게 사용한다. 야구경기에서 발생된 데이터는 통계적인 분석 과정을 거쳐 야구선수를 평가하기 위한 지표나 경기운영을 위한 자료로 사용된다.

미국 메이저리그 오클랜드 팀의 빌리 빈 단장은 선수들의 평가에 통계 데이터 분석을 이용하는 머니볼 이론으로 크게 성공하였다[1]. 빌리 빈 단장은 야구경기의 승패에 밀접한 관계가 있는 야구선수 평가 지표를 분석한 후 연봉에 비하여 평가치가 높은 선수들을 영입하여 메이저리그 최초로 20연승을 이루어 내었다. 그 이후로 메이저리그를 중심으로 야구 데이터 분석이 활발하게 이루어지게 되었다.

타자의 기량을 효과적으로 평가하기 위하여 OPS(On base Plus Slugging), WHIP(Walks plus Hits divided by Innings Pitched), WAR(Win Above Replacement) 등의 다양한 타격 지표를 통계적 기법으로 분석하는 연구가 활발하게 진행되고 있다. 최근에는 상대 투수의 기량을 반영한 타자평가모형이 제안되었다. 이 모형[2]은 기준배점(Run Value), 기여도점수(Contribution Score), 투수가중치를 고려한 투구소모점수(Ball Consumption) 등의 평가요소로 구성되는데, 평가요소들의 비중이 모두 같은 것으로 간주하여 각 평가요소의 점수를 합산하여 계산하였다. 서로 성격이 다른 평가요소들을 단순한 합산을 통하여 하나의 모형에 포함하는 것은 좋은 방법이 아니다. 단순히 합산을 하기 보다는 각 평가요소에 가중치를 부여하여 합산하는 것이 더 합당한 방법이다.

여러 평가요소들의 가중치를 결정하는 문제는 일종의 최적화 문제에 해당한다. 즉, 실제의 데이터에 가장 근접한 가중치

• First Author: Su-Hyun Lee, Corresponding Author: Su-Hyun Lee
*Su-Hyun Lee (suhyun@sarim.changwon.ac.kr), Dept. of Computer Engineering, Changwon National University
**Yerin Jung (wjddpfls0@gmail.com), Research Institute, HiBrain.Net
***Hyung-Woo Moon (hwmoon@changwon.ac.kr), Institute of Industrial Technology Research Center, Changwon National University
*Yong-Tae Woo (ytwoo@changwon.ac.kr), Dept. of Computer Engineering, Changwon National University
• Received: 2019. 01. 21, Revised: 2019. 01. 29, Accepted: 2019. 01. 29.
• This research is financially supported by Changwon National University in 2017~2018.

를 찾는 것이다. 기준이 되는 데이터는 여러 가지가 있으나, 본 논문에서는 한국프로야구 선수들의 연봉을 기준으로 평가요소들의 비중을 계산하였다. 이러한 최적화 문제를 해결하기 위하여 이를 위하여 유전 알고리즘을 이용하였다.

본 논문의 구성은 다음과 같다. 2절에서는 야구데이터를 이용한 연구동향과 유전 알고리즘에 대해서 정리하였다. 3절에서는 기존의 타자평가모형에 대해서 설명하였다. 4절에서는 타자평가모형에서 각 평가요소의 가중치를 계산하기 위하여 유전 알고리즘을 적용하는 방법을 설명하였다. 마지막으로 5절에서는 결론을 맺는다.

II. Preliminaries

1. Related works

타자의 기량을 효과적으로 평가하기 위하여 야구의 타격관련 데이터를 활용하는 연구가 활발히 이루어지고 있다. 많은 연구들이 출루율과 장타율을 합한 지표인 OPS를 활용하고 있다. 이장택[3]은 한국프로야구에 적합한 공격 지표를 분석하고자 주성분분석과 K-평균 군집분석을 이용하였다. 조영석 등[4]은 OPS가 득점에 미치는 영향을 분석하고자 상관분석, 군집분석, 회귀분석을 이용하였다. 김혁주[5]는 OPS를 분해하여 OBP(On Base Percentage) 및 SLG(SLuGging percentage)와 득점간의 상관관계를 분석하여 출루율과 장타율에 가중값을 주는 OPS를 제안하였다. 정진상[6]은 타자가 승리에 기여한 정도에 따라 점수를 부여하기 위하여 타격전후의 주자상태, 아웃상태, 득점 차 등을 비교하는 타자평가모형을 제안하였다.

