선형 퍼셉트론의 부스팅 학습에 의한 텍스트 여과

Text Filtering by Boosting Linear Perceptrons

오장민, 장병탁
서울대학교 컴퓨터공학부

Jangmin O, Byung-Tak Zhang
School of Computer Science and Engineering, Seoul National University
{jmoh, btzhang}@scai.snu.ac.kr

문서 분류나 여과 문제에서 양의 학습 데이터의 부족은 성능 저하의 주요 원인이 된다. 이런 경우 여러 학습 알고리즘이 문제의 특성을 제대로 파악하지 못한다. 본 논문에서는 부스팅 기법을 도입하여 이 문제를 접근해 보았다. 부스팅 기법은 약한 능력을 보유한 학습 알고리즘을 부스팅 과정을 통해 궁극적으로 강력한 성능을 얻을 수 있게 해준다. 간단한 선형 퍼셉트론에 부스팅 기법을 도입하여 문서 여과에 적용하였다. 제안된 알고리즘을 Reuters-21578 문서 집합에 적용한 결과, 재현률 측면에서 다층 신경망보다 우수한 성능을 보였고 특히 양의 학습 데이터가 부족한 문제의 경우 탁월한 결과를 얻을 수 있었다.

Ⅰ. 서론

문서 여과는 입력으로 들어오는 문서를 사용자의 관심사에 해당하는 문서만 선별하여 제공하는 작업이다[2]. 문서 여과 문제의 성능 평가의 주된 척도는 재현률/정확률이다[4, 5].

	+ 정답	- 정답
+ 분류	a	b
- 분류	c	d

재현률은 a/(a+c)이고 정확률은 a/(a+b) 이다. 문서 여과 시스템을 사용하는 사용자는 + 분류되는 문서만 받아보게 되므로 재현률과 정확률을 모두 높이는 것이 중요하다. 문서 여과는 문제의 특성상 양의 학습 데이터(positive examples)와 음의 학습 데이터(negative examples)의 불균형이 심하다. 즉, 대부분의 문서가 특정 관심사에 속하지 않을 가능성이 크므로 양의 학습 데이터가 음의 학습 데이터에 비해

상대적으로 크게 부족한 경우가 발생한다. 불균형을 이룬 학습 데이터에 대해 다층 신경망 같은 일반적인 학습 알고리즘을 적용시키면 썩 뛰어나지 않은 결과를 얻게 된다.

본 논문에서는 이 문제의 접근방법으로 선형 퍼셉트론 기반의 부스팅 기법을 도입하였다. 부스팅 기법을 통하여 높은 재현률을 보이는 여과 시스템을 얻을 수 있었고 불균형 데이터에 대해 부스팅 기법을 도입하는 것이 좋은 대안이 된다는 사실을 알 수 있었다.

논문의 구성은 다음과 같다. 본론에서 선형 퍼셉트론 부스팅 알고리즘을 제안하고 실험에서 Reuters-21578 문서 집합에 대한 실험 결과를 보이고 결론을 맺는다.

Ⅱ. 본론

AdaBoost는 0.5 이하의 에러율 조건을 만족시키는 약학습기(weak learner)를 부스팅 효과를 통해 강한 성능을 얻게 해주는 학습 알고리즘이다. 모델 에러에 대한 바이어스/분산 딜레마(bias/variance dilemma)에 따르면 어떤 모델의 에러는 바이어스와 분산에 관한 항으로 표현할 수 있는데, 모델의 복잡도가 증가할수록 에러의 바이어스를 줄일 수 있는 반면 분산은 증가된다. 그런데 부스팅 기법은 학습이 진행됨에 따라 에러의 분산까지 감소되는 효과가 있다[1, 3].

