The Transactions of the Korean Institute of Electrical Engineers Vol. 61, No. 9, pp. 1336~1339, 2012
http://dx.doi.org/10.5370/KIEE.2012.61.9.1336

SVM-KNN-AdaBoost를 적용한 새로운 중간교사학습 방법

Semisupervised Learning Using the AdaBoost Algorithm with SVM-KNN

이 상 민* · 연 준 상** · 김 지 수*** · 김 성 수†
(Sangmin Lee · Jun-Sang Yeon · Ji-Soo Kim · Sung-Soo Kim)

Abstract - In this paper, we focus on solving the classification problem by using semisupervised learning strategy. Traditional classifiers are constructed based on labeled data in supervised learning. Labeled data, however, are often difficult, expensive or time consuming to obtain, as they require the efforts of experienced human annotators. Unlabeled data are significantly easier to obtain without human efforts. Thus, we use AdaBoost algorithm with SVM-KNN classifier to apply semisupervised learning problem and improve the classifier performance. Experimental results on both artificial and UCI data sets show that the proposed methodology can reduce the error rate.

Key Words : Semisupervised learning, SVM, KNN, AdaBoost

1. 서 론

인공지능을 이용하여 패턴을 인식하는 방법은 기본적으로 데이터에 내재된 정보를 부호화하는 과정을 수반한다. 패턴 인식에서 데이터를 분석하는 방법은 크게 분류와 군집화의 두 가지로 나눌 수 있다. 분류 문제와 군집화 문제의 가장 큰 차이점은 학습에 사용되는 데이터에 원하는 출력정보가 함께 주어지는지의 여부로 볼 수 있다. 출력정보(클래스 정보)가 주어진 데이터를 처리할 때는 분류기법을 사용하고 출력정보가 없는 데이터를 처리할 때에는 군집화 기법을 쓴다[1]. 패턴 인식 분야에서 분류 문제와 군집화 문제는 기계학습에서의 교사학습(supervised learning)과 비교사학습(unsupervised learning)문제로 연결된다. 교사학습의 경우에는 모든 데이터에 대해 출력정보가 요구되고 매우 많은 양의 학습데이터를 필요로 한다. 반면, 비교사학습의 경우 데이터의 출력정보가 없고, 유사한 특성을 가진 정보들을 비슷한 카테고리로 묶는 방법이다. 학습 과정에서 출력정보가 없는 데이터의 활용 과정은 데이터 특성으로부터 만들어진다. 하지만 데이터의 특성을 잘 파악하지 못한다면 학습과정에서 좋은 성능을 기대할 수 없는 문제가 발생한다. 따라서 최근에는 교사학습과 비교사학습을 혼합한 중간교사학습(semisupervised learning)의 연구가 활발하다. 중간교사학습의 기본 개요는 적은 양의 출력정보가 있는 데이터와 함께 많은 양의 출력정보가 없는 데이터들을 활용하는 방법이다.

본 논문에서는 새로운 중간교사학습 방법으로서, SVM분류기의 정확도를 개선하는 SVM-KNN분류기에 AdaBoost알고리즘을 특성을 결합한 새로운 모델을 제안한다. 실험을 통하여 제안된 SVM-KNN-AdaBoost 결합 모델이 기존 분류기에 비해 더 나은 성능을 나타내는 특성이 있음을 확인하였다.

* 준 회 원 : 충북대학교 전기공학과 석사과정
** 정 회 원 : 충북대학교 전기공학과 박사과정
*** 정 회 원 : 충북대학교 지구환경과학과 교수 · 공박
† 교신저자, 정회원 : 충북대학교 전기공학과 교수 · 공박
E-mail: sungkim@chungbuk.ac.kr
접수일자 : 2012년 7월 16일
최종완료 : 2012년 8월 24일

2. SVM-KNN-AdaBoost 결합 모델

2.1 SVM and KNN Classifier

서포트 벡터 머신(Support Vector Machine, SVM)은 일반화 오차를 최소화 할 수 있는 방향으로 학습을 수행하는 선형 분류기이다. 실제 문제에서 많은 분류기는 데이터를 선형적으로 나눌 수 없다. 이 문제를 해결하기 위해 커널 트릭(Kernel Trick)을 이용하여 비선형 경계면을 생성한다[2]. 이 커널의 개념은 비선형적인 훈련 데이터를 커널 함수를 통해 높은 차원의 공간으로 이동하는 것이다. 선형 결정 특징의 구조는 특징 공간이 $\Phi(x_i) \bullet \Phi(x_j) = K(x_i, x_j)$ 의 내적을 통해서만 수행된다. 여기서 함수 $K = R^d \times R^d \rightarrow R$은 커널 함수라고 한다. SVM(Support Vector Machine)은 최적화 문제를 풀어야 한다. 최적화 문제란 마진 $2/||w||$의 최대화와 훈련 오류의 최소화를 말한다. $1/||w||$의 최대화는 $\|w\|^2$의 최소화와 같다. 이것은 2차 항을 가진 convex함수이다. 따라서 다음의 식으로 표현할 수 있다.

