●6월 2,3주차(20170611-20170624)

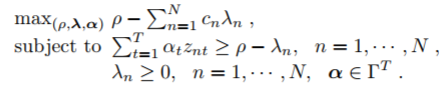
AdaBoostReg는 소프트 마진을 달성하기 위해 마진과 샘플의 영향과 트레이드 오프를 제어하는 직관적인 개념에 기반한 추론 알고리즘이다. 다른 사용되는 표준화된 부스팅 알고리즘과 비교해서 AdaBoostReg는 noisy데이터에 대한 가장 일반화된 결과를 갖고 있다는 것이다. 그러나 정규화 알고리즘 레벨에서 도입되므로, 그 기초가 되는 최적화 계획을 분석하는 것은 곤란하고, 알고리즘의 최종 목표는 명확하지 않다.

LPreg-AdaBoost

분포를 0 ≤ d ≤ c 로 결정함으로써 다음의 최적화 문제가 생긴다.

(1)

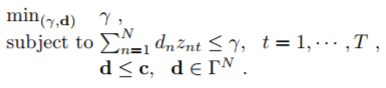
여기서 c는 상수벡터이고, 일반적으로 c=C1의 형식을 취하며 C는 사전 정의된 매개변수이고, 1 ∈ RN은 모두 벡터이다. 식의 최적화 방법은 다음과 같다. 최악의 경우 배전반의 분류 성능이 최대가 되도록 혼합 계수 a set of combination coefficients α 를 찾을 것으로 이해할 수 있다. LP의 방정식은 다음과 같다. 위의 식은

(2)

다음과 같다. LPreg-AdaBoost는 식의 특수한 경우이다. 위의 식은 c1=c2=c3=…..=cN=C로 지정해 놓으면 얻을 수 있다. SVM에서도 분리할 수 없는 데이터의 경우에 위와 같은 방식이 사용된다. 최적화 방식은 위의 식과 같다. 이것은 패턴 xv 소프트 마진 ρs(xn)를 달성하기 위해서 최적화 문제에 음수 이완 변수 λn를 도입한다. 이것은 다음과 같이 정의된다.

(3)

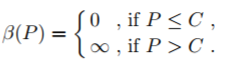
하드 마진의 완화에 의해 몇 가지 패턴은 ρ보다 작은 마진을 가질 수 있다. 그 결과, 알고리즘 관련 클래스 레이블에 따라 패턴을 모두 분류하지 않는다. (2)의 식은 다음에 의해서 주어진다.

(4)

듀얼 도메인에서 직접 작업하면서 마진의 개념을 통해 공식화를 추구한다는 명확성이 손실되게 된다. 그러나 이중성 도메인은 주로 공간이 위의 식에서 정의된 경우를 제외하고 소프트웨어 마진을 지정하는데 적합하지 않기 때문에 유리하다는 것을 알 수 있다. 편의상, 우리는 식을 다음과같이 재정리한다. (1)의 식은

(5)

(5)와 같이 정의되어 있는데, 는 p-norm이고, β(P)는 (6)에 의해 정의된 함수이다.

(6)

에 의해 정의되는 범위는 distribution center d0 = [1/N, · · · , 1/N]의 중앙에 위치한다. 또한, 매개변수 C는 범위 경계와 d0사이의 분포의 비대칭을 어느 정도 반영한다. 식 (5)는 LPreg-AdaBoost가 범위에서 0 penalty에서 penalty 스킴이고, 범위 밖에서는 무한대임을 나타낸다. 이런 의미에서 이 제도는 다소 추론적이며, 너무 제한적일 수도 있다.

LPreg-AdaBoost 구현에 대해 실제로는 H의 기수는 무한이 될 가능성이 있다. 그 결과 gain matrix Z는 명시적인 형태로 존재하지 않아도 된다. 그 결과 linear programming을 직접 실시할 수 없다. 이 문제를 극복하기 위해서 여러가지 알고리즘이 제안되었고, 전형적인 예로 v-Arc, 10, C-Barrier 알고리즘이 있다. 다음 연구노트에서는 |H| 가설 함수 집합의 기수를 나타내기 위해 AdaBoost의 반복 단계의 수로서 T를 확보한다.