전향참조 기법을 이용한 효율적인 선발 라인업 시스템 구현

A Implementation of an Efficient Starting Line-up System using the Forward Channing Technology

박 홍 진* 권 순 환**
Hong-Jin Park Soon-Hwan Kwon

요 약

본 논문은 전문가시스템 기법인 전향추론 기법을 이용하여 프로야구의 효율적인 선발 라인업 시스템을 설계 및 구현한다. 지식베이스는 선수 286명 148개 평가 항목으로 구성되어 있으며, 선수들의 경쟁력을 포인트로 평가 하였다. 1차 포인트는 정적인 정보에 의해 산출된 포인트이며, 2차 포인트는 동적인 정보에 의해 산출된 포인트이다. 1, 2차 포인트를 통해 종합 포인트가 산출된다. 이를 이용하여 주어진 환경과 상대에 대한 효율적인 라인업 시스템 구현한다.

Abstract

This paper introduces how we have designed and implemented an efficient starting line-up system of professional baseball players, using the forward chaining technology which is an expert system technology. The knowledge base adopted holds 286 players and 148 evaluation items. Each player's competitiveness is measured as a numerical value in points through 2 different phases. The 1st phase point represents the static information of the target player and the 2nd his dynamic information. This paper gets a representative value of the player, which is used as the basis on which an efficient line-up system in views of both game environments and counterparts is constructed

☞ keyword : Expert system, Inference engine, Forward chaining technology, Line-up

1. 서 론

스포츠는 선의의 경쟁을 위해 끊임없이 훈련하고 집중함으로써 경기에 대해 최선을 다할 수 있다. 또한 실제 경기에서 기존의 모든 자료를 수집하고 이를 분석하면 다음 경기의 경기력 향상에 크게 도움이 될 수 있다.

스포츠 분야 중 야구는 데이터에 의해 철저히 관리되어야 하는 스포츠로서 실제 경기에서는 이러한 기록이 집중적으로 분석되고 정리하여 투수나 타자와 관련된 상황별 특성을 찾아내는 것은 승리를 위해 매우 중요한 일이다.

[1] 논문에서는 기계 학습 알고리즘에 많이 사용되고 있는 고전적인 알고리즘인 ID3을 이용하였다. 과거 경기기록을 이용하여 학습 트리를 구성하며, 학습된 트리를 기반으로 승패 예측을 가능하게 하였다. 그러나 대부분의 야구 기록은 연속적이므로 트리로 구성되어 있는 ID3 이용 방식은 트리 노드의 개수가 크게 증가되는 문제점을 지니고 있다. [2-4] 논문은 신경망에서 사용되고 있는 역전파 알고리즘을 통해 승패를 예측하고 있다. 역전파 알고리즘은 다층형태로 입력하여 실제 값과의 차이를 줄이지만 수행시간과 복잡성이 증가하게 된다. [5,6] 논문은 휴리스틱(Heuristic)

* 종신회원 : 상지대학교 IT학부 부교수
hjpark1@sangji.ac.kr
** 준 회 원 : 와이즈캣 기획운영팀 근무
ykssah@wisecat.co.kr
[2007/11/27 투고 - 2007/12/18 심사 - 2008/04/08 심사완료]

모델의 기반으로 승패를 예측하는 시스템이다. 이 기법은 적용 대상이 특정한 문제에 한정되어 있고 문제에 따른 해법의 구축도 용이하지 못하다. [7] 논문은 휴리스틱과 역전파 알고리즘을 혼합한 모델로 승패를 예측하고 있다. 기존의 연구들은 주로 경기의 승패 예측에 집중되어 있으며 선발 라인업에 대한 연구는 미진하다.

야구에서 승리를 위해 적시적소에 알맞은 선수가 배정되어야 하며 상황에 따라 가장 강한 선수가 그라운드로 나와야 승리 할 수 있는 확률을 최대화 시킬 수 있을 것이다. 이때 선발 출전하는 선수들의 명단을 선발 라인업(line-up)이라고 한다 [8]. 즉, 선발 라인업의 사전적 의미로는 야구에서 출전 선수의 타격 순서 또는 수비 위치라고 표현되고 있으며, 경기에 선발 출전하는 선수들의 타격(투구)순서를 결정하는 것으로 경기에 나오는 선수들의 순번을 매긴 것이라고도 표현 할 수 있다.

현재의 선발 라인업은 단지 정적인(static) 정보만을 보여주기 때문에 다양한 팬들에게 어필(appeal) 할 수 없을 뿐만 아니라 일반적으로 야구는 수많은 정보 데이터 싸움이지만 데이터 간의 상관관계에 대한 언급이 전혀 고려하지 못하고 있다. 또한 숨겨진 기록(hidden statistic)을 전혀 제공하지 못하고 있기 때문에 주어진 환경에 대한 효율적인의 선발 라인업을 전혀 알 수 없다.

본 논문에서는 이에 대한 해결 방안으로 전문가 시스템 기법의 전향추론 기법을 이용하여 효율적인 선발 라인업 시스템을 설계 및 구현한다. 효율적인 라인업 구성을 위해 기존 정적 정보와 새로운 동적 정보를 비교 수행한다. 정적 정보 비교는 선수들의 직접적인 기록 간의 비교를 뜻한다. 즉 A선수와 B선수가 누가 안타가 몇 개이고 타점이 몇 개 인지를 단순 비교하는 방식이다. 동적 정보 비교는 A선수와 B선수의 정적인 기록에 외부요건(상대팀, 상대투수, 구장등)을 추가하여 그에 해당하는 기록을 산출하여 서로 비교하는 것으로 실제 경기를 위해서는 정적 비교보다 훨씬 더 정확하고 세부적인 비교가 이루어 질수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문의 이론적 기반이 되는 연구로 전문가 시스템과 추론에 대해 설명하고, 3장과 4장에서는 효율적인 선발 라인업 시스템을 설계 및 구현한다. 5장에서는 구현한 시스템과 실제 성적과 비교평가를 한다. 6장에서는 결론 및 향후 연구를 설명한다.

