Journal of the Korean Institute of Industrial Engineers
Vol. 40, No. 1, pp. 8-17, February 2014.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2014.40.1.008

<Original Research Paper>

데이터마이닝을 활용한 한국프로야구 승패예측모형 수립에 관한 연구

오윤학 · 김 한 · 윤재섭 · 이종석†

성균관대학교 시스템경영공학과

Using Data Mining Techniques to Predict Win-Loss in Korean Professional Baseball Games

Younhak Oh · Han Kim · Jaesub Yun · Jong-Seok Lee

Department of Systems Management Engineering, Sungkyunkwan University

In this research, we employed various data mining techniques to build predictive models for win-loss prediction in Korean professional baseball games. The historical data containing information about players and teams was obtained from the official materials that are provided by the KBO website. Using the collected raw data, we additionally prepared two more types of dataset, which are in ratio and binary format respectively. Dividing away-team's records by the records of the corresponding home-team generated the ratio dataset, while the binary dataset was obtained by comparing the record values. We applied seven classification techniques to three (raw, ratio, and binary) datasets. The employed data mining techniques are decision tree, random forest, logistic regression, neural network, support vector machine, linear discriminant analysis, and quadratic discriminant analysis. Among 21(= 3 datasets×7 techniques) prediction scenarios, the most accurate model was obtained from the random forest technique based on the binary dataset, which prediction accuracy was 84.14%. It was also observed that using the ratio and the binary dataset helped to build better prediction models than using the raw data. From the capability of variable selection in decision tree, random forest, and stepwise logistic regression, we found that annual salary, earned run, strikeout, pitcher's winning percentage, and four balls are important winning factors of a game. This research is distinct from existing studies in that we used three different types of data and various data mining techniques for win-loss prediction in Korean professional baseball games.

Keywords: Professional Baseball, Win-Loss Prediction, Winning Factors, Data Mining, Classification Techniques

1. 서 론

현대 사회는 삶의 질에 대한 인식의 변화로 스포츠에 대한 관심이 증가하고 있다. 이러한 시대적 변화 속에 스포츠 마케팅과 국제 대회의 활성화로 스포츠는 전 국가적인 이벤트로 자리 잡고 있다. 특히 한국 야구는 최근 국제대회에서의 괄목할 만한 성과와 국내 선수들의 해외 진출로 많은 관심을 받고 있으며, 프로야구 관중수가 2006년부터 2012년까지 꾸준히 증가하고 있다(Korean Baseball Organization, 2013). 이러한 시민들의 많은 관심과 더불어 특히 야구에서는 '데이터 야구'라는 말이 보편화될 만큼 많은 자료들이 제공되고 있다. 국내에서는 sports2i(http://www.sports2i.com/)라는 업체에서 야구경기 데이터를 독점하여 제공하고 있으며, MLB(Major League Baseball, 미국프로야구)에서는 SABR(Society of American Baseball Re-

이 논문은 2012년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2012R1A1A1012153).
† 연락저자 : 이종석 교수, 440-746 경기도 수원시 장안구 서부로 2066 성균관대학교 자연과학캠퍼스 시스템경영공학과, Tel : 031-290-7608, Fax : 031-290-7610, E-mail : jongseok@skku.edu
2013년 11월 27일 접수; 2013년 12월 29일 수정본 접수; 2014년 1월 9일 게재 확정.

search, http://sabr.org/)라는 단체가 여러 가지 공식들을 제시하여 선수들의 기술에 대한 평가 척도를 제시하고 있다. 열성적인 야구팬들은 이러한 데이터들을 수집해 각 구단 및 선수들을 종합적으로 분석하고 관련 커뮤니티에서 의견을 교환한다(Chea *et al*., 2010). 이러한 현상과 더불어 일반인들도 스포츠토토 등과 같은 투표권을 이용하여 스포츠 경기결과를 예측하는 스포츠문화가 자리 잡고 있다.

일반적으로 스포츠 경기의 예측결과가 주는 의미는 첫째로, 각 팀의 감독에게 경기력 평가를 바탕으로 승리를 위한 작전과 전략을 수립할 수 있도록 도움을 주는 것이다(Kim *et al*., 2007). 둘째로, 선수에게는 기술에 대한 평가 자료를 제시함으로써 개인역량을 향상시켜 팀 승리에 기여할 수 있게 한다(Kim *et al*., 2007). 예를 들어, MLB 2002년 시즌 페이롤(pay roll) 순위 28위에 그쳤던 오클랜드 에슬레틱스는 출루율과 장타율이 승리에 밀접한 관여를 한다는 분석을 바탕으로 팀을 재구성해 아메리칸리그 최다 연승인 20연승을 기록하며 플레이오프에 진출하였다. 또한, 오클랜드 어슬레틱스는 첫 투구에 배트를 휘두를 때는 타율이 1할 4푼밖에 되지 않는다는 정보를 제시하여 타자들의 경기력을 향상시켰다(Lewis, 2004). 셋째로, 스포츠 토토와 프로토 등과 같은 체육진흥투표권의 이용자에게 실제 데이터를 기반으로 한 예측 모형을 통해 보다 과학적인 투자정보를 제공하는 것 등이다(Koo *et al*., 2009; Odachowski and Grekow, 2013).

단체 스포츠경기에서 경기력 평가는 다차원적인 정량적 요인과 함께, 정량화하기 어려운 정성적 요인들이 고려되어야 한다. 따라서 팀의 전략방향, 상대팀의 작전, 그리고 경기의 흐름과 같은 유동적인 상황에 순간적으로 대처할 수 있는 능력 등이 고려되어야 하므로 팀의 경기력을 경기의 기록만을 가지고 분석, 평가하기에는 어려움이 있다. 하지만 야구경기의 경우, 수비팀과 공격팀이 명확히 구분되어 있고 선수들의 위치와 역할이 정해져 있어 다른 스포츠에 비해 비교적 독립적이다. 또한, 경기의 진행 자체가 단일적 상황이 반복되는 특징을 가지고 있어 객관적인 분석이 용이하다(Chea *et al*., 2010). 예를 들어, 국내 프로야구 타자들의 성과를 DEA(Data Envelopment Analysis)와 OERA(Offensive Earned-Run Average)를 이용하여 정량적으로 측정하고 분석한 연구가 수행된 바 있다(Lee and Yang, 2004).

