한국스포츠사회학회지 제16권 제2호, 2003, pp. 295-309
Journal of Korean Sociology of Sport
Vol. 16, No. 2, 2003, pp. 295-309

데이터마이닝을 이용한 한국프로야구 선수들의 연봉에 관한 모형연구

오광모(단국대학교 석사) · 이장택(단국대학교 교수)

Ⅰ. 서 론

스포츠산업은 스포츠와 관련된 재화와 서비스의 상업적 생산 및 전반적인 유통활동을 말하며 그 중에서 프로스포츠는 스포츠산업의 주요 영역이다. 한국의 프로스포츠가 진정한 의미의 프로스포츠 시대를 개막한 것은 1982년 프로야구 출범과 1983년 프로축구가 탄생되면서부터 라고 할 수 있으며, 프로스포츠의 짧은 역사로 인해 아직까지 국내에서는 프로스포츠에 대한 연구결과가 미흡하다고 할 수 있지만 프로야구가 일반대중에게 미치고 있는 영향과 사회적 관심이 매우 큰 현실을 감안한다면 한국 프로야구에 대한 보다 체계적 연구는 반드시 선행되어져야 할 새로운 과제임이 틀림없다.

지금까지 연구되어진 한국프로야구선수의 경기력과 연봉과의 관계에 대한 연구들은 스포츠사회학 및 경영학 측면에서 살펴본 프로야구 연봉제에 대한 논의, 노동법 관점에서의 연봉제, 그

주제어: 회귀분석, 신경망분석, 의사결정나무분석, 데이터마이닝

※ 이 연구는 2003학년도 단국대학교 대학연구비의 지원으로 연구되었음

리고 더 미시적으로 접근된 투수 또는 타자의 연봉에 관한 문제 등이 다루어져왔다(김응식, 2001; 이장영·강효민, 2001). 이러한 주제들에 대한 접근은 프로야구선수들의 연봉과 관련된 문제가 사회적인 문제로 부각됨으로서, 이를 해결하기 위한 노력의 일환으로 더욱 관심을 받고 있다고 하겠다.

투수나 타자의 연봉과 경기수행능력의 관계를 규명하는 연구들은 지금까지 기술통계량과 그래프를 이용한 분석을 수행하거나 연봉에 중요한 영향을 미치는 독립변수를 찾아 회귀분석 또는 요인분석을 통해 새로운 인자를 찾아내어 이들이 지니고 있는 특성을 설명하였다(김응식, 1998; 윤여관, 1990; 윤춘식, 1994; 이군호, 2000; Estensen, 1994; Holbrook & Shultz, 1996; Marburger, 1994). 하지만 본 논문에서는 신경망분석, 의사결정나무분석과 같은 분류규칙을 이용한 데이터마이닝 기법과 회귀분석을 적용하여 연봉에 대한 예측값을 구하고, 한국프로야구 선수들의 실제연봉과 비교하여 가장 적절한 모형을 찾고자 하는데 주력하였으며 그 결과 매년 발생되고 있는 구단과 선수들과의 연봉협상 과정에서 갈등을 줄일 수 있는 연구의 기틀을 마련하고자 한다.

데이터마이닝이란 일반적으로 데이터에서 필요한 부분의 정보를 구하거나, 의사결정을 하게 해주는 지식을 제공하는데 사용되는 기법이라고 할 수 있다. 최근 들어, 데이터마이닝은 경영정보학, 컴퓨터공학, 통계학 등의 학문 분야를 중심으로 커다란 관심을 갖게 되었다. 이러한 관심을 갖게 된 주 이유는 인터넷 시대를 맞이하여 대규모의 실제 데이터 문제를 직접 해결해 줄 수 있는 방법의 절실함과 지금까지 현실의 문제와 동떨어진 학문을 함으로써 생기는 학문과 현장의 괴리감을 극복한다는 각성을 가지기 때문이라고 여겨진다.

본 연구에서는 한국프로야구 선수들의 연봉에 대한 적절한 모형을 설정하기 위해 클레멘타인(ver 7.0)과 answer tree(ver 2.0.1)을 이용한 데이터마이닝 기법을 이용하였다. 그리고 이를 수행하는 방법으로 신경망분석, 의사결정나무분석과 회귀분석을 사용하였는데 세 가지 방법은 몇 개 이상의 변수를 기초로 하여 종속변수의 모형화를 시도하는 기법으로 주로 미래의 결과가 알려지지 않은 경우에 대하여 예측을 하는 모형을 구축하는데 사용되어진다. 이와 같은 접근방법은 프로야구의 연봉문제와 제도화에 크게 기여할 것으로 기대되어지며 특히 복잡한 구조를 가진 자료에서의 예측문제를 해결하기 위한 비선형모형인 신경망모형과 의사결정나무모형을 고려함으로써 선형적으로 설명되지 않는 스타급 선수들의 연봉산정이 보다 합리적으로 가능하다고 판단되어진다.