2. 기반 연구

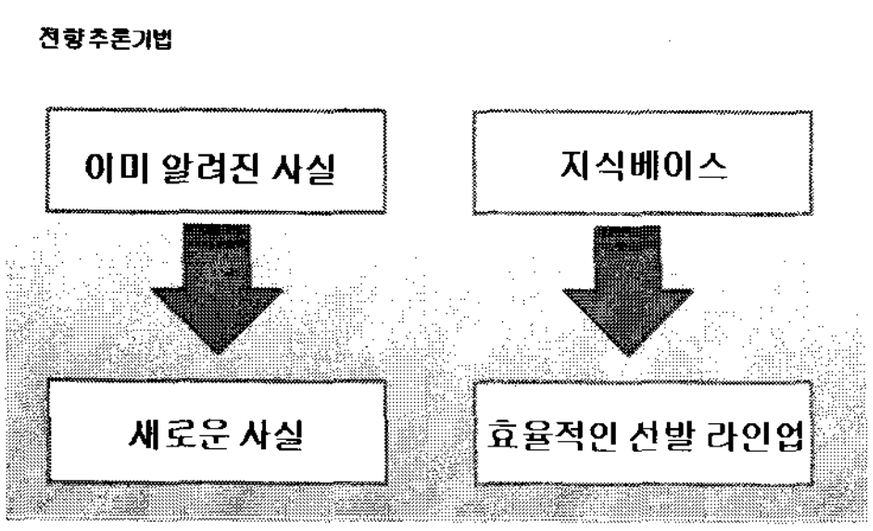
2.1 전문가 시스템

전문가 시스템은 특정 분야의 전문가적 지식 및 경험을 가진 인간(또는 조직)의 판단과 행동을 흉내 내는 컴퓨터 프로그램이다. 예를 들면 잘 알려진 전문가시스템들 중에는, 체스를 두거나 의학 진단을 지원하는 것 등이 있다.

전문가 시스템의 구조는 다음 세가지 요소로 구성되어 있다. 첫 번째는 지식 베이스 부분으로 다루는 분야와 관련된 모든 정보를 가지고 있으며 지식의 표상 언어로 구성되어 있다. 두 번째는 사용자 인터페이스 부분으로 사용자가 편리하고 이해하기 쉽도록 구성되어 있다. 마지막으로 추론 기관은 정보나 자료들과 관련된 다양한 문제들을 해결할 목적으로 지식 베이스의 발견적 방법들을 사용하기 위한 프로그램이다[9-12].

2.2 추론

전문가 시스템의 추론 기법은 크게 3가지로 구분된다. 첫 번째는 전향추론(forward chaining)으로 (그림 1)과 같다. 전향추론은 이용 가능한 정보로부터 출발하여 적절한 결론을 얻는 방법 이다.



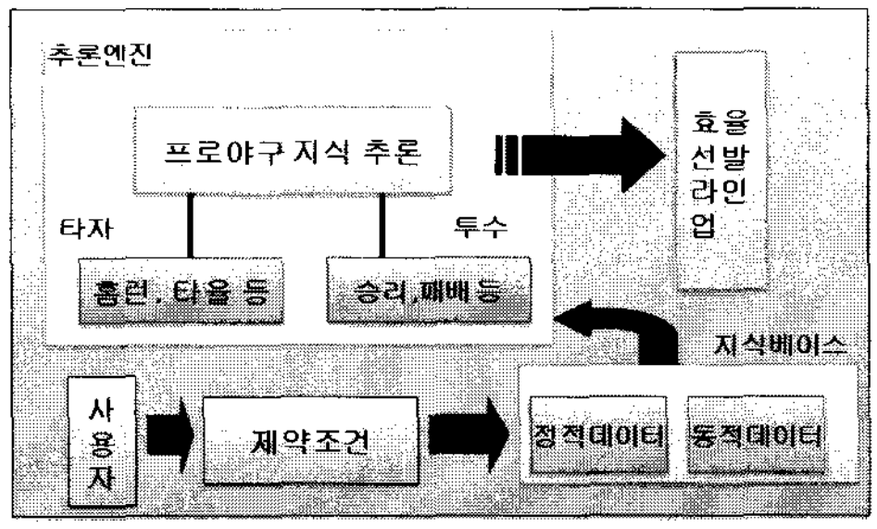
(그림 1) 전향추론기법

즉, 주어진 상황에 해당하는 사실에 의하여 조건부가 만족되는 규칙을 찾아 결론부를 수행하고 다음 단계를 계속 진행 방식이다. 이는 주로 주어진 사실에서 유도되는 모든 것 찾기 위한 문제에 적용된다. 두 번째 방법은 후향추론(backward chaining) 으로 어떤 목표를 하나 정한 후 이 목표가 성립하기 위한 조건들을 하나씩 맞춰 가는 방법으로 보통 목표가 몇 개 되지 않을 때 목표를 하나씩 선택해 가며 선택된 목표가 성립되기 위한 여러 가지 조건(규칙이나 사실)들이 만족 되는가 조사 할 때 효율적이다. 주로 특별히 주어진 결론 입증하기 위한 문제에 적용된다. 마지막 방법은 혼합형 추론(hybrid chaining)으로 이는 전향 추론과 후향 추론을 혼합하여 사용 하는 방식이다[11,12]. 본 논문에서는 전향추론 기법을 사용하여 이미 알려진 사실(지식베이스)을 통하여 새로운 사실(효율적인 선발 라인업)을 유추 하는데 목적이 있다.

3. 효율적인 선발 라인업 시스템 설계

3.1 시스템 구조도

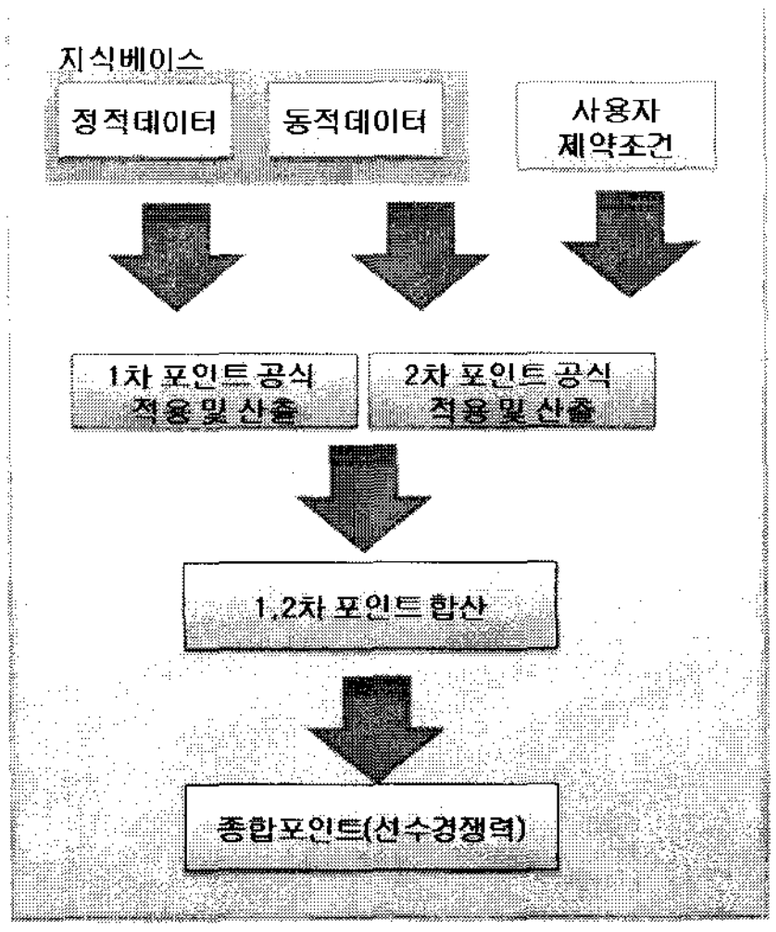
(그림 2)는 효율적인 선발 라인업 시스템 구조도이다. 지식베이스에 저장되어 있는 데이터는 정적(static) 데이터와 동적(dynamic) 데이터로 구분된다. 정적 데이터(변동이 없는 데이터)는 홈런이나 타율처럼 사용자가 입력한 제약조건에 의해 변하지 않는 데이터이며, 동적 데이터(변동이 있는 데이터)는 사용자의 제약조건에 의하여 변하



