혼합형 기계 학습 모델을 이용한 프로야구 승패 예측 시스템

Win/Lose Prediction System : Predicting Baseball Game Results using a Hybrid Machine Learning Model

저자
(Authors)

홍석미, 정경숙, 정태충
SeokMi Hong, KyungSook Jung, TaeChoong Chung

출처
(Source)

정보과학회논문지 : 컴퓨팅의 실제 및 레터 9(6), 2003.12, 693-698(6 pages)
Journal of KIISE : Computing Practices and Letters 9(6), 2003.12, 693-698(6 pages)

발행처
(Publisher)

한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE00618285

APA Style

홍석미, 정경숙, 정태충 (2003). 혼합형 기계 학습 모델을 이용한 프로야구 승패 예측 시스템. 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 9(6), 693-698

혼합형 기계 학습 모델을 이용한 프로야구 승패 예측 시스템

(Win/Lose Prediction System : Predicting Baseball Game Results using a Hybrid Machine Learning Model)

홍 석 미 † 정 경 숙 †† 정 태 충 †††

(SeokMi Hong) (KyungSook Jung) (TaeChoong Chung)

요 약 야구는 매 경기마다 다양한 기록을 생성하며, 이러한 기록을 기반으로 다음 경기에 대한 승패 예측이 이루어진다. 프로야구 승패 예측에 대한 연구는 많은 사람들에 의해 행해져 왔으나 아직 이렇다할 결과를 얻지 못하고 있는 상태이다. 이처럼 승패 예측이 어려운 이유는 많은 경기 기록들 중 승패 예측에 영향을 주는 요소의 선별이 어렵고, 예측에 사용된 자료들 간의 중복 요인으로 인해 학습 모델의 복잡도만 증가시킬 뿐 좋은 성능을 보이지 못하고 있다. 이에 본 논문에서는 전문가들의 의견을 바탕으로 학습 요소들을 선택하고, 선택된 자료들을 이용하여 휴리스틱 함수를 구성하였다. 요소들 간의 조합을 통해 예측에 영향을 줄 수 있는 새로운 값을 산출함과 동시에 학습 알고리즘에 사용될 입력 값의 차원을 줄일 수 있는 혼합형 모델을 제안하였다. 그 결과, 학습 알고리즘으로 사용된 역전파 알고리즘의 복잡도를 감소시키고, 프로야구 경기 승패 예측에 있어서도 정확성이 향상되었다.

키워드 : 기계학습, 휴리스틱, 혼합형모델, 프로야구

Abstract Every baseball game generates various records and on the basis of those records, win/lose prediction about the next game is carried out. Researches on win/lose predictions of professional baseball games have been carried out, but there are not so good results yet. Win/lose prediction is very difficult because the choice of features on win/lose predictions among many records is difficult and because the complexity of a learning model is increased due to overlapping factors among the data used in prediction. In this paper, learning features were chosen by opinions of baseball experts and a heuristic function was formed using the chosen features. We propose a hybrid model by creating a new value which can affect predictions by combining multiple features, and thus reducing a dimension of input value which will be used for backpropagation learning algorithm. As the experimental results show, the complexity of backpropagation was reduced and the accuracy of win/lose predictions on professional baseball games was improved.

Key words : machine learning, heuristic, hybrid model, pro-baseball

1. 서 론

기계학습은 인공지능의 한 분야로 인간의 지능으로 할 수 있는 사고, 학습, 자기 계발 등을 컴퓨터가 모방할 수 있도록 하는 컴퓨터 공학의 한 분야로서, 그 자체로 존재하는 것이 아니라 컴퓨터 과학의 다른 분야와 많은 관련을 맺고 있다. 특히 정보 기술의 여러 분야에서 인공 지능적 요소를 도입하고자하는 연구가 진행되어지고 있다[1]. 이러한 추세에 따라 여러 가지 스포츠 경기에 대한 예측이 많은 인기를 얻고 있는데 그 중 프로야구는 많은 사람들에게 인기있는 스포츠로써 승패 예측 및 다양한 형태의 경기 예측에 대한 연구가 많이 이루어지고 있다. 그러나 이렇다할 좋은 결과는 아직 얻지 못하고 있는 상태이다.