Johnson[7]은 마이너리그 타자를 대상으로 세이버메트릭스 지표와 로지스틱 회귀를 이용한 타자평가모형을 제안하였다. Yates[8]는 깁스 샘플링 알고리즘을 이용하여 좌타자/우타자, 홈경기/원정경기, 야간경기/주간경기 등과 같은 요인이 OPS에 미치는 영향에 대해 연구하였다. McShane 등[9]은 계층적 베이지안 모형을 이용하여 타자들의 공격력을 평가하는 방법을 제안하였다. Rubin[10]은 대체선수 대비 승리기여도를 나타내는 WAR 지표를 사용하여 타자의 공격 능력과 연봉간의 상관관계를 연구하였다.

2. Genetic Algorithm

최적화 문제(optimization problem)는 특정한 집합에서 정의된 값에 대해 그 값이 최대/최소가 되는 상태를 찾는 문제이다. 유전 알고리즘은 최적화문제의 해를 구하기 위하는 한 가지 방법이다. 유전 알고리즘은 생물학적 진화와 자연 선택의 기본 원리에 기반을 둔 확률적 탐색 알고리즘이다. 유전 알고리즘은 Holland[11]에 의해 개발되었고 Goldberg[12]에 의해 체계적으로 구체화되었다.

유전 알고리즘에서는 각각의 해가 하나의 개체가 되어, 이들의 집합인 개체군(population)이 진화를 하는 방식으로 전개된다. 한 개체는 하나 이상의 염색체(chromosome)들로 구성된다. 개체군들이 가진 염색체들이 유전 연산을 통하여 진화하여 좀 더 갱신된 개체군을 생성한다. 개체가 얼마나 좋은 해(solution)인지를 나타내는 최적화 목표함수가 있어 이 함수의 값을 적합도(fitness)라고 한다. 적합도가 클수록 좋은 해로 간주한다. 즉, 진화를 거듭할수록 목표함수의 값이 최대가 되는 방향으로 알고리즘이 실행된다. 이를 위하여 우수한 적합도의 부모 해는 보존시키고, 또한 우수한 부모들의 염색체의 재조합으로 자손 해를 생성한다. 유전 알고리즘에서 사용되는 유전 연산자와 중요 개념은 다음과 같다.

○ 선택

한 세대의 해 중에서 다음 세대의 해를 구성하는데 참여하는 해를 선택한다. 해를 선택하는 방법은 알고리즘의 성능에 큰 영향을 준다. 일반적으로는 적합도가 높은 해의 순으로 선택될 확률을 높게 부여한다. 그러나 현 세대에서 가장 좋은 해는 최종적인 최적해가 될 수도 있지만 지역적으로 좋은 해일 수도 있고, 반대로 현 세대에서는 좋지 않은 해라 할지라도 최종적인 최적해에는 더 가까울 수도 있다. 전역 최적해가 무엇인지를 모르는 상황이므로 가능한 해들의 평균적인 적합도를 높여 가는 것과 동시에 유전자의 다양성을 유지하는 것이 지역적인 최적해에 빠지는 것을 방지하게 된다.

○ 교차

생명체는 한 세대에서 교배를 통해서 다음 세대의 자손을 생산한다. 비슷하게, 유전 알고리즘에서는 선택된 해들을 교배하여 다음 세대의 해들을 생성한다. 두 개의 해를 선택한 후 교차 연산을 통해서 두 해를 섞어서 새로운 해를 구성하게 된다.

○ 변이

생명체에서는 유전자가 직접 변이를 일으켜 주어진 환경에서 살아남을 수 있다. 유전 알고리즘에서 변이 연산은 주어진 해의 염색체 내의 변수들의 순서 또는 값을 변경하여 다른 해로 변형하는 연산이다. 낮은 확률로 변이 연산을 수행하면 세대의 모든 해가 함께 지역적인 최적해에 들어가는 경우를 줄여주며, 해집단의 다양성을 높여 주게 된다.

○ 대치

교차나 변이를 거쳐서 만들어진 해를 새로운 세대의 해집단에 추가하고 현 세대의 해 중에서 적합도가 낮은 해를 제외시키는 연산이다.