이렇게 축적된 방대한 양의 자료 속에 내재된 의미 있는 상관관계, 패턴, 경향 등 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정인 데이터마이닝(Seidman, 2002)을 통해 스포츠분야에서 우승자, 포스트시즌 진출 팀 또는 승・패 예측과 관련한 다양한 연구가 수행되어 왔다(Min and Hyun, 2009; Chea *et al*., 2010; Kim *et al*., 2007; Hong *et al*., 2010; Kim and Park, 2011; Sung and Chang, 2007; Oh and Lee, 2003). Koo *et al*.(2009)은 로지스틱 회귀분석과 인공신경망 모형을 활용하여 국내 남자 프로농구 2007～2008년도 시즌 270경기를 승리집단, 패배집단으로 나눈 540개의 경기결과 자료 사용한 모델과 270개의 경기결과를 그대로 입력 자료로 사용한 모델 등 두 가지 모델로 나누어 경기결과를 예측하였다. 이 밖에도 Kim *et al*.(2007)은 신경망분석을 통해 2006년 독일월드컵 본선 64개 경기를 대상으로 예선 1라운드 16경기 결과를 이용하여 나머지 48경기 결과 예측한 모형과, 해당 경기 직전 경기까지의 결과 데이터의 평균을 바탕으로 결과를 예측한 모델 등 두 가지 모델에 대해 승패와 관련한 기록요인을 점수화하여 변수로 사용해 경기결과를 예측하였다. 특히, 야구 분야에 관해서는 Kim(2001)은 로지스틱 회귀분석과 의사결정나무모형(CHAID 기법) 등을 사용하여 프로야구 승・패 예측모형에 관한 연구를 수행하였고, Hong *et al*.(2003)은 ID3와 통계적 방법, 역전파 알고리즘을 사용하여 승패 예측시스템을 구축하는 연구를 수행하였다. 해외의 연구문헌을 살펴보면, Miljkovic *et al*.(2010)은 NBA 2009～2010 시즌의 778경기결과를 바탕으로 농구경기의 승・패를 예측하기 위하여 Naive Bayes와 선형회귀분석을 도입하는 시도를 하였으나, 정확도는 67% 정도로 그리 높지 않다. Jensen *et al*.(2009)은 MLB 타자들의 홈런 가능성을 추정하기 위하여, 선수들의 과거성적, 나이, 포지션, 홈구장 등의 정보를 바탕으로 Bayesian hierarchical model의 수립을 제안하였다. 홈런 가능성뿐만 아니라, MLB 타자들의 일반적인 배팅능력을 정량적으로 평가하기 위해 중첩 Dirichlet 분포를 사용한 연구도 최근 발표된 바 있다(Null, 2009).

이상과 같이 스포츠 경기의 승・패를 예측하는 국내외 연구들은 대부분 팀 타율, 팀 평균자책점 등의 팀 자체의 데이터를 사용하였으나 해당 경기에 실제 출전하는 선수의 기록을 반영하지 않았다는 점이 단점으로 지적된다. 또한, 기존 연구들은 통산 기록을 이용해 승・패를 예측한다는 점에서 최근의 경기력을 설명하기에 미흡함이 존재한다. 따라서 본 연구는 최근 경기력을 반영하기 위해 전 시즌부터 '직전' 경기까지의 누적 데이터를 기반으로 선발라인업 10명(선발투수 1명, 타자 9명)의 데이터를 종합하는 시도를 하였으며, 과거 연구들이 승・패 예측을 위해 일부의 예측기법만을 사용한 반면, 본 연구에서는 보다 다양한 데이터마이닝 기법을 통해 프로야구 경기의 승・패를 예측함으로써 보다 높은 정확도의 모델을 수립하기 위한 시도를 하였다.

2. 연구 방법

2.1 변인 설정

본 연구에 사용된 변인은 크게 팀, 타자, 투수로 나뉘며 세부적인 변인 및 생성과정은 <Table 1>에 제시하였다. 각 데이터는 선수 및 팀의 통산 데이터를 수집하였으나, 팀의 승패를 예측하는데 있어서 통산데이터보다 가장 최근에 참여한 한 시즌의 데이터가 선수 및 팀에 영향을 준다고 판단하여 이전의 한 시즌부터 2013년 6월까지의 자료를 사용하였다. 각 선수들의

데이터는 KBO(한국야구위원회) 홈페이지에서 제공한 선수자료들과 경기일정을 바탕으로 수집하였다. 몇 가지 변인을 간략히 설명하자면 평균연봉은 2013시즌 개막 시 팀의 외국인선수와 신인선수를 제외한 선수들의 평균연봉을 나타낸 변인이며, 휴식의 경우 2013년 시즌 9구단 체제로 인한 휴식 팀과 경기취소에 의한 영향을 반영시키기 위한 변인이고, 팀의 연승연패는 팀의 분위기를 반영하기 위한 척도로 팀이 승리 시 1부터 1씩 증가하고 패배 시 -1부터 -1씩 감소하도록 하였다. 그리고 상대승률 및 상대팀에 대한 평균자책점은 팀 간 영향력을 측정하기 위한 변인으로 사용하였다. 타자 관련 변인은 경기에 출전하는 9명의 타자들의 수치를 합으로 나타내었으며, 투수 관련 변인은 선발투수의 팀 기여도를 나타낸 지표이다. 다만 자료 수집을 하는 과정에서 각 경기당 선수의 희비를 구하는데 어려움이 있어 출루율 계산과정을 수정하였다.

2.2 데이터 생성 과정 및 변환

본 연구의 목적은 선수 및 팀의 이전 시즌부터 직전경기까지의 기존의 누적 데이터를 바탕으로 다음 경기의 승·패를 예측하는 것이기 때문에 KBO에서 수집한 자료를 바탕으로 <Table 1>에 제시한 방법에 따라 변인을 생성하여 <Table 2>에 의거하여 자료를 생성하였다. 이때, 신인선수와 신생 팀의 자료가 축적되는 최소기간인 처음 3일치의 경기와 무승부가 발생한 7경기를 데이터에서 제외하였고, 그 결과 2013시즌 4월 3일부터 6월 말까지 74일간 각 팀당 경기 수 및 총 경기 수는 <Table 3>에 정리된 바와 같다.