(그림 2) 시스템 구조도

는 데이터로 예를 들어 상대팀 타율이나 주간/야간 타율 등이다. 이러한 데이터를 포함하고 있는 지식베이스를 이용하여 추론엔진 부분에서는 프로야구의 전반적인 지식 즉, 타순별 고려사항, 투수 보직별 고려사항 등을 공식화 하여 포인트화 함으로써 선수들의 우선순위를 매기고 그 후 포인트와 포지션이 겹치지 않도록 배치하여 효율적인 선발 라인업을 작성하게 된다.

3.2 포인트 구조도



(그림 3) 종합포인트(선수 경쟁력) 구조도

(그림 3)은 종합포인트(선수 경쟁력) 구조도 이다. 포인트는 선수들의 경쟁력을 수치로 표현해주는 것으로, 포인트는 1차 포인트와 2차 포인트로

나눈다. 1차 포인트는 정적인 데이터를 기반으로 공식에 대입하여 산출된 포인트이고, 2차 포인트는 동적인 데이터를 기반으로 공식에 대입하여 산출된 포인트이다. 1차와 2차 포인트가 합쳐져 종합 포인트로 선수들의 경쟁력을 평가하게 된다.

3.3 1차 포인트

최근 2년간의 프로야구 감독들의 타순 및 투수 보직 선택에 있어서 어떤 특성 기록에 대한 선호도가 높았고, 그에 따라 어떤 형식으로 타순 및 투수 보직을 선택했는지를 공식화 표 1과 표 3으로 하여 1차적으로 사용자의 제약 조건이 들어오기 전에 포인트화하였다. 사용자의 제약조건이 입력되었을 경우 그 후에 대한 성적을 포인트화 할 수 있다는 장점이 있다.

3.3.1 1차 포인트 공식(타자부분)

(표 1) 1차 포인트 공식(타자)

1번타자	(OBP*20)+(SB*20)+(RC*10)+(SO/2)+(R*10)
2번타자	(OBP*20)+(SB*20)+(RC*10)+(SO/2) +(100-(GDP*10))
3번타자	(RBI*10)+(HR*10)+(AVG*200) +(OPS*100)+(RC*10)+(SLG*200)+ (100-(GDP*10))
4번타자	(RBI*10)+(HR*15)+(OPS*100)+ (RC*10)+(SLG*200)
5번타자	(SLG*200)+(RBI*15)+(OPS*100)+ (AVG*150)+(RC*10)
6번타자	(AVG*200)+(OPS*100)+(OBP*200) +(RBI*10)+(RC*10)+(SO/2)
7번타자	(OBP*200)+(SB*20)+(SO/2)+(RC*10)+ (R*10)
8번타자	(OPS*100)+(HR*10)+(100-(GDP*10))+ (AVG*200)+(RBI*20)+(SO/2)
9번타자	(OBP*300)+(HIT*2)+(SB*20)+ (50-(CS*3))+(R*5)

* 1번 타자

(OBP*20) + (SB*20) + (RC*10) + (SO/2) + (R*10)

= 1번 타자는 출루가 가장 중요하므로 안타와 볼넷 등이 모두 포함된 OBP(출루율)이 중요시 되었으며 출루시 한 베이스를 더 갈수 있는 능력인 SB(도루)도 평가 항목에 들어갔다. 또한 출루의 방해물인 SO(삼진)은 그 수 만큼 포인트서 마니어스 효과를 주었으며 R(득점) 루에 진루시 홈에 들어오는 횟수인 득점을 추가하였다. RC(득점생산력)는 빌 제임스가 고안한 공식으로 득점생산력을 뜻하는데 그에 관한 공식 및 설명은 하단부에 위치하고 있다.

* 2번 타자

(OBP*20) + (SB*20) + (RC*10) + (SO/2) + (100-(GDP*10))

= 2번 타자 역시 1번 타자와는 임무가 그리 다르지는 않기 때문에 OBP과 SB, RC, SO이 평가 대상이 되었으며, 새로 추가된 GDP(병살타)는 1번 타자가 루에 나갔을 경우 같이 죽을 수 있는 확률을 최소화하기 위해 추가된 항목이다.

* 3번 타자

(RBI*10) + (HR*10) + (AVG*200) +(OPS*100) +(RC*10) +(SLG*200) + (100-(GDP*10))

= 3번 타자의 역할은 1번, 2번 타자가 진루시 그를 불러 드릴 수 있는 능력이 중요시 되므로 RBI(타점)과 AVG(타율), OPS(장타율+출루율), RC을 평가항목으로 두었으며 또한 2사 1루 같은 상황에서도 1루 주자를 한 번에 불러 드릴 수 있는 장타를 기대하게 되므로 HR(홈런)과 SLG(장타율) 부분을 평가요소로 추가하게 되었다. 또한 이미 출루한 1번, 2번과 같이 죽을 확률을 최소화하기 위해서 GDP 항목도 포함되었다.

* 4번 타자

(RBI*10) + (HR*15) + (OPS*100) + (RC*10) + (SLG*200)

= 4번 타자는 팀의 중심으로 여러 가지 평가 항목이 필요가 없다. 그저 1.2.3번을 불러 들일 수 있는 능력만이 그의 평가 대상이다. 그러므로 RBI, OPS, RC이 주가 되고 3번타자와 같이 이유에서 장타가 필요하므로 HR과 SLG 부분이 추가로 평가 대상이 되었다

* 5번 타자

(SLG*200) + (RBI*15) + (OPS*100) + (AVG*150) + (RC*10)

= 5번 타자는 4번 타자를 뒷받힘 해주는 역할로 적절한 장타와 타점능력, 그리고 약간의 정교함도 필요로 한다. 그러므로 SLG, RBI, OPS, RC, AVG 부분이 평가 대상이 되었다.