III. A Batter Evaluation Model

본 연구에서 사용한 타자평가모형[2]은 기준배점, 기여도점

수, 투구소모점수, 투수가중치 등의 평가요소로 구성된다. 기준배점(Run Value)은 타자의 공격력을 평가하는 요소로 각 타격의 득점기댓값과 기대득점값을 이용하여 계산한 점수이다. 득점기댓값은 주자상태, 아웃상태별로 얻을 수 있는 기대득점의 평균을 의미하고, 기대득점값은 안타, 2루타, 홈런, 삼진, 병살 등과 같은 항목별로 타격의 결과로 얻을 수 있는 득점의 평균을 의미한다[6]. 기여도점수(Contribution Score)는 같은 타격결과라 할지라도 경기상태에 따라 가치를 차별하여 평가하기 위해 승리기댓값을 이용하여 타격이 승리에 기여한 정도에 따라 부여하는 점수이다. 승리기댓값은 경기상태별 기대 승리 확률을 의미한다[13]. 투구소모점수(Ball Consumption)는 타자가 상대 투수로 하여금 소모시킨 투구수에 따라 주어지는 점수이다. 투수가중치(Pitcher Weight)는 미국 메이저리그에서 사용하는 사이 영 포인트(Cy Young Point)와 같은 방법으로 투수의 성적에 따라 등급을 부여하는 점수이다. 투수가중치는 기준배점, 기여도점수, 투구소모점수에 반영한다.

그림 1은 타자평가모형의 전체 개념이다.

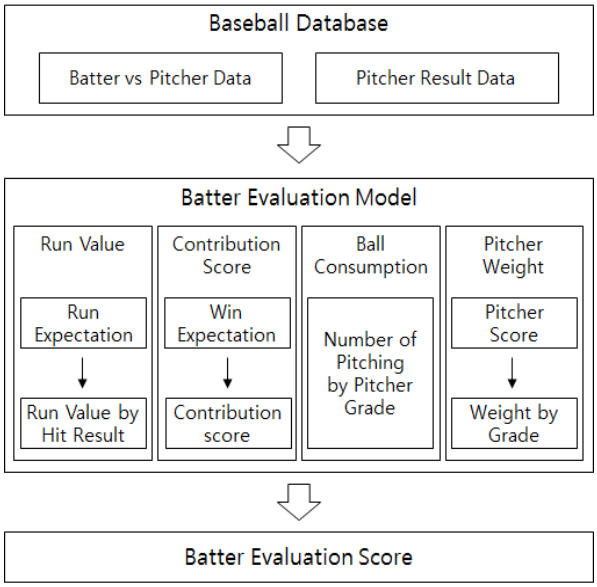

Fig. 1. A Batter Evaluation Model[2]

그림에서 야구 데이터베이스는 투수와 타자간의 상호 전적 데이터와 투수의 성적 데이터로 구성된다. 대결전적 데이터는 특정 투수와 특정 타자의 홈/원정, 이닝, 주자상태 등을 포함하고, 투수 성적 데이터는 특정 투수의 이닝, 자책점, 삼진 등을 포함한다.

(1) 기준배점

기준배점은 타자의 타격결과에 따라 타격결과 항목별로 기대득점값을 부여하는 점수이다. 표 1은 정진상이 제안한 한국프로야구에서의 기준배점이다.

각 타자의 기준배점 계산은 다음 식과 같이 한다.

$$\sum_{i=1}^{count_n}(RV_{c_i} \times PW_{g_i})$$

식에서 $count_n$은 평가대상타자 n의 타석횟수이고, c_i는 i번째 타석의 타격결과이다. RV_{c_i}는 타격결과에 대한 기준배점이다. g_i는 i번째 타석에서의 상대투수이며, PW_{g_i}는 상대투수의 투수가중치이다.

Table 1. Run Values[6]

Item	Run Value	Item	Run Value
Triple play	-82.439	Hit by pitch	11.7775
Uncaught third strike	7.9415	Walks	10.8535
Double play	-29.925	Intentional walk	6.258
Home run	50.449	Stolen bases	7.217
Triple	40.3305	Pickoff	-10.829
Double	29.3055	Sacrifice fly	-2.7615
Hit	16.3415	Strikeout	-10.1605
Error	11.606	Caught stealing	-6.797
Sacrifice bunt	-4.4905	Out	-9.352

(2) 기여도점수

기여도점수는 경기상태별로 타자의 타격결과가 팀의 승리에 기여한 정도를 점수화한 것이다. 예를 들어, 3회에서 5점을 앞선 경우에서의 홈런과 9회에서의 역전 홈런은 동일한 홈런이지만 팀의 승리에 기여한 정도는 차이가 크기 때문에 다른 점수로 평가할 필요가 있다.

