앙상블기법과 그래프샤프닝을 이용한 Semi-Supervised Learning의 학습 성능 향상 및 안정화

최인애, 신현정
아주대학교 공과대학 산업정보시스템 공학부
443-749 경기도 수원시 영통구 원천동 산5

Abstract

기계학습분야의 대부분의 알고리즘들은 모델의 구조 및 학습 파라미터를 어떻게 결정하는가와 주어진 학습 데이터에 노이즈가 어느 정도인가에 따라 그 성능이 변화한다. 모델의 학습에 사용되는 입력값과 목표값이 충분히 주어지는 경우라면, 학습 데이터 셋에서 일부 추출된 validation 셋을 이용하여 알고리즘의 최적 학습 파라미터를 결정할 수 있다. 그러나, 최근 제안된 semi-supervised learning의 경우에는 입력값은 충분히 주어지나 목표값이 있는 데이터의 수가 상당히 제한적이므로 학습 파라미터 결정을 위하여 별도의 데이터 셋을 구성하기가 어렵고, 데이터 셋의 노이즈가 어느 정도인지를 목표값과 비교하여 미리 가늠해보는 것 또한 쉽지 않다. 따라서 본 연구에서는 semi-supervised learning의 학습 파라미터 선택 및 데이터 셋의 노이즈 정도에 상관없이 안정화된 성능을 얻을 수 있는 방법으로서 앙상블 (ensemble)기법과 그래프샤프닝 (graph sharpening)을 사용할 것을 제안한다. 전자는 학습 파라미터 선택 과정을 불필요하게 함으로써 이에 따른 알고리즘의 성능 변화 요인을 제거하고, 후자는 노이즈로 인한 알고리즘의 성능 변동을 안정화시키고 정확도를 향상시킨다. 본 연구의 실험에서는 두 방법의 조합이 semi-supervised learning 알고리즘의 성능 향상 및 안정화에 유의하게 기여함을 보인다.

1 서론

Supervised learning 알고리즘에서는 입력값과 목표값이 레이블 (labeled)된 데이터들을 학습 셋으로 사용한다. 그러나 웹 또는 텍스트마이닝 분야의 스팸 이메일 필터링이나 생명정보학 분야의 단백질 구조 결정과 같은 문제에서는 풍부하게 주어지는 입력값에 비해 그 목표값을 얻기가 좀처럼 쉽지 않다. 실시간으로 쏟아져 나오는 방대한 양의 메일들로부터 스팸메일 여부를 일일이 판단하기도 어려울 뿐만 아니라, 숙련된 기술을 가진 전문가가 하나의 단백질 구조를 결정하는데 소요되는 시간이 적어도 수 개월이 된다는 점을 감안하면 데이터에 레이블을 부여한다는 것이 얼마나 많은 시간과 비용을 소요하는지를 가늠할 수 있다. 따라서 최근 이러한 응용분야들을 중심으로 레이블이 안된 데이터도 학습 셋에 포함시키는 semi-supervised learning이 부각되고 있으며 우수한 성능을 입증하고 있다 (Shin, 2006b; Shin, 2007; Zhu, 2003).

기계학습 (machine learning)분야의 대부분의 알고리즘들은 모델의 구조 및 학습 파라미터를 어떻게 결정하는가와 주어진 학습 데이터에 노이즈가 어느 정도인가에 따라 그 성능이 변화한다. 데이터의 노이즈 문제는 학습 파라미터 설정 문제와 별개의 문제는 아니다. 그러나 노이즈가 제거된 깨끗한 데이터 셋이라 할지라도 문제 자체가 갖는 복잡도로 인하여 학습 파라미터를 결정해야 하는 문제는 여전히 남아 있다. 따라서 이를 두 가지의 경우로 나누어서 접근하고자 한다. 첫째, 전술한 바와 같이 적합한 모델의 구조나 학습 파라미터를 설정하는 일은 문제의 복잡도와 연관된다. 그러나 이를 사전에 알기는 어렵기 때문에 학습 파라미터의 결정은 대체적으로 시행착오를 통해 이루어진다. Supervised learning에서처럼 모델의 학습에 사용되는 입력값과 목표값이 충분히 주어지는 경우라면, 학습 데이터 셋에서 일부 추출된 validation 셋을 이용하여 알고리즘의 최적 학습 파라미터를 결정할 수 있다. 가장 흔히 쓰이고 있는 방법으로는 cross-validation을 들 수 있다. 그러나, semi-supervised learning의 경우에는 문제가 다르다. 즉, 입력값은 충분히 주어지나 목표값이 있는 데이터의 수가 상당히 제한적이므로 학습 파라미터 결정을 위하여 별도의 validation 셋을 구성하기가 어렵다. 둘째, supervised learning 에서는 데이터 셋의 노이즈가 어느 정도인지를 목표값과 비교하여 미리 가늠해보는 것이 가능하다. 앞서 설명하였듯이 데이터 셋의 노이즈 정도는 모델의 복잡도를 어떻게 설정하는지와 연관된다. 노이즈가 많을 경우 모델의 파라미터 수를 감소시키거나 설정값을 조절하여 학습에 제약을 줌으로서 모델이 노이즈에 과적합 (overfitting) 되는 것을 방지할 수 있다. 그러나 이와는 달리 semi-supervised learning에서는 레이블이 된 목표값의 수가 적으므로 데이터 셋의 노이즈 정도를 아는 것 또한 어렵다.