* 6번 타자

(AVG*200) + (OPS*100) + (OBP*200) + (RBI*10) + (RC*10) + (SO/2)

= 6번 타자는 같은 회에 5번 타자 다음으로 나오거나 선두타자로 나올 수 있는 가능성이 많으므로 출루와 더불어 클러치 능력을 함께 평가 하였다. 그러므로 AVG, OPS, OBP, RBI, RC 부분을 중심으로 평가 하였고, 선두 타자의 가능성을 위해 SO에 대한 마이너스를 주므로 한층 더 출루에 대한 중요성을 부각 시켰다.

* 7번 타자

(OBP*200) + (SB*20) + (SO/2) + (RC*10) + (R*10)

= 7번 타자는 제 2의 1번 타자로 평가된다. 그만큼 출루와 한 베이스 더 가는 능력이 중요시되고 출루시 홈에 들어올 수 있는 득점 확률을 극대화 하여야 하므로 그에 맞는 요소들을 평가하였다.

* 8번 타자

(OPS*100) + (HR*10) + (100-(GDP*10)) + (AVG*200) + (RBI*20) + (SO/2)

= 8번 타자는 7번 타자가 출루하였을 때 그를 불러들일 수 있는 능력이 있어야 한다. 흔히들 가장 쉬운 타순이라고들 인식되어 있지만, 8번이 터진다면 게임이 쉽게 풀리는 것을 자주 볼 수 있다. 그러므로 평가요소는 출루와 장타 위주로 평가 하였으면 타점과 홈런에 중점을 두고 병살과 삼진으로 최악의 상황에 대한 마이너스 점수를 부여하였다.

* 9번 타자

(OBP*300) + (HIT*2) + (SB*20) + (50-(CS*3)) + (R*5)

= 상위 타선과 하위를 이어 주는 역할을 하는 선수로 이 선수가 다음 선수로 이어 준다면 그 공격 기회에서는 대량 득점을 노릴 수 있다. 그러므로 출루에 대한 가중치를 최대화 하였으며 출루시 한 베이스 더 가는 도루와 홈에 들어오는 횟수인 득점을 평가요소로 정하였으며 CS(도루실패) 도루실패에 대한 마이너스를 주어 출루 후에 어이없이 주루사 하는 것에 대한 일정치의 감점을 부여하였다.

* RC(Runs Created) 산출 공식

RC 표 2 는 빌 제임스가 쓴 "Win Shares" 에 나타는 공식으로 한 선수가 득점에 공헌한 정도를 수치화 한 것 이다[13].

(표 2) RC(Runs Created)

RC	((A+2.4*C)*(B+3*C))/(9*C)-(0.9*C)	
	A(출루)	안타+볼넷+몸에맞는공-도루자-병살타
	B(전진)	루타+0.26*(볼넷+몸에맞는공)+0.53*(희생타+희생플라이)+0.64*(도루)-0.03*(삼진)
	C(타석)	타석

3.3.2 1차 포인트 공식(투수부분)

(표 3) 1차포인트 공식(투수)

선발투수	(승리*20)+(이닝*2)+(QS*10)+(100-(WHIP*10))+(100-(피안타*10))+ (100-(피OPS*10))+(100-(ERA*10))+(GDP*5)+(100-(BB*5))+(KK*10)+ (100-(피홈런*5))
불펜투수	(홀드*30)+(이닝*2)+(100-(WHIP*20)) +(100-(피안타*5))+(100-(피OPS*5))+(100-(ERA*10))+(GDP*10)+(100-(BB*5))+(KK*10)+(100-(피홈런*5))
마무리투수	(세이브*30)+(100-(WHIP*20))+(100-(ERA*10))+(KK*15)+(GDP*10)+ (100-(피안타*5))+(100-(피OPS*5))+ (100-(피홈런*5))

* 선발 투수

(승리*20) + (이닝*2) + (QS*10) + (100-(WHIP*10)) + (100-(피안타*10)) + (100-(피OPS*10)) + (100-(ERA*10)) + (GDP*5) + (100-(BB*5)) + (KK*10) + (100-(피홈런*5))

= 선발 투수의 가장 큰 목적은 승리에 있다. 또한 불펜 투수의 체력을 아끼기 위해서는 많은 이닝을 소화해야 한다. QS(퀄리티스타트)란 6이닝 동안 선발투수가 3자책점 이하로 막은 게임을 1로 부여하는 수치로 좋은 선발 투수의 기준이 된다. WHIP(한 이닝에 출루를 허용한 타자의 수)은 얼마나 안정적인 투구를 하는지에 대한 평가 요소이다. 선발투수는 다방면에서 뛰어나야지 좋은 선발 투수라고 할 수 있기에 거의 모든 요소가 평가 항목에 추가되었다.

* 불펜 투수

(홀드*30) + (이닝*2) + (100-(WHIP*20)) + (100-(피안타*5)) + (100-(피OPS*5)) + (100-(ERA*10)) + (GDP*10) + (100-(BB*5)) + (KK*10) + (100-(피홈런*5))

= 불펜투수는 홀드(자기 팀이 앞서고 있을 때 투입 돼 상대 팀에 리드를 허용하지 않은 채 다음 투수에게 마운드를 넘겨준 릴리프 투수에게 주어지는 수치)로 불펜투수의 최고의 평가 항목으로 여겨지고 있다. 또한 불펜투수라 함은 안정된 투구로 팀의 승리를 마무리 투수에게까지 넘겨주는 역할을 해야 하므로 WHIP 역시도 중요시되고 있다. 불펜투수도 여러 가지 상황을 겪게 되므로 어느 정도 수준의 평가 항목은 상당수 포함되어 있으며 GDP의 경우 거의 불펜투수는 선발투수가 주자를 남겨둔 상황에 구원 등판하게 되므로 의미 있는 기록으로 가중치를 주었다.

* 마무리 투수

(세이브*30) + (100-(WHIP*20)) + (100-(ERA*10)) + (KK*15) + (GDP*10) + (100-(피안타*5)) + (100-(피OPS*5)) + (100-(피홈런*5))

= 마무리 투수는 세이브가 가장 중요시 되는 요소 이고, 또한 마무리 투수의 불안정은 팀의 패배로 바로 직결되므로 굉장히 안정적인 투구를 해야 하므로 WHIP 이 추가 되었다. 또한 위기 상황에 등판하여 한 두 타자 정도는 SO으로 돌

려 세울수 있는 능력도 있어야 하고, 같은 이유로 GDP의 유도도 좋아야 한다. 한순간에 경기가 역전되는 피 홈런도 마무리 투수가 피해야 할 요소 중 하나이다.

3.4 2차 포인트

2차 포인트는 사용자의 제약 조건에 의해 생성되는 포인트 이다. 표 4와 표 5는 사용자 자신이 응원하는 팀, 상대팀, 상대팀 선발투수, 홈/어웨이 경기장, 주간/야간 의 제약 조건을 입력한다. 여기서 그 제약 조건에 더 우수한 선수를 선발하기 위해 2차 포인트를 계산하여 1차 포인트와 합산하여 각 타순 및 투수 보직에 대한 통합 포인트가 산출한다.