Table 2. Process for data gathering

경기일	데이터	비고
1	직전 시즌	데이터 축적, 승패예측을 하지 않음
2	직전 시즌~1일차 경기	
3	직전 시즌~2일차 경기	
4	직전 시즌~3일차 경기	데이터 축적 및 승패예측을 위해 사용함
	⋮	
t	직전 시즌~(t-1)일차 경기	
	⋮	
77	직전 시즌~76일차 경기	

Table 3. Number of games per team

팀	기아	LG	NC	SK	넥센	두산	롯데	삼성	한화	계
경기수	59	63	62	61	61	61	59	59	61	546

예측 모델 수립을 위한 데이터 생성의 예로서, 53일차인 6월

Table 1. Description of variables for team, batters, and pitcher

	변인	변인생성	비고
팀	홈어웨이	홈 : 1, 어웨이 : 0	
	휴식	전 경기 휴식팀 : 1, 비휴식 팀 : 0	
	연승연패	직전 경기까지의 연승/연패를 수치화	연승 : 양수, 연패 : 음수
	평균연봉	각 구단 별 팀 평균연봉	단위: 억원
	상대승률	직전경기까지 상대 팀 누적 승률	
	상대팀에 대한 평균자책점	(상대팀에 대한 실점×9)/이닝	
타자	출루율	(안타+사구+볼넷)/(타수+사구+볼넷)	희비 제외
	장타율	(단타+이루타×2+삼루타×3+홈런×4)/타수	
	타율	안타/타수	
	사구/타수	(사구+볼넷)/타수	
	타점/타수	타점/타수	
	삼진/타수	삼진/타수	
투수	피안타율	피안타/(타자 수-볼넷-사구)	
	선발승률	승수/경기	
	평균소화이닝	이닝/경기	
	평균자책점	(자책점×9)/이닝	
	볼넷/이닝	(볼넷+사구)/이닝	
	삼진/이닝	삼진/이닝	
	피홈런/이닝	피홈런/이닝	
	타자/이닝	타자 수/이닝	

Table 4. Batting lineup of the team KIA on June 2, 2013 and players' records

라인업	이름	출루율	장타율	타율	사구/타수	타점/타수	삼진/타수
1번타자	이용규	0.370	0.326	0.273	0.154	0.064	0.086
2번타자	김선빈	0.375	0.367	0.289	0.139	0.118	0.100
3번타자	김주찬	0.358	0.410	0.299	0.091	0.104	0.106
4번타자	나지완	0.395	0.427	0.285	0.182	0.162	0.235
5번타자	이범호	0.383	0.390	0.272	0.180	0.148	0.174
6번타자	김원섭	0.396	0.381	0.279	0.194	0.152	0.154
7번타자	김주형	0.245	0.348	0.196	0.065	0.145	0.283
8번타자	차일목	0.381	0.310	0.276	0.169	0.085	0.163
9번타자	박기남	0.344	0.360	0.252	0.140	0.120	0.151
사용 데이터		3.250	3.320	2.420	1.31	1.1	1.45

2일 기아와 LG의 경기에서, 홈팀인 기아의 데이터를 생성하는 과정을 살펴보면 다음과 같다. 팀 데이터는 기아의 누적 데이터 및 LG와의 상대 승률, 상대팀에 대한 평균 자책점을 사용하였고, 당시 기아의 선발투수는 양현종이기 때문에 투수 데이터는 양현종의 2012시즌부터 직전경기인 5월 26일까지의 누적 값을 사용 하였다. 경기 전 발표된 9명의 선발타자 엔트리 및 각 선수의 데이터는 <Table 4>에 나타나 있고, 6월 2일 기아의 최종 데이터는 <Table 5>와 같이 나타낼 수 있다. 이와 같은 방법으로 총 546 경기에 대한 자료를 수집하였다.

Table 5. Data of the team KIA on June 2, 2013 according to the variables in <Table 1>

	변인	KIA 데이터	비고
팀	홈어웨이	1	KIA 홈경기
	휴식	1	지난 경기 휴식
	연승연패	-2	지난 경기까지 2연패
	평균연봉	51.19	KIA선수 평균연봉
	상대승률	0.56	LG에 대한 상대승률
	상대팀에 대한 평균자책점	4.6	LG에 대한 평균자책점
타자	출루율	3.25	Table 4 참조
	장타율	3.32	
	타율	2.42	
	사구/타수	1.31	
	타점/타수	1.10	
	삼진/타수	1.45	
투수	피안타율	0.26	양현종 2012 시즌부터 최근경기(5/26)까지의 누적 데이터
	선발승률	0.16	
	평균소화이닝	2.6	
	평균자책점	3.18	
	볼넷/이닝	0.6	
	삼진/이닝	0.81	
	피홈런/이닝	0.042	
	타자/이닝	4.35	

본 연구에서는 위와 같이 수집된 데이터를 바탕으로, 각 경기에서 팀 간 변인의 차이가 경기 결과에 미치는 영향을 알아보기 위하여, 어웨이 팀의 데이터에서 홈 팀의 데이터를 나눈 나눔데이터, 그리고 홈 팀의 데이터와 어웨이 팀의 데이터를 비교해 홈팀의 값이 크면 1, 작으면 0으로 표시한 이분데이터를 추가로 변환하여 준비하였다. <Table 5>의 값과 같이 수집된 데이터는 원시데이터라 하였다. 데이터 변환에 대해서 <Table 6>에 정리하였으며, 나눔데이터와 이분데이터는 원시데이터 관측치의 절반임을 알 수 있다.

Table 6. Three types of datasets for building predictive models

데이터 종류	설명	관측치
원시데이터	홈 팀과 어웨이 팀 각각의 데이터	546
나눔데이터	어웨이 팀 데이터/홈 팀 데이터	273
이분데이터	If 홈 팀 데이터 ≥ 어웨이 팀 데이터, then 1. Otherwise, 0.	273

2.3 분류 기법

프로야구경기의 승패를 예측하는 것은 예측값이 승 또는 패의 두 가지 값만이 가능하게 되고, 이처럼 종속변인 혹은 예측하고자 하는 값이 범주형이므로 승·패의 예측 모형 수립은 분류문제(classification problem)로 귀결될 수 있다. 분류문제를 해결하기 위해 사용될 수 있는 데이터마이닝 기법들은 그 종류가 다양하다. 하지만 기존의 스포츠분석에서 주로 사용되어 왔던 기법들은 의사결정나무, 로지스틱 회귀분석, 인공신경망 분석, 판별분석 정도로 기법의 종류가 제한적일 뿐만 아니라, 하나의 연구에 여러 가지의 분류기법을 적용한 사례도 드물다. 이에 본 연구에서는 위에 언급된 기법들 외에, 의사결정나무의 메타학습(meta-learning)형태인 랜덤포레스트(random forest)와 패턴인식분야에서 자주 사용되는 지지벡터머신(support

vector machine)을 추가적으로 사용하고자 한다. 본 연구의 모든 자료분석과 예측모형수립은 통계학에서 주로 사용되는 R 소프트웨어(http://cran.r-project.org/)를 이용하여 수행되었고, 각각의 분류기법을 사용하기 위해 설치해야 하는 R 패키지와 사용되는 함수는 <Table 7>에 정리된 바와 같다.