따라서 본 연구에서는 semi-supervised learning의 학습 파라미터의 선택 및 학습 데이터 셋의 노이즈 정도에 상관없이 안정화된 성능을 얻을 수 있는 방법으로서 앙상블 (ensemble)기법과 그래프샤프닝 (graph sharpening)을 사용할 것을 제안한다. 앙상블 (ensemble)기법 또는 앙상블 네트워크에서는 다양한 모델들을 결합시킨 후 멤버 네트워크들의 총론적 값을 취함으로서 개별 네트워크들이 갖는 출력값 에러의 편기 (bias) 또는 오차변동 (variance)을 감소시켜 성능을 향상시킨다 (Briman, 1996; Perrone, 1993; Sharkey, 1997; Tumer, 1996). 앙상블 네트워크의 구성 멤버들 간의 다양성은, 각 멤버가 사용하는 학습 데이터에 변동 (perturbation)을 주거나 또는 멤버가 갖는 학습 파라미터 및 구조에 변동을 줌으로서 얻을 수 있다. 본 연구에서는 동일한 데이터 셋을 사용하되 학습 파라미터에만 변동을 주어 앙상블 네트워크의 멤버들을 구성한다 (Shin, 2001). 이러한 설정은 semi-supervised learning의 학습 파라미터 선택 과정을 불필요하게 함으로써 알고리즘의 성능 변동의 요인 자체를 제거하는 효과를 갖게 된다. 한편, 최근 제안된 그래프샤프닝은 graph-based semi-supervised learning에 속하는 알고리즘들의 성능을 향상시키는 효과적인 방법이다 (Shin, 2006b). Graph-based semi-supervised learning에서는 각 데이터들이 노드들로 표현되고 이들 간의 관계는 가중치 엣지 (weighted edge)로 표현된다. 노드 간 연결관계는 유사행렬 (similarity matrix)로 나타내어지는데 유사도가 있으면 노드 간 엣지가 형성되며 그 유사정도가 클수록 엣지의 연결강도가 증가한다. 따라서 데이터에 노이즈가 많을 경우에는 불필요한 엣지가 많이 형성되게 되며 알고리즘의 성능을 저하시키는 결과를 초래한다. 그래프샤프닝은 노이즈로부터 기인한 연결이나 성능 저하의 요인이 되는 불필요한 엣지들을 제거하는 알고리즘으로서 성능 향상과 안정화의 효과를 준다. 따라서 본 연구에서는 앙상블기법을 통하여 학습 파라미터 선택 과

정을 생략함으로써 학습 파라미터 설정에 따른 성능 변동 요인을 제거하고, 그래프샤프닝을 통해 각 앙상블 멤버의 개별 성능을 향상시키는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 제 2 장에서는 graph-based semi-supervised learning 알고리즘에 대하여 간략히 소개한 후, 본 연구에서 사용하는 그래프샤프닝 (graph-sharpening)과 앙상블 기법 (Ensemble method)을 제시한다. 제 3장에서는 제안한 두 방법의 조합이 semi-supervised learning 알고리즘의 성능 향상 및 안정화에 유의하게 기여함을, 인공 데이터 및 실제 데이터에 대한 정량적 실험 결과로서 보인다. 제 4장에서는 결론을 기술한다.

2 연구 방법

2.1 Graph-Based Semi-supervised Learning

Graph-based semi-supervised learning 알고리즘에서는 l 개의 레이블이 된 데이터 $\{(x_1, y_1), \cdots, (x_l, y_l)\}$와 u 개의 레이블이 안된 데이터$\{x_{l+1}, \cdots, x_{n=l+u}\}$를 사용한다. 레이블, 즉 목표값은 레이블이 된 데이터에서는 $y_l \in \{-1, 1\}$로 레이블이 안된 데이터에서는 $y_u \in \{0\}$로 표기된다. 총 데이터의 수는 u와 l의 합인 $n(n = l + u)$이며 전형적으로 u는 l보다 크다 ($l \ll u$). 데이터들은 노드들로 표현되고 이들 간의 관계는 가중치 엣지 (weighted edge)로 표현된다. 노드 간 연결관계는 유사행렬W 로 나타내어지는데 유사도가 있다고 판단되는 노드들 사이에는 엣지가 형성되어 ($i \sim j$) 그 유사정도가 클수록 엣지의 연결강도가 증가한다.

$$w_{ij} = \begin{cases} \exp\left(-\dfrac{(x_i - x_j)^T (x_i - x_j)}{\sigma^2}\right) & if \quad i \sim j, \\ 0 & otherwise. \end{cases} \tag{1}$$

유사행렬을 만드는 방법으로는 주로 k-nearest neighbor (κNN)가 사용된다. κ 값은 사용자에 의해 정의되며 어떤 값을 사용하느냐에 따라 W행렬이 변할 수 있다. 유사행렬이 주어지면 다음의 최적화 문제를 풀므로서 출력값 f를 얻을 수 있다 (Belkin, 2004).

$$\min_f (f - y)^T (f - y) + \mu f^T L f \tag{2}$$

목표값 y는 $y = (y_1, \cdots, y_l, 0, \cdots, 0)^T$로 표현되며 출력값은 $f = (f_1, \cdots, f_l, f_{l+1}, \cdots, f_{n=l+u})^T$로 표현된다. L은 라플라시안 행렬로서 $L = D - W$로 정의된다. D는 $d_i = \sum_j w_{ij}$와 $D = diag(d_i)$로 얻는다. 출력값 f_i는 레이블이 된 노드에서는 노드의 목표값 y_i와 비슷해야 하고, 자신과 연결된 노드 ($i \sim j$)의 출력값 f_j와 크게 달라지면 안된다. μ는 이러한 두 가지 조건이 학습에 미치는 영향을 조절하는 학습 파라미터로 사용자에 의해 정의된다. 식(2)로부터 다음

의 식(3)과 같이 출력값 f 를 구할 수 있다

$$f = (I + \mu L)^{-1} y \quad . \tag{3}$$

여기서 I 는 단위 행렬을 의미한다.

