불균형 데이터의 효과적 학습을 위한 커널 퍼셉트론 부스팅 기법

오장민 O 장병탁
서 울 대 학 교 컴 퓨 터 공 학 부
{jmoh, btzhang}@scai.snu.ac.kr

Kernel Perceptron Boosting for Effective Learning of Imbalanced Data

Jangmin O O Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

많은 실세계의 문제에서 일반적인 패턴 분류 알고리즘들은 데이터의 불균형 문제에 어려움을 겪는다. 각각의 학습 예제에 균등한 중요도를 부여하는 기존의 기법들은 문제의 특징을 제대로 파악하지 못하는 경우가 많다. 본 논문에서는 불균형 데이터 문제를 해결하기 위해 퍼셉트론에 기반한 부스팅 기법을 제안한다. 부스팅 기법은 학습을 어렵게 하는 데이터에 집중하여 앙상블 머신을 구축하는 기법이다. 부스팅 기법에서는 약학습기를 필요로 하는데 기존 퍼셉트론의 경우 문제에 따라 약학습기(weak learner)의 조건을 만족시키지 못하는 경우가 있을 수 있다. 이에 커널을 도입한 커널 퍼셉트론을 사용하여 학습기의 표현 능력을 높였다. Reuters-21578 문서 집합을 대상으로 한 문서 여과 문제에서 부스팅 기법은 다층신경망이나 나이브 베이스 분류기보다 우수한 성능을 보였으며, 인공 데이터 실험을 통하여 부스팅의 샘플링 경향을 분석하였다.

1. 서론

일반적인 패턴 분류 문제는 입력 벡터 **x** 에 대해 입력 공간(input space)상의 초평면(hyperplane)을 찾는

$$f(\mathbf{x}) = \begin{cases} +1 & \text{if} \quad \mathbf{w}\cdot\phi(\mathbf{x}) > 0 \\ -1 & \text{otherwise} \end{cases}$$

다음과 같은 이진 분류 문제로 표현할 수 있다. 데이터는 '+1', 이나 '–1' 두 클래스로 분류되고, 패턴 분류 알고리의 목적은 우수한 일반화 성능을 얻을 수 있도록 **w** 를 찾는 것으로 볼 수 있다. 이 때 두 클래스에 속하는 데이터 수의 비율이 현격히 차이가 나는 경우를 불균형 데이터(imbalanced data)문제라고 한다.

실세계의 많은 문제는 불균형 데이터 문제에 해당한다. 예를 들어 정보 검색의 한 분야인 정보 여과 문제는 전형적인 불균형 데이터 문제이다. 정보 여과 문제의 성능 비교는 정보 검색에서 일반적인 다음 분할표를 이용한다.

표 1 분할표

		정답	
		+1	-1
예측	+1	TP	FP
	-1	FN	TN

즉, 패턴 분류의 결과를 TP (True Positive), FP (False Positive), FN (False Negative), TN (True Negative) 의 4 영역으로 나눌 수 있다. '+1' 클래스의 경우 사용자의 관심사에 속하는 정보, '-1' 클래스의 경우 사용자의 관심밖의 정보라고 볼 수 있다. 만일 뉴스 기사 여과 시스템인 경우, 뉴스 수집기에 입수되는 문서 중 사용자의 관심사에 속하는 뉴스는 전체 문서 집합 중 아주 적은 양일 것이다. 이 경우, 모든 문서를 '-1'로 분류하는 시스템도 우수한 정확도(accuracy)를 보일 것이다. 따라서 정보 검색에서는 주로 재현률(recall)/정확률(precision)을 통해서 성능을 평가한다.

재현률 = TP/(TP+FN), 정확률 = TP/(TP+FP)

재현률은 '+1' 클래스의 문서들 중 실제로 사용자에게 제공된 문서, 정확률은 사용자에게 제공된 문서 중 실제로 '+1'에 속하는 문서의 비율을 의미한다.

정보 여과 문제처럼 불균형 데이터 문제는 '+1' 클래스의 데이터의 비율이 작은 경우가 많다. 앞으로 본 논문에서 불균형 데이터라 함은 이 경우를 가리킬 것이다. 불균형 데이터의 학습이 힘든 요인으로는 첫째, '-1' 클래스의 많은 데이터 중 노이즈 섞인 데이터가 학습을 방해할 수 있다. 둘째, 문제를 기술하는 중요한 특징을 '+1' 클래스로부터 추출하기가 어렵다. 본 논문에서는 퍼셉트론 기반의 부스팅 기법을 통한 불균형 데이터 문제를 접근해보고, 실험을 통하여 불균형 데이터에 대한 부스팅의 우수성을 살펴 보았다.

