2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호

Heuristic model를 이용한 프로야구 승패예측

김동식*, 홍석미, 정태충
경희대학교 전자계산공학과
e-mail : dskim@iislab.kyunghee.ac.kr

Predication of win/lose of Professional baseball using Heuristic model

Dong-Sik Kim*, Seok-mi Hong, Tae-Chung Jung
Dept of Computer Engineering, Kyung-Hee University

요약

프로야구경기의 승패 예측의 문제는 그리 쉬운 일이 아니다. 왜냐하면 경기에 영향을 미치는 요소가 무한하기 때문이다. 예를 들어, 경기당일의 선수들의 컨디션이나 사기, 경기당일의 날씨, 구장요건, 상대팀에 대한 심리적 요인등 사전에 경기영향을 미치는 요소가 무한하다. 본 연구실에서는 과거 경기기록 자료를 기반으로 유용한 규칙을 찾아내여 분류트리를 만들어 학습하는 ID3 알고리즘을 프로야구 승패예측 시스템 구성에 사용하여 보았으나, 이산적인 자료의 처리로 인해 연속적인 경기자료를 고려하지 못하는 문제로 예측율이 더이상 향상되지 않았다. 따라서, 본 논문에서는 휴리스틱 방법을 이용한 경기전 예측과 경기중 예측을 이닝별 득점으로 세분화하여, 실제 경기상황을 고려한 일반적인 예측 모형을 만들어 예측율을 향상시키고자 한다. 향후에는 더욱 세분화시켜 Case-based에 의한 예측을 하고자 한다.

1. 서론

근래 프로야구는 국내보다는 국외 선수들의 눈부신 활약으로 인해 더욱 흥미가 더해지고 있다. 박찬호의 선발경기에는 야구팬들이 숨을 죽이고 그가 승리하기를 바라고 있다. 그의 승리정도가 어느정도인지를 알 수 있다면, 애태우며 밤을 지세우지는 않을 것이다. 하지만, 프로야구 경기에 영향을 미치는 요소가 무한하기 때문에 어느 정도로 승리할 것인지 예측하기는 쉽지 않다. 예를들어, 경기당일의 선수들의 컨디션이나 사기, 경기당일의 날씨, 구장요건, 상대팀에 대한 심리적요인 등 사전에 경기에 영향을 미치는 요소들이 무수하다.

본 연구실에서는 과거 경기기록 자료를 기반으로 유용한 규칙을 찾아내어 분류트리를 만들어 학습하는 ID3 알고리즘[3]을 프로야구 승패예측 시스템구성에 사용하여 보았다. ID3 알고리즘은 여러가지 형태 중 이산자료에 대한 처리가 용이하며, 생성된 결과를 그대로 규칙으로 표현될 수 있다. 즉, 기존의 자료들을 이용하여 학습트리를 형성하며, 트리를 기반으로 앞으로 경기에 대한 두 팀간의 승패예측을 가능하게 한다. 하지만, 이런 장점에도 불구하고, 다음과 같은 문제점들로 인해 다른 방법의 예측이 필요하게 되었다.

- 현재 경기자료는 연속적인 자료이다
- 데이터의 중복사용으로 인해 계산속도의 저하
- 다양한 경기상황을 반영하지 못한다
- 예측률이 향상되지 않았다
- 구현상의 복잡성 증가

본 논문에서는 과거 경기자료를 통해 경기전 예측으로 활용하고, 경기중 예측을 통해 현재 상황을 반영할 수 있는 휴리스틱 기법에 의한 예측 모형을 만들어 데이터의 중복을 제거하고, 구현상의 복잡성을 현저히 줄이고, 보다 일반적인 예측 모델을 제시하고자 한다.

2. 휴리스틱에 의한 예측 모형

휴리스틱에 의한 예측 기법은 많은 실험을 통해 최적의 해를 찾는 것이다. 다음은 실험에 의한 예측모델을 나타낸 것이다.