Table 7. R packages and functions used for developing predictive models

분류기법	R 패키지	R 함수
의사결정나무(CART)	rpart	rpart()
랜덤포레스트	randomForest	randomForest()
로지스틱 회귀분석	stats	glm(), step()
신경망모형	nnet	nnet()
지지벡터머신	e1071	svm()
판별분석	MASS	lda(), qda()

의사결정나무(decision tree)를 형성하는 데 사용될 수 있는 알고리즘 역시 여러 가지가 있지만, 본 연구에서는 해석의 용이성과 계산의 효율성 측면에서 다른 알고리즘보다 우수한 것으로 알려져 많이 사용되는 CART(classification and regression trees) 알고리즘을 선택하였다. CART 알고리즘은 전체 데이터를 포함하는 뿌리노드(root node)에서 시작하여, 하나의 부모노드(parent node)로부터 두 개의 자식노드(child node)를 재귀적으로 형성하는 과정이다(Breiman *et al.*, 1984). 이때 자식노드를 형성하기 위해 선택되는 변인과 분지기준은 자식노드에 포함되는 데이터들의 불순도를 최대로 감소시킬 수 있는 것들로 선택하게 된다. 데이터의 불순도를 측정하기 위해서 보통 지니계수(Gini index)나 정보엔트로피(information entropy)가 사용되는데, 본 연구에서는 지니계수를 이용하여 의사결정나무를 형성하였다.

랜덤포레스트(random forest)는 의사결정나무의 메타학습형태로써, 다수의 의사결정나무를 형성하고 각각의 예측값들을 조합하여 정밀도가 높은 분류기를 얻는 기법이다(Breiman, 2001). 이때, 각각의 의사결정나무는 전체 데이터로부터 무작위로 선택된 일부의 변인과 표본을 이용하여 형성하므로 예측성능이 낮지만, 다수의 의사결정나무들을 조합함으로써 흔들림이나 잡음이 많은 데이터에 대해서도 좋은 예측성능을 보이는 특징이 있다. 또한, 선택된 변인의 빈도수와 각 의사결정나무의 예측성능을 이용하여, 독립변인들의 중요도를 계산해 낼 수 있는 장점이 있다. 본 연구에서는 랜덤포레스트의 학습을 위하여 100개의 의사결정나무를 형성하였고, 각 의사결정나무에서 사용될 수 있는 독립변인의 수를 세 개로 정하였다.

로지스틱 회귀분석(logistic regression)은 본래 선형확률 모형의 문제를 해결하기 위한 기법으로써, 어떤 객체가 특정 범주에 속할 확률과 독립변인의 관계를 식 (1)과 같이 표현되는 S곡선으로 가정하고, 최우추정법에 의해 회귀계수를 추정한다.

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)} \quad (1)$$

위 식을 특정범주에 속할 확률 p와 그렇지 않을 확률$(1-p)$의 비율(승산비)의 로그에 대하여 전개하면, 독립변인들의 선형함수로 쉽게 바꿀 수 있으며, 이로써 모형에 대한 해석이 용이해지는 장점이 있다(Jun, 2012). 본 연구에서는 프로야구경기의 승패 예측을 위해 이분 로지스틱 회귀분석을 사용하였고, 그와 동시에 중요변수를 선택하기 위해 단계적 변인선택방법(stepwise variable selection)을 도입하였다.

신경망 모형(neural network)은 스포츠경기 승패의 예측, 혹은 스포츠 마케팅 등의 목적을 위해 주로 사용되어 온 기법 중 하나이다. 신경망은 독립변인의 입력층, 종속변인의 출력층, 그리고 은닉노드(hidden node)들의 은닉층으로 구성된다. 모형의 복잡도는 은닉층과 은닉노드의 개수를 정함으로써 결정되는데, 본 연구에서는 세 개의 은닉노드로 구성된 한 개의 은닉층으로써 모형을 구성하였다. 통상 모델의 복잡도가 증가할수록 학습표본에 과적합되는 경향이 있는데, 이를 방지하기 위함이다. 형성된 모형의 학습을 위하여 출력층의 예측값과 실제값의 차이를 최소화하는 방향으로 모형의 가중치를 갱신하는 역전파(back propagation)알고리즘을 사용하였다. 위에 언급한 기법들과는 달리 신경망모형은 학습된 모형의 해석이 불가능하다는 단점이 있다.

지지벡터머신(support vector machine)은 본래 두 개의 범주를 가지는 객체들을 분류하기 위해 개발된 기법으로써, 두 범주 사이의 여분공간을 최대화할 수 있는 분류초평면(separating hyperplane)을 결정한다(Burges, 1998). 이때, 비선형 분류를 위하여 입력변인들을 고차원 공간으로 이동시켜서 새로운 공간에서의 분류초평면이 실제 입력변인들의 공간에서는 복잡한 비선형의 분류경계를 형성하는 효과를 얻게 해주는 커널트릭을 사용하기도 한다. 본 연구에서는 통상 많이 사용되는 가우시안커널(Gaussian kernel)을 사용하였고, 커널함수의 결정계수는 학습표본을 10개로 분할하고 교차타당성(cross-validation)을 이용하여 결정하였다. 지지벡터머신도 신경망모형과 마찬가지로 모형에 대한 해석은 불가능하다.

판별분석(discriminant analysis)에서는 독립변인들의 정규분포가정을 이용하는 의사결정이론과, 범주평균 사이의 거리는 최대로, 범주내의 분산은 최소로 하기 위해 분류경계를 결정하는 피셔(Fisher) 방법으로 판별함수를 유도할 수 있다(Jun, 2012). 이때, 범주에 관계없이 분산-공분산 행렬이 동일하다고 가정하면 판별함수가 선형으로 도출되는데 이를 선형판별분석(linear discriminant analysis : LDA)이라고 하고, 다르다고 가정하면 이차식의 판별함수가 유도되므로 이를 이차판별분석(quadratic discriminant analysis : QDA)이라고 한다. 본 연구에서는 두 가지 방법을 모두 사용하여 프로야구경기의 승패를 예측하는 시도를 하였다.

통상 데이터마이닝에서 예측모형을 수립할 때, 데이터를 학

습 및 평가 표본으로 나누어 학습 표본에서 모형을 생성하고 평가 표본을 이용하여 모델의 성능을 평가한다. 하지만 이러한 방법은 우연히 예측 모델이 그 평가 표본에만 적합한 경우가 발생할 수 있으므로, 본 연구에서는 어떠한 평가 표본에 대해서도 예측률이 좋은 모델을 만들기 위해 100개의 평가 표본을 무작위로 생성하여 평균 오분류율을 구하였다.

