빅데이터 입문
Term Project

체스 rating과 흑백 승리 여부

목차

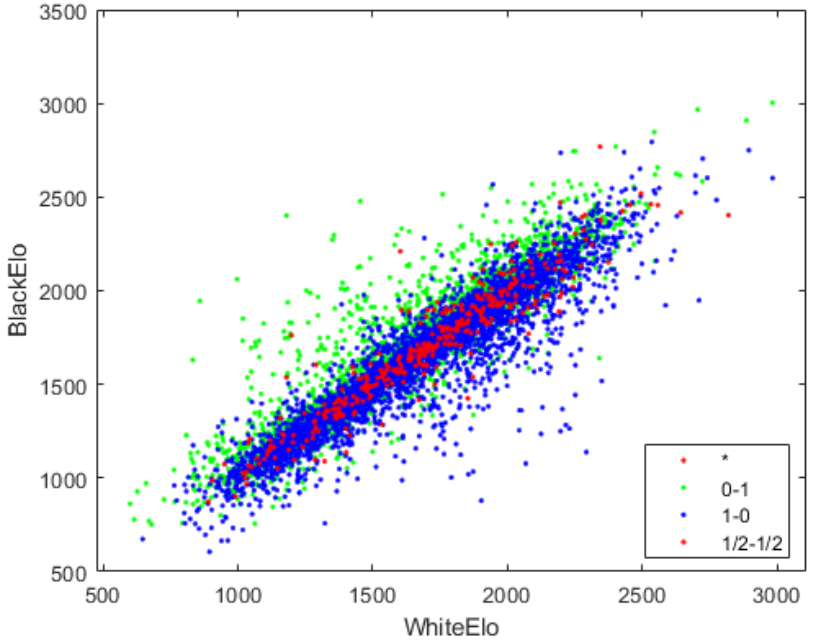
- 요약 및 이전연구
- 실험
- 빅데이터 처리
- 결론

요약 및 이전연구

- 요약
  - 체스의 실력을 나타내는 척도 중 하나인 ELO rating(게임에서 사용되는 MMR과 유사) 등을 이용해 흑, 백 중 승자를 예측
  - 데이터는 kaggle의 5-million-chess-game-results-november-2019를 사용
  - 데이터 패턴을 학습할 모델은 딥러닝이 아닌 기계학습을 사용
- 이전연구
  - 머신러닝 모델의 학습 시간이 길어 데이터 전체의 개수와 특징의 개수를 줄인 결과 데이터의 정확도가 낮게 나옴
  - 향후 과제로 사용언어변경, 특징추가, 모델변경, 데이터 시각화 등을 언급

데이터 시각화

- 기존에는 사용언어변경, 특징 추가 등을 우선순위로 두었으나 현 데이터의 특징이 2차원이라는 점, matlab의 장점이 데이터 시각화라는 점과 앞의 문제를 시각화를 통해 해결할 수 있다는 점 등의 이유로 데이터 시각화를 먼저 진행
- 데이터 시각화 결과 실제로 WhiteElo, BlackElo 두가지 특징으로는 예측하는 것이 힘들 것으로 예상 <- 데이터의 특징을 추가



특징추가

- 기존의 데이터에서 사용하려고 했던 WhiteRatingDifference, BlackRatingDifference, ECO(체스의 오프닝)을 특징으로 포함해 모델 학습 및 예측
- 데이터는 특징만 추가하고 이전과 마찬가지 방법으로 같은 수만큼 추출
- 모델의 검증 또한 마찬가지로 정확도 측정

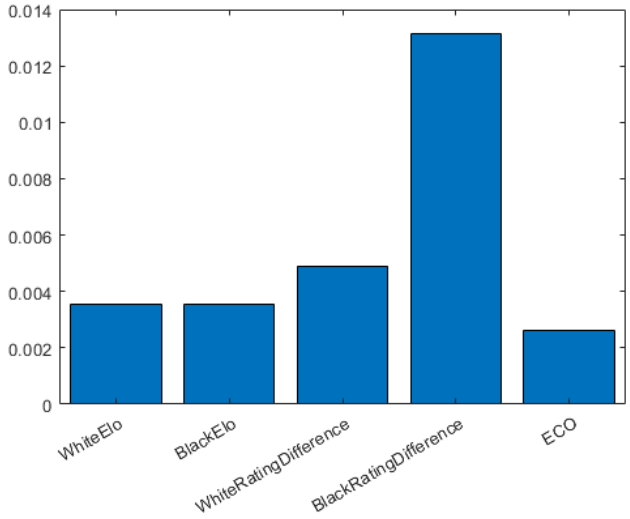
```
svm_result =        RF_result =

    0.9790              0.9890
```

- 이전 0.512, 0.555였던 정확도가 0.98, 0.99로 급격하게 상승함

특징 중요도 확인

- Matlab에서 제공하는 중요도 예측을 통해 특징 중요도를 확인
- 실제로 ELO Rating만으로는 예측하는 것이 힘들지만 예상외로 BlackRatingDifferece가 큰 영향을 미침
- Kaggle에도 RatingDifferece에 대한 자세한 설명은 나와있지 않아 최초에는 WhiteElo – BlackElo, BlackElo – WhiteElo 값으로 생각했음
- 아마 경기 후 흑과 백의 Elo Rating 변화값으로 예상됨
- 만약 경기 결과로 인한 변화라면 사실상 승패가 포함되어 있음 (BlackRatingDifferece가 양수라면 흑승, 음수라면 백승)



