●4월 3주차(20170416-20170422)

*m*개의 범주를 가진 *n*개의 학습 표본과 *B*개의 기저분류자로 구성된       ⋯ 을 가정하면 *b*번째 기저 분류자의 오류율()은 다음과 같이 계산된다.

DRW000024a420e1 DRW000024a42123

여기에서 DRW000024a420e7는 i번째 관측치에 부여되는 가중치로 초기 가중치는 1/n으로 설정된다. 또한 DRW000024a420ed

는 i번째 관측치의 예측변수 벡터이고 DRW000024a420f3는 i번째 관측치의 실제범주를 나타내며 DRW000024a420f9는 예측변수 벡터 DRW000024a420ed에 대한 *b*번째 분류자의 예측결과이다. b+1 번째 분류자에서 i번째 관측치에 부여되는 가중치는 DRW000024a420ff로 조정되어 오분류된 관측치에 더 높은 가중치가 부여된다. 여기에서 b는 분류자의 중요도 또는 정확도의 개념으로 해석되며 DRW000024a42105로 계산된다. b+1번째 분류자의 학습표본을 구성할 때 가중치가 높은 오분류 관측치가 많이 포함되기 때문에 boosting 알고리즘은 오분류 관측치에 초점을 맞춘 학습을 진행할 수 있게 된다. 이러한 방식으로 새로운 분류자 b=1,2,3,B가 생성되며 i번째 관측치의 최종결과는 다음과 같이 각 결과의 가중평균으로 계산된다.

DRW000024a4210f

DRW000024a4211b

부스팅 알고리즘은 단순 평균 개념에 기초한 주요 파라미터로 인하여 다음과 같은 문제가 나타날 수 있다. 첫째, 분류자의 오류율 DRW000024a42129는 단순 평균 오류율로 전체 표본 대비 오분류 표본 비율로 계산된다. 불균형 데이터의 경우 다수 범주의 낮은 오류율로 인하여 단순평균 오류율이 왜곡될 수 있다. 둘째, 분류자의 성과를 나타내는 b역시 단순평균 정확도에 기초한 개념이다. 데이터 불균형 하에서 단순평균 정확도는 성과지표로 유효하지 않기 때문에 범주별 데이터 불균형을 고려한 가중평균 정확도 개념으로 대체할 필요가 있다.