프로야구경기 분석을 통한 승 , 패 예측 모형

A Win , Loss Predicting Model by Analyzing Professional Baseball Game

저자
(Authors)

김차용
Cha Yong Kim

출처
(Source)

한국사회체육학회지 16, 2001.11, 807-819(13 pages)
Journal of Sport and Leisure Studies 16, 2001.11, 807-819(13 pages)

발행처
(Publisher)

한국사회체육학회
KOREAN SOCIETY OF SPORT AND LEISURE STUDIES

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE06242610

APA Style

김차용 (2001). 프로야구경기 분석을 통한 승 , 패 예측 모형. 한국사회체육학회지, 16, 807-819

한국사회체육학회지, 제16호, pp. 807~819
Journal of Sport and Leisure Studies
2001, Vol. 16, pp. 807~819

프로야구경기 분석을 통한 승·패 예측 모형

김 차 용(고려대학교)

●차 례●

Ⅰ. 서 론

우리나라 프로야구는 1982년에 시작되어 벌써 20년째를 맞이하고 있다. 1981년 창설 준비를 시작한 프로야구는 서울의 MBC 청룡, 부산의 롯데 자이언츠대구의 삼성라이온즈, 대전의 OB베어즈, 광주의 해태 타이거즈, 인천의 삼미 슈퍼스타즈 등 모두 6개 팀으로 출발하여, 1982년 3월27일 동대문구장에서 MBC 청룡과 삼성 라이온즈 개막전을 시작으로 한국프로야구는 그 역사적인 닻을 올렸다. 6개팀으로 이어 오던 프로야구는 1986년 빙그레의 합류로 7개팀으로 늘어났다. 그리고 1990년대에 들어서면서 우리나라 프로야구는 드디어 야구의 선진국 수준인 300백만 관중을 돌파하여 명실상부한 국민 스포츠로 자리 잡게 되었다. 그리고 1991년 쌍방울 레이더스가 프로무대에 진출함으로써 프로야구는 현재의 8개구단 진용을 갖추게 되었다(야구연감, 1999). 이렇듯 한국 프로스포츠의 시초가 된 야구는 그동안 프로축구나 농구 등의 다른 프로스포츠가 생겼지만, 여전히 많은 국민들로부터 여가스포츠로서 관심과 사랑을 받아온 인기종목으로 굳건한 자리를 지키고 있다.

이러한 현상은 야구자체가 다른 스포츠에 비해 보다 스릴과 흥미가 있는 운동 때문이기도 하지만 여러 많은 기록분석을 통한 다양한 작전에 의한 과학적 특성이 있기 때문이다.

흔히 야구를 기록의 경기, 확률의 경기라고 하는 이유도 바로 야구의 과학성을 대변하고 있는 말이기도 하다. 야구경기에서 승리하기 위해서는 각 선수들의 기술 및 정확한 판단력과 팀웍도 중요하지

※ 본 연구는 고려대학교 특별연구비에 의하여 수행되었음.

만, 매 순간 전개되는 상대팀의 의도를 간파한 치밀한 감독의 작전, 용병술 등의 전략은 경기의 승·패에 결정적으로 영향을 미치게 되는 것이다.

그러면 이러한 치밀한 감독의 전략은 어디에서 오는가? 요즘 텔레비전 중계 방송을 통해서 흔히 볼 수 있는 광경이 바로 덕아웃에서 컴퓨터를 들고 앉아 기록을 분석 제공하고 있는 야구자료분석가의 모습이다. 그 컴퓨터에는 자신의 팀뿐만 아니라 상대팀의 기록들이 데이터화되어있으며, 이들은 경기 승·패의 주요순간 마다 감독의 작전에 필요한 정보를 제공한다. 감독은 이 정보와 자신의 경험, 경기분위기 등을 종합하여 지시하게 된다.

야구만큼 세밀하고 다양한 기록이 가능한 스포츠는 없다. 이 많은 기록 가운데 어떤 정보에 근거하여 작전을 펼쳐야하는지, 어떤 요인들이 경기에서 어떻게 작용하며 이것들이 얼마나 중요한가? 이에 대한 해답은 어떤 기록이 팀의 승·패와 얼마나 관계가 깊은가에 달려 있는 것이다.

이미 야구 선진국인 미국에서는 미국야구조사협회(SABR)를 중심으로 야구통계학자들로부터 이런 연구가 상당히 활성화되어, 세이버메트릭스(Cyber Matrix)라고 하는 새로운 기록측정분석법이 제시되고 있다. 또한 국내에서도 최근 스포츠복권사업이 시작된 프로 축구나 농구에서 경기 승·패 예측을 위한 모델개발이나 시스템개발에 대한 연구가 발표되었고 또한 계속적으로 진행되고 있다. 즉 과거 경기 결과에 대해 선수들의 경기기록을 통계적 접근 방법을 통하여 모형화하고, 이를 이용하여 경기 결과에 대한 사후 확률을 분석함으로써 앞으로 있을 경기에 대해 승·패를 확률적으로 제공하고 있는 것이다.

이에 경기기록의 중요성과 그 효용가치가 큰 야구에 대한 기록분석을 위한 예측모형개발과 그 활용은 경기결과에 대한 각 선수들과 상대팀에 대한 기록을 탐색하고 이를 다양한 작전을 위한 전술에 이용하게 함으로써, 야구의 과학화를 이루는데 일익을 담당하게 될 것이다.