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{\ell} \xi_i \tag{1}$$

$$\text{s.t } y_i [w^T \Phi(x_i) + b] \geq 1 - \xi_i \quad {}^{\forall} i = 1, \ldots, \ell \tag{2}$$

$$\xi_i \geq 0 \quad {}^{\forall} i = 1, \ldots, \ell \tag{3}$$

w는 법선 벡터로 초평면의 방향을 나타내고 b는 임계값으로 위치를 나타낸다. C는 tradeoff로 여유와 오분류의 관계를 나타낸다. 이 최적화 문제의 라그랑제(Lagrange) 형식은 다음과 같다.

$$g = \frac{1}{2}w^T w + C\sum_{i=1}^{\ell}\xi_i - \sum_{i=1}^{\ell}\alpha_i[y_i(w^T\Phi(x_i)+b)-1+\xi_i] \quad (4)$$
$$-\sum_{i=1}^{\ell}\lambda_i\xi_i$$

모든 $i=1,\ldots,\ell$ 에 대한 라그랑제 승수 $\alpha_i \geq 0$, $\lambda_i \geq 0$ 이다. 이 라그랑제 조건하에 다음과 같은 분류기를 만들 수 있다.

$$g(x) = sign[w^T\Phi(x)+b] = sign[\sum_{i=1}^{\ell}\alpha_i y_i K(x_i,x)+b] \quad (5)$$

이 듀얼(Dual) 문제의 해로 대변되는 α를 구할 수 있다.

$$\text{maximize: } W(\alpha) = \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \quad (6)$$

$$\text{s.t. } :\sum_{i=1}^{\ell}\alpha_i y_i = 0 \text{ 이고, } 0 \leq \alpha_i \leq C, \quad i=1,\ldots,\ell$$

임계값 b는 모든 서포트 벡터 $b = y_j - \sum\alpha_i y_i K(x_i,x_j)$의 평균으로 계산된다.

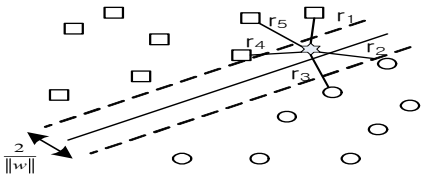


그림 1 SVM과 KNN의 원리
Fig. 1 Principle of SVM-KNN

KNN(K-Nearst Neighbor)알고리즘은 1968년 Cover, Hart에 의해 제안되었다[3]. KNN분류기를 설계할 때에는 고려해야 할 사항이 있다. 우선 K의 값을 설정해주어야 한다. K가 1인 경우에는 바로 이웃한 데이터에만 의존하여 노이즈에 민감한 결과를 초래한다. 반면 K값이 지나치게 커지면 주어진 데이터 주변의 영역만을 중심으로 분류가 수행되는 것이 아니라 전체 데이터 영역에서 영역이 차지하는 비율, 즉 선험 확률에 의존하여 분류를 수행하는 결과를 초래하게 된다. KNN분류기는 주어진 데이터와 학습데이터간의 거리를 바탕으로 이웃 데이터를 찾는 것이다. 그림 1은 SVM과 KNN알고리즘을 결합한 SVM-KNN방식의 원리를 보여준다. 출력정보가 있는 데이터를 이용하여 SVM 분류기를 통해 여유를 최대화하고 출력정보가 없는 새로운 데이터가 주어졌을 때 이웃한 출력정보가 있는 데이터와의 거리를 계산하여 라벨을 부여한다.

2.2 AdaBoost

AdaBoost는 분류기의 학습 방법뿐 아니라 결합 방법도 함께 고려하고 있다[1]. AdaBoost는 같은 데이터 집합을 반복해서 사용하되, 학습할 때마다 각 데이터에 대한 가중치를 적절히 조정하여 학습에 변화를 준다. Weak Learning Classifier는 개별적인 분류기로 그 분류 성능이 각각 다르다. 이 분류기들의 가중치를 조정하여 하나의 강력한 분류기(Strong learning Classifier)를 구축한다. 이러한 결합 방법은 그림 2과 같이 나타낼 수 있다.

