베이즈 분류기에 의한 신문기사의 정치성향 분류

정영재*, 강태원*
*국립강릉원주대학교 컴퓨터공학과
jungyoungjae9@gmail.com, twkang@gwnu.ac.kr

*이 논문은 2015년 강릉원주대학교 서울어코드활성화사업단의 지원을 받아 발표합니다.

Classifying press article's political tendency using Bayesian classifier

Young-Jae Jung*, TaeWon Kang*
*Dept of Computer Engineering, Gangneung-Wonju National University

요 약

베이즈 분류기(Bayesian classifier)를 이용하여 특정 신문기사가 어떤 정치적 성향을 가지는지 분류한다. 이를 위하여 보수 및 진보 성향으로 알려진 언론사의 보도기사를 수집하여 베이즈 분류기를 학습하였다. 즉, 보수 및 진보적 성향을 갖는 기사에서 출현 빈도가 높은 단어의 빈도수를 확인하여 분류기를 구현하였다. 학습에 사용하지 않은 보수 및 진보적 성향의 기사를 사용하여 분류기의 성능을 검증하였다.

1. 서론

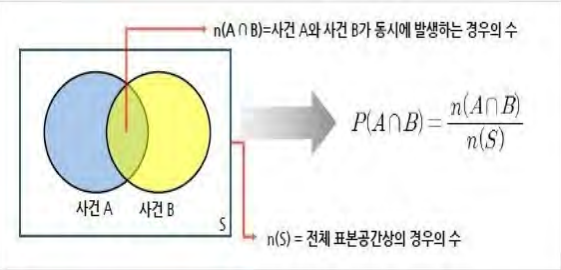
정치적 성향을 말할 때 흔히 보수, 진보라는 이야기를 많이 한다. 보수주의[1]란 백과사전의 정의를 보면 진보주의에 대립되는 개념으로, 일반적으로는 급격한 개혁을 피하고 현재의 체제를 중시하는 사상을 말하고 진보주의[2]란 보수주의에 반대되는 개념으로 주로 이데올로기적인 근대 정치사상의 특정 조류를 가리킨다. 라고 정의 되어있다. 정치적 성향에서 잘못 발전하여 극단적으로 생각하는 극우주의나 극좌주의가 문제를 일으키고 있으며 대표적인 사이트로 '일베'라는 곳이 있다.

인터넷과 스마트폰의 보급에 따라 신문과 뉴스를 통했던 정보가 원하는 시간과 장소에서 얻을 수 있게 되었다. 신문보다는 인터넷 뉴스와 기사를 접하는 것이 현실이다. 이와 같은 상황에서 중립적 정치적 성향을 가지지 못하고 한쪽으로 치우치는 것은 사회문제를 야기할 수 있다.

본 논문에서는 베이즈 분류기를 이용하여 흔히 보수언론과 진보언론의 데이터를 이용하여 특정 기사가 어떤 정치적 성향을 가지는지 분류하여 확인할 것이다.

2. 배경 이론

베이즈 분류기(Bayesian classifier)는 단순한 확률적 분류기를 의미하며 독립가정을 적용한 베이즈 정리에 기반하고 있다. 이는 (그림 1)과 같이 조건부 확률에서 생각해 볼 수 있으며 조건부 확률은 어떤 정보를 미리 알고 있을 때 다른 무언가를 관찰할 가능도(likelihood)를 의미 한다. 본 논문에서는 (그림 1)에서 보수언론(사건 A)에서 빈도수 가 높거나 진보언론 (사건 B)에서 빈도수가 높은 단어를 검색하여 그 차이를 보는 알고리즘을 활용 한다.



(그림 1) 조건부 확률[3]

3. 베이즈 분류기 개발

먼저 분류기 학습 및 시험용 데이터를 모아야 한다. 데이터는 보수언론으로 대표되는 포털 사이트[4]의 기사 데이터와 진보언론으로 대표되는 포털 사이트[5]의 기사 데이터, 중도 성향과 약간의 정치적 성향을 가지는 포털 사이트의 기사 데이터[6]로 구성된다. 학습을 위한 보수, 진보 데이터는 각각 100개, 중도언론의 데이터는 50개, 검증을 위한 데이터 또한 위와 같이 보수, 진보, 중도 데이터 각각 50개 씩 총 400개의 데이터를 확보하였다.

본 논문에서는 각각의 데이터를 포털사이트의 정치면 기사를 통해 수집했으며 이 데이터를 추출하기 위해 프로그래밍 언어 R을 사용하였다. 먼저 보수언론으로 대표되는 포털 사이트의 기사와 진보언론으로 대표되는 포털 사

이트의 기사를 하나의 문자열로 만들어 낸 후 각 데이터에서 나오는 말뭉치들의 빈도수를 확인 하였다. (그림 2)는 기사를 하나의 문자열로 저장한 것을 나타낸다.

