로지스틱 회귀와 서포트벡터머신의 머신러닝 알고리즘 성능 비교 실험

오윤태, *김상철
국민대학교 소프트웨어학부, *교신저자

webking91@naver.com, sckim7@kookmin.ac.kr

Comparative experiment of Machine Learning Performance of Logistic Regression and Support Vector Machines

Yoon-Tae Oh, *Sang-Chul Kim

School of Computer Science, Kookmin University

요 약

본 논문에서는 로지스틱 회귀와 서포트 벡터 머신 2 개의 알고리즘을 선정해 비교 평가하는 실험을 진행하였다. 표본 데이터로 호주의 각 지역 기상관측소에서 다년간 수집된 기상정보를 사용하였고, 내일의 강수 유무를 예측하는 학습을 진행하였다. 표본데이터데 대해 전처리를 진행하였다. 표본 데이터를 사용해 5,000 개, 10,000 개, 30,000 개, 100,000 개의 크기별 데이터를 만들었고, 데이터의 크기마다 학습시간, 분류시간 그리고 분류 정확도를 측정하여 비교 실험하였다.

Ⅰ. 서 론

머신러닝의 알고리즘들은 크게 지도학습과 비지도학습, 강화학습으로 분류가 된다. 본 논문에서는 로지스틱 회귀와 서포트 벡터 머신의 분류기를 제작하여 두 알고리즘을 비교 평가하였다. 각 분류기의 학습시간, 분류시간, 분류정확도를 데이터 크기별로 실험하여 비교 평가하였다. 그 결과를 이용하여 또 다른 데이터들을 분류하는 머신러닝 모델을 만들고자 할 때 알고리즘을 선택하는 기준으로 사용할 수 있는 것을 목적으로 하였다.

Ⅱ. 본론

비교 실험 결과, 로지스틱 회귀 분류기는 학습, 분류시간 모두에서 빠른 성능을 보여주고 있다. 학습시간은 데이터가 커짐에 따라서 늘어나고 있는 모습을 보여준다. 반면 학습을 통해 만들어진 결정 경계를 통해 분류하다 보니, 분류시간은 데이터의 크기가 커져도 크게 증가하지는 않는 현상을 보여준다. SVM 은 초평면을 구성하기 많은 요소를 계산해 분류기를 완성해야 하기에 기본적으로 학습시간이 로지스틱 회귀보다 많이 소요되는 결과가 나온다. 로지스틱 회귀와는 다르게 데이터가 많아질수록 학습시간이 급격히 많이 소요되는 것을 알 수 있다. 분류시간도 학습시간과 동일하게 데이터가 많아질수록 크게 증가하는 모습을 보인다. 데이터 크기별 4 개의 값의 평균을 계산해보면 로지스틱 회귀 84.8%, SVM 84.88%로 SVM 분류기가 조금 더 좋은 정확도를 보여주고 있다. 하지만 차이가 크지 않고, 두 개의 분류기 모두 좋은 성능을 보여주고 있다고 말할 수 있다. 2 개의 분류기 모두 데이터의 크기가 커질수록 정확도가 증가하는 모습을 보여준다

Ⅲ. 결론

비교 실험 결과, 분류기를 학습시키는 시간에서는 로지스틱 회귀가 SVM 보다 월등히 좋은 성능을 보여주고 있다. SVM 알고리즘은 학습시간이 오래 걸리지만, 서포트벡터를 활용해 복잡한 결정 경계를 만들어 과적합문제를 피해 갈 수 있다. 본 논문에서의 실험 결과는 로지스틱 회귀가 SVM 과 비슷한 분류정확도를 보여주면서도 훨씬 빠른 속도로 좋은 성능을 보여주었다. 결론적으로 이항데이터 분류 문제에서는 로지스틱 회귀가 서포트 벡터머신보다 더 좋은 성능을 보여준다. 새로운 표본 데이터를 학습시키기 위해 2 개의 알고리즘 중에서 선택할 때는 실험 결과를 참고하여 우선적으로 데이터의 형태가 선형인지 확인하고, 표본 데이터의 크기와 과적합문제여부에 따라서 선택하면 될 것이다.

참 고 문 헌

[1] 권혁춘, 이병걸, 이창선, 고정우, “로지스틱회귀분석기법과 인공신경망 기법을 이용한 제주지역 산사태가능성분석”, 한국지형공간정보학회지, 2011, pp. 33-40.

[2] A. Gron., “Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems”, O`Reilly Media, 2017.