정확한 예측을 위해서는 보다 학습에 효과적인 자료들이 필요하나, 경기 시 산출되어지는 많은 기록들 중 승패 예측에 영향을 미치는 자료를 선별하는데 있어서 어려움이 있다. 또한 선별된 자료라 할지라도 모든 선수

† 학생회원 : 경희대학교 전자계산공학과
smhong@iislab.kyunghee.ac.kr
†† 비 회 원 : 경희대학교 전자계산공학과
jungks@iislab.kyunghee.ac.kr
††† 종신회원 : 경희대학교 전자계산공학과 교수
tcchung@khu.ac.kr
논문접수 : 2003년 6월 2일
심사완료 : 2003년 8월 23일

들에 대하여 해당 기록을 얻기가 어렵고, 선택된 자료들 간에는 서로 중복된 요인을 갖고 있기도 하다.

지금까지 본 연구실에서는 많은 프로야구 승패 예측 모델을 만들어왔다. 첫 번째로 과거 경기 기록을 기반으로 유용한 규칙을 찾아 분류 트리를 만드는 이산자료 처리에 적합한 ID3 알고리즘을 이용[2]하였으나, 실제 야구 데이타는 연속적인 자료가 대부분이어서 많은 자료의 손실을 가져왔다. 두 번째로 통계적 기법을 이용한 모델[3]에서는 반복적인 실험을 하기 때문에 많은 시간이 걸렸으며, 신경회로망에 의한 예측[4]의 경우 입력 값이 많아짐으로 해서 복잡도가 증가하는 문제점들이 있었다.

이에 본 논문에서는 전문가들의 의견이나 기타 여러 가지 자료들을 통하여 학습 모델 생성에 사용될 자료들을 선택하고, 휴리스틱 함수[5]를 이용하여 선택된 여러 요소(feature)들을 혼합한 새로운 예측용 자료를 생성한다. 그리고 새로 생성된 자료와 원시 자료 중 몇 개를 신경회로망 알고리즘의 입력으로 이용하는 예측모델을 제시하고자 한다. 그림 1은 본 논문에서 제안하고 있는 예측 시스템 모델의 구조를 나타낸다.

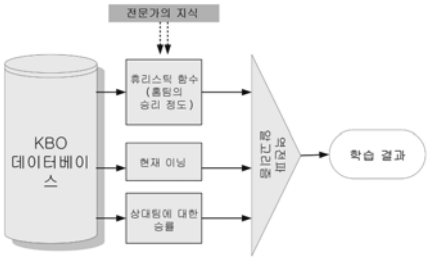

그림 1 제안된 모델

본 논문의 구성은 다음과 같다. 2장에서는 기존의 프로야구 승패 예측을 위해 사용되었던 알고리즘들을 중심으로 예측 시스템의 접근 방법을 살펴보고, 3장에서는 본 논문에서 제안하고 있는 모델에 대하여 보인다. 그리고 4장에서는 구현 및 실험에 대하여 설명하고 5장에서는 결론을 내리도록 하겠다.

2. 기존의 접근 방법

2.1 ID3 알고리즘

ID3는 헌트(Hunt)의 CLS(Concept Learning System) 학습 알고리즘을 확장시킨 방법이다. CLS는 두 개의 클래스를 묘사하는 객체들로부터 클래스를 분류하는 규칙을 생성시키는 상대적으로 간단한 알고리즘이다. ID3는 여러 가지 자료 형태 중 이산자료에 대한 처리에 용이하다. 과거 경기기록(팀타율, 팀승률, 팀방어율, 팀출루율, 팀수비율, 구장별승률, 구장별방어율 등)을 이용하여 학습 트리를 구성하며, 학습된 트리를 기반으로 승패 예측을 가능하게 한다[6]. 그런데 야구 경기기록들은 대부분 연속적이므로 ID3에 그대로 적용할 수가 없다. 연속자료들을 그대로 사용할 경우, 학습 트리의 가지수가 엄청나게 늘어나고, 학습 트리의 폭이 넓어지며 필요없는 노드를 많이 만들게 된다. 학습에 사용될 특징의 수가 늘어나게 되면 이러한 문제는 더욱 심화될 것이다. 그래서 연속자료를 이산자료로 바꾸어주는 이산화과정을 거쳐 자료를 변형시킨 후 ID3을 적용하였다. 그로 인해 기록 자체가 지니고 있는 본래의 특성을 시스템에 완전히 반영할 수 없었다. 또한 학습 예제(traing set)들 중 속성값은 동일한데 결과가 틀린 경우(데이타 충돌)로 인하여 예측률의 감소 문제가 발생하였다.