3. 연구 결과

3.1 집단 간 평균의 비교

분석기법을 적용하기 이전에 각 변인들에 대하여 승리한 팀과 패배한 팀 간 평균 비교(*t*-검정)를 실시하였다. 그 결과를 <Table 8>에 정리하였고, 그 중 유의한 변인들의 평균차이는 <Figure

Table 8. *t*-test results

	변인	집단평균(승)	집단평균(패)	t-value	p-value	통계적 유의성
팀	홈어웨이	0.502	0.498	-0.085	0.932	
	휴식	0.117	0.158	1.367	0.172	
	연승연패	0.007	-0.348	-1.597	0.111	
	평균연봉	50.954	48.223	-2.855	0.004	유의
	상대승률	0.506	0.451	-2.912	0.004	유의
	상대팀에 대한 평균자책점	3.803	4.068	2.406	0.016	유의
타자	출루율	3.041	3.034	-0.348	0.728	
	장타율	3.362	3.316	-2.171	0.030	유의
	타율	2.290	2.274	-0.573	0.567	
	사구/타수	1.243	1.260	0.341	0.734	
	타점/타수	1.141	1.122	-1.078	0.282	
	삼진/타수	1.722	1.728	0.310	0.757	
투수	피안타율	0.247	0.242	-1.210	0.227	
	선발승률	0.278	0.244	-2.299	0.022	유의
	평균 소화이닝	4.794	4.572	-1.525	0.128	
	평균 자책점	3.780	3.906	0.970	0.333	
	볼넷/이닝	0.452	0.486	2.151	0.032	유의
	삼진/이닝	0.719	0.694	-1.411	0.156	
	피홈런/이닝	0.064	0.058	-1.620	0.106	
	타자/이닝	4.227	4.239	0.255	0.799	

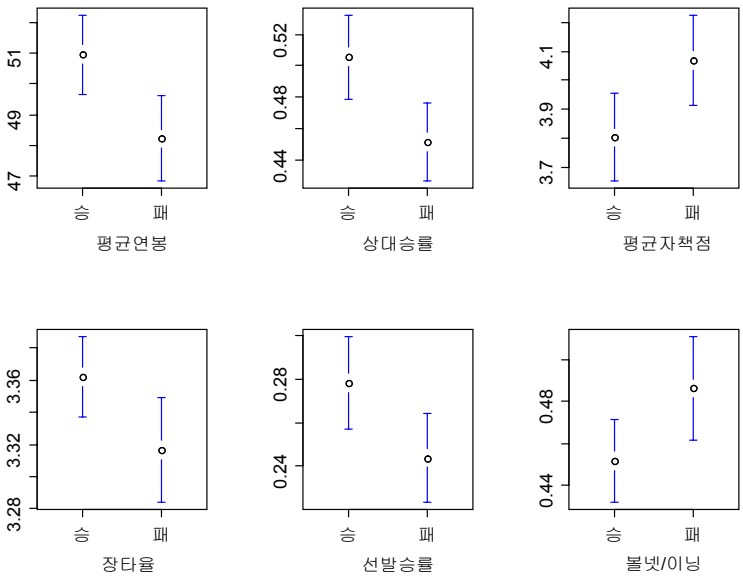

Figure 1. Means and 95% confidence intervals for significant variables

1>과 같다. 평균비교 결과에서 나타나듯이 팀과 관련된 변인 중에서는 평균연봉, 상대승률, 상대팀에 대한 평균자책점이, 타자와 관련된 변인에서는 장타율이, 투수와 관련된 변인에서는 선발투수의 승률과 볼넷/이닝이 통계적으로 유의하게 나타났다($p < .05$). 특히, 팀 관련 변인 중 평균연봉과 상대승률은 다른 유의한 변수들보다 더 큰 차이를 보였다($p < .01$).

3.2 승패예측모형의 수립 및 성능평가

오분류율 : 모델을 생성할 때의 학습표본은 데이터의 60%로, 평가표본은 40%로 구성하였다. 원시데이터의 경우 328개의 학습표본과 218개의 평가표본으로 나뉘었으며, 나눔데이터와 이분데이터의 경우 164개 학습표본과 109개의 평가표본으로 나뉘어졌다. 이렇게 구성된 학습표본에 <Table 7>에서 제시한 분석방법을 적용하여, 분석한 결과는 <Table 9>와 <Figure 2>에 정리하였고, <Table 10>은 <Figure 2>의 가로축에 대한 설명이다. 학습 오분류율은 학습표본에 대한 오분류율이고 평가 오분류율은 무작위로 반복하여 추출된 100개의 평가 표본들의 평균 오분류율이다. 이 결과에 따르면 랜덤포레스트 분석방법은 모든 데이터 종류에서 낮은 오분류율(원시 : 16.82%, 나눔 : 16.13%, 이분 : 15.86%)을 나타내며, 그 다음으로는 이분데이터를 신경망모형으로 예측한 경우(21.70%), 이분데이터를 지지벡터머신으로 예측한 경우(22.88%)가 낮은 오분류율을 보였다. 특히, 랜덤포레스트의 경우 예측성능도 우수할 뿐만 아니라 변인의 중요도 또한 산출해 낼 수 있으므로, 어떤 변수가 경기의 승패에 큰 영향을 끼치는지 알 수 있다. 본 연구의 실험에서는 원시데이터모형의 평가 오분류율보다 두 팀의 상대적인 차이를 나타내는 나눔데이터모형 또는 이분데이터모형의 평가오분류율이 더 낮게 측정되었다.

중요변인 : 본 연구에서 사용된 분석방법들 중에서 의사결정나무, 랜덤포레스트, 단계적 로지스틱회귀분석은 경기의 승패에 영향을 주는 중요변인들을 추출할 수 있다는 장점이 있다. 그 결과를 <Table 11>에 정리하였다. 가장 우측열의 점수는 중요변인으로 선택된 분석방법의 분류 정확도(1-오분류율)를 각각 더하여 나타내었다. 동일한 문제에 대해 데이터의 가공과 사용되는 분석방법에 따라 중요변인이 다르게 나타지만 반복적으로 중요변인으로 선정되는 변인은 점수가 높게 나타나며, 분류를 하는 데 있어 중요한 역할을 한다고 볼 수 있다. 그 결과, 다른 변인들에 비해 중요한 변인은 삼진/이닝, 평균연봉, 평균자책점, 선발승률, 삼진/타수, 사구/타수 순으로 나타났고, 홈어웨이, 연승연패는 상대적으로 중요하지 않은 변인으로 나타났다.