따라서 본 연구는 야구경기의 승·패에 영향을 미칠 수 있는 여러 가지 기록요인들을 분석하여, 이에 따른 타당한 예측모형을 개발함으로서 프로야구의 과학화 및 넓게는 스포츠의 과학성을 제고시키는데 일차적인 목적이 있으며, 이러한 연구결과를 토대로 스포츠현장과 스포츠산업 그리고 학교간의 협력을 통하여 산학간 연계성에도 기여하게 될 것이다.

Ⅱ. 연구방법

1. 연구대상

본 연구는 한국프로야구 경기의 경기 기록을 통한 구단별 특성분석을 통하여 각 팀의 강점과 약점을 분석하기 위하여 '98년 프로야구 패넌트레이스('98년 4월-10월)에 참가한 8개 프로 야구팀의 총 504경기(구단별 126경기)를 대상으로 하였는데, 분석을 위하여 구체적인 연구대상 팀의 주요 기록은 <표 1>과 같다.

2. 연구절차

외국의 스포츠기록관련 분석자료 및 국내외 야구관련논문에서 제시된 내용들을 먼저 탐색하고, 통계적 모형개발에 대한 사전검토를 시도한 후 다음과 같은 절차에 의해 분석하였다.

첫째, 국내프로야구기록 분석을 통한 모형설정 :

표 1. 1998년도 구단별 주요 기록

순 위	팀 명	승	패	무	승 률	안 타	홈 런	도 루	득 점	타 율	장타율	방어율	실 책	병살타	잔 루	희생타
1	현 대	81	45	0	0.643	1123	142	134	633	0.270	0.434	3.03	108	69	839	91
2	삼 성	66	58	2	0.532	1143	143	117	620	0.268	0.428	4.32	103	94	870	77
3	LG	63	61	1	0.504	1131	100	99	609	0.267	0.401	4.18	79	102	914	90
4	OB	61	62	3	0.496	1080	102	123	520	0.256	0.383	3.60	91	87	844	65
5	해 태	61	64	1	0.488	1067	94	58	489	0.258	0.376	3.91	98	125	857	92
6	쌍방울	58	66	2	0.468	1089	101	93	536	0.261	0.391	4.04	95	94	882	95
7	한 화	55	66	5	0.455	1042	123	94	503	0.250	0.392	4.26	96	94	819	99
8	롯 데	50	72	4	0.410	1070	86	76	536	0.255	0.381	4.61	102	76	879	89

실제 최근 프로야구경기 기록표에 나타난 매 경기마다의 기록을 분석하여, 야구승·패예측을 위한 통계적모형을 개발한다.

둘째, 설정된 모형에 대한 타당성검토 : 개발된 야구경기의 승·패예측을 위한 통계적모형의 타당성검토를 통하여 실제적인 실용가능성을 진단하고, 다른 모형과의 비교를 통해 본 모형의 적합성을 제시한다.

셋째, 모형에 따른 각 팀별 특성분석을 시도한다.

3. 측정도구

야구 경기는 기록의 경기라고 할 만큼 경기결과에 영향을 미치는 요인은 무한하다. 팀의 타율, 방어율, 장타율, 홈런수, 삼진수, 루타수, 득점타수, 희생타수, 도루수, 삼진 아웃수, 사사구수, 잔루수, 희생타수, 실책수, 출루율 등의 무수한 경기기록 뿐만 아니라 감독의 경기운영방식이나 위기상황 하에서의 경기대처능력, 선수들의 정신력, 구장의 특성, 관중의 응원 등의 요인에 의해 영향을 받게 된다. 팀의 경기결과는 이러한 제 요인들이 복합된 상호작용으로 나타나게 된다. 그러나 경기기록 외의 다른 요인들도 중요하지만 현실적으로 그 영향력을 분석하기 위하여 수량화하기에 매우 어렵다. 따라서 본 연구에서는 프로야구연감의 기록박스에 제시된 기록 요인만을 분석변인으로 하였다.

위에서 제시한 많은 기록 중 변수선택 과정을 거쳐 최종적으로 홈과 원정경기 여부, 경기 당 총루타수(1루타+2루타×2+3루타×3+홈런수×4), 장타율(총루타수/ 타수), 득점에 대한 집중력(타점/안타수), 희생타수, 병살타수, 잔루수, 도루수, 사사구수, 실책수를 선택하였다. 선택된 요인 중 홈과 원정경기 여부는 홈팀의 잇점이 경기 결과에 미치는 영향력을 알아보기 위함이며, 경기 당 총루타수 및 장타율은 팀의 타격실력의 정도를 측정하고자 하였다. 그리고 득점에 대한 집중력 및 희생타수, 병살타수, 잔루수 등은 찬스에서의 팀의 응집력을, 도루수는 팀의 기동력, 사사구수는 팀의 선구능력, 실책은 팀의 수비능력을 측정하기 위함이다.

따라서 본 연구에서의 프로야구 경기분석을 위한 자료는 '98년도 프로야구 패넌트레이스 경기기록을 이용하였다. 본 연구에 이용된 변수들은 다음과 같다.