(그림 2) 데이터를 하나의 문자형 벡터로 만든 모습

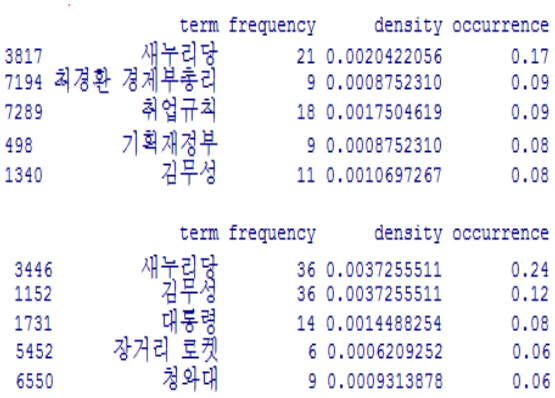
다음 단계는 문자형 벡터에서 텍스트 말뭉치를 만들어 내는 작업이다. 텍스트를 말뭉치로 만든 후, 단어들을 처리해서 분류기를 위한 특정 집합을 구성 한다.

단어 빈도수를 정량화하기 위해 단어 문서행렬(document matrix, TDM)을 구성한다. TDM은 각 단어가 행, 각 문서가 열인 N*M행렬로 [i, j]원소는 단어 i가 문서 j에서 출현한 빈도수를 뜻한다. 말뭉치를 구성하는 옵션으로 불용어(stopword) 455개를 문서에서 제거, 숫자를 제거, 구두점(punctuation) 제거, 말뭉치에서 한 번 이상 나타난 말뭉치들만 고려하는 옵션을 사용했다.

이어서 분류기를 개발할 단계로 보수언론 데이터가 있는 상태에서 각 단어별 발견 확률을 데이터프레임으로 구성한다. 그 후, 학습데이터를 생성하기 위해서 주어진 단어가 나타나는 문서의 비율과 각 말뭉치가 전체 말뭉치에서 나타나는 빈도수를 합산한다. (그림 3)은 각각의 보수언론, 진보언론의 말뭉치 별 빈도수를 계산한 것이다.

3.1 분류기 정의 및 테스트 데이터 검증

학습 데이터세트에 있는 단어와 없는 단어를 자연 언어처리(natural language processing, NLP) 기법을 사용하여 데이터의 성향을 추정한다. 즉, 학습세트에 없는 단어는 매우 작은 확률 값을 가지도록 한다. 본 논문에서는 확률의 기본값을 0.0001%의 값을 주기로 했다. 또한 보수언론 데이터와 진보언론 데이터를 동일하다고 가정하여, 기본 사전 확률을 50%로 가정한다. 하지만 나중에 사전 확률을 바꿀 수 있도록 함수 변경 기능을 추가하였다. 마지막으로 학습 세트의 단어들 중 어떤 단어가 보수언론에 포함되었는지 확인하고, 그 말뭉치를 이용해서 계산한다.



	term	frequency	density	occurrence
3817	새누리당	21	0.0020422056	0.17
7194	최경환 경제부총리	9	0.0008752310	0.09
7289	취업규칙	18	0.0017504619	0.09
498	기획재정부	9	0.0008752310	0.08
1340	김무성	11	0.0010697267	0.08

	term	frequency	density	occurrence
3446	새누리당	36	0.0037255511	0.24
1152	김무성	36	0.0037255511	0.12
1731	대통령	14	0.0014488254	0.08
5452	장거리 로켓	6	0.0006209252	0.06
6550	청와대	9	0.0009313878	0.06

(그림 3) 진보데이터 보수데이터의 문자 빈도 지수

그 결과 아래 (그림 4)와 같은 중도 언론의 데이터가 보수언론 일 확률에 대한 베이즈 추정치가 계산 된다.

```
> summary(test.res)
   Mode   FALSE    TRUE    NA's
logical      21      29       0
```

(그림 4) 중도언론 데이터에 대한 분류기 검증 결과

분류 결과 50개의 데이터 중 21개는 진보언론 데이터로, 29개의 데이터는 보수언론 데이터로 분류하였다. 중도언론의 경우 중립적으로 나오길 원했으니 어느 정도 검증이 되었다고 판단한다.

이어서 보다 정확한 검증을 위해 (그림 5)와 같이 진보언론의 데이터를 검증하여 보았다.

```
> summary(test.res)
   Mode   FALSE    TRUE    NA's
logical      44       6       0
```

(그림 5) 진보언론 데이터에 대한 분류기 검증 결과

분류 결과 긍정 오류율(false-positive rate)이 12%로 결과가 나왔다. 이는 진보언론의 데이터 중 12%가 보수언론 데이터로 식별 된다는 뜻이다.