3.4.1 2차 포인트 공식(타자부분)

타자는 (표 4)에서 나타난 것처럼 총 5개 항목에 대하여 평가가 이루어진다. 항목은 사용자가 입력한 제약조건에 대한 부분으로 상대와 주어진 환경에 대한 효율적인 선수를 뽑기 위한 이유로 상대팀에 대한 타율과 선발투수에 대해 좌, 우 타율이 들어가고 또한 경기의 시작 시간이 주간인지 야간이지에 대해 그에 맞는 타율이 들어간다. 상대팀에 대한 타율과 더불어 홈/어웨이에 대한 타율 역시도 가중치가 포함되어 2차 포인트로 산출되어 1차 포인트와 더해서 종합 포인트로 산출된다.

(표 4) 2차 포인트 공식(타자)

타자	상대팀 타율 *100
	상대선발 좌.우 투수타율 *100
	주간/야간 타율 *100
	상대팀 OPS *100
	홈/어웨이 타율 *100

3.4.2 2차 포인트 공식(투수부분)

(표 5) 2차 포인트 공식(투수)

투수	홈/어웨이 방어율 (100-(방어율*10))
	주간/야간 방어율 (100-(방어율*10))
	상대팀 방어율 (100-(방어율*10))
	상대 승리 승리*20 (선발투수만 해당)
	상대 홀드 홀드*20 (불펜투수만 해당)
	상대 세이브 세이브 *20 (마무리투수만 해당)

총 6개 부분 표 5으로 평가되며 투수들에게 공통적으로 상대팀 방어율과 주간/야간 방어율 그리고 홈/어웨이 방어율에 대한 가중치가 2차 포인트로 적용되며, 선발투수에게는 승리, 불펜투수에게는 홀드, 마무리 투수에게는 세이브가 가중치가 적용되어 산출된다. 이와 같이 산출된 2차 포인트는 1차 포인트와 합산되어 종합 포인트로 산출된다.

4. 효율적인 선발 라인업 구현

4.1 구현상의 제약조건

프로야구란 체력적인 조건과과 함께 선수들의 정신적인 상태가 크게 작용되는 부분이 있다. 또한 선수들의 부상이나 날씨, 감독과의 불화등 수치적으로 표현 할 수 없는 조건들 또한 많이 존재하기 때문에 효율적인 선발 라인업 구현에 있어서 구현상의 다음과 같은 제약 조건이 있다.

- 2006년 한국 프로야구 선수들의 기록만 사용한다.
- 기후상황이나 그라운드 조건등 정확히 알수 없는 조건에 대해서는 제외한다
- 모든 선수는 항상 최상의 몸 상태를 유지한다.(체력문제, 슬럼프 및 부상 제외)

- 그 외에 수치적으로 표현할 수 없는 정신적인 측면도 제외한다.
- 모든 평가는 공격적인 부분에 대해서만 한다.(수비적 요소 제외)

4.2 지식베이스 구현

팀명	이름	포지션	AVG	OBP	SLG	OPS	AB	HIT	2B	3B	HR
삼성	강명구	2B	0.053	0.25	0.053	0.303	19	1	0	0	0
삼성	강봉규	LF	0.234	0.306	0.391	0.696	128	30	6	1	4
삼성	김대익	RF	0.243	0.3	0.301	0.602	239	58	6	1	2
삼성	김재걸	SS	0.211	0.266	0.257	0.524	171	36	6	1	0
삼성	김종훈	LF	0.202	0.266	0.245	0.511	163	33	7	0	0
삼성	김창희	RF	0.221	0.297	0.345	0.642	307	68	12	1	8
삼성	김한수	3B	0.254	0.34	0.359	0.699	343	87	15	0	7
삼성	박정환	2B	0.165	0.2	0.176	0.376	91	15	1	0	0
삼성	박종호	2B	0.238	0.325	0.297	0.622	273	65	13	0	1
삼성	박진만	SS	0.283	0.38	0.432	0.812	382	108	22	1	11
삼성	박한이	CF	0.285	0.393	0.376	0.769	471	134	21	2	6
삼성	심정수	LF	0.141	0.229	0.188	0.417	85	12	1	0	1
삼성	양준혁	DH	0.303	0.445	0.477	0.922	413	125	31	1	13
삼성	이정식	C	0.229	0.297	0.361	0.658	83	19	5	0	2
삼성	조동찬	3B	0.259	0.323	0.376	0.698	370	96	13	0	10
삼성	조영훈	1B	0.283	0.342	0.389	0.731	180	51	11	1	2
삼성	진갑용	C	0.288	0.343	0.399	0.743	358	103	22	0	6

(그림 4) 지식베이스

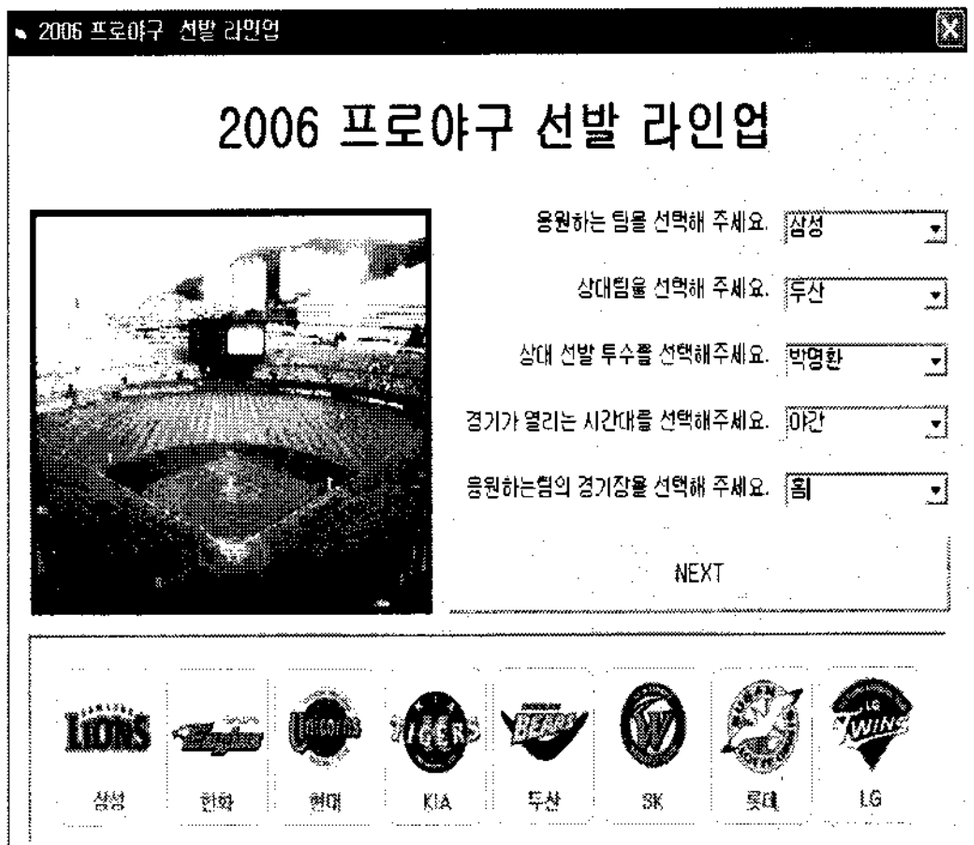
지식베이스에서는 정적인 정보(타율,안타,홈런,타점등)과 동적인 정보(상대팀 타율,홈/어웨이 타율,주간/야간타율 등)으로 구성되었으며 타자부분에서는 총 66개의 부분에 대한 158명의 선수에 대한 기록을 확보 구축하였으며, 투수 부분에서는 총 82개의 부분에 대한 128명의 선수에 대한 기록을 확보 구축 하였다[14-17].

