야구 기록과 뉴스 데이터를 활용한 타격 예측 모델

- **야구 기록과 뉴스 데이터를 활용한 타격 예측 모델**

요약)
정량지표 뿐 아니라 정성 지표를 함께 활용하여 타격 예측 모델 구축
기계학습을 통해 분류된 뉴스데이터를 활용한 정성적인 지표 제안 -> 타율의 상승하강 여부를 분류하는 모델을 제안하기

1. 서론

1) 관련 이론

- 세이버 메트릭스 : 게임이론과 통계학적 방법론을 도입한 객관적인 평가 지표
  But, 게임의 상황에 대한 예측은 불가능함.
- SVM (Support Vector Machine) : 데이터의 카테고리를 판단하는 분류 모델, 분류/회기 석에 사용됨.

2) 본 연구의 필요성

- 야구기록, SVM을 사용한 뉴스 데이터 분석을 통해 타자의 타율 상승하강을 예측할 수 있는 분류 모델 제안
- 기록 (객관적 지표) 이외의 뉴스 등 정성적인 지표도 추가 => 분류기의 정확도 올리기
  => 기존의 통계가 보여주지 못한 당일 타자의 타격 예측결과를 도출
  => 뉴스에서 나타나는 선수 및 팀의 상태, 흐름이 경기에 미치는 영향 확인해보기

2. 관련 연구

1) 로지스틱 회귀분석 모델

: 가장 안타를 칠 확률이 높은 타자를 예측하는 연구

- 입력값 : 최근 몇 시합에서의 타율, 최근 몇 시합에서 타자가 공을 배트에 맞춘 평균 횟수, 투수의 최근 피안타율, 타자의 경기장에 따른 타율

2) SVM

: 분류와 회귀분석에 사용되는 지도학습 모델
: 주어진 집합을 바탕으로 두 개의 카테고리를 나누는 선형 분류 모델 형성
: 지도학습 모델, 분류기준은 사용자가 정함.

3) TF-IDF

: 텍스트마이닝에서 사용되는 가중치
: 해당 값이 높을수록 '특정 문서 내'에서 단어 빈도수가 높고 '전체 문서들 중'에서

는 그 단어를 포함한 문서가 적은 단어(흔하지 않은 단어)로 판단할 수 있음.

3. 모델 제안

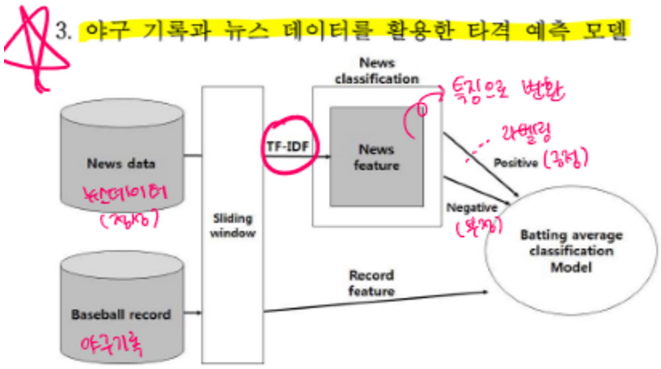

그림 1. 분류를 사용한 타격 예측 모델 구조

1) 뉴스데이터와 야구기록을 슬라이딩 윈도우 크기만큼 가져오기
2) 수집된 뉴스데이터는 TF-IDF값을 사용하여 특징으로 변환됨.
(유의미한 단어들을 추출, 단어들에 가중치를 부여)

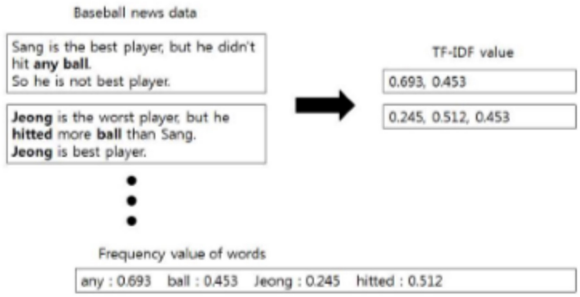

그림 2. TF-IDF를 사용한 뉴스 특징 분석

3) 긍정, 부정 라벨링 (긍정 : 1, 부정 : -1)
4) SVM 분류 모델 만들기
   a. 뉴스 분석 결과 : 뉴스가 나온 날짜의 긍정, 부정 뉴스 개수
   b. 야구 기록
      : 타자의 최근 경기 데이터 사용 (타자의 타격확률에 가장 큰 영향)
      : 최근 경기의 타율 평, 타석 수, 안타, 홈런 개수, 장타율, 출루율

=> 시간에 따라 변화하는 지표이기 때문에, n일 단위로 슬라이딩 윈도우 만들기
=> '정해진 시간'만큼의 뉴스 데이터 분석 결과, 기록 : 독립변수
매일의 타율 : 종속변수
=> 시간의 변화에 따른 타자의 타율 변화 고려 가능

4. 결론

1) 이전 연구와의 차이점
: 이전에는 단순히 투수의 피안타율, 경기장 별 타자의 타율과 같은 기록 사용
하지만, 선수의 타율에 더욱 직접적으로 영향을 주는 것은 '최근 몇 경기 동안의 타자의 기록', '팀과 선수의 환경/상태변화'
=> '정성 요소'를 야구 뉴스 데이터를 활용해 추가