부스팅 기법에 의한 불균형 데이터 문제 해소

Boosting Linear Perceptrons for Unbalanced Data

오장민, 장병탁, 서울대학교 컴퓨터공학부 인공지능연구실, {jmoh,btzhang}@scai.snu.ac.kr

초록: *선형 퍼셉트론에 기반한 부스팅 기법을 고안하여 불균형 데이터로 구성된 문서 여과 문제에서 우수한 성능을 보여주었다.*

연구 배경: 문서 여과는 입력으로 들어오는 문서를 사용자의 관심사에 해당하는 문서만 선별하여 제공하는 작업이다. 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 얻을 수 있다.

선형퍼셉트론 부스팅:

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

$D_1(i) = 1/N$;

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

그림 1: 선형 퍼셉트론 부스팅 알고리즘

알고리즘은 AdaBoost 기반으로서 선형퍼셉트론을 약학습기로 사용하였다. AdaBoost 는 0.5 이하의 에러율 조건을 만족시키는 약학습기(weak learner)를 부스팅 효과를 통해 강한 성능을 얻게 해주는 학습 알고리즘이다. AdaBoost 는 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[1].

실험: 정보검색에서 성능 평가는 일반적으로 재현률과 정확률의 측정으로 이루어 진다[2]. 본 논문에서의 실험 문서 집합은 Reuters-21578 을 대상으로 하였고 이중 “earn”, “crude”, “grain” 클래스에 대하여 문서 여과 실험을 하였다.

	재현률			정확률		
	earn	crude	grain	earn	crude	Grain
MN	98.1	43.6	77.6	97.7	93.2	91.5
LA	98.1	**80.8**	**80.0**	97.2	92.9	92.1

표 1: 다층신경망(MN)과 선형 퍼셉트론 부스팅(LA)의 실험결과

실험결과: “crude”의 경우 학습 데이터 8,762 개 중 양의 예제는 370 개 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.

[1] *R.E. Schapire, et al., “Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods”, 1998*

[2] *W.B. Frakes, et al.,* Information Retrieval Data Structures & Algorithms, *1997*