2.2 통계적 방법

통계적 기법을 이용한 경우는 예측이 가능한 여러 가지 자료들의 확률 값을 이용하여 실제 경기와 같이 게임을 수행시켜 결과를 얻는 방법이다. 예를 들면 현재 등판 투수의 피안타율과 현재 등판한 타자의 타율을 산출하여 타자의 안타 여부(특정 투수에 대한 특정 타자의 타율)를 예측하고, 만약 타자가 안타를 치는 경우 타자의 타율을 세분화한 루타별 타율을 이용하여 타자의 타격 상황을 예측하는 것이다. 이러한 방법을 통하여 선수의 타격, 진루상황, 득점 등을 예측함으로써 경기 결과를 예측하게 된다. 또한 경기 출장 횟수가 적은 타자의 경우, 현재 자신의 타율을 시스템에 그대로 사용하는 것은 부적절하며, 이로 인해 잘못된 예측이 될 수도 있다. 이에 대부분의 선수들은 전체 선수들에 대한 평균 타율 정도의 타격 능력을 갖는다는 가정 하에서 평균 타율을 기준으로 일정 신뢰구간 내에서 가감하여 보다 객관성을 띤 개인 타율을 생성하였다. 하지만 이 방법은 오프라인에서 실행되므로 현재 경기 정보를 적용할 수 없으며, 반복적인 시뮬레이션을 통해 경기 승패를 예측해야 함으로 많은 시간이 소요되는 단점을 가지고 있다.

2.3 역전파 알고리즘(Backpropagation Algorithm)

신경회로망 알고리즘에는 여러 가지 종류가 있는데 그 중 역전파 알고리즘(이하 BP)은 다층형태로 입력 데이타에 따른 결과 예상치와 각 테스트 데이타의 실제값 간의 차를 줄이기 위해 노드들간의 연결 강도를 조절하면서 학습하는 교사 학습(supervised learning)[7-9]의 한 형태이다. 야구 경기의 자료들은 연속적인 형태이므로, 이러한 자료를 처리하는데 효과적이다. 그러나 BP를 이용한 학습 네트워크 생성 시 입력 노드의 수가 많을 수록 적당한 분류 성능을 얻기 위해 더 큰 네트워크 사이즈를 요구한다. 만약 BP가 복잡해지면, 수행 시간과 복잡성이 증가한다. 그러므로 가능한 네트워크의 부하를 줄이면서 학습 능력을 높이기 위해 입력의 수를

줄이는 것과 사용될 요소를 어떻게 선택하느냐가 주요 문제이다.

2.4 휴리스틱(Heuristic) 모델

휴리스틱이란 반복적인 시행착오와 경험을 통한 경험적 학습을 말한다. 휴리스틱 방법으로는 유전자 알고리즘(GA), 시뮬레이티드어닐링, 타부 탐색 그리고 전문가의 경험에 의한 의견 수렴 과정을 거쳐 그 문제 영역에 접근하는 방법들이 있다[10]. 휴리스틱 기법은 그 적용 대상이 특정한 문제에 한정되고, 문제에 따른 해법의 구축도 용이하지 않다는 문제점을 지니고 있지만 실제 경기 자료를 바탕으로 휴리스틱 학습 모형을 만들면, 자료의 상관 관계와 계층적 구조를 고려하여 입력으로 활용할 수 있는 수식을 이끌어 낼 수 있다.

3. 제안된 모델 : 혼합형 모델(휴리스틱 기법+역전파 알고리즘)

과거 본 연구실에서는 프로야구 승패 예측을 위해 ID3, 통계적 기법, BP에 의한 예측을 수행하였다. 그 결과 ID3는 연속적인 형태를 갖는 경기 자료의 특성을 그대로 반영하지 못했으며, 통계적 방법은 반복적인 수행을 통한 예측으로 많은 시간이 걸렸다. 또한 BP를 이용한 경우 많은 자료를 학습에 이용하게 되므로 생성된 학습 네트워크의 복잡도를 증가시키는 등의 문제점을 가지고 있다. 이러한 기존의 방법들이 갖는 문제를 해결하기 위하여 휴리스틱 함수와 신경 회로망 모델을 혼합하는 예측 모형을 만들고자 한다. 그림 2는 본 논문에서 구현한 예측 모듈의 구성도이다.