논문의 구성은 다음과 같다. 2 절에서 불균형 데이터 문제에 대한 접근법을 들었다. 3 절에서는 퍼셉트론 부스팅 기법을 소개하고, 4 절에서는 실험 및 결과를 보이고 결론을 맺는다.

2. 불균형 데이터 문제 접근법

불균형 데이터에 대한 접근법은 다음과 같이 정리해 볼 수 있다[4, 5, 12].

1. '+1' 클래스의 데이터를 업샘플링(up-sampling)하여 '-1' 클래스와 데이터의 비율을 맞추는 방법

2. '-1' 클래스의 데이터중 일부를 제거하는 다운샘플링(down-sampling)을 통해 '+1' 클래스와의 비율을 맞추는 방법
3. '+1' 클래스에 대한 에러 비용(error cost)을 '-1'에 대한 에러 비용을 달리 하는 방법

방법 1, 2 는 샘플링을 통하여 클래스간의 균형을 맞추는 방법이다. 방법 1 은 두 클래스의 학습 기회가 비슷하도록 학습 알고리즘에게 '+1'의 클래스의 데이터의 학습 횟수를 더 부여하는 기법이다[7]. 방법 2 는 '-1' 클래스를 대표하는 소수의 데이터를 선별하여 클래스간 비율을 맞추는 방법이다[6].

두 방법의 샘플링 전략에 따라 랜덤 샘플링(random sampling) 과 집중 샘플링(focused sampling)으로 세분할 수 있다. 랜덤 샘플링은 방법 1,2 에서 추가되거나 제거될 데이터의 샘플링을 임의로 하는 것이다. 집중 샘플링은 특정 전략을 통해 샘플링을 하는 방법이다. 한가지 방법은 능동 학습(active learning) 기법이다[8]. [8]은 데이터의 일부분으로 학습한 모델이 제일 큰 에러를 보이는 데이터를 학습 데이터로 추가하여 점진적으로 학습해 나가는 기법을 도입하였다. 다른 한가지 방법은 결정 경계(decision boundary) 근처의 데이터를 샘플링 하는 방법이 있다.

방법 3 은 '+1' 클래스 데이터의 에러 비용을 '-1' 클래스보다 상대적으로 크게 하여 학습시키는 방법으로, 학습 알고리즘은 '+1' 클래스의 에러에 민감하게 반응하여 학습하게 된다.

3. 퍼셉트론 부스팅

AdaBoost 는 전체 학습 데이터에 대해서 0.5 이하의 에러율 조건을 만족시키는 약학습기(weak learner)로 앙상블 머신(ensemble machine)을 구축시키는 부스팅 효과를 통해 강한 성능을 얻을 수 있는 기법이다[10, 11]. 바이어스/분산 딜레마(bias/variance dilemma)에 따르면 앙상블 머신의 분산은 단독 머신의 경우보다 줄어들게 된다. AdaBoost 는 바이어스와 분산을 동시에 감소시키는 작용을 한다[10].

AdaBoost 는 데이터의 확률 분포를 유지한다. 현 단계에서 학습된 약학습기는 전체 학습 데이터에 대한 에러를 구하고 에러를 바탕으로 약학습기의 앙상블 가중치를 계산한다. 낮은 에러를 보이는 약학습기는 높은 가중치를 받는다. 또한 약학습기가 잘못 분류한 데이터는 확률을 높이고, 옳게 분류한 데이터는 확률을 낮춘다. 다음 단계에서는 이 확률 분포로부터 샘플링을 통해 학습 데이터를 재구성한다.

그림 1 은 AdaBoost 에 기반한 퍼셉트론 부스팅 알고리즘이다.