2.2 그래프샤프닝 (Graph Sharpening)

Graph-based semi-supervised learning에서는 각 데이터들이 노드들로 표현되고 이들 간의 관계는 가중치 엣지 (weighted edge)로 표현된다. 노드 간 연결관계는 유사행렬 (W행렬)로 나타내어지는데 유사도가 있으면 노드 간 엣지가 형성되며 그 유사정도가 클수록 엣지의 연결강도가 증가한다. 따라서 데이터에 노이즈가 많을 경우에는 불필요한 엣지가 많이 형성되게 되며 알고리즘의 성능을 저하시키는 결과를 초래한다. 그래프샤프닝은 노이즈로부터 기인한 연결이나 성능 저하의 요인이 되는 불필요한 연결엣지를 제거하기 위하여 W행렬을 변화시켜 성능을 향상 시키는 알고리즘이다 (Shin, 2006). 대부분의 graph-based semi-supervised learning 알고리즘에서는 W행렬은 고정되어 있고 대칭적인 구조를 띄고 있다. 즉, 엣지는 방향성에 대한 고려없이 데이터간의 유사도만을 반영하는 무방향성 엣지이다. 하지만 W행렬을 레이블이 된 데이터와 레이블이 안된 데이터의 관계로 설명할 때는 모든 관계가 대칭적일 필요는 없다. 즉, 지금까지 모든 연결엣지의 가중을 동일하게 준 것과는 달리 정보 흐름의 중요도에 따라 가중을 차별화하여 줄 수 있다. 첫째, 레이블이 된 데이터에서 레이블이 안된 데이터로의 엣지는 반대의 경우보다 좀더 유용한 정보를 전달한다고 볼 수 있다. 왜냐하면 레이블이 안된 데이터에서 레이블이 된 데이터로의 전달은 불확실한 정보를 포함할 확률이 높다고 보여지기 때문이다. 둘째, 서로 다른 레이블을 가진 데이터들간의 엣지는 양방향으로 도움이 되지 않는 정보를 전달할 수 있는 가능성이 있다. 마지막으로 레이블이 안된 데이터들간의 정보 전달은 다르게 생각해야 한다. 레이블이 안된 데이터들간의 정보 흐름도 그 중요도에 차이가 있을 수 있으나 어떠한 정보가 더 중요한지 미리 알기 어렵기 때문에, 이들 사이의 연결 엣지에는 방향성을 고려하지 않는다. 그래프샤프닝에 대한 자세한 내용은 (Shin, 2006)을 참조하기 바란다. 일반적인 graph-based semi-supervised learning 알고리즘의 W행렬을 다음과 같은 블록 행렬로 표현 한다면

$$W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix},$$

그래프샤프닝에 의하여 변화된 유사행렬은 다음의 식(4)와 같다.

$$W_s = \begin{bmatrix} diagonalmatrix & 0 \\ W_{ul} & W_{uu} \end{bmatrix} \tag{4}$$

여기서 W_{lu} 는 레이블 안된 데이터에서 레이블 된 데이터로의 엣지의 가중치를 의미하며 나머지도 마찬가지 방법으로 읽으면 된다 ($u \to l$). 그래프샤프닝에 의한 출력값은 다음과 같다.

$$f_u = \mu(I + \mu(D_{uu} - W_{uu}))^{-1} W_{ul} y_l . \quad (5)$$

식(5)는 식(3)과 비슷하지만 레이블 된 데이터를 제외했다는 것이 다르다. 그래프샤프닝에서는 레이블 된 데이터의 출력값에는 정보의 손실이 없으므로 $f_l = y_l$ 의 관계가 만족된다. 다음의 그림1과 그림2는 그래프샤프닝 전후의 변화를 간단히 도식화한 것으로서, 그래프샤프닝에 의해 일부 엣지가 제거되거나 방향성이 생기게 됨을 보여준다.

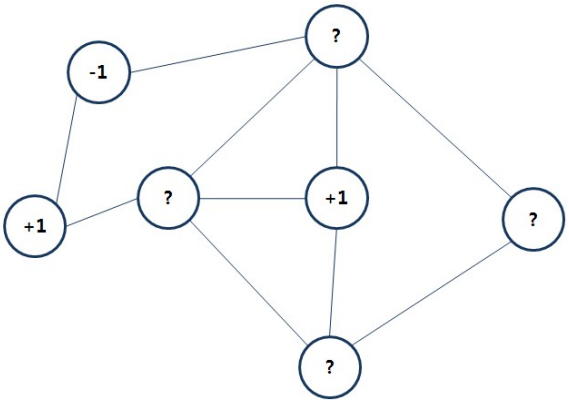


그림1. 일반적인 유사행렬 W에 의한 그래프 : 레이블이 된 노드는 '+1'과 '-1'로 레이블이 안 된 노드는 '?'로 표기되어 있으며 엣지에 방향성이 없다.