3.1 휴리스틱 함수의 적용

예측을 위해 많은 자료를 이용할 경우 복잡한 구조의 신경망이나 결정 트리를 요구하게 되며 학습 모델을 생성하는데도 많은 시간이 필요하다. 또한 자료의 일부만 학습에 활용할 경우, 해결할 문제에 대한 충분한 자료를 학습 모델 생성에 활용하지 못하므로 좋은 예측을 기대하기 어렵다. 그러므로 충분한 정보를 제공하면서도 학습용 자료의 수를 줄일 수 있다면 적은 비용으로 더 나은 해를 얻을 수 있다. 본 논문에서는 예측에 사용될 자료의 수를 줄이는 방법으로 휴리스틱 함수를 사용하였다.

휴리스틱 함수는 경기 전 예측 함수와 경기 중 예측 함수로 구성되어진다. 경기 전 예측은 예정된 실제 경기가 수행되기 전 선수 개인 기록 데이타베이스를 이용하여 습득한 예측용 데이타베이스를 기반으로 두 팀의 경기 승패 가능성을 미리 가늠해 보는 것이다. 경기 전 예측을 위해 사용한 자료는 각 이닝별 과거 해당 이닝 이후의 평균 득점, 과거 현재 이닝 전까지의 안타수, 현재 이닝까지의 평균 안타수이다. 경기 중 예측은 경기가 진행되는 과정에서 앞으로의 승패를 예측해 보고자 하는 것으로 현재 상황에서 투수의 평균 방어율과 상대 타자에 대한 방어율 등의 경기 중의 종합적인 정보 및 과거 기록을 조합하여 경기 상황을 표현한다. 학습에 사용할 정보를 산출하는 수식은 다음과 같다.

(가) x 이닝에서 홈팀의 승리 정도

$$HomeWinRate(x) = \frac{HomeScore(x) - AwayScore(x)}{\max\{HomeScore(x), AwayScore(x)\}}$$

HomeWinRate(x) : x 이닝에서 홈팀의 승리 정도 (범위 : 1 ~ -1)

HomeScore(x) : x 이닝에서 홈팀의 예상 득점

AwayScore(x) : x 이닝에서 원정 팀의 예상 득점

x 이닝에서 홈팀이 승리할 수 있는지 여부를 판단할 수 있는 값이다. 이 값은 BP의 입력 값으로 사용되어진다. 많은 경기 자료들을 조합하여 HomeWinRate(x)로 표현함으로써 BP의 입력자료의 수를 현저히 줄이는 효과를 가져왔다. 결과가 음수이면 경기에 패할 가능성이 높아지고, 0이면 무승부, 양수면 승리할 가능성이 높아짐을 의미한다.

(나) HomeScore(x), AwayScore(x)

$$HomeScore(x) = CurRealScore(x) + RemPreScord(y) + RemPlayerPreScore(x)$$

CurRealScore(x) : 현재 이닝까지의 실제 점수

RemPreScore(y) : y 이닝(다음 공격 이닝) 이후의 남은 이닝 동안의 예상 득점

RemPlayerPreScore(x) : 현 이닝 중 남은 선수의 공격에 대한 예상 득점

현재의 홈팀(원정팀)의 득점과 과거 경기 기록을 바탕으로 산출한 경기 전, 경기 중 예측 점수를 합하여 홈팀(원정팀)의 경기를 통한 총득점을 표현한다

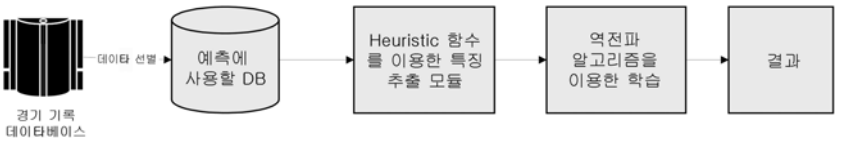

그림 2 예측 모듈 구성도

(다) RemPreScore(y)

$$\begin{aligned}\mathrm{Re}\,m\,\mathrm{Pr}\,eScore(y) &= InGame(y) * \min\left\{\frac{y}{4}, 1\right\} \\ &+ \mathrm{Pr}\,eGame(y) * \max\left\{\frac{4-y}{4}, 0\right\}\end{aligned}$$

InGame(y) : 경기 중 예측 점수

PreGame(y) : 경기 전 예측 점수

각 팀의 예상득점을 산출하는 부분으로 경기 전 자료와 경기 중 자료를 이용하여 예상 득점을 산출한다. 모든 선수들이 타석에 들어서는 4이닝 이후부터는 과거 기록에 의한 가중치를 감소시킴으로써 현재 경기에서 산출된 실제 기록을 경기 예측에 반영하도록 하였다.