문형우[13]는 2007~2012년의 한국프로야구경기 데이터를 사용하여 마르코프 체인(Markov chain)으로 승리기댓값을 계산하였다.

예를 들어, 표 2는 6회초 상황에서 승리기댓값을 계산한 결과이다. 표에서 주자의 상태는 1루/2루/3루 각각에 대하여 주자가 있는 경우에 1, 주자가 없는 경우에는 0으로 표시한 것이다. 예를 들어, 주자가 1루에만 있는 경우는 100으로 표시하였다. -5~-1은 5점~1점 차이로 뒤지고 있는 상황을 나타내고, +1~+5는 1점~5점 차이로 이기고 있는 상황을 나타낸다.

표에서 마지막 행은 3아웃으로 이닝이 끝난 후의 승리기댓값을 의미한다. 6회초를 동점으로 마치면 경기를 승리할 확률은 0.451이고, 1점 차이로 이기고 있는 상황에서 이닝을 종료하면 승리할 확률은 0.618이다. 6회 초 주자가 없이 0아웃 상태에서 2점을 앞서고 있다면, 이 상황에서의 승리할 확률은 0.802이다.

6회 초 주자 1루 0아웃 상태에서 동점이라면 타격 전 승리기댓값은 0.558이다. 이 때 홈런을 쳤다고 가정하면 주자가 없는 상태에서 2점을 앞서게 되어 타격 후 승리기댓값은 0.802이 된다. 이 타격에 의한 기여도점수는 타격 후 0.802에서 타격 전 0.558의 차이인 0.244가 된다.

Table 2. Win Expectation Values[13]

6th Inning Top												
Out	Runner	-5	-4	-3	-2	-1	tie	+1	+2	+3	+4	+5
0	000	0.031	0.057	0.102	0.191	0.302	0.484	0.660	0.802	0.891	0.935	0.967
0	100	0.053	0.093	0.156	0.260	0.386	0.558	0.714	0.835	0.909	0.948	0.975
0	010	0.058	0.101	0.174	0.282	0.426	0.597	0.746	0.855	0.919	0.955	0.981
0	001	0.065	0.113	0.199	0.312	0.476	0.646	0.786	0.880	0.931	0.964	0.988
0	110	0.094	0.154	0.242	0.363	0.499	0.652	0.781	0.875	0.932	0.963	0.984
0	101	0.103	0.169	0.270	0.396	0.551	0.702	0.821	0.900	0.944	0.972	0.991
0	011	0.109	0.183	0.287	0.429	0.584	0.729	0.838	0.908	0.950	0.977	0.992
0	111	0.166	0.256	0.369	0.501	0.637	0.761	0.857	0.920	0.958	0.979	0.992
1	000	0.020	0.039	0.072	0.153	0.254	0.441	0.628	0.783	0.880	0.928	0.962
1	100	0.031	0.057	0.102	0.193	0.305	0.486	0.662	0.803	0.892	0.936	0.967
1	010	0.034	0.062	0.114	0.208	0.331	0.513	0.684	0.817	0.898	0.940	0.971
1	001	0.040	0.074	0.142	0.242	0.395	0.577	0.737	0.850	0.914	0.952	0.981
1	110	0.051	0.089	0.151	0.253	0.374	0.546	0.704	0.829	0.906	0.945	0.973
1	101	0.059	0.103	0.182	0.291	0.442	0.613	0.760	0.864	0.923	0.958	0.983
1	011	0.067	0.120	0.200	0.322	0.468	0.632	0.770	0.867	0.927	0.961	0.983
1	111	0.093	0.151	0.235	0.347	0.479	0.634	0.768	0.868	0.927	0.959	0.982
2	000	0.014	0.029	0.055	0.131	0.223	0.412	0.606	0.769	0.873	0.922	0.958
2	100	0.018	0.036	0.067	0.148	0.245	0.432	0.621	0.778	0.878	0.926	0.961
2	010	0.020	0.038	0.075	0.156	0.264	0.452	0.637	0.789	0.883	0.930	0.964
2	001	0.020	0.040	0.078	0.161	0.272	0.460	0.644	0.793	0.885	0.931	0.965
2	110	0.026	0.049	0.091	0.178	0.285	0.468	0.648	0.795	0.887	0.932	0.965
2	101	0.027	0.052	0.096	0.184	0.295	0.478	0.656	0.800	0.890	0.934	0.966
2	011	0.030	0.059	0.104	0.202	0.312	0.491	0.664	0.803	0.893	0.937	0.967
2	111	0.036	0.070	0.122	0.221	0.337	0.514	0.681	0.814	0.898	0.940	0.970
Inning End		0.026	0.038	0.060	0.138	0.238	0.451	0.618	0.756	0.863	0.924	0.948