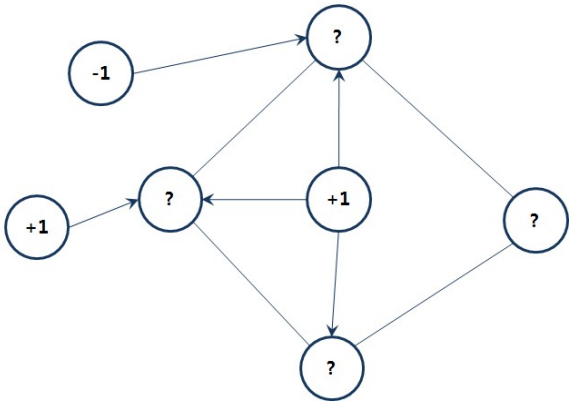


그림2. 그래프샤프닝의 W_s 에 의한 Sharpened 그래프 : 정보흐름의 중요도에 따라 일부 엣지가 방향성을 띄게 되거나 제거된다.

2.3 앙상블 기법 (Ensemble Method)

앙상블 (ensemble) 기법 또는 앙상블 네트워크는 다양한 모델들을 결합시킨 후 멤버 네트워크들의 총론적 값을 취함으로서 개별 네트워크들이 갖는 출력값 에러의 편기 (bias) 또는 오차변동 (variance)을 감소시켜 성능을 향상시키는 방법이다 (Briman, 1996; Perrone, 1993; Sharkey, 1997; Tumer, 1996). 앙상블 네트워크의 구성 멤버들간의 다양성은 bagging (Briman, 1996) 또는 boosting (Freund, 1996)과 같이 각 멤버가 사용하는 학습 데이터에 변동 (perturbation)을 줘서 얻을 수도 있으나, 데이터 변동이 없더라도 멤버가 갖는 학습 파라미터 및 구조에만 변동을 줌으로서 얻을 수도 있다 (Shin, 2001).

본 연구에서는 graph-based semi-supervised learning 알고리즘을 기본 학습 알고리즘으로 하여 하나의 앙상블 네트워크를 구성한다. 멤버 네트워크들은 동일한 데이터 셋을 사용하며 학습 파라미터인 κ 와 μ 의 값에만 변동을 준다. 그 과정은 다음과 같다.

1. 그림 3과 같이, 유사행렬 W의 파라미터인 κ ($\kappa = 1, \cdots, K$)와 식(2) 또는 식(5)의 파라미

터인 μ ($\mu = 1, \cdots, M$)에 다양한 값을 주어, 총 $|K| \times |M|$ 의 멤버 네트워크를 가진 하나의 앙상블 네트워크를 구성한다.

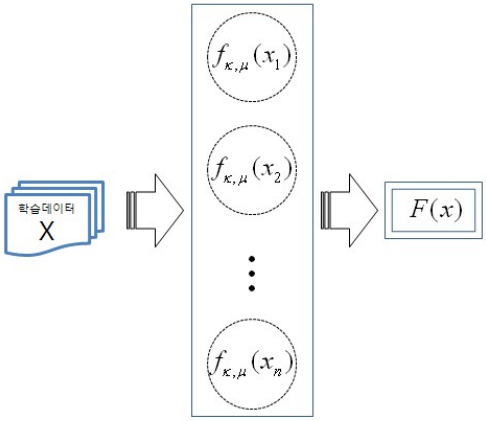


그림3. 앙상블 네트워크의 구조

2. 각 멤버 네트워크들의 출력값을 단순평균하여 식(6)과 같이 앙상블 네트워크의 최종 출력값을 계산한다.

$$F(x) = \frac{\sum_{i=1}^{|K| \times |M|} f_{\kappa,\mu}(x_i)}{|K| \times |M|} \qquad \kappa = 1, \cdots, K \quad \mu = 1, \cdots, M \tag{6}$$

3 실험 결과

본 연구에서 제안된 방법은 인공데이터와 벤치마크 데이터를 이용하여 실험, 비교되었다. 인공데이터에 대한 실험은 본 논문에서 제기된 문제, 즉 학습 파라미터 변동이 graph-based semi-supervised learning 알고리즘의 성능에 얼마나 영향을 주는지 가시화하기 위하여 실행되었다. 벤치마크 데이터는 검증된 실제데이터를 사용하여 본 연구에서 제안된 방법의 성능 및 안정성을 비교, 파악하기 위하여 사용되었다. 알고리즘의 비교척도로는 Area Under the ROC curve (AUC)를 사용하였다. AUC는 레이블이 되지 않은 데이터로부터 얻은 출력값이 원래의 레이블과 일치하는가를 표현한 일종의 기대값으로 0에서 1까지의 값을 가지며 1에 가까울수록 정확하다. 설명의 편의를 위하여, 데이터가 레이블이 된 경우를 labeled라 표기하고 레이블이 안된 경우를 unlabeled라 하겠다. 또한 실험의 비교를 위해 graph-based semi-supervised learning 알고리즘만 기본 학습 알고리즘으로 사용한 경우는 single의 original 즉, single-original로 표기하고 기본 학습 알고리즘에 앙상블 네트워크를 구성한 경우는 ensemble의 original 즉, ensemble-original로 표기한다. 그래프 샤프닝에 의해 구성된 경우는 single-sharpened와 ensemble-sharpened로 표기한다.

3.1 인공 데이터

실험은 그림 4의 인공적으로 만들어진 Two-moon 데이터 셋으로 실험하였다. 데이터는 각각 250개씩 두 개의 클래스로 구성되어 있다. 각 클래스는 랜덤하게 선택된 245개의 unlabeled 데이터와 5개의 labeled 데이터를 사용하였다. 유사행렬 W는 κNN 방법을 바탕으로 노드간 연결엣지 $(i \sim j)$가 결정되었으며 식(1)에 의하여 가중치가 부여되었다. κNN 의 κ 는 $\kappa \in \{3,5,10,20,30\}$ 으로 설정하였고 smoothing 파라미터 μ 는 $\mu \in \{0.01,0.1,1,10,100,1000\}$ 으로 설정하였다.