표 2. 분석에 사용된 변수

변 수 명	변수설명
승패여부	목표변수 (종속변수)
홈팀여부, 희생타, 병살타, 잔루, 삼진, 사사구, 총루타, 도루수, 실책수, 장타율, 집중력(타점/안타수)	예측변수 (독립변수)

4. 분석 방법

프로야구연감의 기록박스에 제시된 매 경기의 각 팀별 기록요인을 코딩하여, SAS(ver 6.12)과 Answer Tree프로그램을 이용하여 분석하였는데, 구체적인 분석 방법은 다음과 같다.

첫째, 경기기록요인을 통한 승·패 예측모형을 설정하기 위하여, 로지스틱회귀(Logistic regression) 분석을 이용하였으며, 모형선택을 위하여 단계적 회귀분석법을 이용하였다.

둘째, 승·패에 미치는 영향력을 분석하기 위하여 비모수적 의사결정나무(Answer Tree)분석을 실시하였다. 분석에 이용된 알고리즘은 CHAID(Chi-squared Automatic Interaction Detection)를 이용하였다.

셋째, 팀별 예측모형을 통한 특성분석을 위하여 로지스틱회귀분석을 이용하였다.

Ⅲ. 연구결과

야구만큼 많은 기록을 가진 경기는 없다. 따라서 경기내용 하나하나는 경기의 승리 및 패배에 직결되므로 어느 하나 소홀히 할 수 없는 것이다. 그러나 이러한 많은 기록 요인 중에서도 팀의 승·패에 결정적으로 영향을 미치고, 어느 요인이 더 중요한가 하는 정보는 지도자나 선수 개인에 있어서도 매우 중요하다 하겠다. 따라서 본 연구에서는 이러한 야구 기록 요인 중에서 승·패를 결정짓는 요인이 무엇이며 또한 승·예측의 적합한 모형은 어떠하며 그리고 이들 요인들은 어떠한 조합으로 구성되어 경기 승·패에 영향을 주고 있는지 지를 규명하고자 한다.

1. 로지스틱회귀분석을 통한 승·패 예측모형

로지스틱회귀분석은 판별분석과 그 분석 목적이 유사하나, 판별분석의 경우 독립변수의 형태가 등

표 3. 경기내용 요인을 통한 승·패에 대한 로지스틱회귀분석(전체모형)

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	1	-2.5095	0.4211	35.5122	0.0001	.
홈과원정	1	-0.1635	0.1682	0.9448	0.3310	-0.045
희 생 타	1	0.3860	0.0874	19.5294	0.0001	0.225
병 살 타	1	-0.1006	0.0695	2.0969	0.1476	-0.068
잔 루	1	-0.1111	0.0533	4.3467	0.0371	-0.155
삼 진	1	-0.1157	0.0324	12.7406	0.0004	-0.165
사 사 구	1	0.1067	0.0604	3.1194	0.0774	0.126
총 루 타	1	0.1710	0.0280	37.2203	0.0001	0.520
도 루 수	1	0.1610	0.0878	3.3632	0.0667	0.087
실 책 수	1	-0.6558	0.1040	39.7447	0.0001	-0.301
장 타 율	1	0.0404	0.5775	0.0049	0.9442	0.004
집 중 율	1	3.5804	0.5466	42.9089	0.0001	0.597

간척도 이상의 연속형 변수이어야 하지만, 로지스틱회귀분석은 명목척도 또는 서열척도와 연속형변수 등의 변수도 포함되고, 변수들이 다변량정규분포를 한다는 가정하기 힘들 때 사용하는 기법이다(김충련, 1993).

<표 3>은 11개의 야구경기 기록요인이(홈팀여부, 희생타, 병살타, 잔루, 삼진, 사사구, 총루타, 도루수, 실책수, 장타율, 집중력)이 경기 승·패에 미치는 영향력을 분석하기 위한 로지스틱회귀분석 결과를 제시한 것이다.

표준회귀계수의 크기를 고려 해 볼 때, 경기 기록 중 집중력(β=0.597)요인이 가장 중요한 영향요인으로 나타났으며, 다음이 총루타(β=0.520), 실책수(β=-0.301), 희생타(β=0.225), 삼진(β=-0.165), 잔루(β=-0.155), 사사구(β=0.126), 도루수(β=0.087), 병살타(β=-0.068), 홈팀여부(β=-0.045), 장타율(β=0.004)순으로 영향을 미치는 것으로 나타났다. 그리고 이들 요인 중 집중력(p=0.0001), 총루타(p=0.0001), 실책수(p=-0.301), 희생타(p=0.225), 삼진(p=0.0004), 잔루(p=-0.0371)요인이 통계적으로 유의한 경기기록임을 보여주고 있다.

위에서 제시된 통계적회귀모형에 의한 분류결과, 승리의 경우 385경기(77.7%), 패배의 경우 390경기(78.8%)는 제대로 승리한 것으로 분류하였고, 110경기(21.2%)는 승리를 패배로, 105경기(22.20)는 패배를 승리로 잘못 분류하여 분류정확률이 78.3%정도로 나타났음을 제시하고 있다.