각 타자의 기여도점수의 계산은 다음 식과 같이 한다.

$$\sum_{i=1}^{count_n}(CS_{s_i} \times PW_{g_i})$$

식에서 s_i는 i번째 타석의 경기상태로 이닝상태 9종류, 이닝초/말 2종류, 아웃상태 3종류, 주자상태 8종류, 득점 차 -5점~5점까지 11종류, 이닝종료상태로 총 4,950종류가 있다. CS_{s_i}는 타격결과에 대한 기여도점수이다.

(3) 투구소모점수

투구소모점수는 타자가 상대 투수로 하여금 던지게 한 투구의 수를 타자평가 지표로 반영하기 위한 점수이다. 기존의 투구소모점수는 상대 투수의 평균 투구수만을 반영하였기 때문에 상대 투수의 기량을 반영하지 않았었다. 따라서 상대 투수의 기량이나 등급에 따라서 타자가 소모시킨 공의 가치를 다르게 계산할 필요가 있다.

표 3은 2011년 한국프로야구에서 경기당 투수가 던진 투구수를 등급별로 평균한 자료이다.

Table 3. Average Pitches per Game[2]

Pitcher Grade	1	2	3	4	5	6	7	8	9
Average Pitches	63.5	48.1	40.8	34.8	30.3	29.1	23.0	28.0	35.6

타자 개인별 투구소모점수 계산에 표 3의 투수 등급별 평균 투구수를 이용한다. 예를 들어, 타자가 1등급 투수를 상대로 5개의 공을 던지게 하였다면 해당 타자는 투구수에서 평균투구수를 나누어 5/63.5 = 0.079점을 받게 된다. 또한 타자가 7등급 투수를 상대로 5개의 공을 던지게 하였다면 5/23.0 = 0.217점을 받게 된다. 투구소모점수는 상대 투수의 등급별로 던지게 한 공 하나의 가치를 점수로 계산한 것이므로 상대 투수의 평균 투구수가 작을수록 하나의 공에 대해서 높은 점수를 부여한다.

각 타자의 투구소모점수의 계산은 다음 식과 같이 한다.

$$\sum_{i=1}^{count_n}(BC_i \times PW_{g_i})$$

식에서 BC_i는 i번째 타석에서의 투구소모 횟수이다.

(4) 투수가중치

투수가중치는 전체 투수를 기량에 따라 구분하여 9개의 등급으로 나누어 등급별로 부여한 가중치이다. 표 4는 2011년도 한국프로야구 투수들의 경기 기록 데이터를 이용하여 등급별 투수가중치를 계산한 결과이다.

Table 4. Pitcher Weights by Grade[2]

Grade	Percentage	Weight	Reverse-weight
1	0~4	0.817	0.023
2	4~11	0.541	0.052
3	11~23	0.376	0.069
4	23~40	0.230	0.085
5	40~60	0.129	0.129
6	60~77	0.085	0.230
7	77~89	0.069	0.376
8	89~96	0.052	0.541
9	96~100	0.023	0.817

타자를 평가할 때 투수가중치를 고려하는데, 기여도점수와 기준배점은 양의 값 또는 음의 값을 가질 수 있다. 양의 값에 대해서는 가중치를 적용하고, 음의 값에 대해서는 역가중치를 적용한다.

IV. Calculation of Weights of Evaluation Elements

타자평가모형의 각 평가요소를 이용하여 타자의 평가점수(Batter Evaluation Score)를 계산하는 식은 다음과 같다.

$$BES = \alpha_1 CS + \alpha_2 RV + \alpha_3 BC$$

각 평가요소의 가중치를 의미하는 α_1, α_2, α_3을 유전 알고리즘으로 구하였다. 가중치를 구하는 절차는 그림 2와 같다.