Ⅱ. 연구방법

1. 자료의 구성

본 논문에 이용된 자료는 한국야구위원회에서 발간한 2001년 한국프로야구연감에

기록된 통산기록 및 2000년 개인별 성적과 각 구단 홈페이지에서 수집된 연봉(단위: 만원)이다. 타자의 경우는 2000년 한 경기이상 뛴 214명 중 투수가 타격 혹은 타자로 등록된 경우의 14명, 2001년 은퇴 및 미 계약자 그리고 용병의 경우 퇴출되어 2001년 연봉이 책정되지 않은 31명을 제외한 169명이고, 투수의 경우에는 2000년 한 경기이상 뛴 154명 중 은퇴 및 미 계약 또는 해외진출로 2001년 연봉이 책정되지 않은 8명을 제외한 146명이다. 그러나 일정한 타수 및 투구이닝 이외의 자료는 분석에서 제외하는 것이 분석의 안정성에 바람직하다는 판단으로 타자의 경우 2000년 34타수 이상의 기록을 가진 126명, 투수의 경우 2000년 1이닝 이상의 기록을 가진 144명을 대상으로 분석하였다.

표 1. 변수의 구성

타자에 대해 사용된 변수		투수에 대해 사용된 변수	
연차	실책	연차	총완투
소속	총타율	소속	총완봉
포지션	총경기수	방어율	총승리
타율	총타수	경기수	총패전
경기수	총득점	완투	총세이브
타수	총안타	완봉	총타자수
득점	총이루타	승리	총이닝수
안타	총삼루타	패전	총피안타
이루타	총홈런	세이브	총피홈런
삼루타	총루타	타자수	총볼넷
홈런	총타점	이닝수	총데드볼
루타	총도루	피안타	총삼진
타점	총도루자	피홈런	총실점
도루	총희생타	볼넷	총자책점
도루자	총볼넷	데드볼	
희생타	총데드볼	삼진	
볼넷	총삼진	실점	
데드볼	총병살타	자책점	
삼진	총실책	총방어율	
병살타		총경기수	
총 39개		총 34개	

2. 변수의 구성

2001년 한국프로야구연감에 나타난 통산기록 및 2000년 개인별 성적을 토대로 본 연구에서 사용할 변수를 표 1과 같이 구성하였는데, 타자의 경우는 총 39개의 변수, 투수의 경우는 총 34개의 변수를 사용하였다.

3. 모형선택을 위한 방법

본 연구에서 사용한 모형구축도구를 약술하면 다음과 같다.

1) 신경망분석(Neural Network)

신경망분석은 인간의 두뇌가 문제를 해결하는 구조를 이용한 분석이다. 전형적인 신경망은 망을 생성하는 층들이 배열된 뉴런이란 것으로 구성되어져 있다. 각각의 뉴런은 전체에서 어느 한 부분을 작용하는 과정 요소로 생각되어질 수 있다. 모든 뉴런을 연결하게 되면, 어떤 패턴을 학습하고, 데이터간의 상호관계를 학습하는 네트워크로 발전하게 된다.

신경망을 이용하여 예측 모형을 구축할 때, 입력층은 예측에 필요한 모든 변수들이 포함된다. 다음 출력층에서는 종속변수와 비교하여 나온 결과 값들이 들어있다. 은닉층은 입력층의 변수들을 결합하여 연결하는 뉴런들로 구성되어 있다. 일반적으로 은닉층은 그 수가 많지는 않지만 여러 개를 가질 수가 있다. 그리고 하나의 층에 있는 모든 뉴런들은 다음 층과 모두 연결되어져 있다. 클레멘타인 신경망분석에서 훈련방법은 여러 가지 존재하나 본 연구에서는 기본설정인 빠른(Quick) 방법을 사용하였다. 빠른 방법은 신경망의 은닉층이 1개 뿐인 방법으로 가장 간단한 몇 개의 뉴런을 가지고 있어 가장 빠른 시간 내에 신경망분석의 결과를 제공하여 준다.

2) 의사결정나무분석(CART 알고리즘)

CART(Classification and Regression Trees)는 이산형 목표변수인 경우 적용하는 지니지수(Gini Index)와 연속형 목표변수인 경우 적용하는 분산의 감소량을 이용하여 부모마디로부터 자식마디가 2개만 형성되게 한다는 이지분리(binary split)를 수행하는 알고리즘이다(Breiman et al., 1984). 여기서 지니지수는 각 마디에서의 불순도 또는 다양도를 재는 측도 중의 하나이며, n개의 원소 중에서 임의로 2개를 추출하였을 때, 추출된 2개가 서로 다른 그룹에 속해있을 확률을 의미하며 심프손의 다양도지수로도 알려져 있다.

CART이외에도 의사결정나무분석을 위한 다른 알고리즘은 CHAID, C5.0 등이 있는데, 최근에는 이들의 장점을 결합하여 보다 개선된 알고리즘들이 제안되고 상용화되고 있으며, 점점 더 알고리즘의 구별이 모호해지고 있다.

3) 회귀분석(Regression Analysis)

종속변수가 다른 변수들에 의해서 어떻게 설명 또는 예측되는 지를 알아보기 위하여 자료를 적절한 함수식으로 표현하여 분석하는 통계적 방법을 회귀분석이라 한다. 선형회귀모형에서 모수 회귀계수들을 추정하기 위해서 일반적으로 최소제곱법을 사용한다. 최소제곱법은 오차들의 제곱합이 최소가 되도록 회귀계수를 추정하는 방법이다. 또한 각 회귀계수들에 대한 검정은 t-검정을 이용하여 수행할 수 있다.