2.1 홈팀이 x이닝에서 승리할 확률

$$R(x) = \frac{C_h(x) - C_a(x)}{\max(C_h(x),\ C_a(x))}$$

① x 이닝에서 홈팀의 승리정도 :
$R(x)$ (단, 확률범위: 0 ~ 1)

② x 이닝에서 홈팀의 예상득점 : $C_h(x)$

③ x 이닝에서 원정팀의 예상득점 : $C_a(x)$

위의 경우 홈팀이 승리할 확률을 표현한 것으로 x이닝에서 홈팀이 그 순간에 이길수 있는지를 판단할 수 있는 $R(x)$값의 형태로 표현된다.

2.2 팀의 득점 예측

2.2.1 총점

$$C1(x) = Rp(x) + G(y) + A(x)$$

현재의 득점과 경기전, 경기중의 예측점수에 공격팀의 경우 현재 경기에 대한 결과로 표현할 수 있다.

① 현재 play에서 획득한 실제 점수 : $Rp(x)$

② y(다음공격) 이닝 이후의 남은 이닝 동안의 예상득점 : $G(y)$

③ 공격팀의 경우 현 이닝 중 남은 선수의 play에 대한 예상득점 : $A(x)$

$$G(y) = G_p(y) \times \min\left\{\frac{y}{4}, 1\right\} + G_f(y) \times \min\left\{\frac{4-y}{4}, 0\right\}$$

2.2.2 $G(y)$

① 경기 전 예측점수 : $G_f(y)$

② 경기 중 예측점수 : $G_p(y)$

위의 $G(y)$의 경우, 경기중의 예측을 위한 방법으로서 경기전 예측점수는 $G_f(y)$는 현재의 경기의 전반 기록이 없으므로, 타자가 한번씩 타석에 들어온 것을 가정하여, y이닝까지의 결과를 반영하고, 그 이후에 경기한 결과를 더 많이 반영하여 다음공격 이닝 이후의 남은 이닝 동안의 예상득점을 표현하고 있다.

2.2.2.1 $G_p(y)$

$$G_p(y) = \left(Opp_{ar} \times \frac{Opp_{era}}{OppAll_{era}} \times 0.5 \right) + \left(Opp_{ar} \times 0.5 \right)$$

① 상대팀에 대한 평균득점 : Opp_{ar}

② 현재팀에 대한 상대팀 투수의 평균 방어율 : $OppAll_{era}$

③ 상대팀의 현재 투수 방어율 : Opp_{era}

$$G_f(y) = PastI_{ar} \times \frac{CI_{hit}}{PastI_{hit}}$$

2.2.2 $G_f(y)$

① 과거 해당 이닝 이후의 평균득점 : $PastI_{ar}$

② 과거 현재 이닝 전까지 안타수 : $PastI_{hit}$

③ 현재 이닝까지의 평균 안타수 : CI_{hit}

2.3 $A(x, r, oc)$를 구하는 공식

x이닝에서 주자상황이 r이고, 남은 타자수가 최소 oc(3-outcount)일때, 예상득점을 표현한다.

주자상황(r)	아웃	기대점수
없음(0)	0	0.516
	1	0.277
	2	0.099
한명	0	1.149
	1	0.751
	2	0.309
득점권	0	1.462
	1	0.982
	2	0.432
만루	0	2.296
	1	1.500
	2	0.741

자료 : 1997년 내셔널리그 대상의 통계자료(KBIS)

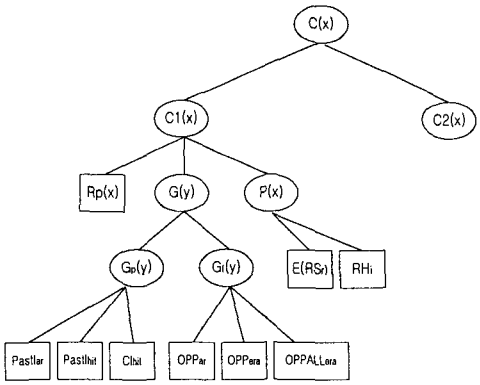
$$P(x) = \sum_{i=0}^{oc} \left\{ RH_i \times \sum_{r=0}^{3} E(RS_r) \right\} = RS_0 \times RH_i + RS_1 \times RH_i + RS_2 \times RH_i + RS_3 \times RH_i$$

① $E(RS_r)$: 주자상황에 따른 아웃카운트별 평균득점

② RH_{oc} : 남은 타자수(oc : 3-outcount)

주자상황에 따른 아웃카운트별 평균득점의 기대치는 $E(RS_r)$은 각 아웃카운트에서 평균득점을 어느정도 했는가를 계산하여 반영한다.

