
저자 (Authors)	김준우, 조다설 Jun-Woo Kim, Da-Seol J
출처 (Source)	한국컴퓨터정보학회 학술발표논문집 26(1) , 2018.1, 241-242(2 pages) Proceedings of the Korean Society of Computer Information Conference 26(1) , 2018.1, 241-242(2 pages)
발행처 (Publisher)	한국컴퓨터정보학회 The Korean Society Of Computer And Information
URL	http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07303239
APA Style	김준우, 조다설 (2018). 데이터마이닝 기법을 이용한 프로야구 경기 승패 예측. 한국컴퓨터정보학회 학술발표논문집, 26(1), 241-242
이용정보 (Accessed)	DGIST 114.71.101.*** 2020/08/11 12:56 (KST)

저작권 안내

DBpia에서 제공되는 모든 저작물의 저작권은 원저작자에게 있으며, 누리미디어는 각 저작물의 내용을 보증하거나 책임을 지지 않습니다. 그리고 DBpia에서 제공되는 저작물은 DBpia와 구독계약을 체결한 기관소속 이용자 혹은 해당 저작물의 개별 구매자가 비영리적으로만 이용할 수 있습니다. 그러므로 이에 위반하여 DBpia에서 제공되는 저작물을 복제, 전송 등의 방법으로 무단 이용하는 경우 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

Copyright Information

Copyright of all literary works provided by DBpia belongs to the copyright holder(s) and Nurimedia does not guarantee contents of the literary work or assume responsibility for the same. In addition, the literary works provided by DBpia may only be used by the users affiliated to the institutions which executed a subscription agreement with DBpia or the individual purchasers of the literary work(s) for non-commercial purposes. Therefore, any person who illegally uses the literary works provided by DBpia by means of reproduction or transmission shall assume civil and criminal responsibility according to applicable laws and regulations.

데이터마이닝 기법을 이용한 프로야구 경기 승패 예측

김준우*, 조다설[○]

^{○*}동아대학교 산업경영공학과

e-mail: kjunwoo@dau.ac.kr*, ektjf264@naver.com[○]

Predicting Win-Loss of Professional Baseball Game by Using Data Mining Techniques

Jun-Woo Kim*, Da-Seol J[○]

^{○*}Dept. of industrial and Management Systems Engineering, Dong-A University

● 요약 ●

야구 관람객들은 주로 자기가 선호하는 팀의 경기나 이길 가능성이 높은 경기를 관람하고자 한다. 때문에 시중에 지난 경기, 당일의 경기, 미래 경기에 대한 정보를 얻을 수 있는 KBO 사이트와 경기 승/패를 예측하기 위한 정보를 얻을 수 있는 사이트에서 경기 기록에 대한 정보를 얻어 관람 일을 결정하는데 도움을 얻는다. 따라서 본 연구에서는 데이터마이닝을 통하여 프로야구 팬들이 특정 팀의 승/패를 예측하는데 사용할 수 있는 유용한 규칙과 패턴을 도출해보고자 한다.

키워드: 데이터마이닝(data mining), 야구(baseball), 분류(classification)

I. Introduction

야구 관람객들이 승/패를 위해 찾아보는 정보는 주로 홈팀 상대팀의 최근 서로 팀에 대한 전적, 선발 투수의 최근 경기 내용(투구수, 방어를 등), 타자 들의 최근 경기 내용(타율, 안타율, 장타율 등) 등이다.

따라서, 본 연구에서는 국내 프로야구팀 중 L팀의 경기 승패를 클래스로 하는 분류 분석을 실시하기 위해 과거 경기별로 각각의 변수들에 대한 데이터를 수집, 데이터 전처리 후 weka[1]를 통하여 유의한 변수를 선정하고, 데이터마이닝[2] 기법 중 의사결정나무, 단순 베이저안 분석, 인접 이웃 분류기 등을 적용하여 가장 성능이 높은 방법을 선별해보고자 한다.

2) 날씨 변수

날씨 데이터의 경우 날씨, 풍속, 기온 총 3개의 변수를 선정하여 데이터를 수집 하였다.

III. Classification Analysis

2016년 경기를 대상으로 수집한 분석용 데이터에는 데이터마이닝 기법 중 범주형 목표 변수의 값 추정에 사용되는 분류 기법들을 적용하였으며, 세부적으로는 의사결정나무, 단순 베이저안 분류기 및 인접이웃분류기를 이용한 분석 결과를 소개하고자 한다.

II. Dataset

본 연구에서는 특정 프로야구 팀의 경기 승패에 영향을 주는 요인으로 크게 경기력(상대팀, 홈팀의 선발 투수에 대한 정보) 및 날씨의 두 가지 영역의 예측변수들을 사용하였으며, 이러한 원시 데이터는 KBO 웹사이트 및 기상청 자료 개방 포털 사이트를 통해 수집하였다. 세부적으로 각 영역에 포함되는 변수들은 다음과 같다.