4. 분석방법

연봉예측을 위한 확률모형을 설정하기 위하여 데이터분석의 초기단계에서 데이터를 면밀히 검토할 필요가 있다. 따라서 탐색적 데이터 분석을 시행하여 종속변수에 입력의 오류는 있는지 혹은 데이터의 분포가 어떠한지를 개략적으로 판단하기 위해 막대그림, 상자그림, 기술통계, 히스토그램, 그리고 개별변수들과 종속변수간의 상관계수 등을 통해 파악한 후 이산형 변수의 경우 평균비교를 통하여 모형에 특정변수의 포함여부를 결정하였다. 그리고 SPSS 클레멘타인과 answer tree를 이용하여 신경망분석, 의사결정나무분석, 신경망분석후의 의사결정나무분석, 의사결정나무분석후의 신경망분석, 회귀분석과 같은 5가지 모형을 만든 후 구한 예측값과 종속변수를 이용하여 평균절대편차(Mean Absolute Deviation, MAD)와 평균제곱오차(Mean Squared Error, MSE)의 두 가지 모형평가 기준 아래에서 가장 최적의 모형을 찾아내었다.

Ⅲ. 분석결과

1. 탐색적 데이터 분석

데이터에 내포된 구조와 특징을 파악하거나 통계적인 모형 선택 및 진단을 하기 위해 막대그림과 상자그림을 사용하였으며 자료가 가지고 있는 정보를 요약한 기술통계량, 두 양적 변수간의 선형관계를 알아보기 위한 상관계수를 통하여 탐색적 데이터 분석의 초기단계를 수행하였다.

그림 1과 그림 2를 통하여 타자들의 평균연봉에서는 현대와 LG가, 투수들의 평균연봉에서는 두산과 삼성이 높은 연봉을 받고 있음을 알 수 있으며, 상자그림을 통해서 타자와 투수의 경우 모두 오른쪽으로 꼬리가 긴 형태의 분포를 가지며 이상값이 존재한다는 것을 확인할 수 있다. SPSS에서는 이상값이 아닌 최대값, 3사분위수, 중위수, 1사분위수, 이상값이 아닌 최소값과 같은 5개의 숫자로 상자그림을 기본적으로 그리며 3사분위수로부터 상자길이의 1.5배 이상 떨어진 관측값인 이상값을 o, 3사분위수로부터 상자길이의 3배 이상 벗어난 관측값인 극단값을 *로 표기한다. 예를 들어 타자인 경우에 현대와 두산은 이상값이 존재하지 않지만 SK인 경우에는 47번째

선수가 이상값 임을 보여주고 있다.

한편 타자연봉에 대한 상관계수는 타점(0.831), 포볼(0.828), 득점(0.810), 루타(0.808) 등의 순으로 주로 2000년 페넌트레이스 성적의 변수들이 높은 상관관계를 나타냈으며, 투수연봉에 대한 상관계수는 총삼진(0.728), 총승리(0.722), 총이닝수(0.675), 총패전(0.624) 등의 순으로 주로 통산성적의 변수들이 높은 상관관계를 나타냈음을 알 수 있었다.