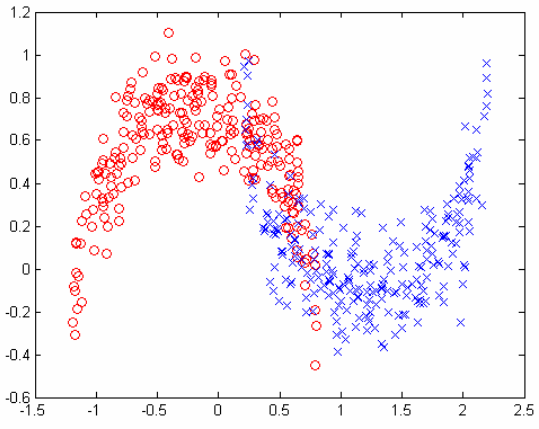
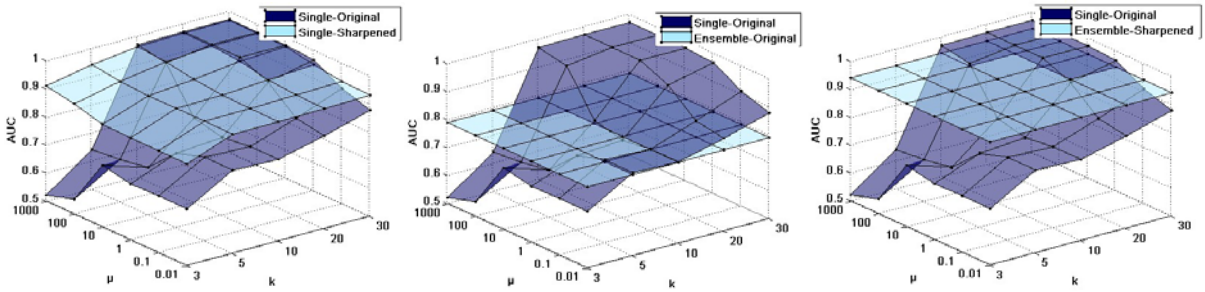


그림 4. Two-moon

그림 5는 학습 파라미터 κ 와 μ 의 변화에 따른 AUC의 변화를 보여주고 있는 것으로 single-original과 나머지 다른 세가지 방법의 비교를 하였다. Single-original의 경우, 학습 파라미터 변동에 따라 성능이 민감하게 변화하는데 반해 그래프샤프닝을 적용한 single-sharpened의 경우는 그림5(a)에서처럼 안정적이면서도 향상된 성능을 보여주고 있다. 그림5(b)는 앙상블의 효과를 보여주는 것으로 성능은 single original의 평균치에 해당한다. 마지막으로 그림5(c)에서는 앙상블 기법과 그래프샤프닝의 조합효과를 보여준다. Ensemble-sharpened의 AUC가 앞의 두 가지 비교의 경우보다 향상되었음을 알 수 있다.



(a) Single-original Vs. Single-sharpened (b) Single-original Vs. Ensemble-original (c) Single-original Vs. Ensemble-sharpened

그림 5. 학습 파라미터 κ 와 μ 의 변화에 따른 AUC의 변화

3.2 실제 데이터

실제데이터는 http://www.kyb.tuebingen.mpg.de/ssl-book/benchmarks.html에서 얻을 수 있는 벤치마크 데이터를 사용하였다. 각 데이터 셋은 레이블이 10개만 주어진 경우와 100개가 주어진 경우로 나누어지며 각각은 또한 12개의 데이터 셋으로 분할되어있다. 데이터의 기본적인 특성은 표 1에 나타나있다. 벤치마크 데이터의 실험은 κ 를 $\kappa \in \{3,5,7,10,20,30\}$ 으로 하고 μ 는 $\mu \in \{1,1000\}$ 으로 하였다.

표 1. 벤치마크 데이터 셋

Data set	Classes	Dimension	Points	Comment
Digit1	2	241	1,500	Artificial
USPS	2	241	1,500	Imbalanced
BCI	2	117	400	small, noisy
g241c	2	241	1,500	Artificial
g241n	2	241	1,500	-

표2는 각각의 데이터 셋에 대해, “single대 ensemble” 그리고 “original대 sharpened”로부터 가질 수 있는 모든 조합의 경우에 대한 AUC 비교결과를 보여준다. 각 경우, 두 알고리즘 A, B의 성능을 비교하기 위한 검정방법으로서 Wilcoxon signed-ranks test를 사용하였으며 p값이 작을수록 A와 B가 유의적인 차이를 보이는 것을 의미한다 (Demšar, 2006). 표에 기재된 값은 AUC 평균값으로 12개로 분할된 데이터 셋들의 AUC 평균을 의미한다. 표준 편차도 동일한 방법으로 계산되었다.

표2의 ‘C’로 표기된 열의 P-value를 참고하여 single과 ensemble을 비교해 볼 때, 앙상블 네트워크가 성능향상에 통계적으로 유의한 영향을 주었음을 알 수 있다. 즉, ensemble-(original 또는 sharpened)의 AUC가 single-(original 또는 sharpened)에 비해 증가하였고, 특히 $\mu = 1$ 일 때 그 효과가 더 커짐을 알 수 있었다. 한편, ‘A’로 표기된 열과 ‘B’로 표기된 열의 P-value는 각각 single 또는 ensemble의 경우에 대한 그래프샤프닝의 효과를 보여준다. (Single 또는 ensemble)-sharpened의 AUC가 (single 또는 ensemble)-original에 비해 각각 증가하였고 특히 $\mu = 1,000$ 일 때 보다 큰 효과를 볼 수 있었다. 본 연구에서 제안하는 두 가지 방법의 조합, 즉 ensemble-sharpened의 경우 $\mu = 1$ 일 때와 $\mu = 1,000$ 일 때 모두 전반적으로 가장 높은 AUC를 보여주고 있다.