3.2 모든 데이터 종류에 대한 분류기 검증

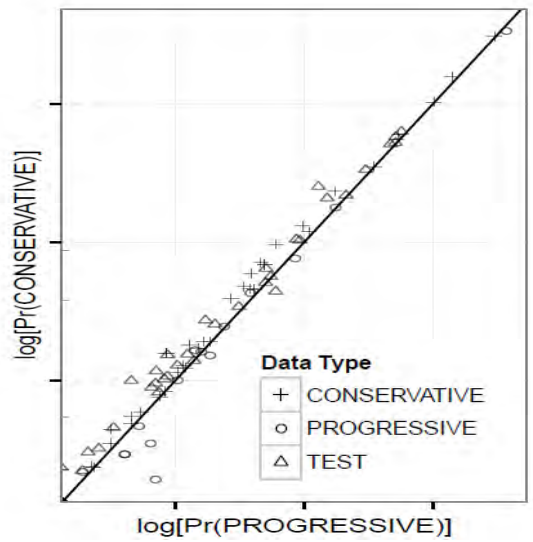
(그림 6)은 3.1에서 했던 확률 비교를 전체 데이터에 적용하는 함수를 만들어 분류기를 검증한 것이다.

```
> print(class.res)
                  PROGRESSIVE CONSERVATIVE
progressive2.col         0.88         0.12
test2.col                0.38         0.62
conservative2.col        0.22         0.78
```

(그림 6) 각각의 데이터에 대한 분류기 검증 결과

각각의 데이터가 보수언론 데이터인지 아닌지 판단하는데 위의 결과를 보면 진보언론의 데이터의 경우 0.88 비율로 로 바르게 분류하고 보수언론의 경우 0.78, 중도언론의 경우 보수언론0.62, 진보언론 0.38로 보수언론으로 분류하고 있다.

(그림 7)은 분류기 성능이 얼마나 되는지 더 알아보기 위해서, x축은 진보언론으로 예측할 확률, y축은 보수언론으로 예측하는 확률인 산포도를 그린 것이다.



(그림 7) log-log 척도에서 정치성향 예측 확률에 대한 산포도

그래프에는 y=x로 표시되는 성형 결정 경계가 그려져 있는데 이 선은 분류기가 보수언론 데이터일 확률과 진보언론 데이터의 확률을 비교하여 데이터의 종류를 판별하기 때문이다. 따라서 (그림 7)의 결정 경계 위의 점들은 모두 보수언론 데이터, 아래쪽의 점은 모두 진보언론의 데이터야 한다. 그러나 그래프를 보면 그렇지 않기도 하지만 대체로 같은 종류의 데이터가 뭉쳐있다.

긍정 오류율과 부정 오류율 문제를 해결하기 위해서 분류기를 수정하여 결과를 확인 하였다. 위의 분류기는 언론데이터가 보수데이터일 경우와 진보데이터일 경우를 같다고 가정하였으나 실제로는 각각의 데이터의 사전확률은 다를 것이다. 즉 위의 분류기는 보수언론 데이터 50%, 진보언론 데이터가 50%인데 실제로 언론사별 데이터는 위와는 다를 것이다. 이에 초점을 맞추어 사전 확률을 변경하여 결과 비율에 반영하도록 하는 개선 방법을 사용하여 결과를 확인 하여 보았으나 데이터가 부족하여 결과에 변화는 없었다(그림 8).

```
> print(class.res)
                  PROGRESSIVE CONSERVATIVE
progressive2.col          0.88         0.12
test2.col                 0.38         0.62
conservative2.col         0.22         0.78
```

(그림 8) 사전확률을 변경한 후 결과

4. 결론

언론은 빠르고 정확하게 정보를 전달한다. 언론 자체가 편향된 정치적 성향을 가지고 있다고 말하진 않지만 어느 정도 성향은 띄기 마련이다. 최근 SNS의 발달을 통해 무분별한 극좌, 극우주의적 뉴스를 접할 기회가 많은데 개개인의 사용자는 이에 충분히 주의를 가지고 기사를 읽어야 할 것이고 기자또한 무분별한 자극적인 기사가 아닌 올바르고 정확한 정보를 전달하는데 노력하면 국내의 언론은 더욱 발전할 수 있을 것이라 생각한다.

정치적 이념과 성향을 단순히 이것 아니면 저것으로 분류하는 이항분류로 데이터를 분류하는 것은 어렵지만 본 논문에서 제작한 분류기는 제법 괜찮은 성능으로 분류를 해내었다. 좀 더 많은 양의 데이터와 우선순위에 관련된 특성들을 이용한 자동분류방법, 회귀를 이용한 입력과 출력값 사이의 관련성을 이용해 뚜렷한 가설을 세우는 방법을 이용한다면 더욱 분류성능이 좋은 분류기를 만들 수 있을 것이라 생각한다.

참고문헌 및 출처

[0]Machine Learning for Hackers(O'reilly - 2013)

[1]http://terms.naver.com/ (보수주의)

[2]http://terms.naver.com/ (진보주의)

[3]http://blog.naver.com/james_parku/220192779851 (그림)

[4]http://www.chosun.com/ (조선일보)

http://joongang.joins.com/ (중앙일보)

http://www.donga.com/ (동아일보)

[5]http://www.hani.co.kr/ (한겨레신문)

http://www.khan.co.kr/ (경향신문)

http://www.pressian.com/ (프레시안)

[6]http://www.hankookilbo.com/ (한국일보)

http://www.seoul.co.kr/ (서울신문)