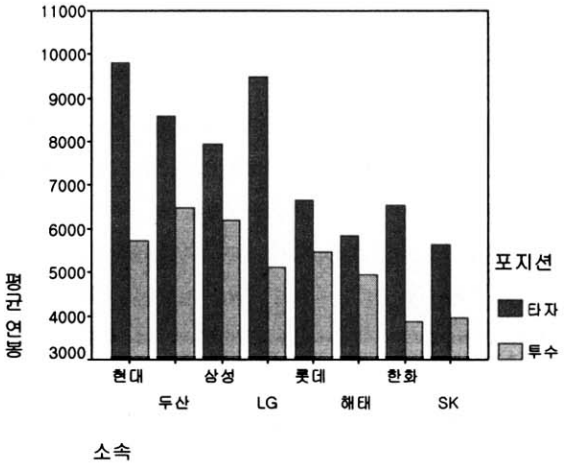

그림 1. 타자, 투수의 소속별 연봉의 막대그림

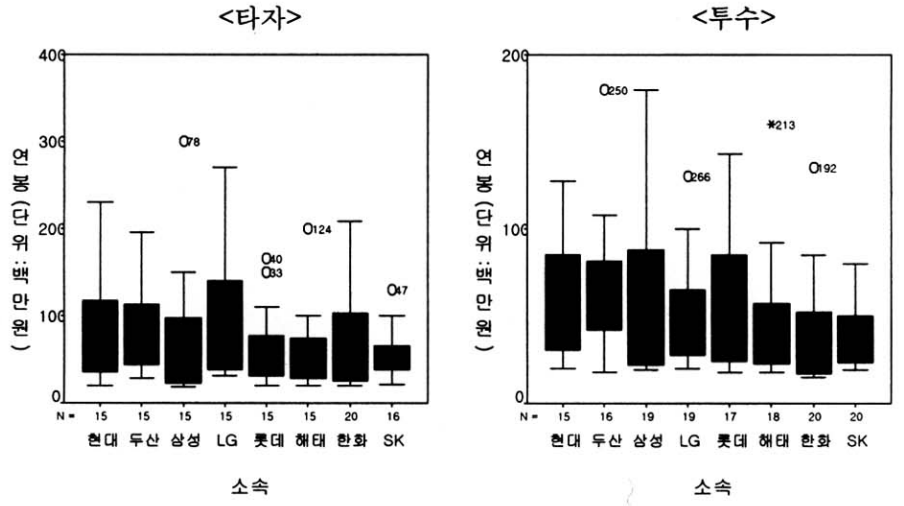

그림 2. 타자, 투수의 소속별 연봉의 상자그림

2. 자료의 재표현 및 종속변수의 정규성 검정

회귀분석을 사용하는 경우에 종속변수인 한국프로야구선수들의 연봉이 정규분포를 따른다는 정규성 가정을 만족시키기 위해 로그변환으로 자료를 재표현 하였다. 로그변환을 실시한 결과 상자 그림 내에 이상값이 존재하지 않았고, 정규성 검정을 이용하여 로그변환을 취한 타자, 투수의 연봉자료가 정규분포를 따른다는 가설을 채택할 수 있었다.

3. 팀간, 포지션별 평균비교

팀간, 포지션별의 평균차이를 검정하여 그룹 간 특성의 차이를 추후 분석에 포함할지의 여부를 결정하기 위하여 일원배치분산분석을 실시하였다. 그 결과 타자의 소속에 따른 유의확률은 0.313, 타자의 포지션에 따른 유의확률은 0.516, 투수의 소속에 따른 유의확률은 0.171로서 유의수준 10%에서도 모두 유의하지 않기 때문에 추후의 분석에서 타자의 경우 소속과 포지션, 투수의 경우 소속변수는 제외하였다.

4. 타자에 관한 분석

1) 신경망분석

클레멘타인을 이용하여 신경망분석을 실시하였다. 기존의 39개 변수 중에서 평균비교결과 유의한 차이를 보이지 않은 소속과 포지션 변수를 제외한 37개의 변수를 입력변수로 놓고 로그변환을 한 연봉자료를 종속변수로 지정하였다.

클레멘타인의 출력결과는 신경망모형의 예측정확도는 96.459%였으며, 첫 번째 은닉층에는 3개의 뉴런, 출력층은 1개의 뉴런으로 구성되었다. 또한 입력변수간의 중요도는 0과 1 사이의 값을 취하는데 총볼넷(0.366), 총삼진(0.309), 총타율(0.246), 총회생타(0.230), 득점(0.216), 타점(0.202)의 순서로 중요성을 가지고 있음을 확인할 수 있었다.

2) 의사결정나무분석

answer tree의 CART 알고리즘을 이용하여 의사결정나무분석을 시행하였다. 정지규칙은 최대트리 깊이를 5, 부모마디와 자식마디의 개수를 각각 5와 1로 정하였고 불순도의 최소향상도는 디폴트값인 0.0001로 지정하였다. 로그변환을 행한 연봉자료에 대한 뿌리마디의 결과는 평균 8.6643, 표준편차 0.7162이고 뿌리마디만 있을 경우에 비해 96.117% 정도가 회귀나무에 의하여 추가적으로 설명되어지는 분산의 비율임을 알 수 있었다.

로그를 취한 연봉자료의 첫 번째 분리는 득점에 의하여 이루어졌음을 확인할 수 있었으며 득점으로 분리했을 때 분산의 감소정도가 0.2924로서 전체분산의 57% 이상을 줄여주고 있음을 알

수 있었다. 아울러 득점, 총루타, 총타율, 총타점, 총홈런 순으로 로그변환을 통한 연봉 데이터에 영향력을 갖고 있음을 알 수 있었다.

3) 의사결정나무분석 후의 신경망분석

신경망분석은 사용되는 변수가 많을수록 훈련에 소요되는 시간은 더 길어지고 종속변수에 영향을 주지 않는 변수도 여전히 모형에 남게 된다는 문제점을 가지고 있으므로 신경망에 사용될 변수의 수를 줄이기 위한 전처리 과정으로 CART 알고리즘을 사용하였다.

CART 알고리즘을 사용한 후의 신경망모형의 예측정확도는 96.07%이고 기존 변수 중 병살타, 희생타, 총병살타가 제외된 까닭으로 34개의 뉴런이 입력층에 사용되었고 첫 번째 은닉층에는 3개의 뉴런으로 구성되어 있으며 출력층은 1개의 뉴런으로 구성되었다. 그 결과 입력변수간의 중요도로부터 총희생타(0.374), 총삼진(0.293), 총홈런(0.281), 총볼넷(0.273), 타점(0.242) 등의 순으로 중요도를 가지고 있음을 확인할 수 있었다.

4) 신경망분석후의 의사결정나무분석

신경망분석의 가장 큰 단점인 그 과정의 불투명성, 즉, 모형이 구축이 되었을 때 이것이 어떻게 해서 구축이 되었는지에 대한 과정과 이유가 부족하다는 단점을 보완하여 신경망 해석의 이해를 돕기 위해 신경망을 통해 형성된 예측값을 가지고 CART알고리즘을 수행하였다.