표 4. 로지스틱회귀 예측모형에 의한 분류결과(전체모형)

분류		예측 승	예측 패	전체
실제	승	385	110	495
		77.70	22.30	
	패	105	390	495
		22.20	78.80	
전체		490	500	990
		49.49	50.51	
적중률		78.3%(오류율=21.6%)		

예측을 위해서는 가능한 한 많은 입력변수를 포함시키는 것이 합리적인 전략이기는 하나 만약 부적절하거나 관련성이 없는 입력변수를 포함시키는 것은 모형의 설명(일반화)를 떨어뜨리는 역기능을 가져올 수 있고, 모형의 불안정의 원인이 될 수 있다. 따라서 사전에 충분한 탐색을 통해서 이들 변수를 제거하거나 변수선택방법 등을 통하여 이러한 단점을 극복할 수 있는데, 본 연구에서는 이러한 점을 고려하여 변수선택을 이용하여 보다 간결한 축소 모형을 분석하였다.

변수선택법에 의한 예측모형은 E(logit)=-3.0541+0.3820*희생타-0.1150*삼진+0.1443*총루타+0.1484*도루-0.6547*실책+4.2716*집중력으로 나타났는데, 표준화회귀계수를 볼 때 집중력(β=0.606), 총루타(β=0.439), 실책수(β=-0.301), 희생타(β=0.223), 삼진(β=-0.164), 도루수(β=0.080) 순으로 그 영향력이 크게 나타나 있음을 보여 주고 있다.

<표 5>에서 도출된 통계적 회귀모형에 의한 분류결과, 승리의 경우 386경기(78.5%), 패배의 경우 389경기(78.1%)는 제대로 승리한 것으로 분류하였고, 109경기(21.2%)는 승리를 패배로, 106경기(21.50%)는 패배를 승리로 잘못 분류하여 분류정확률이 78.3%정도로 나타나서, 전체모형과 거의 유사한 예측정확률을 보이고 있다.

2. 의사결정나무분석을 통한 승·패 예측모형

로지스틱회귀분석의 단순성과 분석의 편리함은 선형성(linearity)을 가정함으로서 생긴다. 이는 분석상의 장점으로 받아들여질 수 있으나, 변수들간

표 5. 경기내용 요인을 통한 승·패에 대한 단계적 로지스틱회귀분석(축소모형)

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	1	-3.0541	0.3432	79.1986	0.0001	.
회 생 타	1	0.3820	0.0835	20.9360	0.0001	0.223
삼 진	1	-0.1150	0.0317	13.1607	0.0001	-0.164
총 루 타	1	0.1443	0.0190	57.8114	0.0001	0.439
도 루 수	1	0.1484	0.0854	3.0175	0.0412	0.080
실 책 수	1	-0.6547	0.1031	40.3399	0.0001	-0.301
집 중 력	1	4.3716	0.4328	102.0371	0.0001	0.606

의 복잡한 비선형성을 가지는 경우는 예측의 유용성 측면에서 문제가 있게 된다. 그리고 일부변수들간의 교호작용을 모형에 포함시켜 분석이 필요한 경우, 유용한 교호작용을 탐색하는데는 실제적으로 매우 어려운 일이다.

표 6. 로지스틱회귀 예측모형에 의한 분류결과(축소모형)

분 류		예 측 승	예 측 패	전 체
실 제	승	386	109	495
		78.50	21.90	
	패	106	389	495
		21.50	78.10	
	전 체	492	498	990
		49.70	50.30	
적중률		78.3%(오류율=21.6%)		

이러한 점을 고려한다면 보다 해석상 용이하며, 그리고 변수들간 교호작용효과의 탐지 및 선형성에 덜 민감한 의사결정나무분석의 필요성이 제기된다. 의사결정나무는 의사결정규칙을 나무구조로 도표화하여 분류와 예측을 수행하는 분석방법이다. 이 방법은 분류 또는 예측의 과정이 나무구조를 통한 추론규칙에 의해서 표현되기 때문에, 판별분석이나 회귀분석 등에 비해서 분석자가 그 과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다

의사결정나무는 의사결정규칙을 직관적인 다이어그램과 도표 및 테이블화 하여 분류 및 예측하기 때문에, 다른 통계적 방법들에 비해서 분석자가 그 과정을 빠르고 쉽게 이해하고, 식별하게 하여 그 관계를 쉽게 설명할 수 있어서, 해석이 용이하다는 장점을 가지고 있다. 그리고 어떤 입력변수가 목표변수를 설명하기 위해서 더 중요한지를 쉽게 파악할 수도 있게 해 준다. 그리고 두 개 이상의 변수가 결합하여 목표변수에 주는 영향력의 효과를 쉽게 알 수 있게 해주며, 또한 의사결정나무는 자료의 분포형태에 대한 가정을 필요치 않는 비모수적 방법으로, 자료가 순서형 또는 연속형 변수라 하더라도 단지 순위(rank)만 분석의 대상이 되기 때문에 이상치(outlier)에 민감하지 않다는 장점이 있다(최종후 외, 1999).