3.1 커널 퍼셉트론

Rosenblatt 의 퍼셉트론 수렴 증명은 기계학습 분야의 신기원을 열어주었다. 퍼셉트론의 가중치 갱신 알고리즘은

$$\text{if} \quad y_i(\mathbf{w}_k \cdot \mathbf{x}_i) \le 0 \quad \text{then}$$
$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta(k) y_i \mathbf{x}_i$$

입력: $(\mathrm{x}_1, y_1),\ldots,(\mathrm{x}_N, y_N)$, $y_i=\{-1,+1\}$

$D_1(i) = 1/N;$

t=1,…,T 에 대해 다음 루프를 반복

- 분포 D_t에 따라 N 개의 학습 데이터 생성
- 학습 데이터에 대한 퍼셉트론 NN_t학습
- 약학습 모델 h_t: $X \rightarrow \mathrm{sgn}(NN_t(X))$에 대하여 다음을 계산

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

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

- 분포 D_t를 갱신

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

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

그림 1 퍼셉트론 부스팅 알고리즘

이다(primal form). 그런데, 최종 학습 가중치는 다음처럼

$$\mathbf{w} = \sum_{i=1}^{l} y_i \alpha_i \mathbf{x}_i$$

데이터의 조합으로 표현할 수 있으므로 퍼셉트론의 학습 알고리즘을

$$\text{if} \quad y_i(\mathbf{w}_k \cdot \mathbf{x}_i) \le 0 \quad \text{then}$$
$$\alpha_i = \alpha_i + \eta(k)$$

과 같이 데이터의 가중치 변경으로 생각할 수 있다(dual form). 이 때, 학습된 퍼셉트론은 다음과 같이

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} = \sum_{i=1}^{l} y_i \alpha_i (\mathbf{x}_i \cdot \mathbf{x})$$

처럼 데이터의 조합으로 표현할 수 있다. 퍼셉트론의 식이 학습데이터에 대한 가중치 α_i 와 데이터와 $\mathbf{x}$ 의 내적으로 표현되므로 자연스럽게 커널 함수로의 확장을 생각할 수 있다. 즉,

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} = \sum_{i=1}^{l} y_i \alpha_i K(\mathbf{x}_i \cdot \mathbf{x})$$

의 커널 퍼셉트론을 얻을 수 있다. 커널 퍼셉트론은 Support Vector Machine (SVMs)의 아이디어와 유사하다[1, 2, 3, 9]. 퍼셉트론을 약학습기로 사용하기에 곤란한 문제의 경우 예를 들어 커널 퍼셉트론을 사용하면 효과적일 것이다. 본 논문에서 사용된 커널 함수는

$$K(\mathbf{x}_i \cdot \mathbf{x}) = (1 + \mathbf{x}_i \cdot \mathbf{x})^d$$

의 다항식 커널로서, d 는 차수를 의미하며 d = 1 인 (정상 퍼셉트론)과 d = 3 인 커널 퍼셉트론을 사용하였다.

4. 실험

4.1 Reuters-21578 문서 여과

문서 분류에서 널리 쓰이는 Reuters-21578 문서 집합 중 3 가지 토픽에 대해 문서 여과 실험을 하였다. 3 토픽의 데이터 구성비('+1' 클래스/'-1' 클래스)를 표 2 에 보였다.

표 2 '+1'데이터 비율

	train	test
earn	32.4%	34.6%
grain	4.7%	4.2%
crude	4.2%	5.2%

이 불균형 데이터에 대한 실험 결과를 표 3 에 보인다.

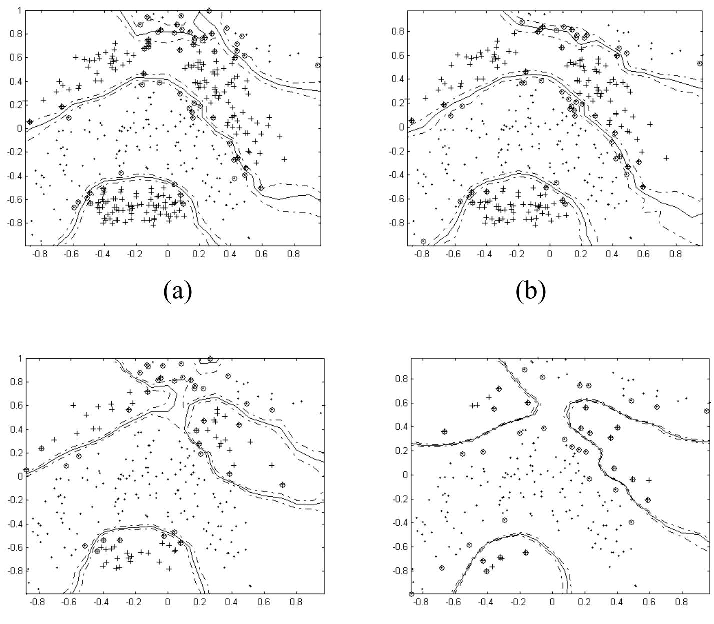
표 3 *F*1 (재현률/정확률 채산점)

