

프로야구 승/패 예측 SVM과 소자승법(등급결정모형) 적용

프로야구 승/패 예측 SVM과 소자승법(등급결정모형)적용

: 2009년 자료라 참고만!

요약)

- 프로야구 경기 예측을 위해 2가지 모델을 적용하는 시스템 개발
 - a. 데이터 마이닝 기법 중 하나인 svm 모델을 적용함.
 - b. 최소승자법 이용해 등급을 결정함.
- 각 팀의 경기력을 8가지 요소로 수치화

1. 프로젝트 개요

1) SVM 모델

: 최적화 기법에 바탕을 둔 분류 기법

: 프로야구 승패예측에 적용 (각 팀 당 70경기에서 얻어진 데이터를 바탕으로 215 경기 예측)

2) 등급결정 모형

: 최소자승법에 의한 rating 결정 모형 (106 경기에서 얻어진 등급을 바탕으로 145 경기 결과 예측)

1. 데이터 구성

1) 요소 (8가지)

- 장타율 = 안타수/타수
(누타수 = 단타 + (2루타*2)+(3루타*3)+(홈런*4))
- 타율 = 안타수/타수
- 출루율 = (안타수+베이스온볼스) / (타수+베이스볼스+희생플라이)
- 방어율 = 자책점/투구이닝수
- 사사구 = 베이스온볼수 + 4구 / 투수이닝수
- 피안타율 = 피안타수/타수
- 득점 = 득점/타자이닝수
- 탈삼진 = 탈삼진/투구이닝수

2) 분리초평면

(홀팀기준)

데이터 = 홀팀 요소 값 - 원정 팀 요소 값

$X_i = X_{ai} - X_{bi}$

분리초평면

구분	장타율	타율	출루율	득점	방어율	사사구	탈삼진	피안타
홈	Xa1	Xa2	Xa3					Xa8
원정	Xb1	Xb2	Xb2					Xb8

	원정	Xb1	Xb2	Xb2	Xb8
(데이터)	결과	X1	X2	X3		X8
(변 수)	계수	W1	W2	W3		W8

분리 초평면 계수(w_i)가 최적화모형에 의해 결정되면 분리초평면으로 구분되는 한쪽구역($(Y=1)$ 은 승), 다른 쪽 구역($(Y=-1)$ 은 패)

분리초평면을 결정하는 최적화 모형(SVM)

Minimize $||w||^2$

Subject to $y_i (w x^i + b) - 1 \geq 0, i=1,2,\dots,k$

여기서 data의 요소가 8개 이므로
 $w=(w_1, w_2, \dots, w_8)$ 분리초평면 계수

학습 data 사용된 경기 수가 281개인 경우
 $x_i=(x_1^i, x_2^i, \dots, x_8^i), i=1,2,\dots,281$
 $y_i=1$ or -1 (학습데이터 사용된 경기 결과), $i=1, \dots, 281$

: 이 부분 이해 잘 못함.