아래 <그림 1>과 <표 7>은 야구경기기록이 승·패에 미치는 영향력을 알아보기 위하여, 의사결정나무 모형에서 CHAID 방법(Kass:1980)을 이용한 다중 나무구조(multi-tree structure)의 분류결과이다. 총 16개의 최종마디(leaf mode)로 이루어진 나무구조가 형성되었다. 본 나무구조분류 결과 중 주요내용을 살펴보면, 야구경기의 승·패를 결정짓는 제일 중요한 변수는 타격의 집중력이다. 타격집

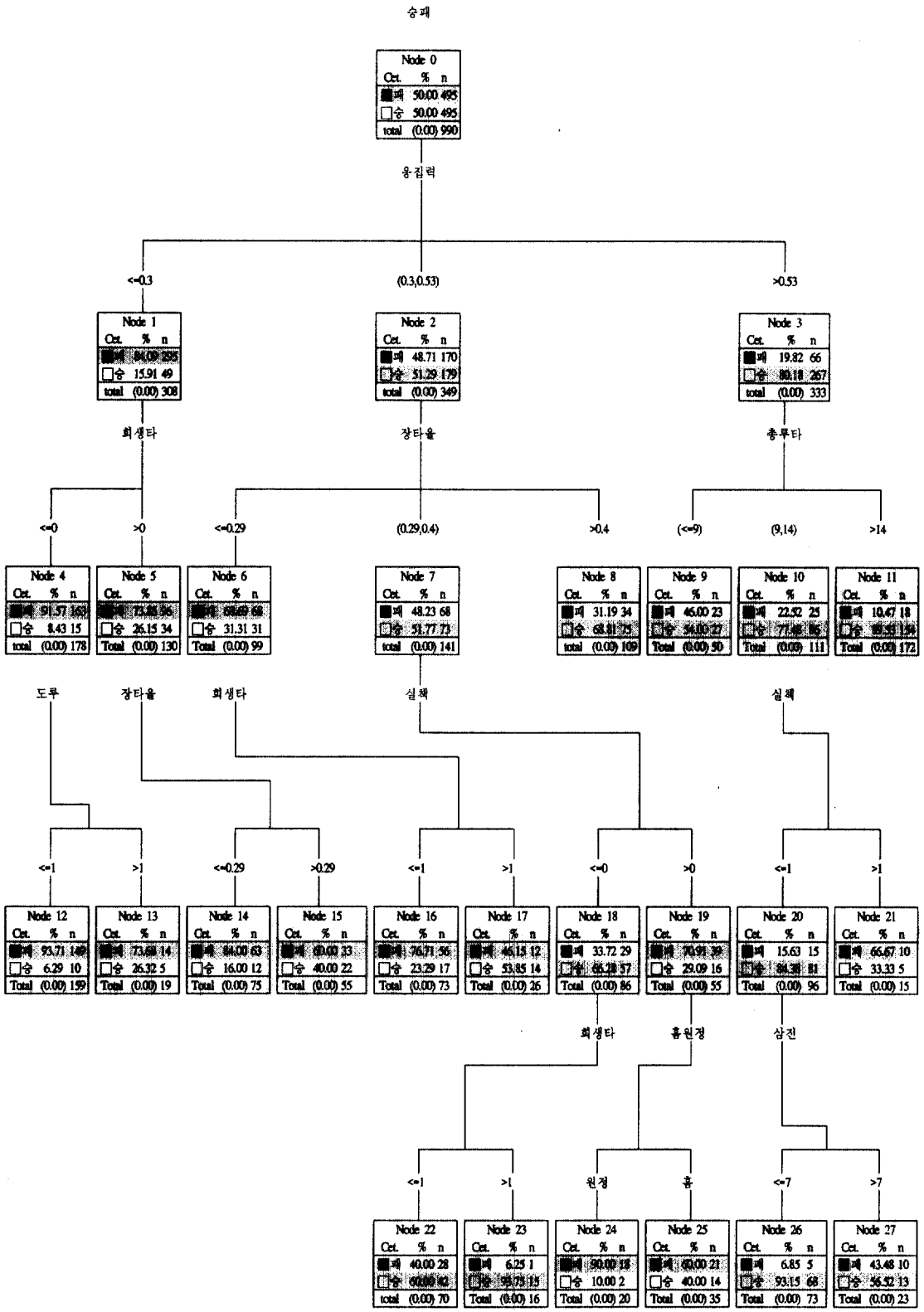

그림 1. 다중 나무 구조

표 7. 나무구조에 의한 분류결과

Nodes	Node(n)	Node(%)	Gain(n)	Gain(%)	Resp(%)	Index(%)
23	16	1.6	15	3.0	93.8	187.5
26	73	7.4	68	13.7	93.2	186.3
11	172	17.4	54	31.1	89.5	179.1
8	109	11.0	75	15.2	68.8	137.6
22	70	7.1	42	8.5	60.0	120.0
27	23	2.3	13	2.6	56.5	113.0
9	50	5.1	27	5.5	54.0	108.0
17	26	2.6	14	2.8	53.8	107.7
15	55	5.6	22	4.4	40.0	80.0
25	35	3.5	14	2.8	40.0	80.0
21	15	1.5	5	1.0	33.3	66.7
13	19	1.9	51.0	26.3	52.6	
16	73	7.4	17	3.4	23.3	46.6
14	75	7.6	12	2.4	16.0	32.0
24	20	2.0	2	0.4	10.0	20.0
12	159	16.1	10	2.0	6.3	12.6

Node(n): 마디에 속하는 개체의 수
Node(%): 마디에 속하는 개체의 수/전체개체의 수
Resp(n): 마디에 속하는 목표범주의 개체수
Resp(%):마디에 속하는 목표범주의 개체수/전체에서 목표 범주의 개체수
Gain(%):마디에 속하는 목표범주의 개체수/마디에 속하는 개체의 수
Index(%): 마디에서의 목표범주의 비율/전체에서의 목표 범주의 비율

중력이 작은 경우(≤0.3)에는 승리할 확률이 15.91%(49/306경기)가 되고, 큰 경우(>0.53)에는 80.18%(267/333경기)의 높은 승률을 나타내고 있다. 다음으로는 타격의 집중력의 크기에 따라서 영향을 주는 변수가 다소 다르다. 집중력이 작은 경우에는 희생타, 큰 경우에는 총루타에 큰 영향을 받으며, 집중력이 그 중간인 경우는 장타율에 좌우된다.