3. 각 함수의 계층도 및 관련 data



[그림 1 Hueristic 모형의 계층도]

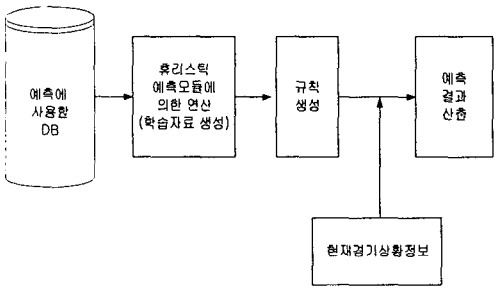
◯ 표시된 부분은 □로부터 계산된 결과를 합하여 얻을 수 있는 정보로 전체 예상득점을 표현하고 있다.

4. 학습 형태 및 예측 방법

4.1 예측방법

프로야구는 기록경기라 할만큼 많은 자료들이 산출되며, 이 기록들은 다음 경기의 예측에 사용될 수 있다. 예측 모듈은 이러한 다양한 기록들을 이용하여 실제 경기 이전이나 경기 수행중에 두 팀간의 승패 및 여러가지 정보를 예측할 수 있는 시스템이다.

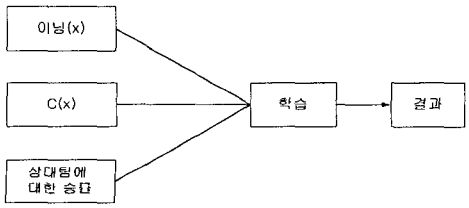
통계 패키지를 이용한 예측, Nueral에 의한 방법등이 있는데, 본 논문에서는 휴리스틱에 의해 일반적인 모형을 만들어, 하나의 식에 많은 정보를 함축적으로 표현하고 있고, 실제 경기 중이나 경기 전 상황을 경기정보를 받아 승패 여부 및 다양한 정보를 예측할 수 있는 Nueral에 의한 예측방법을 사용하고 있다[4][5].



[그림 2] Nueral에 의한 예측방법

예측을 위해서는 DATA ENGINE 패키지를 이용하였으며, 모델링을 하는데는 신경회로망을 사용하였고, 학습방법으로는 역전파 알고리즘을 사용하였습니다. 그 이유는 본 논문에서 제시한 경기 승패 예측은 과거 기록 데이터에 대한 결과 데이터가 있으므로 그 결과가 발생하는 규칙을 찾는 것인데, 이러한 규칙은 모든 사례에 대해 규칙을 만들기 어렵고 승패 예측에 이용되는 데이터가 수치이므로 사람이 이해 할 수 있는 명확한 규칙을 유도하기도 어렵다. 그러므로 근사 추정능력과 일반화 능력에 기반을 두고 예측(Forecasting)과 추정(Estimation) 등의 분야에서도 많은 연구가 진행되고 있는 신경회로망을 이용하는 것이 바람직하고, 학습방법으로는 지도학습(Supervised Learning)에 해당되는 역전파 알고리즘을 사용하였다[4][5].

4.2 학습 형태



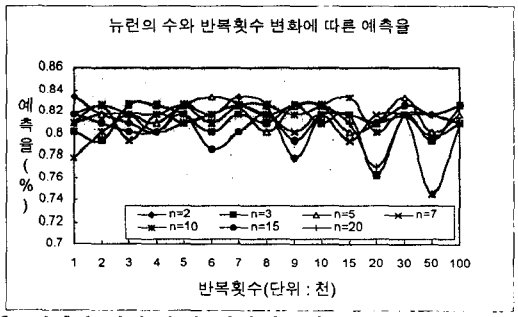
[그림 3] 예측을 위한 자료의 학습 형태

[그림 3]은 이닝, C(x), 상대팀에 대한 승률을 입력하여 학습데이터로 이용하고, 출력결과로 나온 예측값과 실제 경기결과를 비교하여 예측을 하게 된다.

