●4월 4주차(20170423-20170429)

-Adaboost의 여러 개선 방법이 있는데, 그 중에서 어떤 문제점을 갖고 어떻게 개선을 할 수 있는지에 대해서 조사를 해보았다.

-이번 논문에서는 에이다부스트가 훈련 샘플에 노이즈가 존재하는 경우 약한 분류기의 수가 늘어날수록 overfitting현상이 발생하는 문제점에 대해서 어떻게 개선할 것인가에 대해 생각해본다.

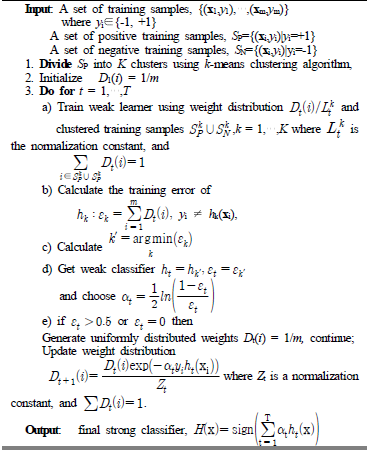
-overfitting현상이란 분류기가 훈련 샘플에 특화되어 일반화 성능이 저하되는 현상으로 작은 훈련오차에도 불구하고 테스트 오차가 크게 나타나게 되는 현상이다.

-Adaboost는 이전 단계의 약한 분류기가 잘못 분류한 자료의 가중치를 증가시켜서 다음 약한 분류기에서 잘못 분류된 자료를 더 잘 분류하게 하도록 한다.=>이 방법은 노이즈 샘플에 대한 가중치를 지나치게 증가시키게 되어서 약한 분류기의 훈련 횟수가 증가할수록 강한 분류기의 overfitting현상을 발생시키고, 일반화 성능을 저하시키게 된다. 이러한 문제점에 대해서 어떻게 개선을 할 수 있는지에 대해서 알아본다.

===해결책===

-약한 분류기의 성능이 약하기 때문에 약한 분류기가 일반 샘플들과 노이즈 샘플을 동시에 잘 분류하는 분류 경계를 생성하지 못하는 것에 문제점을 두고 본다. 따라서, 두 개의 클래스를 가지는 이진 분류 문제에 있어서 훈련 샘플들 중에 긍정, 부정 샘플들과 선형적으로 잘 분류될 정도의 군집들로 분할한 후에 약한 분류기가 하나의 것만을 분류하게 되면, 각 군집을 잘 분류하는 분류 경계가 생성되어서 약한 분류기의 훈련 횟수가 증가하여도 overfitting현상을 막을 수 있다.

그것 뿐만 아니라, 각각의 군집들을 분할하는 약한 분류기들의 결합으로 강한 분류기의 형성이 가능하게 된다.



이는 개선된 Adaboost의 pseudocode이고, 우선, 긍정 샘플들을 k-means 알고리즘을 통해서 Krodml 군집으로 분할한 후에 각 군집별로 하나의 긍정, 부정 샘플 전체를 나누는 약한 분류기의 gnqhemf을 구한다. 그 이후에 오차를 최소로 만드는 군집의 해당 약한 분류기 후보를 현재 훈련 단계의 약한 분류기로 결정한다. 약한 분류기 결정 후 샘플들의 가중치 변경은 기본적인 Adaboost 알고리즘과 동일하게 된다.