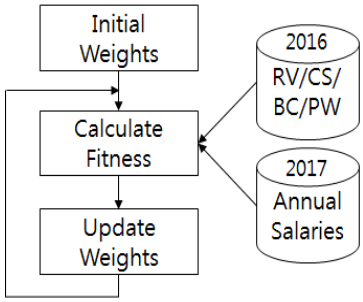

Fig. 2. Procedure for Calculating the Weights

본 연구에서 사용한 데이터는 2016년 한국프로야구의 경기 데이터이다. 프로야구 선수들의 연봉은 지난해의 성적으로 결정되므로 연봉 데이터는 2017년을 기준으로 하였다.

유전 알고리즘에 의해서 구해지는 가중치는 적합도에 의해서 그 값의 좋고 나쁨이 결정되는데, 반복적인 계산에 의해서 적합도가 좋아지는 방향으로 가중치 값이 구해지게 된다. 즉, 계산된 BES와 연봉 데이터의 차이를 최소화하도록 가중치가 반복 계산된다. 오차를 일반화시키기 위한 척도로는 절대값오차, 평균제곱 오차, 제곱근평균제곱오차 등이 있으나 본 연구에서는 제곱근평균제곱오차(RMSE)를 사용하였다. RMSE는 개별 관측값들이 중심에서 떨어져 있는 정도를 나타내는 척도로서 값이 작을수록 오차가 적기 때문에 적합도가 높다는 것을 의미한다. RMSE(Root Mean Square Error)의 식은 다음과 같다. 식에서 E_i는 모형에서 계산한 값이며, O_i는 실제의 값이다.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i - O_i)^2}$$

이와 같은 작업을 수행하는 유전 알고리즘을 R 코드로 작성하였으며, 그림 3에 표시하였다.

```
41  #2 with annual salaries
42  bat <- read_excel("batter2016.xlsx")
43  sal <- read_excel("yearsalary2017.xlsx")
44
45  dt <- data.table(sal)
46  sal_ <- dt[dt$TYPE == "연봉"]
47  df <- merge(bat, sal_[,c(1,4)], by="NAME")
48
49  dfo <- df[order(-df$MONEY), ]
50  dfo200 <- dfo[1:200, ]
51
52  cs <- scale(dfo200$CS, scale = T)
53  rv <- scale(dfo200$RV, scale = T)
54  bc <- scale(dfo200$BC, scale = T)
55  obs <- scale(dfo200$MONEY, scale = T)
56
57  f <- function(x) {
58    sim <- cs*x[1] + rv*x[2] + bc*x[3]
59    -sqrt(sum((sim-obs)^2)/nrow(dfo200))
60  }
61
62  lb <- 0; ub <- 1.0
63  GA <- ga(type="real-valued", fitness=f,
64           lower=c(lb,lb,lb), upper=c(ub,ub,ub),
65           suggestions=c(0.333,0.333,0.333))
66  summary(GA)
67  plot(GA)
```

Fig. 3. R Code for Genetic Algorithm

그림 3의 42, 43번 줄은 엑셀에 저장된 한국프로야구데이터를 읽는 코드이다. 49, 50번 줄은 연봉 순위로 200개의 데이터를 추출하는 과정이다. 연봉이 낮은 선수는 신입선수이거나 부산 등으로 전년도 데이터가 없는 경우가 많아 연봉상위 200명으로 제한하였다. 57~60번 줄은 적합도를 계산하는 함수이다. 58번 줄에서 가중치를 적용하여 BES를 구하고, 59번 줄에서 RMSE를 계산한다. 63~65번 줄에서 유전 알고리즘을 실행한다. 유전 알고리즘에서 가중치의 초기값은 1/3로 설정하였다. 이 알고리즘은 기본적으로 개체군은 50, 반복횟수는 100으로 하여 실행된다.

그림 4는 유전 알고리즘이 최적해를 찾아가는 과정을 나타낸 것이다. 세로축은 적합도 값이며, 가로축은 세대를 의미한다. 세대가 진행되면서 적합도 값이 증가하다가 어느 순간에 최고점에 도달하는 것을 볼 수 있다.