Table 9. Error rates of predictive models

예측모형	데이터	학습 오분류율	평가 오분류율
의사결정나무	원시데이터	0.2043	0.2853
	나눔데이터	0.1951	0.2780
	이분데이터	0.2927	0.3301
랜덤포레스트	원시데이터	0.1768	0.1682
	나눔데이터	0.1585	0.1613
	이분데이터	0.1646	0.1586
로지스틱 회귀분석	원시데이터	0.4421	0.4322
	나눔데이터	0.4146	0.4092
	이분데이터	0.4146	0.4124
신경망모형	원시데이터	0.3689	0.3617
	나눔데이터	0.3049	0.3183
	이분데이터	0.2500	0.2170
지지벡터머신	원시데이터	0.2805	0.3314
	나눔데이터	0.3110	0.3525
	이분데이터	0.1159	0.2288
선형판별분석	원시데이터	0.4024	0.4059
	나눔데이터	0.4329	0.4061
	이분데이터	0.3720	0.3688
이차판별분석	원시데이터	0.2988	0.3434
	나눔데이터	0.3476	0.3572
	이분데이터	0.1829	0.2585

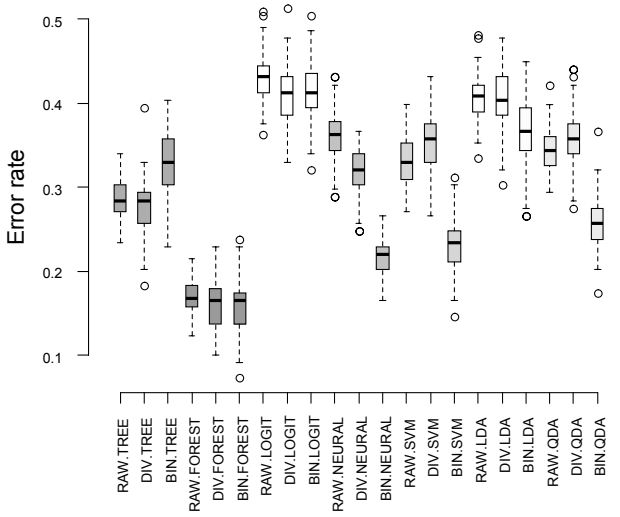

Figure 2. Boxplots of test error rates

Table 10. Horizontal axis labels of <Figure 2>

데이터 모형		분석방법	
RAW	원시데이터	TREE	의사결정나무
		FOREST	랜덤포레스트
DIV	나눔데이터	LOGIT	로지스틱회귀분석
		NEURAL	신경망 모형
BIN	이분데이터	SVM	지지벡터머신
		LDA	선형판별분석
		QDA	이차판별분석

Table 11. Important variables selected by decision tree, random forests, and stepwise logistic regression

	변인	의사결정나무			랜덤포레스트			로지스틱회귀분석			점수
		원시	나눔	이분	원시	나눔	이분	원시	나눔	이분	
팀	홈어웨이										0
	휴식			O						O	1.3
	연승연패										0
	평균연봉			O		O	O	O		O	3.5
	상대승률	O	O	O							2.1
	상대팀에 대한 평균자책점	O			O						1.5
타자	출루율	O	O								1.4
	장타율		O						O		1.3
	타율			O			O				1.5
	사구/타수	O	O		O						2.3
	타점/타수	O		O	O						2.2
	삼진/타수			O		O	O				2.3
투수	피안타율	O						O	O		1.9
	선발승률		O		O	O					2.4
	평균소화이닝		O			O					1.5
	평균자책점	O	O		O			O	O		3.4
	볼넷/이닝		O								0.7
	삼진/이닝	O		O		O	O		O		3.6
	피홈런/이닝	O		O				O			2.0
	타자수/이닝		O								0.7

3.3 결과해석

예측 모형의 해석이 가능한 분석방법은 의사결정나무, 로지스틱 회귀분석이 있는데, 이들 중 오분류율이 가장 작은 의사결정나무-나눔데이터 모형의 예측모델을 통해 결과를 해석하였다. <Figure 3>은 나눔데이터를 의사결정나무 기법을 통해 학습한 예측모형이다. 최종 노드들 중 개체 수 10개 이상, 분류정확도 80% 이상인 3, 8, 11, 24번 노드들의 해석을 통해 데이터의 어떠한 변인이 경기결과에 영향을 미치는지 알아보았다. 그 결과는 <Figure 4>에 표시된 바와 같다. <Figure 4>는 홈팀의 값을 기준으로 한 어웨이 팀 값의 비율로써, 경기 결과에 어떠한 영향을 미치는지를 보여준다. Node 3에서는 홈팀에 비해 어웨이 팀의 사구/타수가 작을수록(약 86% 이하) 홈팀이 승리할 확률이 높아짐을 알 수 있다. 다음으로 Node 8은 홈팀에 비해 어웨이팀의 사구/타수(약 86% 이상)와 선발승률(약 73% 이상)이 높고 볼넷/이닝이 특정 구간내에 있을 때(94% 이상 101.8% 미만) 홈팀이 패배한다고 예측하였다. Node 11에서는 사구/타수 및 선발승률이 Node 8과 같은 경향을 보이는데도 홈팀이 승리할 것으로 예측하였는데, 이는 상대적으로 볼넷/이닝과 평균자책점의 영향력이 더 크기 때문이라고 볼 수 있다. 마지막으로 Node 24는 볼넷/이닝과 장타율이 기대와 반대로 나타났으나, 사구/타수, 선발승률, 평균소화이닝의 영향력이 상대적으로 더 크게 작용함을 확인할 수 있다.

4. 결 론

본 연구의 목적은 2013년도 시즌 국내 프로야구 팀과 선수들의 누적데이터를 통해서 다음 경기의 승패를 예측하는 것이다. 이를 위해 홈 팀과 어웨이 팀 각각의 경기기록으로 생성한 원시데이터, 어웨이 팀의 데이터를 홈 팀의 데이터로 나눈 나눔데이터, 홈 팀 데이터와 어웨이 팀 데이터의 차이를 0, 1로 나타낸 이분데이터를 생성하였다. 본격적으로 예측 모델을 수립하기 이전에, 원시데이터 모형에 대해 *t*-검정을 실시하였고, 팀과 관련된 변인에서는 평균연봉, 상대승률, 상대팀에 대한 평균자책점이, 타자와 관련된 변인에서는 장타율이, 투수와 관련된 변인에서는 선발투수의 승률과 볼넷/이닝이 통계적으로 유의하게 나타났다. 분류기법으로는 의사결정나무, 랜덤포레스트, 로지스틱 회귀분석, 신경망분석, 지지벡터머신, 판별분석을 사용하였고 연구의 결론은 다음과 같다.