F1	earn	crude	grain
MLP	**97.80**	59.41	80.59
KPB	97.65	**86.43**	**85.62**
Naïve	97.70	57.72	77.46

MLP 는 다층 퍼셉트론(은닉 뉴런수는 30), KPB 는 커널 퍼셉트론 부스팅(d = 1), Naïve 는 나이브 베이스 분류기를 의미한다.

4.2. 인공 데이터(banana)

부스팅 기법의 결정 경계와 샘플링 경향을 분석하기 위해 banana 데이터를 이용하여 실험을 하였다. 전체 502 개의 데이터를 생성한 후, '+1' 클래스의 비율이 90%, 70%, 30%, 10% 되도록 다시 샘플링하였다. 이렇게 생성된 데이터 집합을 80%를 학습 데이터로 나머지를 테스트 데이터로 사용하였다. 결과를 그림 2 에 보였다.



(a) (b) (c) (d)

그림 2 결정 경계 및 샘플링 효과

그림 2 에서 '*'는 '+1', '.'는 '-1'의 클래스이며, 동그라미는 부스팅 회수 T 의 1.5 배 이상 샘플링 된 데이터를 의미한다.

4.3. 실험 결과

그림 1 의 문서 여과 문제에서 'earn' 같은 균형 데이터에 대해서는 실험에 사용된 알고리즘간의 성능차가 거의 없지만, 'crude'의 경우 부스팅 기법이 매우 우수함을 알 수 있다. 그림 2 를 보면 부스팅의 효과를 알 수 있다. 두 클래스에서 동그라미에 해당하는 데이터의 비율이 거의 같고, 이들은 결정 경계 부근에 집중되어 있다. 이는 부스팅이 결정 경계 근처의 데이터를 샘플링하는 경향이 있음을 의미한다.

5. 결론

본 논문에서는 불균형 데이터 문제를 위한 커널 퍼셉트론 부스팅 기법에 대해 살펴 보았다. 퍼셉트론이 약학습기로 부족한 경우 커널 퍼셉트론을 사용하는 방법을 도입하였다. 실세계 문제에서 부스팅 기법은 만족할만한 성능을 보였으며, 부스팅은 결정 경계 근처의 데이터에 집중한다는 것을 알 수 있었다.

감사의 글

본 연구는 과학기술부 뇌연구개발사업(BR-2-1-G-06)과 교육부 BK21 프로그램에 의하여 일부 지원되었음

참고 문헌

[1] N. Cristianini and J. S. Taylor, "*An Introduction to Support Vector Machines and other kernel-based learning methods*", Cambridge Press, 2000

[2] Y. Freund and R.E. Schapire, "Large Margin Classification Using the Perceptron Algorithm", Machine Learning, 37(3):277-296, 1999

[3] Guyon and D.G. Stork, "Linear Discriminant and Support Vector Classifiers", *Advanced in Large Margin Classifiers*. MIT Press, 1999

[4] N. Japkowicz, "Learning from Imbalanced Data Sets: A Comparision of Various Strategies", AAAI-2000 Workshop, 2000

[5] N. Japkowicz, "The Class Imbalance Problem: Significance and Strategies", In *Proceedings* of ICAI-00, 2000

[6] M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection", In *Proceedings* of ICML-97, 1997

[7] C.X. Ling and C. Li, "Data Mining for Direct Marketing: Problems and Solutions", In *Proceedings* of KDD1998, 1998

[8] S.W. Park and B-T. Zhang, "Learning Constructive RBF Networks by Active Data Selections". In *Proceedings* of ICONIP-00, 2000

[9] C. Saunders, A. Gammermann, and V. Vovk, "Ridge Regression Learning Algorithm in Dual Variables", In *Proceedings* of ICML-98, 1998

[10] 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

[11] R.E. Schapire, "Theoretical views of boosting and applications", In *Proceedings* of COLT-99, 1999

[12] G. M. Weiss and H. Hirsh, "Learning to Predict Rare Events in Event Sequences", In *Proceedings* of KDD-98, 1998