즉 희생타가 없는 경우는 승률이 8.43%(15/173경기), 희생타가 있는 경우는 26.15%(34/130경기)이며, 장타율이 0.29이하이면 31.31%(31/99경기), 0.29-0.4사이면 51.77%(73/141경기), 0.4 이상이면 68.61%(75/109경기)의 승률을 나타내며, 총루타 수가 9개 이하면 54%(27/50경기)이고, 9-14개 사이면 77.48%(86/111), 14보다 많으면 89.53%(154/172경기)의 높은 승률을 보이게 되는 것으로 나타났다. 그리고 그 다음 승·패 예측에 대한 영향력 있는 변수는 위에서 나타난 변수 외에 도루, 실책, 홈팀 여부 그리고 삼진의 과소에 따라 다르게 분류되고 있다.

위의 의사결정나무의 규칙을 간단히 요약하면 다음과 같다. 16마디 중 승률이 높은 마디는 23, 26, 11번째 마디로 거의 90% 이상의 승률을 보이고 있고, 반대로 승률이 낮은 마디는 12, 24, 14번째 마디로 15% 이하의 승률을 나타내고 있다.

의사결정나무의 결과 중 높은 승률과 낮은 승률을 나타낸 결과는 다음과 같이 요약할 수 있다<표 8>. 가장 승률이 높은 경우는 93.8% 의 승률을 나

표 8. 분류결과의 요약

Node	집중력(득점/안타수)	장타율(총루타/타격)	총루타	삼 진	실 책	도 루	희생타	홈원정	승리할 확률
23	0.3-0.53 이하	0.29-0.4 이하	-	-	0	-	2 이상	-	93.8
26	>0.53	-	9-14 이하	7개 이하	0	-	-	-	93.2
11	>0.53	-	15 이상	-	-	-	-	-	89.5
14	0.3 이하	0.29 이하	-	-	-	-	1 이상	-	16.0
24	0.3-0.53 이하	0.29-0.4 이하	-	-	1 이상	-	-	원정	10.0
12	0.3 이하	-	-	-	-	1 이하	0	-	6.3

타낸 23번 째 마디로서, 집중력이 0.3-0.53사이 이면서, 장타율이 0.29-0.4이며, 실책이 없으며, 희생타가 2개 이상인 경우다. 다음이 승률 93.2%의 26번째 마디로 나타났는데, 집중력이 0.53보다 높고, 총루타수가 9-14 이하이며, 삼진은 7개 이하면서 실책이 없는 경우이다. 세 번째는 집중력이 0.53보다 높고, 총루타수가 15 이상인 경우로 승률 89.5%를 보이고 있다.

가장 승률이 낮은 경우는 12번째 마디로서 6.3%의 승률을 나타내고 있는데, 집중력이 0.3이하로 낮으면서, 희생타가 없으며, 도루가 1개 이하일 때이며, 다음이 집중력 0.3-0.53사이이면서, 장타율이 0.29-0.4 이하이고, 실책이 한 개 이상의 원정경기인 경우 승률이 10%를 보여주고 있다. 그리고 집중력이 0.3이하이면서 장타율이 0.29이하로 낮은 반면에, 희생타는 한 개 이상인 경우로 승률 16%로 나타났다. 이와 같은 결과로 보아 야구경기에서의 승리는 팀의 집중력과 장타율 및 총루타수 등과 같은 타력과 같은 주요 요인과 선구능력의 척도인 삼진수, 수비력인 실책수, 기동력을 나타내는 도루, 팀 응집력의 희생타 그리고 홈팀여부 등과 같은 다양한 요인들과 복합적으로 관계되어 나타나는 현상으로 파악되어 진다하겠다.

<표 9>는 의사결정나무의 평가를 위해 분류결과에 대한 위험도표이다. 결과를 살펴보면 채택된 모형의 오분류율이 22.0%임을 알 수 있다. 따라서 본 의사결정나무에 의한 분류 정확 율이 78% 정도로 나타났으며, 승리에 대한 예측률이 패배에 대한 것보다 약간 낮게 나타나고 있음을 보여주고 있다.

표 9. 구축된 모형에 대한 위험도표(Risk chart)

분 류		실 제		전 체
		승	패	
예 측	승	364(73.5%)	87(19.3%)	451
	패	131(26.5%)	408(82.4%)	539
전 체		495(50.0%)	495(50.0%)	495
적중률		78%(오류율=22%)		

3. 팀별 예측모형을 통한 특성분석

앞의 분석에서는 야구경기에서 경기내용 요인을 통한 승·패 결정모형의 구축과 그 타당성을 검토하였다. 이러한 분석은 일반화된 모형으로 그 의의가 있으나, 각 개별 팀 별로 상대팀과의 비교를 통한 보다 실제적인 차원에서의 분석 자체도 그 필요성이 있다하겠다. 이를 위하여 팀별 예측모형을 통한 팀별 특성분석을 시도한 결과를 종합한 것이, 아래 <표 10>에 제시되어있다.