AdaBoost는 기본적으로 데이터의 확률 분포를 가지고 학습이 진행된다. 각 학습 단계마다 현재의 데이터의 확률 분포로부터 학습된 약학습기를 얻게 된다. 약학습기의 학습이 끝난 후 올바르게 분류된 데이터에 대해서는 확률을 낮추고 올바르지 않게 분류된 데이터에 대해서는 확률을 높인다. 이런 식으로 반복해서 나중 단계에서는 학습이 어려운 데이터에 집중해서 학습이 된 약학습기를 얻게 된다. 최종 학습 모델은 각 단계에서 생성된 약학습기의 조합으로 구성된다.

선형 퍼셉트론 부스팅은 AdaBoost의 약학습기로 선형 퍼셉트론을 사용한 것이다. 선형 퍼셉트론은 이진 선형 분리 함수(binary linear separator)이며 대부분의 학습 데이터에 대하여 0.5 이하의 에러율을 갖는다.

입력: $(x_1, y_1),\ldots,(x_N, y_N)$, $y_i=\{-1,+1\}$

$D_1(i) = 1/N$;

$t=1,\ldots,T$ 에 대해 다음 루프를 반복

- 분포 D에 따라 N 개의 학습 데이터 생성
- 이 학습 데이터를 입력으로 퍼셉트론 NN_t 학습
- 약학습 모델 h_t: $X\rightarrow\{-1: NN_t < 0, +1: NN_t > 0\}$에 대하여 다음을 계산

$$\varepsilon_t = \sum_{h_t(\mathbf{x}_i)\neq y_i} D_i$$

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

- 분포 D를 갱신

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t}\times\begin{cases} e^{-\alpha_t}, & \text{if } h_t(\mathbf{x}_i) = y_i \\ e^{+\alpha_t}, & \text{if } h_t(\mathbf{x}_i) \neq y_i \end{cases}$$

출력: $H(\mathbf{x}) = \mathrm{sgn}\left(\sum_{t=1}^{T}\alpha_t h_t(\mathbf{x})\right)$

선형 퍼셉트론 부스팅 알고리즘

학습의 종료 조건은 T번째 iteration 후이지만 ε_t=0.5이면 약학습기의 조건도 만족시키지 않을뿐더러 이 경우 α_t=0이 되어

분포 D가 더 이상 갱신되지 않는다. 따라서 이경우 학습을 종료시키고 ε_t=0인 경우에도 종료시킨다.

Ⅲ. 실험

실험에 사용된 문서 집합은 Reuters-21578 이다. 이 문서 집합은 세가지 학습/테스트 데이터 구성이 있는데 ModApte 구성을 따랐다. 문서는 tfidf의 벡터형으로 표현하였다. 문서를 표현하는 단어 수는 8,754가 되도록 불용어 목록, 스테밍기법등을 사용하여 줄였다. 최종적으로 실험에 사용된 데이터는 총 학습 데이터가 8,762 개이고 테스트 데이터는 3,009 개였다. 문서 여과에 사용된 범주와 양의 예제 비율은 다음과 같다.

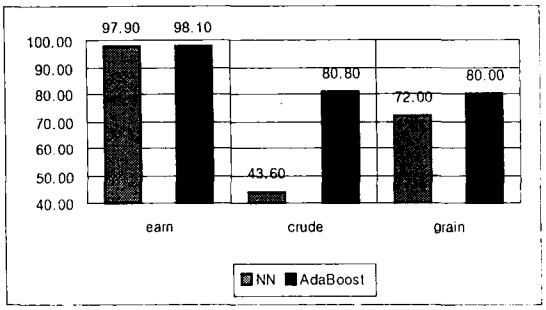
범주	학습	테스트
earn	32.4%	34.6%
grain	4.7%	4.2%
crude	4.2%	5.2%

표에서 grain, crude는 양의 예제가 매우 적음을 알 수 있다.