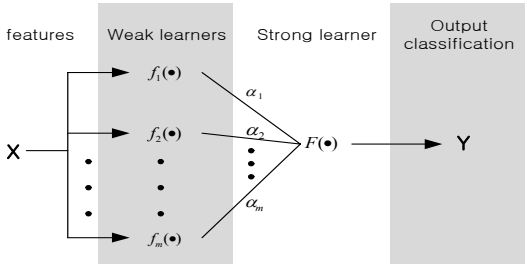


그림 2 부스팅에 의한 결합 방법
Fig. 2 Boosting Method

AdaBoost알고리즘은 이전 단계의 분류기의 학습 결과를 활용하여 다음 단계의 학습에 사용될 데이터에 가중치를 부여함으로써 분류기들 간의 차별성을 부여하고, 지수 오차함수의 측면에서 최적화된 결합 가중치를 찾아 분류기들을 결합한다. Fruend와 Schapire는 각각의 Weak learning Classifier의 오분류율이 0.5보다 작은 조건만 만족하면 분류기의 결합을 통해 학습데이터에 대한 오차를 기하급수적으로 감소시킬 수 있음을 보였다[4].

2.3 SVM-KNN과 AdaBoost 결합 모델의 개요

학습기들을 결합함으로써 성능을 향상시키기 위해서는 결합되는 각각의 학습기들이 상호보완적인 역할을 할 수 있도록 해야 한다.

그림 3은 본 연구에서 제안한 SVM-KNN-AdaBoost 모델을 나타낸다. 라벨이 있는 데이터를 SVM알고리즘을 사용하여 학습을 한 후, 이 출력정보를 사용하여 KNN 알고리즘을 적용하여 출력정보가 없는 데이터의 라벨을 부여하였다. 마지막으로 라벨을 부여받은 데이터와 기존의 데이터를 함께 AdaBoost알고리즘에 적용한다.

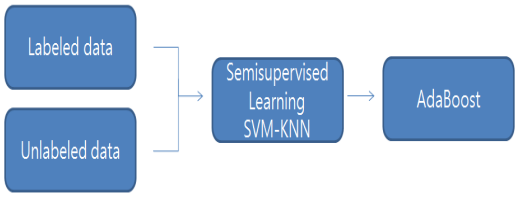


그림 3 SVM-KNN-AdaBoost 결합 모델
Fig. 3 SVM-KNN- AdaBoost Model

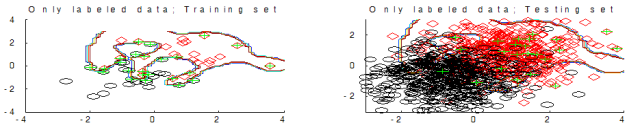
3. 실 험

본 논문에서는 인위적으로 생성한 데이터와 UCI 데이터를 사용하였다[5]. 각각의 실험에 SVM의 커널 함수의 매

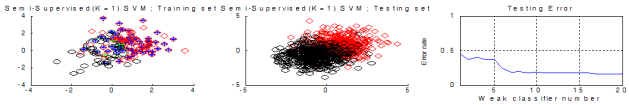
개변수 $\sigma=0.8$을 선택하였고, KNN은 유클리디안 거리를 이용하였으며 K의 값은 1과 5로 하였다. AdaBoost의 가중치는 초기화를 시킨 후 오분류에 따른 가중치를 계산하여 사용하였다. 비교 대상으로는 제한된 출력정보가 있는 데이터를 사용한 SVM분류기와 SVN-KNN분류기를 제안된 알고리즘과 비교하였다. 각각의 실험결과는 그래프에 나타내었다.