<표 10>은 구단별 로짓모형 결과에서 변수들의 표준화된 회귀계수를 비교한 것으로, 집중력 변수가 모두 선택되어 팀의 승·패에 많은 영향을 준

표 10. 구단별 로짓모형의 선택된 변수의 표준화된 회귀계수비교

요인 \ 팀		현 대	삼 성	LG	OB	해 태	쌍방울	한 화	롯 데
홈팀여부				-0.708					
실 책	자기팀		-0.725					-2.609	-2.352
	상대팀	1.70	0.668		1.409				
사사구	자기팀								
	상대팀	-1.337						2.380	-0.850
삼 진	자기팀				1.474			-3.301	
	상대팀								
도 루	자기팀								
	상대팀								
잔 루	자기팀							-2.922	
	상대팀								
병살타	자기팀							-2.414	
	상대팀								
희생타	자기팀								
	상대팀								
응집력	자기팀	4.200	1.135	1.541	3.588	3.275	3.064	5.401	1.944
	상대팀	-3.319	-2.484	-1.453	-4.778	-2.382	-1.676	-6.732	-1.641
장타율	자기팀			1.447		1.111	2.309	6.086	
	상대팀		-1.191		-1.604	-1.811			-2.376
총루타	자기팀	2.603	1.998		3.599				1.322
	상대팀	-4.230		-1.459			-1.785		

것을 알 수 있다. 그리고 자기팀의 집중력이 모두 양수이고, 상대팀의 집중력이 모두 음수이므로 자신의 팀은 집중력이 강하고 상대팀은 약한 경우에 승리 할 수 있음을 보여 주고 있는데, 이는 야구경기의 특성상 당연한 결과로 보여진다.

'98시즌 패넌트시리즈 각 상하위팀을 중심으로 팀별 특성을 비교하면, 현대팀은 상대팀 실책이 많고, 사사구는 적게 내주고, 총루타수가 많은 경기에서 승리를 많이 한 것으로 예측되어 졌는데, 실제 패넌트레이스 결과와 종합해 비교하여 보면, 1위팀인 현대는 전체 안타수에 비하여 득점력이 가장 높아 집중력이 높은 팀이었으며, 발빠른 선수가 많아(도루율 1위) 상대팀의 실책을 유도한 경기가 많았고, 선발 5인이 모두 10승 이상을 올린 막강한 마운드를 보유하여 승리의 원동력이 되었다.

삼성은 팀의 실책은 작은 반면 상대팀의 실책은 많으며, 총루타수가 많고 상대방의 장타율이 작은 경우에 승리를 많이 한 것으로 예측되어졌다. 실제 이 팀은 전체 안타수 및 홈런수가 가장 많았고, 전체 구단 중 짜임새 있는 내·외야진으로 가장 탄탄한 수비력을 바탕으로 한 팀이었다.

LG는 팀의 장타율이 높고 상대팀은 총루타수가 작으며, 특히 홈 경기인 경우 승리를 많이 한 것으로 예측되어졌는데, 실제 홈경기의 승률이 원정경기에 비하여 상대적으로 높게 나타난 팀이었다.

한화는 팀의 실책, 삼진, 잔루가 작은 반면, 장타

율은 높고, 상대팀의 병살타가 적은 경우에 승리할 가능성이 높은 것으로 예측되었는데, 실제 경기 결과를 보면 잔루수는 가장 작았고 병살타도 그렇게 많지 않았으며, 높은 장타율을 보유한 팀 이어서 모형 예측과 상이함을 보여 주었다. 그러나 최하위 팀 롯데는 팀의 실책이 적고, 상대팀의 사사구 및 장타율을 적으며, 팀의 총루타수가 많은 경우 승리를 한 것으로 예측되어졌는데, 실제는 많은 실책과 낮은 장타율 및 방어율을 기록하여 승률이 낮은 주요 원인으로 파악되어, 모형예측과 일치하는 것으로 나타났다.

Ⅳ. 결 론

본 연구는 야구경기의 기록을 팀 전략에 활용하여 보다 수준 높은 경기와 야구의 과학성을 제고시키기 위한 방편으로서 프로 야구경기의 승·패요인 분석을 위한 모형을 추정하였다.

분석대상자료는 '98시즌 프로야구 경기기록자료를 이용하였으며, 경기 승·패예측 모형분석을 위한 통계적 방법은 로지스틱회귀분석 및 의사결정나무분석을 이용하였다. 주요 결과는 다음과 같다.

첫째, 총 11개의 야구기록요인이 경기 승·패에 미치는 영향력을 분석하기 위한 로지스틱회귀분석 결과 집중력, 총루타, 실책수, 회생타, 삼진, 잔루, 사사구, 도루수, 병살타, 홈팀여부, 장타율순으로 나타났다. 이들 요인 중 집중력(p=0.0001), 총루타(p=0.0001), 실책수(p=-0.301), 회생타(p=0.225), 삼진(p=0.0004), 잔루(p=-0.0371)요인이 유의한 경기기록으로 나타났다.