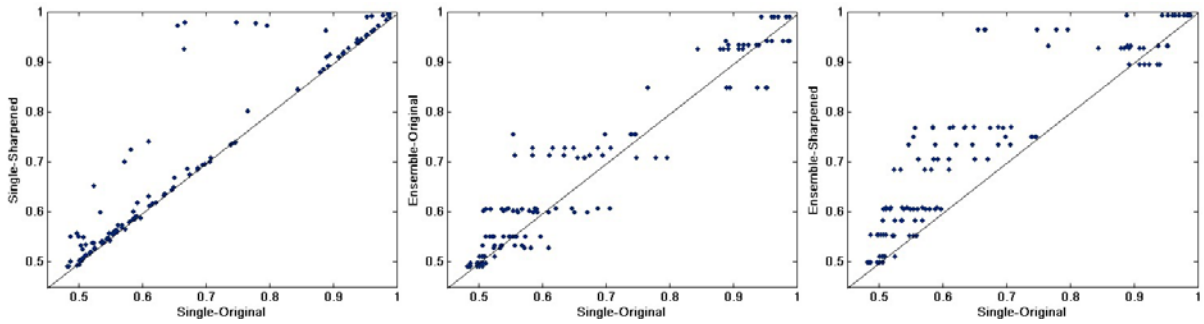
표 2. 벤치마크 데이터 셋의 결과

(A : Single-original Vs. Single-sharpened, B : Ensemble-original Vs. Ensemble-sharpened, C : Single-sharpened Vs. Ensemble-sharpened)

Data set			Single													A p-value	Ensemble		B p-value	C p-value
			Original							Sharpened										
			K=3	K=5	K=7	K=10	K=20	K=30	평균	K=3	K=5	K=7	K=10	K=20	K=30	평균		Original	Sharpened		
μ=1	(1) Digit1	10 Labeled	0.84	0.88	0.89	0.89	0.91	0.92	0.89±0.05	0.84	0.88	0.88	0.89	0.91	0.91	0.89±0.05	0.0001	0.92±0.04	0.93±0.04	0.0027	0.0000
		100 Labeled	0.94	0.96	0.96	0.97	0.98	0.99	0.97±0.02	0.94	0.96	0.96	0.97	0.98	0.99	0.97±0.02		0.99±0.01	0.99±0.01		
	(2) USPS	10 Labeled	0.55	0.70	0.74	0.74	0.75	0.75	0.70±0.10	0.55	0.70	0.74	0.73	0.74	0.74	0.70±0.10	0.0000	0.75±0.07	0.75±0.06	0.0025	0.0000
		100 Labeled	0.89	0.91	0.92	0.92	0.94	0.94	0.92±0.02	0.89	0.91	0.92	0.93	0.94	0.94	0.92±0.02		0.93±0.01	0.90±0.01		
	(3) BCI	10 Labeled	0.53	0.51	0.51	0.50	0.51	0.51	0.51±0.03	0.53	0.51	0.51	0.50	0.51	0.50	0.51±0.03	0.0000	0.51±0.03	0.51±0.03	1.0000	0.0000

		100 Labeled	0.56	0.55	0.56	0.55	0.52	0.52	0.54±0.02	0.56	0.56	0.56	0.56	0.54	0.54	0.55±0.02		0.55±0.02	0.55±0.01		
	(4) g241c	10 Labeled	0.51	0.52	0.54	0.55	0.57	0.59	0.55±0.06	0.51	0.52	0.54	0.55	0.57	0.59	0.55±0.06	0.0000	0.60±0.05	0.61±0.05	0.0003	0.0000
		100 Labeled	0.56	0.59	0.61	0.63	0.68	0.70	0.63±0.06	0.56	0.58	0.61	0.63	0.67	0.69	0.63±0.06		0.71±0.03	0.77±0.02		
	(5) g241n	10 Labeled	0.51	0.54	0.55	0.56	0.58	0.59	0.56±0.04	0.51	0.54	0.55	0.56	0.58	0.59	0.56±0.04	0.1601	0.60±0.04	0.61±0.04	0.0003	0.0000
		100 Labeled	0.59	0.62	0.64	0.65	0.69	0.71	0.65±0.05	0.58	0.62	0.64	0.65	0.69	0.71	0.65±0.05		0.73±0.04	0.77±0.05		
μ=1,000	(1) Digit1	10 Labeled	0.77	0.89	0.89	0.94	0.95	0.95	0.90±0.09	0.80	0.91	0.91	0.94	0.95	0.95	0.91±0.08	0.0000	0.85±0.10	0.93±0.05	0.0000	0.0000
		100 Labeled	0.89	0.96	0.95	0.98	0.99	0.99	0.96±0.04	0.96	0.99	0.99	0.99	0.99	0.99	0.99±0.01		0.94±0.03	0.99±0.01		
	(2) USPS	10 Labeled	0.42	0.53	0.52	0.57	0.58	0.61	0.54±0.13	0.45	0.60	0.65	0.70	0.72	0.74	0.64±0.20	0.0000	0.53±0.10	0.68±0.13	0.0000	0.6817
		100 Labeled	0.67	0.66	0.67	0.75	0.78	0.80	0.72±0.10	0.93	0.97	0.98	0.98	0.98	0.97	0.97±0.02		0.71±0.08	0.96±0.01		
	(3) BCI	10 Labeled	0.51	0.48	0.49	0.48	0.50	0.50	0.49±0.02	0.52	0.49	0.50	0.49	0.49	0.49	0.50±0.03	0.0000	0.49±0.02	0.50±0.03	0.0004	0.0002
		100 Labeled	0.51	0.49	0.50	0.50	0.50	0.51	0.50±0.02	0.55	0.55	0.56	0.55	0.53	0.53	0.55±0.02		0.50±0.02	0.55±0.02		
	(4) g241c	10 Labeled	0.51	0.52	0.54	0.55	0.57	0.60	0.55±0.06	0.51	0.52	0.54	0.54	0.57	0.59	0.54±0.06	0.0000	0.55±0.05	0.61±0.05	0.0000	0.3557
		100 Labeled	0.55	0.59	0.62	0.65	0.69	0.71	0.63±0.06	0.55	0.59	0.62	0.64	0.68	0.70	0.63±0.06		0.61±0.04	0.73±0.03		
	(5) g241n	10 Labeled	0.51	0.53	0.54	0.54	0.56	0.57	0.54±0.04	0.50	0.53	0.54	0.54	0.56	0.57	0.54±0.04	0.0000	0.53±0.03	0.58±0.04	0.0000	0.6839
		100 Labeled	0.56	0.59	0.59	0.61	0.65	0.67	0.61±0.05	0.57	0.60	0.62	0.63	0.67	0.69	0.63±0.05		0.60±0.04	0.70±0.06		