(라) PreGame(y)

$$\begin{aligned}\mathrm{Pr}\,eGame(y) &= OppAverageScore * \frac{OppCurPitcher}{OppPitcher} \\ &* 0.5 + OppAverageScore * 0.5\end{aligned}$$

OppAverageScore : 상대팀에 대한 평균 득점

OppCurPitcher : 상대팀의 현재 투수 방어율

OppPitcher : 현재 팀에 대한 상대팀 투수의 평균 방어율

경기 중 예측 점수를 산출하는 부분으로, 야구 승패 예측에 가장 큰 영향을 미치는 투수와 타자의 기록을 이용하였다. 투수가 경기에 영향을 미치는 기준을 50% 정도로 가정하여 실험하였다.

(마) InGame(y)

$$InGame(y) = PassAverageScore * \frac{CurInning}{PassCurInning}$$

PassAverageScore : 과거 해당 이닝 이후의 평균 득점

CurInning : 현재 이닝까지의 평균 안타수

PassCurInning : 과거 현재 이닝 전까지의 안타수

과거 기록과 현재 상황을 이용하여 경기 중 예상 득점을 산출하는 공식이다.

경기 기록 데이타베이스로부터 경기에 영향을 주는 기록 12개를 추출하고, 이 중 중복 요소를 가지고 있는 9개의 자료를 조합하여 새로운 값(홈팀이 승리할 확률)을 생성하였다. 새로 생성된 값과 나머지 2개(현재 경기 이닝, 상대팀에 대한 승률)를 BP의 입력으로 이용하고 나머지 한 개(홈팀의 실제 기록상 승률)는 BP에서 연결강도 조정을 위해 활용될 목표 값으로 활용되었다. 그러므로 혼합형 시스템에서 사용된 자료의 수는 4개지만 실제 예측에 활용된 자료는 12개이다.

3.2 역전파 알고리즘의 적용

휴리스틱 함수를 이용한 예측은 경험을 통하여 학습용 자료를 도출해 냄으로써 신뢰성에 대한 문제가 제기된다. 그러므로 휴리스틱 함수를 통해 산출된 자료들을 역전파 알고리즘을 이용하여 일반화시킴으로써 보다 안전한 예측 시스템을 구현할 수 있다. BP는 처리 노드가 많기 때문에 몇 개의 노드나 연결이 가진 결함이 비교적 시스템 전체에 크게 영향을 주지 않으며(결함 내구성, fault tolerance), 새로운 환경에 즉각적으로 프로그램을 갱신하고 유지(적용성, adaptability)하는데 용이하다. 잘 선별된 자료들을 BP에 이용할 경우 높은 분류 성능 발휘할 수 있는 이점을 가지고 있다.

경기 예측을 위한 학습 모델은 다음과 같은 순서에 의해 생성되어진다. 첫 번째, 휴리스틱 함수 생성과 BP의 입력으로 사용할 값들을 기록지로부터 추출하여 예측용 DB를 생성한다. 즉, 매 경기마다 생성되는 기록지에는 해당 경기에 대한 모든 상황이 기록되므로 기록지로부터 얻은 값들에 대한 평균이나 합을 구하여 과거 경기기록을 생성한다. 또한 기록지 내의 특정 이닝에서의 상황들(현재 이닝, 안타수 등)은 학습 모델 생성에 사용할 경기 중 정보로 활용한다. 두 번째, 예측용 DB의 값을 휴리스틱 함수에 적용시켜 새로운 예측용 입력 자료인 홈팀의 승리 정도를 산출한다. 세 번째, BP를 이용하여 학습 모델을 생성한다. BP의 입력 값으로 기록지로부터 얻은 특정 상황에 대한 이닝(현재 이닝), 홈팀의 승리 정도, 상대팀에 대한 평균 승률, 실제 승률(실제 경기 상에서 해당 이닝까지의 승률)등을 이용하여 학습 모델을 생성한다.

이러한 과정을 거쳐 생성된 경기 승패 예측용 학습 모형은 실제 경기에서 이미 알고있는 과거 기록과 현재 경기에서 발생하는 상황(이닝, 현재 이닝까지의 평균 안타수, 실제 점수 상황)을 학습 모델에 적용시킴으로써 경기 중 예측이 가능하도록 하였다.