CART 알고리즘의 정지규칙은 의사결정나무분석 만을 수행할 때와 같은 정지규칙을 사용하였으며, 그 결과 회귀나무가 형성된 후의 분산비율은 0.03090으로 뿌리마디만 있을 경우에 비해 96.91%가 회귀나무에 의하여 추가적으로 설명되어지는 분산의 비율임을 알 수 있었다. 또한 변수의 중요성 순서는 총루타, 안타, 총안타, 총홈런, 타점 순이었다. 타자연봉인 경우의 신경망분석후의 의사결정나무분석을 통해 그려진 최종나무구조의 첫 번째 분리는 총루타에 의하여 이루어졌으며 총루타로 분리했을 때 분산의 감소정도가 0.3018로서 전체분산의 61%이상을 줄여주고 있음을 알 수 있었다. 아울러 총루타, 안타, 총안타, 총홈런, 타점 순서로 예측연봉에 대해 영향력을 갖고 있음을 확인할 수 있었다.

5) 회귀분석

로그변환을 한 타자연봉을 y로 두고, 나머지 변수를 독립변수로 하여, 변수선택의 방법으로 단계선택방법을 이용하였다. 그 결과 추정된 선형회귀식(결정계수=0.892)은 다음과 같다.

$$\hat{y}=6.722+0.013\times\text{타점}+0.004\times\text{총희생타}+5.397\times\text{총타율}+0.007\times\text{포볼}+0.035\times\text{연차}-1.17\times\text{타율}$$

5. 투수에 관한 분석

1) 신경망분석

투수모형에 대한 신경망분석은 33개의 뉴런이 입력층에 사용되었고 첫 번째 은닉층에는 3개의 뉴런, 출력층은 1개의 뉴런으로 구성되었다. 또한 신경망모형의 예측정확도는 94.991%이고 변수의 중요도는 총삼진(0.331), 총이닝수(0.239), 승리(0.238) 등의 순서였다.

2) 의사결정나무분석

타자의 경우와 마찬가지로 CART 알고리즘을 이용하여 의사결정나무분석을 시행하였다. 정지규칙, 최대 트리 깊이, 부모마디와 자식마디의 개수 및 불순도의 최소향상도는 타자인 경우와 동일한 조건을 사용하였으며, 로그변환된 연봉자료에 대한 뿌리마디의 결과는 평균 8.344, 표준편차 0.6276이고 위험 추정치의 값은 0.39111이다.

또한 로그변환을 통한 연봉에 대한 첫 번째 분리는 총승리에 의하여 이루어졌으며 총승리로 분리했을 때 분산의 감소정도가 0.238로서 전체분산의 60%이상을 줄여주고 있음을 알 수 있다. 아울러 나무구조그림으로부터 총승리, 방어율, 총피안타, 볼넷, 총방어율 순으로 로그변환된 연봉에 대해 영향력을 갖고 있음을 알 수 있었다.

3) 의사결정나무분석 후의 신경망분석

변수 중에서 데드볼, 완투, 완봉이 제외된 30개의 뉴런이 입력층에 사용되었고 클레멘타인의 출력결과 첫 번째 은닉층에는 3개의 뉴런, 출력층은 1개의 뉴런으로 구성되었다. 아울러 신경망모형의 예측정확도는 94.389%이고 투수연봉에 대한 의사결정나무분석 후의 신경망분석에서는 총승리(0.325), 총삼진(0.323), 승리(0.260) 등의 순서로 중요도를 가졌다.

4) 신경망분석 후의 의사결정나무분석

정지규칙, 최대 트리 깊이, 부모마디와 자식마디의 개수 및 불순도의 최소향상도는 타자인 경우의 의사결정나무분석와 동일한 조건을 사용하였으며, CART 알고리즘을 이용하여 시행한 결과 신경망분석을 통한 예측연봉자료에 대한 뿌리마디의 결과는 평균 8.3451, 표준편차 0.5996이었다.

신경망분석 후의 의사결정나무분석 결과 최종 나무구조그림으로부터 예측연봉에 대한 첫 번째 분리는 총승리에 의하여 이루어졌음을 알 수 있었으며 총승리로 분리했을 때 분산의 감소정도가 0.2388로서 전체분산의 66% 이상을 줄여주고 있음을 확인할 수 있었다. 또한 총승리, 삼진, 총삼진, 총방어율, 총세이브 순서로 예측연봉에 대해 영향력을 갖고 있음을 알 수 있었다.

5) 회귀분석

로그변환을 한 투수연봉을 y로 두고, 나머지 변수를 독립변수로 하여 단계선택방법을 이용하

여 추정된 중회귀모형(결정계수=0.849)은 다음과 같다.

$$\hat{y}=7.673+0.0007\times\text{총실점}+0.0221\times\text{세이브}+0.0774\times\text{승리}+0.0016\times\text{총경기수}-0.208\times\text{완봉}$$

6. 모형에 대한 비교 및 결과

5가지 모형을 통해서 각 모형에 대한 예측값들을 구하였고, 종속변수가 연속변수인 경우에 최적모형의 선택기준으로 가장 보편화되어 있는 두 가지 판정기준인 평균절대편차와 평균제곱오차를 통해 가장 예측력이 뛰어난 모형을 찾고자 하였다. 이 경우 i-번째 선수의 연봉을 y_i, i-번째 선수 연봉의 예측값을 $\hat{y}_i$ 라고 두면, 총관측치의 수가 n인 경우에 평균절대편차(MAD)는 $MAD = \sum_{i=1}^{n} |y_i - \hat{y}_i|/n$, 평균제곱오차(MSE)는 $MSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/n$ 으로 각각 정의된다.