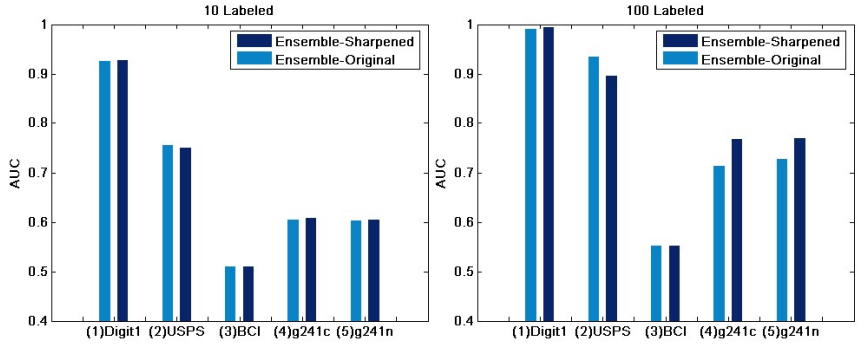
그림 6 은 single-original 과 나머지 다른 세가지 방법의 비교로 대각선 위쪽으로 점들이 많으면 y 축의 알고리즘이 x 축의 알고리즘보다 성능이 좋다는 의미를 갖는다. 그림 6(a)에서는 single-sharpened 의 성능이 single-original 보다 좋거나 같음을 보여준다. 그림 6(b)에서는 ensemble 효과를 보여주는 것으로 ensemble 의 효과가 있기는 하나, 그림 6(a)에서만큼 확연하지는 않다. 그림 6(c)에서는 본 연구에서 제안하는 방법, 즉 ensemble 과 sharpened 조합이 월등히 성능을 향상시킴을 보여준다.



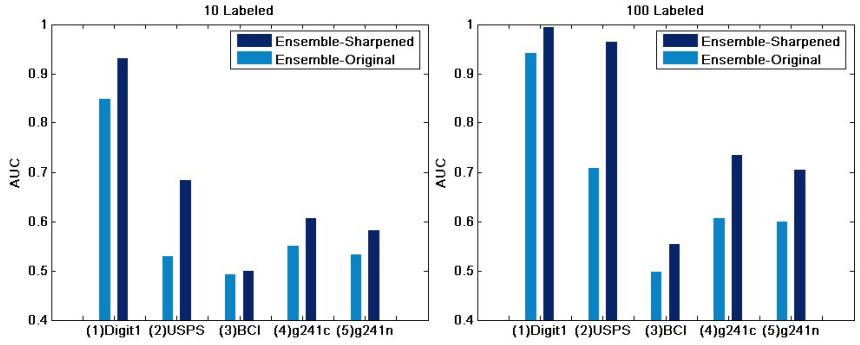
(a)Single-original Vs. Single-sharpened (b)Single-original Vs. Ensemble-original (c)Single-original Vs. Ensemble-sharpened

그림 6. Single-original 과 나머지 다른 세 가지 방법의 비교로 대각선 위쪽으로 점들이 많으면 y 축의 알고리즘이 x 축의 알고리즘보다 성능이 좋다.

그림 7 은 앙상블 네트워크에 적합한 그래프샤프닝의 효과를 보여주기 위한 것으로 ensemble-original 과 ensemble-sharpened 의 AUC 를 막대 그래프로 비교한 것이다. $\mu = 1$ 일 때 10 labeled 경우 ensemble-original 과 ensemble-sharpened 가 동일하다고 볼 수 있다. 100 labeled 의 경우는 USPS 를 제외하고 ensemble-sharpened 가 ensemble-original 보다 높거나 동일하다고 볼 수 있다. $\mu = 1,000$ 일 때는 10 labeled 와 100 labeled 모두 ensemble-sharpened 가 높게 나왔다. 따라서, 그래프샤프닝은 labeled 된 데이터가 많을수록, smoothing 파라미터 값이 클수록 효과적임을 알 수 있다.