4. 시스템의 실험 및 구현

4.1 실험 환경 및 결과

본 논문에서 제시한 프로야구 경기의 승패 예측 시스템의 성능 평가를 위해 KBO(한국 야구위원회)의 데이타베이스 자료를 사용하였다. 1998년 KBO 경기 자료를 기반으로 1022개의 초기 자료를 생성하였다. 그 중 동일한 패턴이 나타나지 않도록 중복 자료를 제거하였고, 각 이닝별 데이타의 수를 일정하게 맞추기 위하여 각 이닝별로 70개씩 모두 630개의 자료를 추출하였다. 630개의 자료 중 학습 자료로 504개, 테스트 자료로 126개를 사용하였다. BP에서 은닉층은 1계층인 경우와 2계층인 경우에 대하여 실험하였다. 반복횟수로는 5000, 7000, 10000, 20000, 50000까지 수행하였다. 표 1은 은닉층의 수와 반복 횟수의 변화에 따른 혼합형 모델의 예측률을 보여준다.

실험 결과 뉴런의 수가 10개, 은닉층이 2개, 전달함수는 tanh, 반복 횟수가 7000일 때, 84.92%의 가장 높은

표 1 은닉층의 수와 반복 횟수에 따른 예측률(n : 뉴런의 수)

반복횟수	은닉층이 1개인 경우				은닉층이 2개인 경우			
	n=2	n=5	n=7	n=10	n=2	n=5	n=7	n=10
5000회	0.8175	0.7937	0.8333	0.8016	0.7937	0.8175	0.8095	0.8016
7000회	0.8175	0.8254	0.8254	0.8095	0.8016	0.8016	0.8254	0.8492
10000회	0.8175	0.8175	0.8254	0.7937	0.7937	0.8333	0.8016	0.7857
20000회	0.8254	0.7937	0.8175	0.7937	0.6349	0.8254	0.8333	0.8254
50000회	0.8016	0.8016	0.8095	0.7937	0.6349	0.8333	0.7063	0.8095

예측률을 보임을 알 수 있었다. 표 2는 기존 알고리즘들과 제안된 모델의 결과를 비교한 것이다.

표 2 예측 결과 비교표

사용한 알고리즘	예측률
ID3 알고리즘에 의한 예측	81.0%
통계적 시뮬레이터에 의한 예측	80.1%
역전파 알고리즘을 이용한 예측	76.0%
제안한 모델(혼합형 모델)	84.9%

BP를 이용한 예측의 경우, 본 논문에서 제시하고 있는 혼합형 모델 생성에 사용된 12개의 자료들이 입력으로 사용되었다. 이 경우에는 혼합형 모델에 비해 입력의 수는 많으나 자료들 간의 중복 요소로 인해 학습 네트워크의 복잡도만 증가할 뿐 예측률 향상에는 큰 영향을 주지 못하였다. 혼합형 모델은 휴리스틱 함수를 이용하여 다양한 자료들을 혼합한 학습 요소를 산출해 냄으로써 학습 자료의 수도 줄이고 예측률도 높이는 결과를 보일 수 있었다.

4.2 구현된 결과 화면

기존의 많은 시스템들처럼 단순한 게임 형태의 경기를 수행하는 것이 아니라 컴퓨터를 통해 사용자가 팀을 구성하고, 실제 기록을 기반으로 경기를 수행한다. 또한 사용자가 직접 팀을 구성할 수 있어서 보다 흥미로운 경기 진행이 가능하도록 하였다. 경기 환경설정 모드를 이용하여 팀을 구성하고, 간단한 이미지를 이용하여 경기의 진행상황을 보이도록 하였다. 경기가 진행되면서 현재 입력 상황에 따라 예측된 경기 승률을 보여준다.

그림 3 경기 실행 화면

5. 결 론

현재 프로야구가 많은 사람들에게 있어서 각광받는 스포츠 경기이나 직접 야구장에 찾아가서 경기를 즐기는 것은 그리 쉽지 않다. 하지만 인터넷을 통하여 프로야구 스타들의 활약상이나 홈페이지의 방문과 자기가 응원하는 팀의 전적을 보기 위해 사이트를 찾는 것은 아주 흔한 일이 되었다. 본 연구에서는 이러한 사람들의 욕구를 보다 흥미 있게 해결하고자 인공지능 기법을 야구 승패 예측 시스템 구현에 활용하여 보았다.