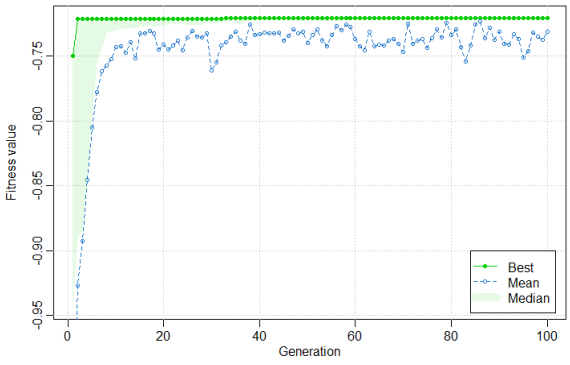

Fig. 4. Fitness Values by Iteration

그림 5는 유전 알고리즘의 실행 결과이다. 각 평가요소의 가중치를 의미하는 α_1, α_2, α_3는 x1, x2, x3로 각각 표시되어 있다. 구해진 값은 x1은 0.04, x2는 0.40, x3은 0.25이다. 즉, 타자평가점수는 다음과 같은 식에 의해서 구해질 수 있다.

$$BES = 0.04CS + 0.4RV + 0.25BC$$

```
-- Genetic Algorithm ----------------------

GA settings:
Type                  =  real-valued
Population size       =  50
Number of generations =  100
Elitism               =  2
Crossover probability =  0.8
Mutation probability  =  0.1
Search domain =
      x1 x2 x3
lower  0  0  0
upper  1  1  1
Suggestions =
     x1    x2    x3
1 0.333 0.333 0.333

GA results:
Iterations             = 100
Fitness function value = -0.7204435
Solution =
             x1        x2        x3
[1,] 0.0423503 0.4010507 0.259834
```

Fig. 5. Result of Genetic Algorithm

그림 5의 결과는 기여도점수(CS)의 가중치인 α_1이 매우 낮다. 이는 본 논문에서 제시하는 기여도점수가 실제 연봉에서는 거의 반영되지 않는다는 의미이다.

컴투스 프로야구 포인트[14]는 한국야구위원회(KBO)와 MBC SPORTS+가 시행중인 프로야구 선수 포인트 제도이다. 컴투스 프로야구 포인트는 경기 기록으로 프로야구 선수들의 순위를 결정하는 제도로 매 경기의 데이터를 기준으로 각 선수들에게 매일 점수가 누적되어 시즌 후에 최종적인 순위가 결정된다. 이 포인트는 2017는 이전까지는 카스포인트로 불리워졌다.

2016년 한국프로야구 타자들의 카스포인트를 이용하여 BES와의 적합도를 유전 알고리즘으로 구하였다. 그림 6은 결과이다.

```
-- Genetic Algorithm ----------------------

GA settings:
Type                  =  real-valued
Population size       =  50
Number of generations =  100
Elitism               =  2
Crossover probability =  0.8
Mutation probability  =  0.1
Search domain =
      x1 x2 x3
lower  0  0  0
upper  1  1  1
Suggestions =
     x1    x2    x3
1 0.333 0.333 0.333

GA results:
Iterations             = 100
Fitness function value = -0.4284113
Solution =
             x1        x2        x3
[1,] 0.02243318 0.4963026 0.4137324
```

Fig. 6. Result of Genetic Algorithm (CassPoint)

카스포인트를 기준으로 한 가중치는 α_1이 0.02, α_2가 0.50, α_3은 0.41로 나타났다. 기본배점이 많이 반영되고 있음을 볼 수 있다. 카스포인트 역시 기여도점수는 고려하지 않음을 보여주고 있다.

V. Conclusions

본 연구에서는 상대 투수를 고려한 타자평가모형에서의 평가요소의 가중치를 유전 알고리즘에 의하여 구하여 보았다. 연구에서 사용한 타자평가모형은 동일한 타격이라도 경기의 결과에 영향을 주는 정도인 승리기여도를 반영하고 투수의 등급별로 가중치를 반영하여 동일한 타격결과라도 상대 투수의 기량에 따라 평가점수가 달라지는 모형이다.

한국프로야구의 2016년 기록 데이터를 이용하여 실험한 결과, 타자의 연봉은 타격결과에 따른 기대득점값을 많이 반영하고 있는 것으로 나타났다. 반면에 경기상태별로 타격결과가 팀의 승리에 기여한 정도를 나타내는 기여도점수는 연봉이나 포인트 제도에 반영되지 않고 있는 것으로 나타났다. 야구에서는 개인의 성적도 중요하지만 팀의 성적 역시 매우 중요하다. 팀의 성적에 기여한 선수에게 많은 연봉이 주어지도록 하여야 할 것이며, 이를 위해서는 기여도점수가 연봉이나 포인트 제도에 도입되어야 할 것이다.