1) 경기력 변수

경기력의 경우 상대팀/홈팀 선발 투수, 평균 이닝, 투구 수, 피안타, 파홈런, 삼진, 4사구, 평균 자책점, 날짜, 구장, 경기 시간, 관중 수 총 19 개의 데이터를 변수로 설정하여 데이터를 수집 하였다.

1. 의사결정나무

Weka를 이용하여 C4.5알고리즘을 적용한 결과, 적중률 약 70%의 트리를 획득할 수 있었으며, 이 트리의 구조는 Fig.1에서 확인할 수 있다. 이를 보면, 첫 번째 분기 노드는 상대 선발 투수의 평균 이닝이 선정되었으며, 이닝 소화력이 낮은 상대 선발을 만난 경기에서 승리할 확률이 높은 점을 볼 수 있다. 아울러, 두 번째 분기 노드는 자팀 선발 투수의 평균 이닝으로 상대 선발의 이닝 소화력이 높더라도 자팀 선발의 이닝 소화력이 높으면 역시 승리할 가능성이 크게 나타났다.

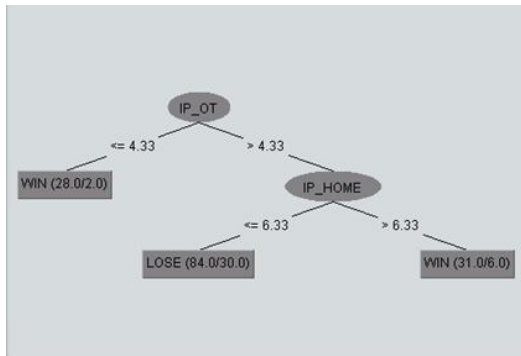


Fig. 1. Decision Tree

다만, 본 연구는 프로야구팀의 경기력으로 투수의 역량만을 예측변수에 반영하였다는 한계를 갖는다. 이에, 저자들은 향후 팀의 타격 역량까지를 고려하는 보다 정교한 분류 모델을 개발하고자 한다.

REFERENCES

- [1] Weka 3; Data Mining Software in Java, <http://www.cs.waikato.ac.nz/ml/weka>
- [2] P.N. Tan, M. Steinbach and V. Kumar, "Introduction to Data Mining," Addison-Wesley, 2005.

2. 단순 베이저안 분류기

동일한 데이터에 단순 베이저안 분류기를 적용했을 경우, 적중률 약 60%의 분류 모형이 얻어졌으며, Fig. 2는 Weka를 이용한 분석 결과를 보여준다.

```

=== Summary ===
Correctly Classified Instances 95      56.4336 %
Incorrectly Classified Instances 48      33.5664 %
Kappa Statistic 0.3519
Mean absolute error 0.2312
Root mean squared error 0.3956
Relative absolute error 68.1794 %
Root relative squared error 96.3866 %
Total Number of Instances 143

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.713	0.365	0.713	0.713	0.713	0.347	0.754	0.791	WIN
	0.000	0.000	0.000	0.000	0.000	0.000	0.145	0.014	LOSE
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	DRAW

```

=== Confusion Matrix ===
 a b c <-- classified as
57 23 0 | a = WIN
23 38 0 | b = LOSE
0 2 0 | c = DRAW

```

Fig. 2. Neighborhood Classifier Result (KNN=21)

3) 인접 이웃 분류기

2016년도 데이터에 인접 이웃 분류기를 적용했을 때는 의사결정나무와 유사하게 약 70%의 적중률이 산출되었으며, Fig. 3는 세부적인 분석 결과이다.

```

=== Summary ===
Correctly Classified Instances 101      70.1389 %
Incorrectly Classified Instances 43      29.8611 %
Kappa Statistic 0.3951
Mean absolute error 0.463
Root mean squared error 0.4759
Relative absolute error 93.2077 %
Root relative squared error 95.4882 %
Total Number of Instances 144

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.636	0.244	0.689	0.636	0.661	0.396	0.693	0.646	WIN
	0.756	0.364	0.711	0.756	0.733	0.396	0.693	0.668	LOSE

```

=== Confusion Matrix ===
 a b <-- classified as
42 24 | a = WIN
19 59 | b = LOSE

```

Fig. 3. Neighborhood Classifier Result (KNN=21)

IV. Conclusions

데이터에 대한 분석 결과 전체적으로는 인접 이웃 분류기의 적중률이 가장 높았으며, 의사결정나무 분석에서는 자팀과 상대팀 선발 투수의 역량이 경기의 승패에 가장 많은 영향을 주는 것으로 나타났다.