지식베이스의 기록들 중에는 일반인들이 잘 사용하지는 않지만 방송이나 신문 기사 등에서 아주 많이 활용되는 여러 기록들이 포함 되어 더욱 더 세부적이고 심층적인 비교가 가능하도록 구현하였다.

4.3 사용자 인터페이스 구현

(그림 5)는 사용자 인터페이스 부분으로 각종 제약조건을 설정하는 부분이다. 이 부분에서 2차 포인트가 설정되는 요소를 정하는 부분으로 맨 처음 자신이 응원하는 팀을 고를 수 있다. 두 번째로는 상대팀을 고르는데 상대팀을 고르면 타자와 투수의 지식베이스부분에서 그 선수의 상대팀에 대한 성적이 1차 포인트로 추가된다. 이때 응원팀과 상대팀은 같은 팀이 될수 없다. 3번째로는 상대 선발을 고르게 되는데 이때는 상대팀 선발만이 선택요소로 나오게 된다.

상대팀 선발을 고르는 이유는 상대팀 선발의 투구유형에 따라서 즉 좌투수인지 우투수 인지에 따라 타자들의 좌투수.우투수 타율이 1차 포인트에 추가되기 때문이다. 그 다음 제약조건은 주간/야간을 고르게 되는데 그 이유 역시도 2차 포인트를 구하여 1차 포인트에 합산하기 위함이다. 마지막으로는 홈/어웨이를 고르게 된다. 사용자 인터페이스의 모든 제약조건은 2차 포인트와 관련된 것으로 이미 나와 있는데 1차 포인트에 합산되어 최종 포인트가 되는 것이다.

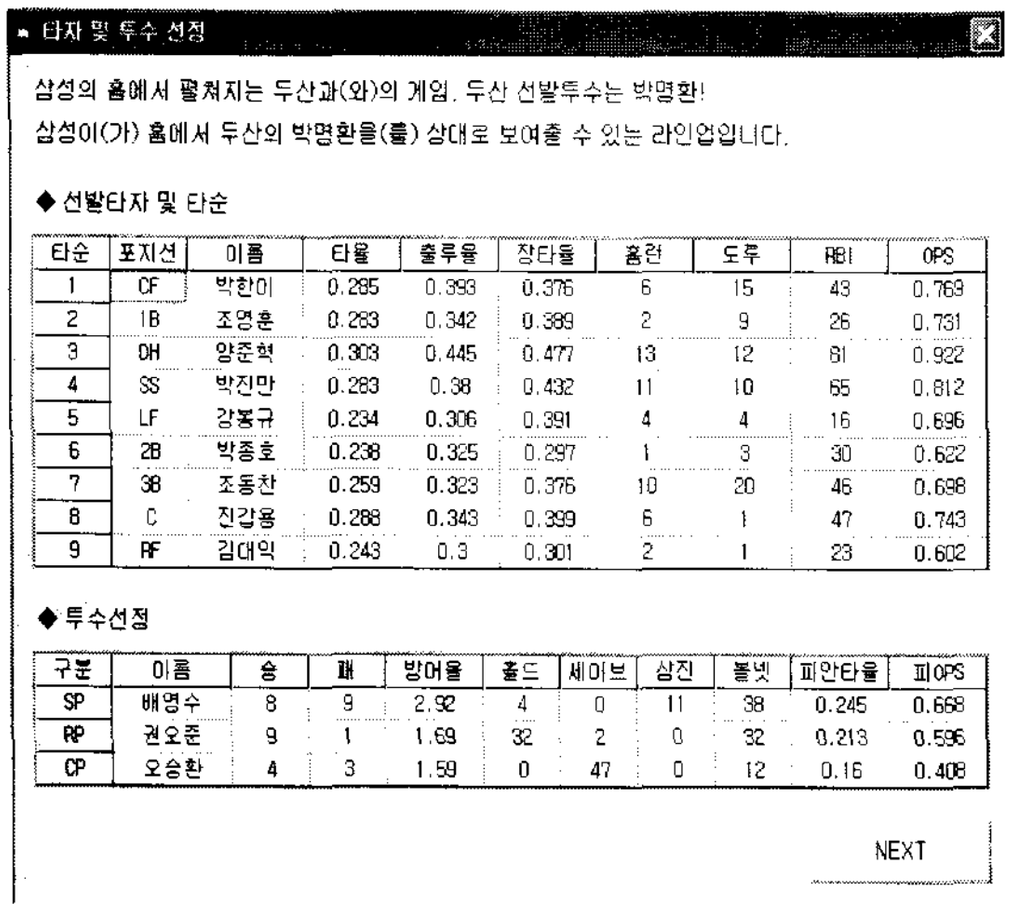


(그림 5) 사용자 인터페이스

4.4 효율적인 라인업 작성 및 배치도

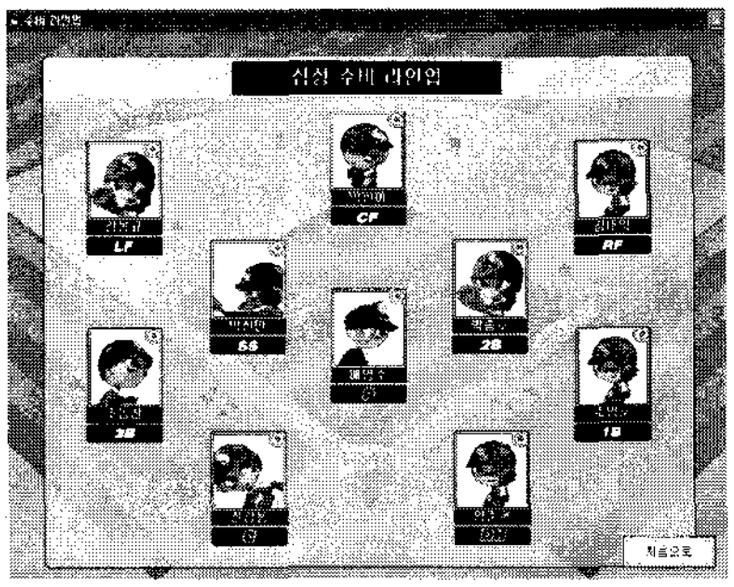
(그림 6)은 첫 번째 화면에서 입력된 제약조건을 토대로 만들어진 종합 포인트가 높은 순서대

로 타순이 매겨진 상대에 대한 효율적인 선발 라인업이다.