표 2. 추정된 예측값의 평균절대편차와 평균제곱오차 (단위: 백만원)

	판단기준	신경망분석	의사결정나무분석	나무분석 후 신경망분석	신경망분석 후 나무분석	회귀분석
타자	MAD	7.643*	8.032	8.363	11.556	14.938
	MSE	14121.627*	17814.360	17057.090	31736.690	54884.847
투수	MAD	6.991	4.916*	8.044	7.819	10.430
	MSE	13957.917	6951.911*	17592.562	18339.091	34041.782

표 2는 적합된 모형들에 대한 두 가지 평가척도들의 값이다. 타자연봉의 경우는 값을 구해본 결과, 신경망분석의 MAD와 MSE 값이 가장 작다. 그리고 투수연봉의 경우는 MAD와 MSE 값이 의사결정나무분석인 경우에 가장 작게 나타났다. 이 경우 분류규칙을 이용한 데이터마이닝 방법들은 전통적인 통계분석 방법인 회귀분석보다 훨씬 예측의 질을 높이며, 전처리 과정을 거친 데이터마이닝 방법은 그다지 효율성이 높지 않음을 확인할 수 있었다.

또한 상관분석을 이용하는 경우에도 타자인 경우에 선수들의 연봉과 예측연봉의 상관계수 값은 신경망분석 모형이 0.985, 의사결정나무분석 모형이 0.982, 의사결정나무분석 후의 신경망분석 모형이 0.982, 신경망분석 후의 의사결정나무분석 모형이 0.971, 회귀분석 모형이 0.945로 신경망분석 모형이 가장 바람직한것으로 나타났으며, 투수인 경우에는 상관계수의 값이 신경망분석 모형이 0.962, 의사결정나무분석 모형이 0.980, 의사결정나무분석 후의 신경망분석 모형이 0.956, 신경망분석 후의 의사결정나무분석 모형이 0.953, 회귀분석 모형이 0.921로 의사결정나무분석 모형의 상관계수가 가장 큰 값을 나타내었으며 상관분석의 결과는 MAD와 MSE의 판정기준에 따른 모형선택과 같은 결론을 내릴 수 있었다.