첫째, 프로야구의 경기기록과 정보를 이용해 누적된 자료를 바탕으로 생성한 세 가지 데이터 종류에서 원시데이터 모형을 사용하였을 때의 오분류율보다 나눔데이터 모형 또는 이분데이터모형의 오분류율이 더 낮음을 확인할 수 있었다. 이를 통해 한 경기에서 두 팀의 데이터를 각각 사용하는 것 보다 두 팀의 상대적인 차이를 사용한 데이터 모형이 경기의 승패를 예측하는데 효과적이라는 것을 확인할 수 있었다.

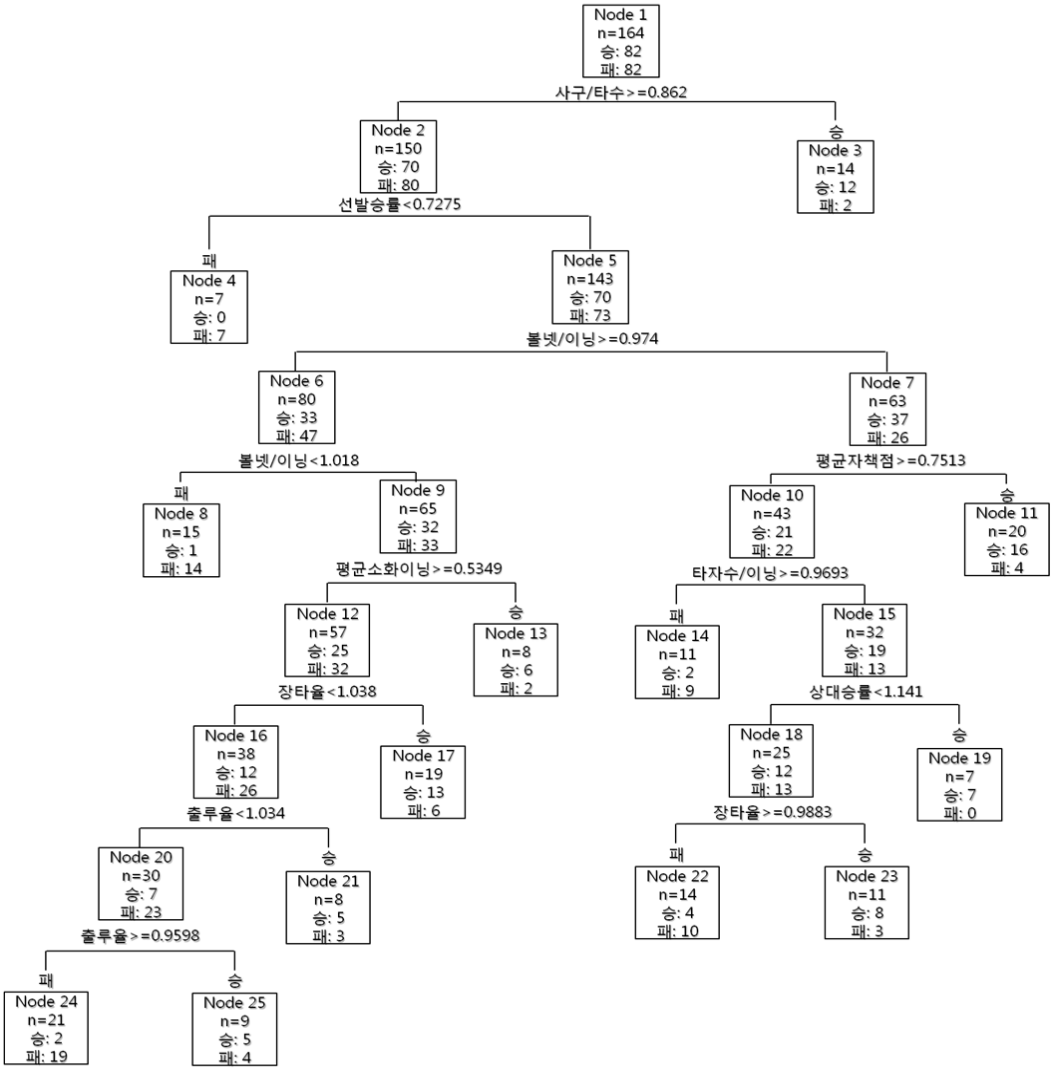

Figure 3. Trained decision tree based on DIV dataset

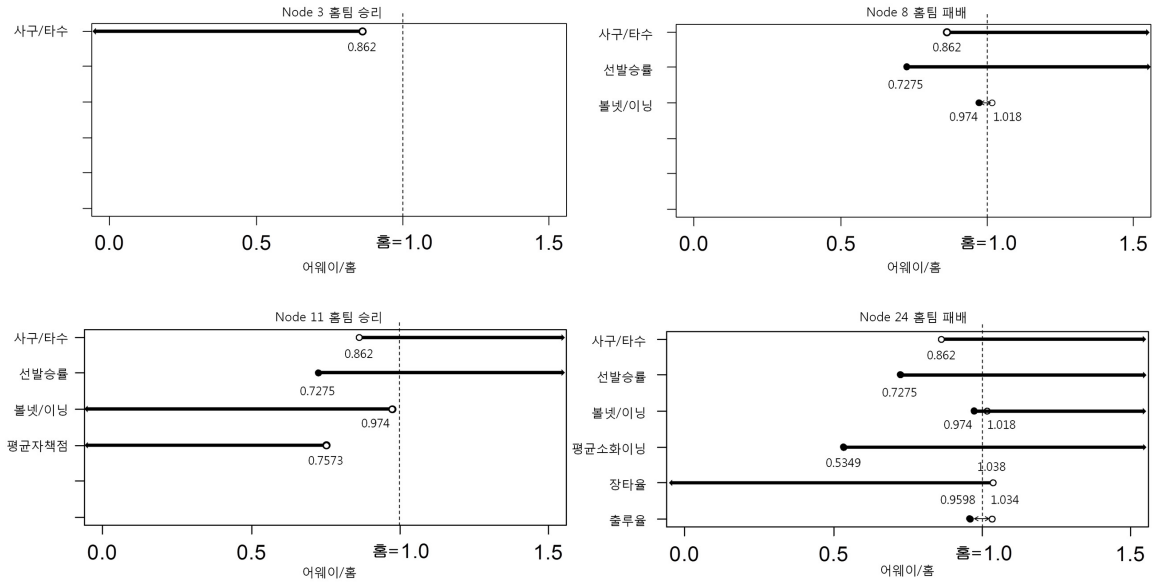

Figure 4. Interpretation of four leaf nodes of the decision tree in <Figure 3>

둘째, 세 가지의 데이터 종류에 대해 일곱 가지 분석기법들을 사용하여 예측 모델을 수립한 결과, 랜덤포레스트를 사용한 경우(원시 : 16.82%, 나눔 : 16.13%, 이분 : 15.86%), 이분데이터에 신경망모형을 사용한 경우(21.70%), 이분데이터에 지지벡터머신을 사용한 경우(22.88%)순으로 오분류율이 낮게 측정되었고, 특히 이분데이터 모형에 랜덤포레스트를 사용하였을 때 오분류율이 가장 낮은 것으로 나타났다. 따라서 데이터 종류나 분석기법들에 따라 예측 결과가 차이가 나는 것을 확인할 수 있었다.