3.1 인공의 데이터에 관한 실험

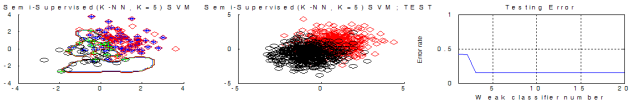
이진 분류 문제에 대해 50개의 출력정보가 있는 데이터와 150개의 출력정보가 없는 데이터를 생성하였다. 또한 1000개의 테스트 샘플도 생성했다. Weak learner는 20개를 사용하였다. 영역 1의 가우시안 평균은 (0.9 , 0.9), 영역 2의 가우시안 평균은 (-0.9 , -0.9)로 설정하였고 가우시안 분포는 단위공분산행렬을 사용하였다. 출력정보가 없는 데이터의 라벨을 부여하기 위해 KNN알고리즘을 사용하였고, K의 값을 1인 경우와 5인 경우를 고려하였다. 그림 4(a)의 첫 번째 그림은 출력정보가 있는 데이터를 SVM분류기를 통해 분류한 그림이고 Testing error rate는 0.417이었다. 두 번째 1000개의 테스트 데이터를 분류한 그림이다. (b)와 (c) 그림은 KNN알고리즘을 사용하여 각각 K=1, K=5로 설정하여 출력정보가 없는 데이터에 라벨을 부여한 후 SVM분류기를 통해 분류하였다. 그 후 테스트 데이터에 적용 후 AdaBoost알고리즘을 결합하였다. (b)의 경우 SVM-KNN을 한 Testing error rate는 0.4010이었고 AdaBoost모델과 결합하였을 때 M이 20인 경우 0.16이었다. (c)의 경우 SVM-KNN Testing error rate는 0.4030이었고 M이 20인 경우에는 0.146이었다. 이는 출력정보가 제한적인 SVM 분류기보다 출력정보가 없는 데이터를 이용한 중간교사학습 SVM-KNN의 오류율이 낮음을 알 수 있고, 중간교사학습 SVM-KNN분류기와 AdaBoost알고리즘을 결합한 모델이 오류율이 개선되는 것을 알 수 있다.



(a) 제한된 출력정보를 가진 데이터의 SVM 분류



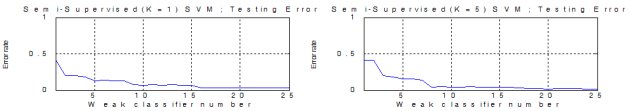
(b) 중간교사학습 SVM-KNN과 AdaBoost의 결합, K=1 m = 20



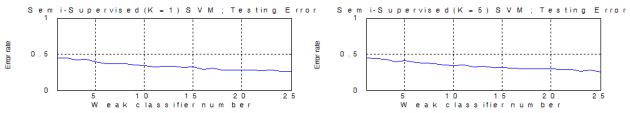
(c) 중간교사학습 SVM-KNN과 AdaBoost의 결합, K=5 m = 20

그림 4 결합모델의 비교 그래프

Fig. 4 Comparing the combining model



(a) 가우시안 분포 밀도의 오류율



(b) 가우시안 혼합 분포 밀도의 오류율

그림 5 50차원의 오류율 비교

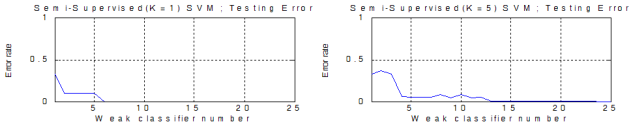
Fig. 5 Comparing error rate Using 50 Dimension

다음으로 50차원의 이진 분류 문제를 고려하였다. 앞선 실험과 마찬가지로 출력정보가 있는 50개의 데이터와 출력정보가 없는 150개의 데이터를 사용하였고 1000개의 테스트 샘플을 생성하였다. 영역 1에서의 가우시안 평균은 (0.23 ,…,0.23)이고 영역2에서의 가우시안 평균은 (-0.23,…,-0.23)이며, 가우시안 분포는 단위 공분산 행렬을 사용하였다. 그림 5의 (a)에서 출력정보가 있는 데이터만을 사용한 SVM분류기의 Testing error rate는 0.416이고 K가 1인 SVM-KNN Testing erorr rate는 0.411이고 M이 25일 때 Testing error rate는 0.026이다. K가 5인 Testing erorr rate는 0.411이고 M이 25일 때 Testing error rate는 0.01이다. 또 다른 인공의 실험으로는 중복이 많은 가우시안 혼합 분포 밀도에 대한 실험이다. 각 영역은 같은 사전 확률을 가지며 각 영역의 조건 분포는 $P(x|y=1)=0.49\ N(\mu_1, I) + 0.51N(\mu_2, I)$와 $P(x|y=-1)=0.49\ N(-\mu_1, I)+0.51N(-\mu_2, I)$이고 여기서 μ_1 =(0.25, 0.25, … ,0.25)와 차원의 반이 음의 값을 포함한 μ_2=(0.25, 0.25, … , -0.25, -0.25)이다. 그림 5의 (b)에서 출력정보가 있는 데이터만을 사용한 SVM분류기의 Testing error rate는 0.47이고 K가 1인 SVM-KNN Testing erorr rate는 0.45이고 M이 25일 때 Testing error rate는 0.267이다. K가 5인 Testing erorr rate는 0.447이고 M이 25일 때 Testing error rate는 0.25이다. 그림 5(b)의 결과에서 보듯 중간교사학습 SVM-KNN분류기의 경우 중복도가 많은 데이터의 경우 K의 값에 상관없이 비슷한 오류율을 나타내는 반면 AdaBoost알고리즘을 결합한 경우 오류율이 작아지는 것을 확인할 수 있다.