(그림 6) 효율적인 선발 라인업

이때 타순의 우선순위는 4.3.1.2.5.6.7.8.9 로 하였다. 보이는 요소가 모든 요소가 아닌 만큼 보이는 화면에 나타난 기록은 그리 중요하지 않다. 그 안에 수많은 기록들이 비교 포인트화 되어 나왔기 때문이다. 다만 보이는 요소는 우리들이 친숙하게 알고 있는 성적들로 보는 사람들로 하여금 친숙함을 느끼게 하기 위해 어려운 요소들은 비주얼 적인 화면에서 감추었다. 우선은 타순별로 가장 종합 포인트가 높은 선수가 위 순서로 배치되는데 이때 가장 중요한 점은 포지션이 중복 돼서는 안된다는 점이다. 투수 부분에서는 선발과 중간 마무리 3부분의 선수를 선발하였다. 타선에서는 실제 야구와 약간은 다를수도 있겠지만 투수 부분은 선발은 거의 각 팀의 에이스가 중간은 셋업맨이 마무리는 클로져가 나오게 된다. 사실 타순의 유동성은 많지만 투수쪽은 워낙 확실한 선수들이 있으므로 어떠한 방법을 써서 비교해도 큰 변화는 없다. 그 이유는 한 팀에 마무리가 2명 이상인 경우는 매우 드물기 때문이다. 선수 배치도 (그림 7)은 앞서 보여준 효율적인 라인업을 사용자에게 이해하기 쉽도록 보여주는 화면으로 포지션별로 배치되어 있다.



(그림 7) 선수 배치도

5. 비 교

본 장에서는 본 논문에서 개발한 효율적인 선발 라인업 시스템과 실제 팀의 라인업간에 비교를 한다. 비교 대상은 투수부분으로 비교한다. 투수만 비교 하게 된 이유는 야구는 흔히 투수놀음이라고 불리울 만큼 투수의 역할이 팀 승리에 미치는 영향이 그 만큼 크기 때문이다. 또한 투수 보직에 대한 분업화가 잘 되어 있어서 그에 따른 결과값이 명확하게 나오기 때문이다.

(표 6) 1차 포인트 비교결과

	승+QS 최다선발투수	1차 포인트 최다선발투수	홀드 최다 투수	1차 포인트 최다불펜투수	세이브 최다 투수	1차 포인트 최다세이브투수	정확률 (%)
두산	리오스	리오스	김승회	김승회	정재훈	정재훈	100%
기아	그레이 싱어	그레이 싱어	윤석민	한기주	윤석민	윤석민	66%
롯데	이상목	장원준	가득염	이왕기	나승현	나승현	33%
삼성	브라운	하리칼라	권오준	권오준	오승환	오승환	66%
SK	채병룡	채병룡	정우람	정우람	정대현	카브레라	66%
LG	정재복	이승호	카라 이어	카라 이어	우규민	우규민	66%
한화	류현진	류현진	권준헌	권준헌	구대성	구대성	100%
현대	켈러웨이	장원삼	이현승	신철인	박준수	박준수	33%

타자부분의 비교는 단기전에 대한 타순을 비교해야 하는데 그에 유동성이 너무 심하며 한선수의 부상으로 타순과 포지션에 도미노 현상이 일어나 그 명확성이 매우 떨어진다. 또한 타자들은 경기날 컨디션에 영향을 많이 받으므로 보통 로테이션 시스템을 쓰는 감독들도 있고, 또한 감독들의 성향에 따라 타순변동이 심하므로 예측에 대한 어려움이 매우 크다. 그 단적인 예로 NPB (일본프로야구)의 05시즌 우승팀인 지바 롯데의 감독인 바비 발렌타인 감독은 언론에서 흔히 바비매직이라고 불리우는 라인업을 작성하는데 그는 132경기에서 120개의 다른 라인업을 작성하므로서 언론의 주목을 받기도 하였다. 이렇듯 타자들의 타순에 대한 예측비교는 현실적으로 여러 변수들의 영향으로 그 어려움이 있다.

선발투수는 선발승에 QS(6이닝3실점이하,퀄리티스타트) 를 더하여 계산하였고 불펜투수는 홀드 마무리 투수는 세이브를 평가 항목으로 선정 실제 데이터와 1차 포인트 최다 득점자를 비교해 본 결과 (표 6)과 같다. 100%가 나온 두산과 한화는 확실한 1선발과 함께 불펜과 마무리 역시 안정된 팀이란 것을 보여주고 있다. 그러나 33%가 나온 롯데나 현대는 투수력이 매우 불안정 한것을 확인할수 있다. 그 예로 100%의 두산과 삼성은 06시즌 기준 팀 방어율이 1위와 4위이고, 33%인 현대와 롯데는 각각 5위와 7위로 하위권에 머무른 것을 확인할 수 있다. 또한 선발 포인트 최다 득점자 류현진은 06시즌 투수부분 골든글러브 수상자로 최고의 투수임을 입증했고, 다승왕, 방어율왕 까지 투수부분 트리플 크라운은 달성하였다. 또한 홀드왕 권오준 역시 불펜 포인트 최다 득점자이고 세이브왕 오승환도 마무리 투수부분 최다 득점자로 자리 잡고 있었다.

(표 7), (표 8), (표 9)는 실제로 제약조건에 근거하여 최고 성적을 올린 선수와 작성된 효율적인 선발 라인업 시스템으로 계산한 선수와 얼마나 일치하는지를 임의의 조건에서 실험한 결과이다.

(표 7) 두산 대 LG 비교결과(주간)

	팀	상대팀	홈/어웨이	주간/야간	선발	불펜	마무리	정확률(%)
실제 최고성적 올린선수	두산	LG	홈	주간	리오스	김상현	정재훈	-
효율적 라인업선수	두산	LG	홈	주간	리오스	김승회	정재훈	66%

(표 7)과 (표 8)은 66%의 정확률을 보이고 있다. 표 7에서 예상이 빗나간 불펜 부분은 김상현이 시즌에는 그렇게 많은 이닝과 좋은 결과를 보여주지 못했으나 LG전에서 잠시 나와 좋은 결과를 보여서 실제 최고 선수가 된 것 인데 반해 김승회는 시즌 중 많은 이닝과 좋은 결과를 보인투수로 1차 포인트의 차이가 너무커서 2차 포인트에서 미처 따라 잡지 못한 결과 인것 같다. 표 8의 결과도 비슷한 관점으로 해석할 수 있다.