(a) μ =1



(b) μ =1,000

그림 7. Ensemble-original 과 ensemble-sharpened 의 AUC 비교

실험결과를 요약하면 다음과 같다. 첫째, 그래프샤프닝의 효과는 original과 sharpened의 비교함으로써 측정할 수 있었다. 즉, sharpened를 사용했을 경우, original 보다 AUC가 최대 0.25 (USPS, $\mu = 1,000$, 100 labeled), 최소 -0.03 (USPS, $\mu = 1$, 100 labeled)이 증가함을 알 수 있었고 이는 통계적으로 유의한 수준이다. 그래프샤프닝의 효과는 레이블이 된 데이터의 수가 많을수록, μ 의 값이 클수록 보다 유의하게 나타났다. 둘째, 앙상블기법의 효과는 original과 ensemble을 비교함으로써 알 수 있었는데, ensemble을 사용할 경우 original에 비해 AUC가 최대 0.14 (g241c, $\mu = 1$, 100 labeled), 최소 -0.05 (Digit1, $\mu = 1,000$, 10 labeled)의 통계적으로 유의한 증가 효과가 있었다. 셋째, 두 방법을 조합한 효과는 single-original과 ensemble-sharpened를 비교함으로써 측정할 수 있었다. 시너지 효과로 인해 ensemble-sharpened를 사용할 경우 AUC가 최대 0.24 (USPS, $\mu = 1,000$, 100 labeled), 최소 -0.02 (USPS, $\mu = 1$, 100 labeled)의 증가 효과가 있었다.

4 결론

본 연구에서는 학습 파라미터 선택 및 데이터 셋의 노이즈 정도에 상관없이 graph-based semi-supervised learning 알고리즘이 안정된 성능을 얻을 수 있게 하기 위한 방법으로서 앙상블기법과 그래프샤프닝의 사용을 제안하였다. 다양한 학습 파라미터 값들로 학습된 모델들을 앙상블 네트워크의 멤버 네트워크들로 사용함으로써, 시행착오를 거쳐서 한 개의 학습 파라미터 값을 선택하는 모델선정 과정이 불필요하게 되었다. 또한 앙상블 네트워크는 개별 네트워크에 비해 에러의 분산과 편기를 줄여주는 효과가 있으므로 학습 성능이 보다 안정화되었고 향상되었다. 한편, 데이

터의 노이즈에 기인한 노이즈 엣지나 학습에 불필요한 엣지들을 그래프샤프닝을 적용하여 제거함으로써 개별 네트워크의 성능이 유의하게 향상되었다. 제안하는 방법의 효과는 인공데이터 셋 및 벤치마킹 데이터 셋에 대하여 측정되었다. 실험결과, 앙상블기법과 그래프샤프닝 각각 semi-supervised learning 알고리즘의 성능을 향상시키는데 유의한 영향을 미침을 알 수 있었다. 또한, 두 가지 방법의 조합을 사용할 경우, 즉 그래프샤프닝을 적용한 모델들로 앙상블 네트워크를 구성하였을 경우, 그 시너지 효과는 각각을 독립적으로 사용할 때 보다 월등히 효과적임을 보였다. 전반적으로 결론을 내리자면, 제안하는 방법을 적용하였을 경우 알고리즘의 성능이 일관되게 향상되는 것을 알 수 있었으며, 개별 네트워크의 성능 대비 그 효과가 많게는 AUC가 0.24나 증가하는 것을 알 수 있었다.

Acknowledgements

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD), by the grant for Post Brain Korea 21, and by the grant from Ajou university.

참고문헌

M. Belkin, I. Matveeva, and P. Niyogi. Regularization and regression on large graphs. Lecture Notes in Computer Science(In COLT), 624-638,2004

L. Breiman. Bagging Predictors. Machine Learning, 24, 123-140, 1996

J. Demšar. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7,1-30,2006

Y. Freund, and R.E. Schapire. Experiments with a new boosting algorithm. In Machine Learning: Proceedings of the Thirteenth International Conference, 148-156, 1996

M.P. Perrone. Improving Regression Estimation: Averaging Methods for Variance Reduction with Extension to General Convex Measure Optimization. PhD Thesis, Department of Physics, Brown University, Providence, RI, 1993

A.J.C. Sharkey. Combining Diverse Neural Nets. The Knowledge Engineering Review, 12(3), 231-247, 1997

H. Shin, and S. Cho. Pattern selection using the bias and variance of ensemble. Journal of the Korean Institute of Industrial Engineers, 28(1), 112-127, 2001

H. Shin, N.J. Hill, and G. Raetsch. Graph-based semi-supervised learning with sharper edges. Lecture Notes in Artificial Intelligence (LNAI 4212), 402-413, 2006a

H. Shin and K. Tsuda. Prediction of Protein Function from Networks. in Book: Semi-Supervised Learning, Edited by O. Chapelle, B. Schoelkopf, A. Zien. Chapter 20. pp. 339–352. MIT press, 2006b.

H. Shin, A.M. Lisewski, and O. Lichtarge. Graph Sharpening plus Graph Integration: A Synergy that Improves Protein Functional Classification. Bioinformatic. Oxford University Press (to appear),

2007

K. Tumer, and J. Ghosh. Error Correlation and Error Reduction in Ensemble Classifiers. Connection Science, 8, 385-404, 1996

L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Transactions on Systems, Man and Cybernetics, 22(3), 1-19, 1992

X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. ICML, 20, 912-919, 2003

X. Zhu. Semi-supervised learning with graphs. Ph.D. dissertation, Carnegie Mellon University, Pennsylvania, USA, 2005

http://www.kyb.tuebingen.mpg.de/ssl-book/benchmarks.html