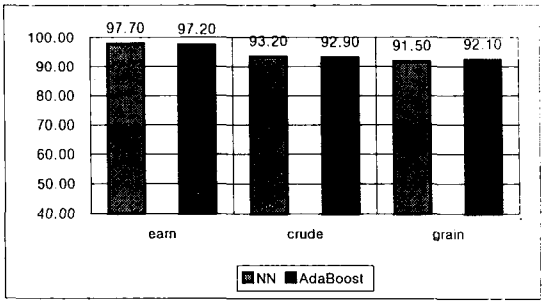
선형 퍼셉트론 부스팅과 성능 비교를 위해 다층 신경망을 구성하였다. 다층 신경망은 뉴런 5개의 은닉층, 뉴런 1개의 출력층으로 구성하였다. 출력층은 선형 뉴런이, 은닉층에는 하이퍼탄젠트 뉴런이 사용되었다. 학습률(learning rate)은 0.2, 모멘텀은 0.5로 설정했다. 학습 데이터에 대한 LMS (least mean square) 에러가 0.005 미만이 되거나 학습회수가 500이 될 때까지 진행시켰다.

선형 퍼셉트론 부스팅의 약학습기인 선형 퍼셉트론의 학습률과 모멘텀은 다층 신경망과 동일하게 설정했다. LMS 에러가 0.01 미만이거나 학습 회수가 25가 될 때까지 약학습기를 학습시켰다. 부스팅은 100번까지 진행시켰다.

테스트 데이터에 대한 실험 결과를 아래에 보인다.



재현률



정확률

그림에서 NN은 다층 신경망을, AdaBoost는 선형 퍼셉트론 부스팅을 뜻한다.

패턴 분리 문제에 있어 다층신경망은 우수한 정확률을 갖는다. "earn" 데이터처럼 양의 예제(positive example)와 음의 예제(negative example)가 적절하게 있을 때 다층신경망은 재현률에 있어서도 우수한 성능을 보인다. 하지만 "crude"의 경우처럼 양의 예제가 부족한 경우에는 현격히 저조한 재현률을 보인다. 하지만 선형 퍼셉트론 부스팅은 이런 문제에도 훌륭히 적응한다.

"crude"의 경우 학습 데이터 8,762개 중 양의 예제는 370개 밖에 되지 않는다. 학습 데이터의 심한 불균형 때문에 다층신경망은 음의 예제가 지닌 특징에 편중되어 학습이 이뤄진다. 반면에 선형 퍼셉트론 부스팅의 경우 재현률이 80.80%로 43.60%의 다층 신경망에 비해 매우 높다.선형 퍼셉트론 부스팅은 부스팅이 진행됨에 따라 상대적으로 학습이 어려운 양의 예제에 학습을 집중시키므로 데이터가 지닌 특징을 최대한 학습하게 된다.

Ⅳ. 결과

본 논문에서는 문서 여과 문제를 위해, 부스팅 기법을 적용한 선형 퍼셉트론 부스팅을 제안하였다.일련의 정보검색 문제에서 양의 예제가 적음은 항상 낮은 재현률에 대한 부담을 준다. 이 경우에 부스팅 기법은 이에 대한 해결책으로 상당한 의미가 있을 것이다. 실험을 통하여 선형 퍼셉트론 부스팅이 부족한 양의 예제를 가지고도 뛰어난 재현률을 얻을 수 있다는 것을 알 수 있었다.

Ⅴ. 참고문헌

[1] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee, "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods", *The Annals of Statistics*, 26(5): 1651-1686, 1998.

[2] R.E. Schapire, Y. Singer, and A. Singhal, "Boosting and Rocchio applied to text filtering", In Proc. SIGIR-98, pp.251-223, 1998

[3] S. Haykin, *Neural Networks* second edition, Prentice-Hall, Inc., 1997

[4] Y. Yang ,"An Evaluation of Statistical Approaches to Text Categorization", Technical Report CMU-CS-97-127

[5] W.B. Frakes and R. Baeze-Yates, *Information Retrieval Data Structures & Algorithms*, Prentice-Hall, Inc., 1997