승패 예측 시스템 구현에 있어서 가장 중요한 문제인 예측 자료 선택에 있어서는 휴리스틱 함수를 이용하여 보다 많은 의미를 가지면서도 실제 예측 알고리즘에는 적은 수의 자료가 활용되도록 특징의 차원을 감소시킴으로써 예측 모델의 복잡도를 감소시킬 수 있었다. 그리고 휴리스틱 함수를 통해 산출된 자료들을 역전파 알고리즘에 의해 일반화시킴으로써 보다 안정적인 예측 시스템을 구현할 수 있었다.

제안된 모델은 프로야구 경기 승패 예측기 생성에 활용하였다. 기존의 역전파 알고리즘 외에 ID3 알고리즘이나 통계적 방법을 이용한 예측 시뮬레이터보다 더 나은 예측률을 보였다. 이러한 예측 모듈을 게임 시 적극 활용하여 생동감 있는 경기를 할 수 있게 함으로써 프로야구 게임이 활성화 되도록 할 수 있을 것이며 그러기 위해서는 실제 경기 상황과 유사한 상황을 전제로 더 많은 자료를 포함할 수 있는 공식을 이끌어 내는 것이 필요하다. 또한 휴리스틱 모델 생성 시 사용된 고정된 값들을 특정 상황에 맞게 변형하는 문제와 더욱 세밀한 홈/원정별, 구장별, 수비별 상황을 고려한 사례기반에 의한 예측이 이루어져야 할 것이다.

참 고 문 헌

[1] H. Almuallim and T. G. Dietterich. Efficient

algorithm for identifying relevant features. In Proc. of 9th Canadian Conf. on Artificial Intelligence, Vancouver, Britich Columbia, pages 38-45. Morgan Kaufmann, 1992.
[2] 서재순, "귀납적 추론을 이용한 프로야구 승패 예측 시스템 개발에 관한 연구", 경희대학교, 1994.
[3] 홍석미, "프로야구 승패 예측을 위한 게임 시뮬레이터 개발에 관한 연구", 경희대학교, 1997.
[4] 허준회, "프로야구 경기 시뮬레이터에서 데이타마이닝을 이용한 투수 선정 및 투수 교체 시기 선택에 관한 연구", 경희대학교, 1999.
[5] P. S. Bradley, O. L. Managasarian, and W. N. Street. Feature selection via mathematical programming. INFORMS Journal on Computing, 10(2):209-217, 1998.
[6] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. Artificial Intelligence, pages 245-271, 1997.
[7] W. S. Sar;e. Neural networks and statistical models. In Proc. of 19th Annual SAS Users Group International Conference, pages 1538-1550. SAS Institute, 1994.
[8] M. Riedmiller. Advanced supervised learning in multi-layer perceptrons-from backpropagation to adaptive learning algorithms. International Jpurnal of Computer Standards and Interfaces, 16(5): 265-278, 1994.
[9] R. Battiti. Using mutual information for selecting features in supervised neural net learning. IEEE Transaction on Neural Networks, 5(4):537-550, July 1994.
[10] C. Guerra-Salcedo, S. Chen, D. Whitley, and S. Smith. Fast and accurate feature selection using hybrid genetic strategies. In Proc. of Genetic and Evolutionary Computation Conference, pages 177-184, Piscataway, NJ, 1999. IEEE Service Center.

홍 석 미
1994년 상지대학교 전자계산학과 졸업(이학사). 1997년 경희대학교 대학원 전자계산공학과(공학석사). 1998년~현재 경희대학교 대학원 전자계산공학과 박사과정. 관심분야는 기계학습, 데이타마이닝, 에이전트, 정보보호, 최적화

정 경 숙
1995년 경희대학교 수학과 졸업(이학사) 1997년 경희대학교 대학원 전자계산공학과(공학석사). 1999년~현재 경희대학교 대학원 전자계산공학과 박사 과정. 관심분야는 인공 지능, 정보보호, 암호화, 데이타마이닝

정 태 충
1980년 서울대학교 전자공학과(학사) 1982년 한국 과학 기술원 전자공학전공(공학석사). 1987년 한국 과학 기술원 전자공학전공(공학박사). 1987년~1988년 KIST 시스템 공학센터 선임연구원 1988년~현재 경희대학교 컴퓨터공학과 교수 관심분야는 기계학습, 정보보호, 최적화, 에이전트