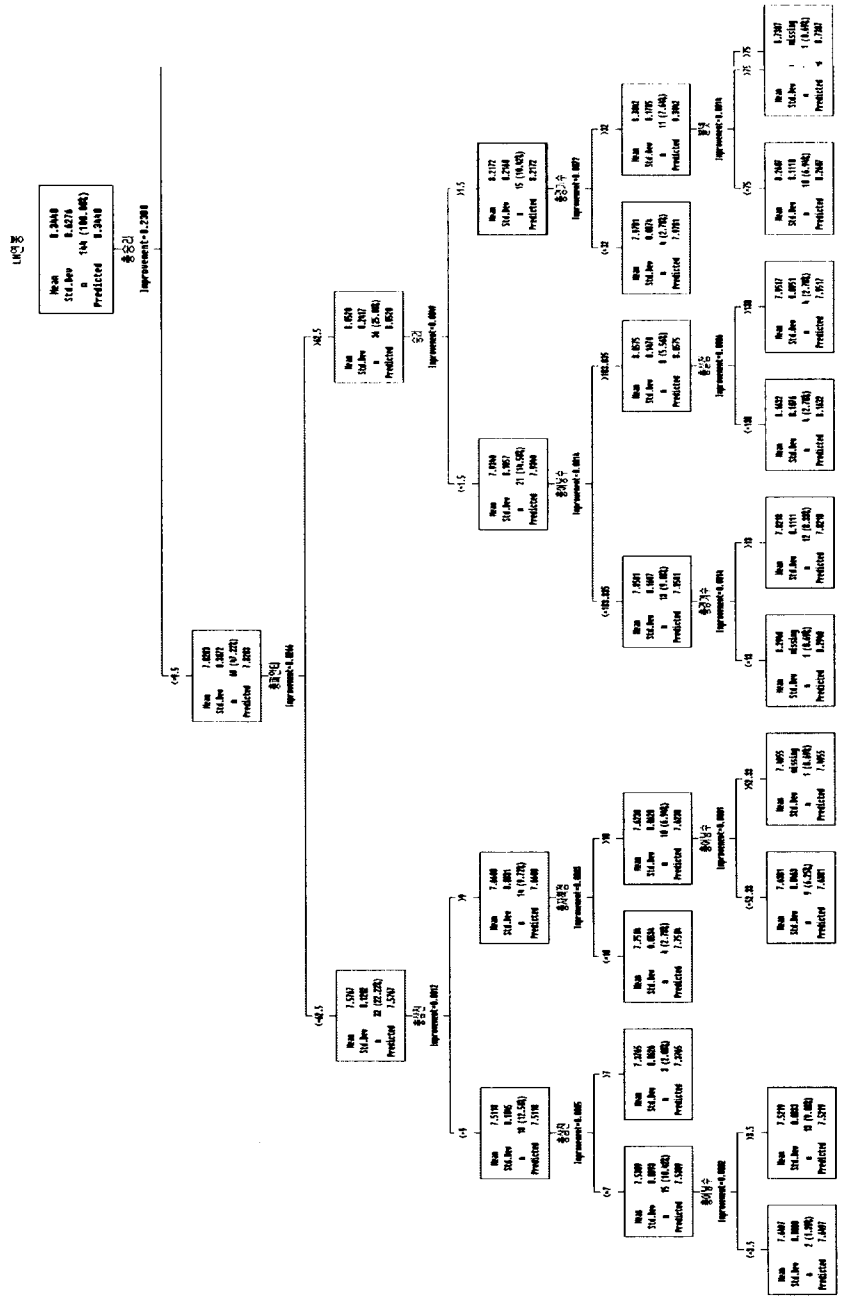

그림 3. 투수연봉 의사결정나무분석의 최종 나무구조

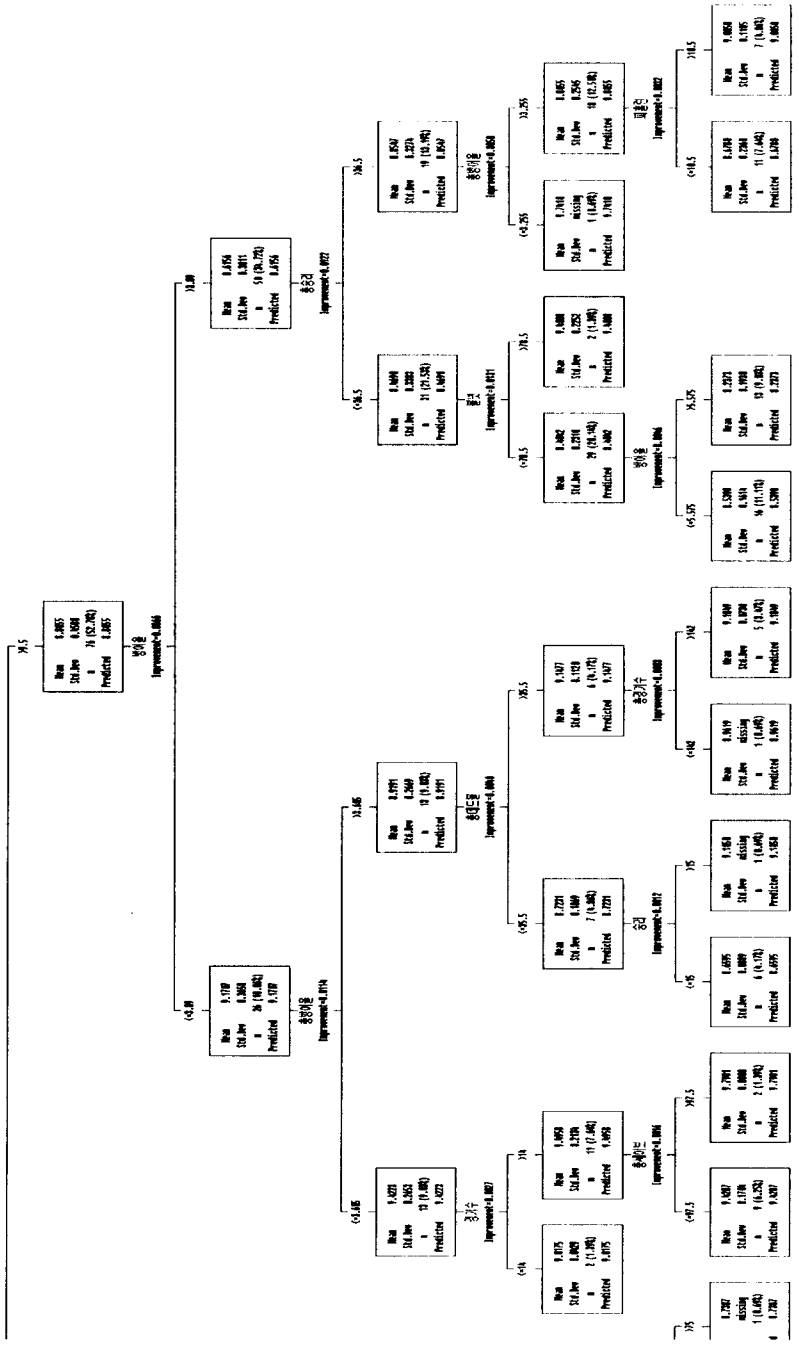

그림 3. 투수연봉 의사결정나무분석의 최종 나무구조

이상을 종합하면 2001년 한국프로야구선수들의 연봉에 관한 모형은 타자인 경우는 신경망분석이, 투수인 경우는 의사결정나무분석이 가장 적절함을 알 수 있다. 신경망분석은 모형의 구체적 모양을 제시하지 않기 때문에 본 연구에서는 결과를 싣지 못하였으며, 반면 투수모형에 대한 의사결정나무 분석결과는 그림 3과 같다. 두 개의 그림으로부터 가장 큰 분류기준으로는 총승리에 의하여 이루어졌음을 알 수 있다. 총승리로 분리했을 때 분산의 감소정도가 0.238로서 전체분산의 60%이상을 줄여주고 있음을 알 수 있고 다음 순으로 방어율이 0.0366, 총피안타가 0.0266으로 각각 9%, 6%이상을 줄여주고 있음을 알 수 있다. 총승리의 수가 9.5를 초과하고 방어율이 3.89를 초과하면 총승리에 의해 분리가 발생하며 방어율이 3.89 이하이면 총방어율에 의해 분리가 발생한다. 한편 총승리수가 9.5 이하이면 총피안타에 의해 분리가 발생하고 총피안타가 62.5를 초과하면 승리에 의해, 62.5 이하이면 총삼진에 의해 분리가 발생함을 알 수 있다. 따라서 그림 3을 통해서 최고연봉을 받기 위한 투수의 조건은 총승리, 방어율, 총방어율, 경기수, 세이브 수의 순서로 능력을 인정받아야 됨을 알 수 있다.