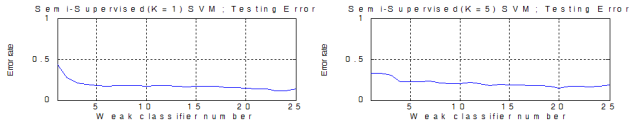
3.2 UCI 데이터에 관한 실험

실험 데이터로 UCI 데이터의 IRIS와 IONOSPHERE를 사용하였다. IRIS는 150개의 데이터로 구성되어있고, 15개의 데이터를 출력정보와 함께 사용하고 30개의 데이터를 출력정보를 지우고 사용하였다. 남은 105개의 데이터를 테스트 데이터로 사용하였다. 그림 6(a)는 IRIS의 출력정보가 있는 데이터만을 사용한 SVM분류기의 Testing error rate는 0.413이고 K가 1인 SVM-KNN Testing erorr rate는 0.33이고 M이 증가함에 따라 오류가 감소하는 것을 볼 수 있다. K가 5인 Testing erorr rate는 0.33으로 K가 1일 때와 같으며 M이 증가할수록 오류가 감소하는 현상을 보인다. IONOSPHERE데이터는 351개의 34차원 데이터로 이루어져

있으며 이중 36개의 데이터를 출력정보를 가지게 설정하고, 74개의 데이터의 출력정보를 없게 설정하였다. 남은 261개의 데이터를 테스트 데이터로 사용하였다. 그림 6(b) IONOSPHERE의 출력정보가 있는 데이터만을 사용한 SVM 분류기의 Testing error rate는 0.359이고 K가 1인 SVM-KNN Testing erorr rate는 0.43이고 M이 25일 때 Testing error rate는 0.15이다. K가 5인 Testing erorr rate는 0.33, M이 25일 때0.33이다.



(a) IRIS 데이터의 오류율



(b) IONOSPHERE 데이터의 오류율

그림 6 UCI 데이터의 오류율 비교

Fig. 6 Comparing error rate Using UCI data

인공의 데이터와 UCI데이터의 실험결과에서 제한된 양의 출력정보 데이터를 이용하여 SVM분류기를 단독으로 사용한 것보다 출력정보가 없는 데이터를 중간교사학습 SVM-KNN분류기를 사용한 방법이 오류율이 낮다는 것을 알 수 있다. 하지만 K값의 변화에 따라 성능이 좋아질 때도 있고 그렇지 못할 때도 있음을 알 수 있다. 제안된 알고리즘을 사용하였을 때 인공의 데이터와 UCI데이터 모두에서 오류율이 감소한 것을 볼 수 있다.

4. 결 론

중간교사학습의 성능개선을 위하여 본 연구에서는 SVM 분류기와 KNN알고리즘을 적용하여 출력정보가 없는 데이터에 라벨을 부여한 후 이들 데이터를 바탕으로 AdaBoost 알고리즘을 결합하여 오류를 줄여가는 방법을 제안하였고 실험을 통해 제안된 방법의 성능을 보였다.

인공의 데이터와 UCI데이터의 실험결과에서 SVM분류기 단독으로 사용한 것보다 SVM-KNN분류기가 상대적으로 오류율이 낮다는 것을 알 수 있다. 또한 제안한 알고리즘을 사용하였을 때 오류율이 지속적으로 감소하는 것을 보였다.

추후 연구로는 중간교사학습문제에서 분류하고자 하는 데이터에 적합한 SVM분류기의 매개변수, KNN알고리즘의 K값과 AdaBoost의 Weak Learning Classifier의 수에 대해 연구가 필요하다.

참 고 문 헌

[1] Hae Y. Park, Kwan Y. Lee, "Pattern and Machine Learning from Fundamental to Applications, Ihan Press, 2011

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in Proc. 5th Annu. Workshop Comput. Learn. Theory, 1992, pp. 144-152

[3] Dasarathy, B. V., Nearst Neighbor (NN) Norms, NN Pattern Classification Techniques. IEEE Computer Society Press, 1990.

[4] FREUND, Y. and SCHAPIRE, R. E. 1996. Experiments with a new boosting algorithm. In Machine Learning: Proceedings of the Thirteenth International Conference 148 156. Morgan Kauffman, San Francisco.

[5] A. Patel, S. Sundararajan, and S.Shevade, “Semisupervised classification using sparse Gaussian process regression,” in Proc. 21st Int. Jint Conf. Artif. Intell., 2009, pp. 1193-1198.