●5월 5주차(20170528-20170603)

Adaboost의 overfitting

Adaboost 알고리즘은 최근의 분류 방법론에 있어서 가장 중요한 발전 중 하나이다.

이 알고리즘은 많은 응용 프로그램에서 큰 성공을 거두고 있다. Low noise 영역에서는 Adaboost에 과도한 문제가 있는 것이 거의 없다. 이 현상을 이해하기 위한 하나의 중요한 설명은 margin 개념으로 간주된다. Adaboost가 효과적으로 마진을 늘릴 수 있는 것이 관찰되고, 큰 마진은 일반적으로 좋은 일반화에 도움이 된다. 즉 주어진 데이터에 대해 큰 마진을 달성할 수 있으면 일반화 오차는 작다는 것이다.

그러나 매우 큰 패턴을 이용한 최근의 연구에서는 overfitting이 발생할 수 있다는 것을 보여주고 있다. 그래서 Adaboost의 인기가 높아지고 있는 것에 비추어 overfitting현상을 확인하고 효과적인 솔루션을 찾고, noisy환경에서 Adaboost의 성능을 향상시키는 것이 중요하다. High-noise 영역에서 Adaboost 분류 결과의 빈약함의 주된 이유는 습득하기 어려운 몇 가지 예에 너무 많은 비중을 할당하여 알고리즘이 왜곡된 데이터 분포를 생성하는 것이다. 따라서 알고리즘에서 정규화 항을 도입해서 데이터 분포의 비대칭을 제어하는 것이 하나의 전략이다. 이 원리에 따라 가장 초기에 제안된 알고리즘 중 하나가 AdaboostReg알고리즘이다. 이것은 소프트 마진을 실현하기 위해 마진과 샘플의 영향 사이에서 tradeoff를 제어하는 직관적인 개념의 기반 추론 알고리즘이다. 다른 사용 가능한 정규화 boosting 알고리즘과 비교하여 AdaboostReg 알고리즘은 noisy 데이터에 대한 가장 일반화된 결과가 있다. 그러나 정규화 알고리즘 레벨에서 도입되기 때문에 기본적인 최적화 계획을 분석하는 것은 곤란하고, 알고리즘의 최종 목표도 명확하지 않다.

정규화 용어를 Adaboost에 포함하는 것이 쉽지 않기 때문에 새로운 정규화 boosting 알고리즘을 설계할 수 있는 경우는, Adaboost와 잘 알려진 minimax 문제의 밀접한 관계를 이용할 수 있다. 이 방법의 장점으로 잘 연구된 수학적 프로그래밍 기술의 일부를 직접 이용할 수 있다는 것이다. 이 전략의 대표적인 예로는 LPreg-Adaboost와 C-Barrier 알고리즘인데, 이것들은 v-Arc의 기본 체계를 구성한다. LPreg-Adaboost는 비 분리형 데이터의 경우 SVM과 마찬가지로 기본 도메인의 최적화 문제에 완만한 변수가 도입되고 있다.