셋째, 일곱 가지 분석기법들 중에서 중요변인을 확인할 수 있는 의사결정나무, 랜덤포레스트, 로지스틱회귀분석을 통해서 각각 중요변인을 점수로 변환하여 비교한 결과 삼진/이닝, 팀의 평균연봉, 투수의 평균 자책점, 선발승률, 타자의 삼진/타수, 사구/타수 순서로 중요한 변인으로 나타났다. 이를 통해, 팀의 평균연봉이 선수들의 기량을 충분히 반영하고, 선발투수의 성적이 타자의 성적보다 상대적으로 중요하다는 것을 추론할 수 있다.

기존의 스포츠 분석 관련 연구들은 한정된 몇 가지의 분석기법만을 사용한 반면, 본 연구는 기존에 사용되지 않았던 기법들을 포함하여 여러 가지 기법들을 적용한 점, 그리고 각 기법들로부터 승패에 영향을 미치는 중요 요인들을 산출하여 그 의미를 파악하였다는 점에서 의의가 있다. 특히, 랜덤포레스트 분석기법은 성능도 우수할 뿐만 아니라, 변수의 중요도 역시 산출해 낼 수 있으므로, 현재까지 스포츠분석에서 사용된 사례는 없지만, 본 연구에서 사용한 결과 다른 분석기법들에 비해 비교적 낮은 오분류율을 나타내 추후 스포츠 분석에서 사용한다면 좋을 예측결과를 기대할 수 있을 것이다. 또한, 경기에서 팀 간의 상대적인 차이를 반영한 새로운 종류의 데이터를 가공하여 사용했다는 점, 선수들의 누적데이터를 이용하여 경기의 승・패를 예측했다는 점이 기존의 연구와 다르다고 할 수 있다.

본 연구의 한계점으로는, 분석에 사용된 데이터가 6월까지의 경기이었기 때문에 7월 이후의 경기 승패를 예측하기 위해서는 직전경기까지의 누적데이터가 필요하다는 점과, 검증되지 않은 신인선수들과 NC와 같은 신생팀의 자료가 부족하였다는 점이다. 향후 몇 년간 자료가 누적된다면 더 정확한 예측모형을 얻을 수 있을 것이라고 기대한다.

참고문헌

Breiman, L. (2001), Random forests, *Machine Learning*, **45**(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and regression trees*, Wadsworth, CA, USA.

Burges, C. J. C. (1998), A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, **2**, 121-167.

Chea, J.-S., Cho, E.-H., and Eom, H.-J. (2010), Comparisons of the outcomes of statistical models applied to the prediction of post-season entry in Korean professional baseball, *The Korean Journal of Measurement and Evaluation in Physical Education and Sport Science*, **12**(1), 33-48.

Hong, C., Jung, M., and Lee, J. (2010), Prediction model analysis of 2010 South Africa world cup, *Journal of the Korean data and information science society*, **21**(6), 1137-1146.

Hong, S., Jung, K., and Chung, T. (2003), Win/Lose prediction system : Predicting baseball game results using a hybrid machine learning model, *Journal of Korea Information Science Society : Computing Practices*, **9**(6), 693-698.

Jensen, S. T., McShane, B. B., and Wyner, A. J. (2009), Hierarchical Bayesian modeling of hitting performance in baseball, *Bayesian Analysis*, **4**(4), 631-652.

Jun, C.-H. (2012), *Data Mining Techniques and Applications*, Hannarae, Seoul, Korea.

Kim, C. (2001), A win-loss predicting model by analyzing professional baseball game, *Journal of Sport and Leisure Studies*, **16**, 807-819.

Kim, D., Lee, S., and Kim, Y. (2007), Prediction for 2006 Germany world cup using Bradley-Terry model, *The Korean journal of applied statistics*, **20**(2), 205-218.

Kim, J. H., Ro, G. T., Park, J. S., and Lee, W. H. (2007), The development of soccer game win-lost prediction model using neural network analysis : FIFA world cup 2006 Germany, *Korean Journal of Sport Science*, **18**(4), 54-63.

Kim, N.-K. and Park, H.-M. (2011), Predicting the score of a soccer match by use of a Markovian arrival process, *IE Interfaces*, **24**(4), 323-329.

Koo, S., Kim, H., and Chang, S. (2009), A comparative study on win-loss prediction models for Korean professional basketball, *Korean Journal of Sport Science*, **20**(4), 704-711.

Korean Baseball Organization (2013), *2013 KBO Annual Report*, Korean Baseball Organization, Seoul, Korea.

Lee, D.-J. and Yang, W. M. (2004), Performance evaluations of professional baseball players using DEA/OERA, *IE Interfaces*, **17**(4), 440-449.

Lewis, M. M. (2004), *Moneyball : The Art of Winning an Unfair Game*, W. W. Norton and Company, NY, USA.

Miljkovic, D., Gajic, L., Kovacevic, A., and Konjovic, Z. (2010), The use of data mining for basketball matches outcomes prediction, *Proceedings of the 8th International Symposium on Intelligent Systems and Informatics*, 309-312.

Min, D. K. and Hyun, M. S. (2009), Prediction of a winner in PGA tournament using neural network, *Journal of the Korean data and information science society*, **20**(6), 1119-1127.

Null, B. (2009), Modeling baseball player ability with a nested Dirichlet distribution, *Journal of Quantitative Analysis in Sports*, **5**(2), 1-36.

Odachowski, K. and Grekow, J. (2013), Using bookmaker odds to predict the final result of football matches, *Lecture Notes in Artificial Intelligence*, **7828**, 196-205.

Oh, K.-M. and Lee, J.-T. (2003), A model study on salaries of Korean pro-baseball players using data mining, *Journal of Korean Sociology of Sport*, **16**(2), 295-309.

Seidman, C. (2002), *MS SQL server2000 data mining* (Technical Reference).

Sung, H. and Chang, W. (2007), Forecasting the results of soccer matches using poisson model, *IE Interfaces*, **20**(2), 133-141.