둘째, 변수선택법에 의한 예측모형은 E(logit)=-3.0541+0.3820*회생타-0.1150*삼진+0.1443*총루타+0.1484*도루-0.6547*실책+4.2716*집중력으로 나타났는데, 집중력, 총루타, 실책수, 삼진, 회생타 순으로 경기 승·패에 대한 영향력이 크게 나타났다.

셋째, 의사결정나무분석을 이용한 승·패 예측결과, 승률이 높은 경우는 93.8% 의 승률을 나타낸 경우로서, 타격집중력이 0.3-0.53사이 이면서, 장타율이 0.29-0.4사이 이며, 실책이 없으며, 회생타가 2개 이상인 경기에서였다. 다음으로 93.2%의 승률을 나타낸 경우인데, 집중력이 0.53보다 높고, 총루타수가 9-14이하이며, 삼진은 7개 이하면서 실책이 없는 경우이다.

가장 승률이 낮은 경우는 6.3%로써 집중력이 0.3 이하로 낮으면서, 회생타가 없으며, 도루가 1개 이하일 때이며, 다음이 집중력 0.3-0.53사이이면서, 장타율이 0.29-0.4이하이고, 실책이 한 개 이상의 원정경기인 경우 승률이 10%로 나타났다.

● 참 고 문 헌 ●

김충련(1993). SAS라는 통계상자. 데이타리서치.

박상순·원태현 공저(1995). SAS를 이용한 통계자료분석. 중앙대학교 출판부.

박성현(1994). 회귀분석. 민영사.

박제영 외(2000). '99-2000 시즌 한국프로농구 경기의 승·패 요인. 한국사회체육학회 지, 제14호, 327-338.

윤형기(1998). '97-98 FILA배 프로농구 경기의 승·패에 영향을 미치는 기술 요소 분석. 경희대 체육과학논총 제11호, 143-164.

정동군 외(2000). 야구경기의 구질변화와 타격결과의 관계. 한국사회체육학회지, 제13호, 495-504.

최종후 외(1999). 데이터마이닝 의사결정나무분석. 자유아카데미.

프로야구연감(1999). 한국야구위원회.

허명회(1990). SAS회귀분석. 자유아카데미.

한국통계학회 대학생논문 경연대회 논문집(1996). 98-109. 통

계청

http://http://www.seoul.co.kr/baseball(스포츠서울프로야구정보)

http://www.koreabaseball.com(한국야구정보시스템)

http://www.kss.or.kr(한국통계학회)

Albert, J.(1994). Exploring baseball hitting data: what about those breakdown statistics?. Journal of the American Statistical Association , 89, 1066-1074.

Albright, S. C.(1993). A statistical analysis of hitting streaks in baseball. Journal of the American Statistical Association , 88, 1175-1183.

Barry, D., and Hartigan, J. A.(1993). Choice Models for Predicting Divisional Winners in Major League Baseball. Journal of the American Statistical Association , 88, 766-774.

Breiman, L., Friedman, J. H., Olsen, R. A., and Stone, C. J.(1984). Classification and Regression Trees. California.

Kass, G.(1980), An exploratory technique for investigating large quantities of categorical data, Applied Statistics, 29:2, 119-129.

Michael, J. A., Berry, and Gordon, S. L.(1997). Data Mining Techniques For Marketing, Sales, and Customer Support. John Wiley & Sons SPSS Marketing Department(1998), Answer Tree 1.0 User's Guide.

Sten W. Bergman and John C. Gittins(1985). Statistical Methods for Pharmaceutical Research Planning, Marcel Dekker, New York and Basel

ABSTRACT

A Win · Loss Predicting Model by Analyzing Professional Baseball Game

Kim, Cha-Yong

The purpose of this study is to estimate a model of professional baseball which is for predicting whether win or lose the game. These were made by collecting baseball records and would were used for team strategy. Baseball teams, therefore, would have more scientific strategies and have better games. The data used for this study were professional baseball game records for 98 year. The statistical methods used for predicting a game were logistic regression analysis and decision tree analysis and major results are as follows.

First, in order to analyze effects on total 11 records factors to the game which win or lost, logistic regression analysis was performed. Concentration, total base hit, the number of error, sacrifice hit, struck out, left on base, base on balls, stolen base, double play out, playing in home field, and long hit affected on the game in order. Among these factors, concentration(p=.0001), total hit(p=.0001), the number of error(p=-.301), sacrifice hit(p=.225), struck out(p=.0004) and left on base(p=-.307) significantly affected the game.

Second, the predicted model through selecting variable method was E(logit)=−3.0541+0.3820×sacrifice hit−0.1150×struck out+0.1443×total hit+0.1484×stolen base−0.6547×error+4.2716×concentration. Concentration, total hit, the number of error, struck out, sacrifice hit had their effects in order.

Third, the analysis of decision making tree showed that .3-.53 of concentration, .29-.4 of long hit, no error and over 2 sacrifice hit in the case of 93.8% win. Next, over .53 concentration, below 9-14 of total hit, below 7 of struck out and no error were come out in 93.2% win. The lowest winning percentage was 6.3% which had below .3 of concentration, no sacrifice hit, and below 1 of stolen base. Next .3-.53 of concentration, below .29-.4 of long hit, over 1 errors and having visit game were resulted in 10% winning percentage.

접 수 일 : 2001. 10. 5
게재확정일 : 2001. 10. 19