5. 실험결과

실험에 사용된 자료는 1998년 KBO의 경기자료를 바탕으로 1022개의 초기자료를 생성하였다. 자료중 패턴이 같으면, 제대로 된 학습 결과를 얻을 수가 없으므로, 동일한 패턴이 승패에 나타나지 않도록 중복 자료를 제거하여 936개의 자료를 얻었다. 그러나, 936개의 자료가 이닝별로 일정하지 않아, 많은 자료를 가진 이닝은 학습이 많이 되어 한쪽으로 편중된 예측결과를 가질 수 있다. 따라서, 각 이닝별 수를 일정하게 맞추기 위해 각 이닝별 70개씩 모두 630개를 추출하였다.

Test 자료는 이 데이터가 제대로 학습됐는지를 알아보기 위해 전체 자료의 20%에 해당되는 126개를 각 이닝별로 14개씩 추출하였다. 실험은 Hidden Layer 갯수와 뉴런의 수, 전달함수에 대해 반복횟수를 늘려가면서 측정하였다.



[그림4] 뉴런의 수와 반복횟수에 따른 예측 변화률

뉴런의 수에 따라 반복횟수를 증가시키더라도 예측율이 일정치 않은 것은 초기 뉴런의 Weight값을 랜덤하게 생성하므로, 반드시 높은 예측값을 구할 수 있는 것은 아니다.

뉴런의 수(N)=2일 때, 반복횟수가 1000, 7000에서 최고치 83.3%을 보였고, N=5, N=10일때도 최고치를 보였으며 Hidden Layer의 비교를 위해 N=7도 포함하였고, 반복회수의 평균 예측율은 5000, 7000에서 가장 높은 예측율을 보였으나, 실험을 위해 그 이후의 10000, 20000, 50000회도 포함하여 Hidden Layer에 따른 비교분석을 실시하였다.

반복횟수에 따른 예측률					
뉴런수	5000(회)	7000(회)	10000(회)	20000(회)	50000(회)
n=2 ①	0.8175	0.8175	0.8175	0.8254	0.8016
n=5 ②	0.7937	0.8254	0.8175	0.7937	0.8016
n=7 ③	***0.8333***	0.8254	0.8254	0.8175	0.8095
n=10 ④	0.8016	0.8095	0.7937	0.7937	0.7937

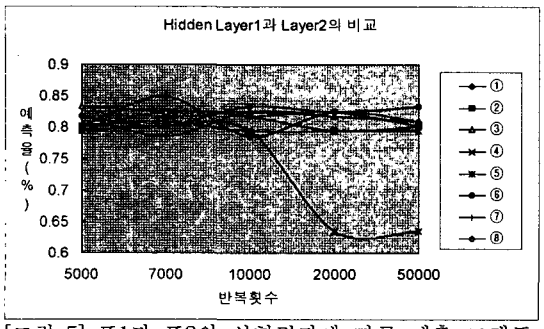
[표1] Hidden Layer 1의 예측율

[표1]의 Hidden Layer 1에서는 뉴런의 수가 7이고, 반복횟수가 5000일 때, 높은 예측율을 보이고 있다.

반복횟수에 따른 예측률					
뉴런수	5000(회)	7000(회)	10000(회)	20000(회)	50000(회)
n=2 ⑤	0.7937	0.8016	0.7937	0.6349	0.6349
n=5 ⑥	0.8175	0.8016	0.8333	0.8254	0.8333
n=7 ⑦	0.8095	0.8254	0.8016	0.8333	0.7063
n=10 ⑧	0.8016	***0.8492***	0.7857	0.8254	0.8095

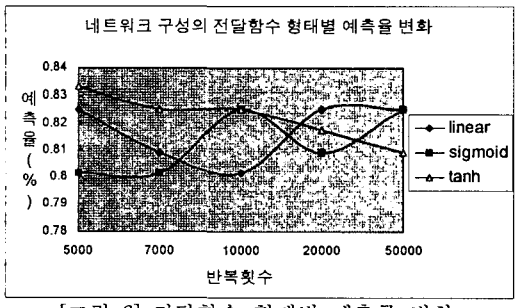
[표2] Hidden Layer 2의 예측율

[표2]에서는 뉴런의 수가 10이고, 반복횟수가 7000회 일때, 가장 높은 84.9%을 보이고 있다.