Ⅳ. 결 론

본 논문은 한국프로야구선수의 경기력과 연봉 간에 있어서 회귀분석 및 데이터마이닝 기법을 통해 가장 적절한 통계적 모형을 찾아 그 관계를 규명하는데 목적이 있었다. 그 결과 한국프로야구 타자연봉은 신경망분석, 투수연봉은 의사결정나무분석이 가장 적절한 예측값을 제시하는 모형임을 알 수 있었다.

비록 신경망분석이나 의사결정나무분석을 수행할 때 요구되는 여러 가지 조건의 변화에 따라 본 논문의 결과와 상반된 결론이 나올 수도 있지만 제안된 모형들을 이용하여 구한 연봉예측값과 실제데이터 사이의 상관계수들이 타자와 투수 모두 0.98 이상이기 때문에 한국프로야구 선수들이 받는 연봉은 많고 적음을 떠나서 경기수행능력을 잘 반영한다고 볼 수 있으며, 모형의 개선도는 기껏해야 1% 정도인 까닭에 따라서 본 논문에서 제안된 모형이면 설명력이 충분하다고 볼 수 있다.

또한 관중동원능력, 미래잠재능력과 같이 연봉에 충분히 영향을 미칠 수 있으나 측정 불가능한 변수들이 존재하기에 프로야구선수의 경기력과 연봉과의 관계를 본 논문의 분석만으로는 완전하게 규정짓기 어려울 것으로 간주되어지지만 본 연구의 결과는 구단과 선수의 연봉협상과정에서 발생될 수 있는 갈등을 줄여 각각의 입장에서 객관적 자료를 제시할 수 있는 근거를 마련하는데 기초를 제공할 수 있다고 믿어진다. 이 논문에서 제시한 모형을 가지고 향후 데이터에 적용하는 문제와 프로야구 선진국인 메이저리그의 데이터와 비교해 보는 일도 의미 있는 일이라 하겠다.

참고문헌

김용식(1998). 한국 프로야구선수의 연봉정산 모형. 미간행 박사학위논문. 성균관대학교 대학원.

김용식(2001). 한국 프로야구선수의 경기력과 연봉과의 관계. 한국스포츠사회학회지, 14(1), 15-24.

윤여관(1990). 한국 프로야구 연봉자료에 관한 통계적 분석. 미간행 석사학위논문. 고려대학교 대학원.

윤춘식(1994). 한국프로야구 연봉에 관한 확률모델 개발에 대한 연구. 미간행 석사학위논문. 단국대학교 대학원.

이군호(2000). 한국 프로야구 선수의 연봉결정에 관한 분석. 미간행 석사학위논문. 단국대학교 대학원.

이장영, 강효민(2001). 한국 프로야구 투수의 경기수행과 연봉책정의 관계. 한국스포츠사회학회지, 14(1), 115-124.

최종후, 한상태, 강현철, 김은석, 김미경(2002). AnswerTree 3.0을 이용한 데이터마이닝 예측 및 활용. 서울: SPSS아카데미.

허준, 최병주(2001). 클레멘타인을 이용한 데이터마이닝. 서울: SPSS아카데미.

2001 한국프로야구 연감(2001). 서울: 한국야구위원회.

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

Estensen, P. S.(1994). Salary determination in major league baseball: Classroom exercise. *Managerial and Decision Economics*, 15.

Holbrook, M. B. & Shultz, C. Z.(1996). An updating model of salary adjustments in major league baseball: How much is a home run worth?. *Journal of Sport Management*, 10, 2, 131-148.

Marburger, D. R.(1994). Bargaining power and the structure of salaries in major league baseball. *Managerial and Decision Economics*, 15.

ABSTRACT

A Model Study on Salaries of Korean Pro-Baseball Players Using Data Mining

Oh, Kwang-Mo·Lee, Jang-Taek(Dankook university)

The structure of Korean pro-baseball salaries is of considerable interest to teams, players, and fans alike. The purpose of this paper is to analyze the relationships between 2001 annual salaries of Korean pro-baseball players and the players performance. We suggest a predictive model of Korean pro-baseball salaries in terms of single season and career performance statistics. Data mining techniques using SPSS, clementine, and answer tree are used to display the results of the model - building process. Hitters and pitchers were studied separately. In conclusion, neural network analysis model is most appropriate for batter's annual salary and decision tree analysis model is most suitable for pitcher's annual salary under criterion of MAD and MSE after building statistical model.

· 논문제출일 : 2003. 10. 14
· 논문심사일 : 2003. 10. 28
· 심사완료일 : 2003. 11. 17

keyword: regression analysis, neural network analysis, decision tree analysis, data mining