빅데이터 적용

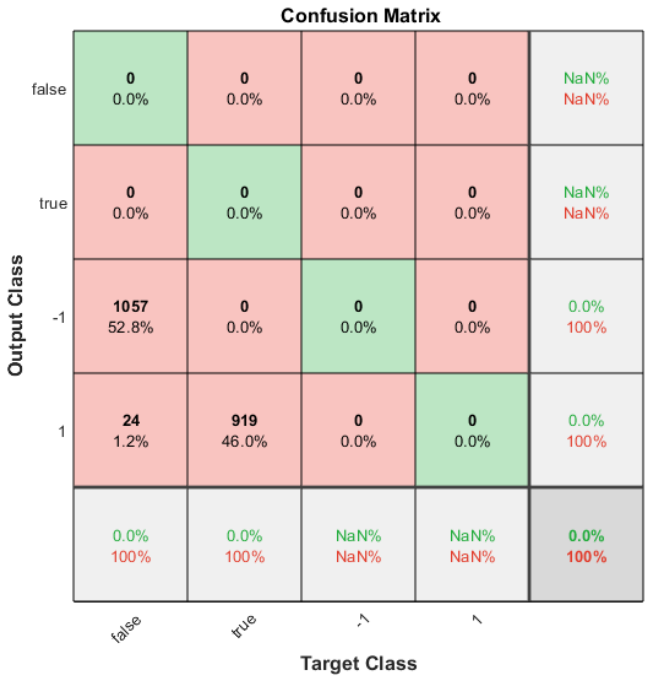
- 빅데이터의 모델은 SVM을 기반으로 한 SGD를 사용했고 학습에는 테스트 데이터 2000개를 제외한 500만개의 데이터 전체사용
- 테스트 데이터는 500만개의 데이터 중 2000개 무작위 복원추출(앞선 방법과 동일)
- Epoch는 100회 진행
- 기존에는 흑승, 백승, 무승부 모두 나눠 학습했지만 무승부의 비율이 작고 학습 시간의 간소화를 위해 흑승과 흑승이 아닌것 두가지로 나눠 구분
- 모델 검증은 마찬가지로 정확도 사용
- 언어는 그대로 matlab을 사용하였고 SGD는 툴박스를 받아 사용

빅데이터 결과

- 모델의 툴박스 그대로 정확도를 나타낼 경우 45.95가 나오지만 입력은 True, False로 모델은 -1, 1로 출력하기 때문에 True와 1, False와 -1을 서로 매치해서 확인하면 약 98.8%의 정확도가 나오는 것을 볼 수 있다.

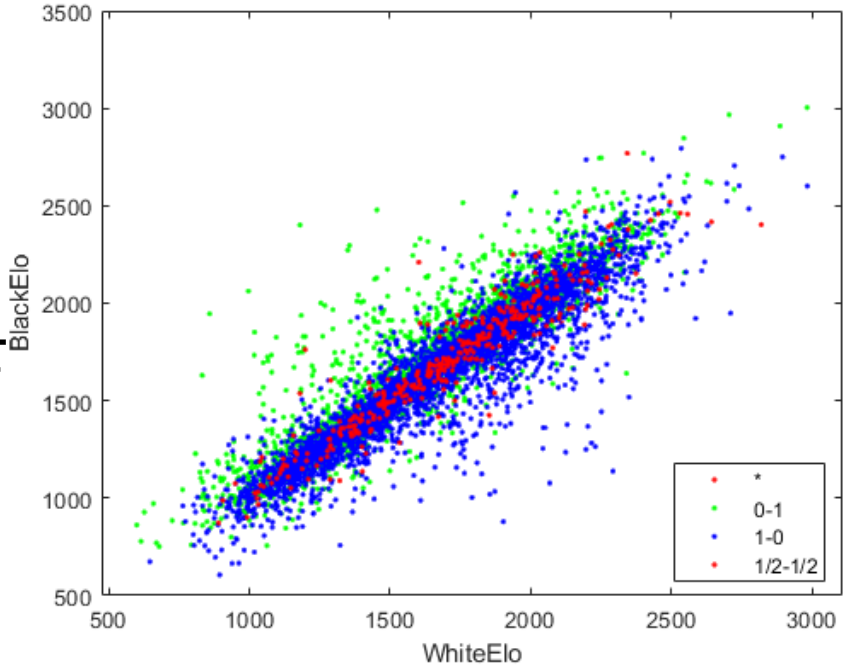
```
test_accuracy =

   45.9500
```



결론

- ELO rating만을 이용해서는 실제로 게임의 승패를 알기 힘들다.
- 실제 체스 경기는 공정한 경기가 치뤄진다는 것을 확인할 수 있다.
- 특징에 대한 자세한 이해가 부족했던 것이 아쉽다.
- 여기서 더 나아가 클러스터링 등을 통해 부정행위 같은 outlier detection에 활용하는 것이 가능할 것 같다.
- 오히려 흑이 승리한 게임이 의외가 많다.



참고자료

- 최대진 교수님 빅데이터 입문 수업
- Kaggle 5 Millon Chess Game Result – November 2019
  https://www.kaggle.com/timhanewich/5-million-chess-game-results-november-2019
- 매트랩 서포트 벡터 머신 분류
  https://kr.mathworks.com/help/stats/support-vector-machine-classification.html?s_tid=CRUX_lftnav
- 매트랩 분류 앙상블
  https://kr.mathworks.com/help/stats/classification-ensembles.html?s_tid=CRUX_lftnav
- 기타 매트랩 툴박스 매뉴얼 등