[그림 5] 표1과 표2의 실험결과에 따른 예측 그래프

[그림 5]에서 알수 있듯이, Hidden Layer 2 중 N=2일 때을 제외하고는 거의 유사한 예측율을 보였다. 따라서, Hidden Layer 2 에서 가장 높은 예측을 보였으나, 복잡한 구조와 계산속도의 저하로 인한 문제를 고려하면, Hidden layer 1으로 예측을 해도 무리가 없을 것이다. 마지막으로 예측에 변화을 줄 수 있는 전달함수에 의한 변화을 Hidden Layer 1 에서 표현해 보았다.



[그림 6] 전달함수 형태별 예측률 변화

[그림 6]은 전달함수 linear, sigmoid, tanh에 의한 예측율을 비교하였다. 전달함수의 형태중 가장 높은 예측값을 나타낸 것은 tanh로 나타나고 있는데, 반복횟수가 증가할수록 예측율이 떨어지는 경향을 보이는 것으로 나타나고 있다.

지금까지 실험을 통해 예측에 변화를 줄 수 있는 Hidden Layer의 수, 뉴런의 수, 전달 함수의 형태에 의해 실험을 해 본 결과 Hidden Layer는 Layer 2, 전달함수는 tanh, 뉴런의 수는 10에서 84.9%의 예측률 보였다.

6. 결론 및 향후 연구과제

프로야구 승패를 예측하는데 있어서 중요한 것은 적절한 자료의 추출과 변형일 것이다. 본 논문에서는 경기상황에서 중요한 요소를 가능한 모두 활용하고, 이를 하나의 통합된 공식으로 일반화하여 경기 전 예측과 경기 중 예측을 하고, 예측모듈에 있어서 간단한 형태의 자료를 제공함은 물론, 실 시간적인 형태의 방송에서 활용할 수 있을 것이다.

과거 ID3에 의한 예측[3]에 있어서 최고 81%을 보였던 것에 비해 보다 일반적인 모델을 형성하였으며, 과거의 전체자료를 이용한 것에 비해 간단한 수식의 형태로 표현함으로써 계산속도의 향상은 물론 예측율의 성능개선을 가져 올 수 있었다.

본 논문의 실험결과에서도 보왔듯이, 실제 경기 상황과 유사한 상황을 전제로 공식을 이끌어 낸다면 보다 나은 예측을 할 수 있다는 것이 입증 되었다. 향후에는 더욱 세밀한 볼카운트별의 상황, 수비별 상황을 고려한 case-based에 의한 예측을 하고자 한다[8].

7. 참고문헌

[1] 홍석미, 프로야구 승패예측을 위한 게임 시뮬레이터 개발에 관한 연구, 1997
[2] 허준회, 프로야구 경기 시뮬레이터에서 데이터 마이닝을 이용한 투수선정 및 투수교체시기 선택에 관한연구, 1999
[3] 서재순, 귀납적 추론을 이용한 프로야구 승패예측 시스템 개발에 관한 연구, 1994
[4] MIT GmbH, Aachen Germany, "DataEngine Overview and User Manual", 1997
[5] MIT GmbH, Aachen Germany, "DataEngine Tutorials and Theory",1997
[6] Adiraans, Zantinge 저, 용환승 역, 데이터 마이닝, 그린,1998
[7] 하일성야구정보연구소, 하일성 없이도 프로야구 10배 재미있게 즐기는 책, 하늘 출판사, 1995
[8] 김다윗, 신경망 분리모형과 사례기반추론을 이용한 기업 신용 평가, 1997