(표 8) 두산 대 LG 비교결과(야간)

	팀	상대팀	홈/어웨이	주간/야간	선발	불펜	마무리	정확률(%)
실제 최고성적 올린선수	LG	두산	홈	야간	심수창	카라이어	우규민	-
효율적인 라인업선수	LG	두산	홈	야간	이승호	카라이어	우규민	66%

(표 9) 한화대 삼성 비교결과

	팀	상대팀	홈/어웨이	주간,야간	선발	불펜	마무리	정확률(%)
실제 최고성적 올린선수	한화	삼성	원정	야간	류현진	권준헌	구대성	-
효율적인 라인업선수	한화	삼성	원정	야간	류현진	권준헌	구대성	100%

(표 9)는 100%의 정확률을 보이는데 그 이유로는 한화의 분업화가 잘된 안정된 투수진을 이유로 들 수 있겠다. 투수 부문 골든 글러브의 류현진과 메이저리그 출신 구대성의 마무리는 그 누가 평가하여도 이의가 없는 한화의 대들보 역할을 하는 투수들로 실제 성적과 효율적인 라인업에서 계산된 성적이 일치함을 표 9를 통하여 알 수 있다.

5. 결론 및 향후 연구 방향

본 논문은 전향추론 기법을 이용하여 프로야구의 효율적인 선발 라인업 구현하였다. 기존의 변동이 없는 정적인 데이터 중심을 선발 라인업 선출하였다. 본 논문은 정적인 데이터 뿐만아니라 상대팀의 타율이나 주간/야간등의 변동이 가능한 데이터를 고려하여 보다 효율적인 선발 라인업 시스템을 설계및 구현하였다. 2006년 한국 프로야구 선수들의 기록을 기반으로 타자 부분에서는 66개 부분에 대한 158명의 선수와 투수 부분에서는 82개의 부분의 128명에 대한 선수의 기록을 기초로 하여 구축하였다.

본 논문에서 구축한 효율적인 라인업 시스템과 실제 라인업과 비교한 결과 팀 성적이 높으면 높을수록 라인업이 일치되었으며, 팀 성적이 낮을수록 본 논문에서 제안한 시스템과 실제 상황과는 다르게 나왔다. 또한, 비교 대상이 되는 선수들이 너무 적은 게임에 투입된 경우 예기치 못한 선수들이 나오는 것도 확인 할 수 있었다.

향후 연구로는 더 많은 제약 조건(심판, 날씨 등)과 폭 넓은 선수 데이터 사용(07, 08년도 기록 업데이트등) 그리고 팀 간의 예상 산출점수 공식화, 승패 예측, 선수와 선수간의 관계 적립등 보다 많은 정보를 세분화 시키면 보다 정확성 높은 선발 라인업 시스템을 구축할 수 있을 것이다.

참 고 문 헌

[1] A. L. Blum and P. Langlet. "Selection of relevant features and examples in machine learning", Aritificail Intelligence, Vol. 97, pp.245-271, 1997

[2] Warren S. Sarle, "Neural Networks and Statistical Models", Proceedings of the 19th Annual SAS Users Group International Conference, pp. 1538-1550 April, 1994

[3] Martin Riedmiller, "Advanced Supervised Learning in Multi-layer Perceptrons From Backpropagation to Adaptive Learning Algorithms", International Journal of Computer Standards and Interfaces, pp. 265--278, 1994

[4] Roberto Battiti, "Using mutual information for selecting features in supervise neural net learing", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 5, NO. 4, JULY 1994

[5] P. S. Bradley, O. L. Managasarian, and W. N. Street. "Feature selection via mathematical programming.", INFORMS Journal on Computing, Vol. 10, No 2, pp. 209-217, 1998

[6] 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, pp. 177-184, 1999

[7] 홍석미, 정경숙, 정태충, " 혼합형 기계 학습 모델을 이용한 프로야구 승패 예측 시스템", 한국정보과학회논문지, 제9권 16호, pp. 693-698, 2003

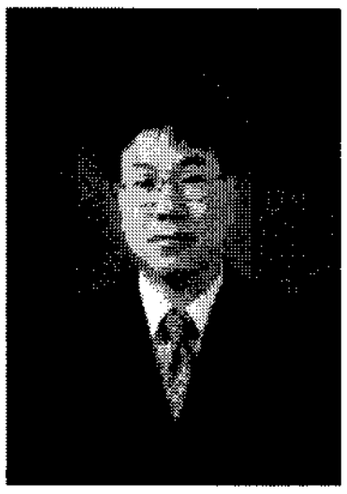
[8] "한국프로야구관전가이드북", 스포츠미디어, 2007

[9] 양기철, "인공지능의 이해", 생능출판사, 2003

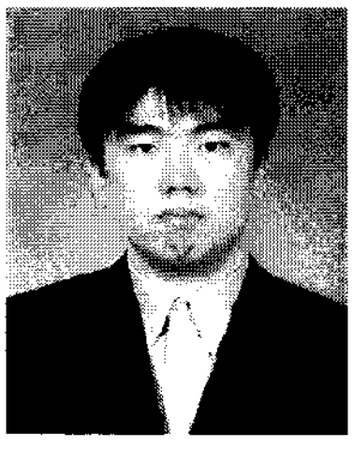
[10] George F. Luger, Artificial Intelligence:

Structures and Strategies for Complex Problem Solving,Addison-Wesley; 2005
[11] Peter Jackson Harlow, Introduction to Expert Systems, Addison Wesley, 1999
[12] 정일주, "전문가시스템",시그마프레스, 1996
[13] Bill James , "Win Shares", STATS Inc ,2002
[14] 박노준, "프로야구스카우팅리포트", 스포츠넷, 2007
[15] "한국프로야구 기록대백과", KBO, 2007
[16] "한국프로야구 레코드북", KBO, 2007
[17] www.koreabaseball.co.kr

◐ 저 자 소 개 ◑



박 홍 진(Hong-Jin Park)
1993년 원광대학교 컴퓨터공학과 졸업(학사)
1995년 중앙대학교 대학원 컴퓨터공학과 졸업(석사)
2001년 중앙대학교 대학원 컴퓨터공학과 졸업(박사)
2001년 ~ 현재 상지대학교 IT학부 부교수
관심분야 : 분산시스템, 시스템 프로그래밍. USN
E-mail : hjpark1@sangji.ac.kr



권 순 환(Soon-Hwan Kwon)
2007년 상지대학교 컴퓨터 공학과 졸업(학사)
2007년~현재 와이즈캣 기획운영팀 근무
관심분야 : 데이터베이스, 전문가시스템.
E-mail : ykssah@wisecat.co.kr