기여도점수를 연봉에 얼마만큼 반영하는 것이 좋은지에 대해서는 추가의 연구가 필요하다. 적절한 가중치가 정해진다면 상대 투수의 성적을 반영하여 타순을 구성하는데 도움을 줄 것이며, 타자의 객관적인 가치를 평가하는 기준을 제시할 수 있어 연봉 협상에도 활용할 수 있을 것이다.

REFERENCES

[1] E. Wassermann, D. R. Czech, M. J. Wilson and A B. Joyner, “An Examination of the Moneyball Theory: A Baseball Statistical Analysis,” The Sports Journal, January 2005.

[2] Yerin Jung, “Korean Professional Baseball Batter Estimation Model Reflecting Big Data Analysis and Pitcher Ability,” Master dissertation, Changwon National University, 2016. (in Korean)

[3] Jang Taek Lee, “Measurements for hitting ability in the Korean pro-baseball,” Journal of The Korean Data Analysis Society, Vol. 25, No. 2, pp. 349-356, March 2014. (in Korean)

[4] Young Suk Cho and Young Ju Cho, “A study on OPS and runs from Korean baseball league,” Journal of The Korean Data Analysis Society, Vol. 7, No. 1, pp. 221-231, February 2005. (in Korean)

[5] Hyuk Joo Kim, “Effects of on base and slugging ability on run productivity in Korean professional baseball,” Journal of The Korean Data & Information Analysis Society, Vol. 23, No. 6, pp. 1165-1174, December 2012. (in Korean)

[6] Jin-Sang Jung, “Efficient estimation model of hitter using Big Data analysis in Korea Baseball League,” Master dissertation, Changwon National University, 2014. (in Korean)

[7] G. B. Johnson, "Evaluation and ranking of minor-league hitters using a statistical model," Doctoral dissertation, Kansas State University, 2006.
[8] P. A. Yates, "Estimating Situational Effects on OPS," Journal of Quantitative Analysis in Sports, Vol. 4, No. 2, February 2008.
[9] B. B. McShane, A. Braunstein, J. Piette and S. T. Jensen, "A Bayesian Variable Selection Approach to Major League Baseball Hitting Metrics," Journal of Quantitative Analysis in Sports, Vol. 7, No. 4, November 2009.
[10] S. L. Rubin, "Market Efficiency of Major League Baseball Player Salaries: A Look at the Moneyball Hypothesis Ten Years Later," Doctoral dissertation, University of Delaware, 2013.
[11] J. H. Holland, "*Adaptation in natural and artificial system*," University of Michigan Press, 1975.
[12] D. E. Goldberg, "*Genetic algorithm in search, optimization & Machine Learning*," Addison Wesley, 1989.
[13] Hyung Woo Moon, Yong Tae Woo and Yang Woo Shin, "Run expectancy and win expectancy in the Korea Baseball Organization(KBO) League," The Korean Journal of Applied Statistics, Vol. 29, No. 2, pp. 321-330, February 2016. (in Korean)
[14] Com2uS Pro-Baseball Point, *http://cpbpoint.mbcplus.com*

Authors

Su-Hyun Lee received the B.S. in Computer Science from Kwangwoon University, Korea in 1987. He received the M.S. and Ph.D. degrees in Computer Science from Korea Advanced Institute of Science and Technology(KAIST), Korea, in 1989, 1994, respectively. Dr. Lee is a Professor in the Department of Computer Engineering, Changwon National University since 1996. He is interested in computer algorithm, programming languages, compiler, and blockchain.

Yerin Jung received the B.S. in Computer Engineering from Changwon National University, Korea in 2015. She received the M.S. degrees in Computer Engineering from Changwon National University, Korea, in 2017. Ms. Jung is currently a researcher in Hibrain.Net. She is interested in big data analysis, data mining.

Hyung-Woo Moon received the B.S. in Computer Engineering from Kosin University, Korea in 2007. He received the M.S. and Ph.D. degrees in Computer Engineering from Changwon National University, Korea, in 2009, 2014, respectively. Dr. Moon is a Researcher in the Institute of Industrial Technology Research Center, Changwon National University since 2014. He is interested in sports data analytics and sports big data architecture.

Yong-Tae Woo received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Kyungpook National University, Korea, in 1982, 1984 and 1995, respectively. Dr. Woo is a Professor in the Department of Computer Engineering, Changwon National University since 1987. He is also CEO of Hibrain.net Co. He is interested in data modeling, internet business, and big data analysis.