속성분할이 없는 향상된 협력학습 방법
(An Improved Co-training Method without Feature Split)

이 창 환 † 이 소 민 ††
(Chang-Hwan Lee) (So-Min Lee)

요 약 분류학습에서 높은 정확도를 유지하기 위해서는 충분한 분류 데이타가 필요하게 되는데 분류 데이타는 미 분류 데이타보다 생성하기가 어려운 경우가 많다. 따라서 미 분류 데이타를 활용하여 분류의 정확도를 향상 시키는 것은 큰 효용성을 가지며 이러한 미 분류 데이타를 활용하는 대표적인 학습방법 중의 하나는 협력학습(co-training) 알고리즘이다. 이는 데이타를 두 개의 독립적인 속성그룹으로 나누어 두 개의 분류자로 학습한 후 미 분류 데이타를 분류하고 그중 가장 신뢰성이 높은 데이타를 분류 데이타에 포함하고 이를 반복하는 학습모델이다. 하지만 이 방법은 전체 데이타의 속성을 독립적인 두개의 집합으로 분할하여야하는 제약이 있다. 따라서 본 연구에서는 이와 같은 문제점을 개선하여 보통의 데이타베이스에 적용시킬 수 있는 새로운 협력학습방법을 제시 하고자한다. 즉, 두 개의 독립적인 속성 그룹으로 나누는 가정을 따르지 않고 전체 속성을 사용할 수 있으며 두 개 이상의 분류자를 사용하는 새로운 협력학습방법을 제안하였다.

키워드 : 기계학습, 인공지능, 협력학습

Abstract In many applications, producing labeled data is costly and time consuming while an enormous amount of unlabeled data is available with little cost. Therefore, it is natural to ask whether we can take advantage of these unlabeled data in classification learning. In machine learning literature, the co-training method has been widely used for this purpose. However, the current co-training method requires the entire features to be split into two independent sets. Therefore, in this paper, we improved the current co-training method in a number of ways, and proposed a new co-training method which do not need the feature split. Experimental results show that our proposed method can significantly improve the performance of the current co-training algorithm.

Key words : Machine learning, Artificial Intelligence, Co-training

1. 서 론

기계 학습 연구는 많은 방면에서 진보를 해왔으며 데이타 마이닝, 인터넷 및 전자상거래기술, 웹 마이닝, 텍스트 마이닝 등의 다양한 분야에서 핵심 기술로서 활발히 연구되고 있다. 이러한 기계학습 방법은 현재 크게 나누어서 분류 데이타(labeled data)를 이용하여 학습하는 감독자 학습(supervised learning)방법과 미 분류 데이타(unlabeled data)를 이용하여 학습하는 비감독자 학습(unsupervised learning)방법으로 나뉜다[1].

감독자 학습 방법에 있어서 지금까지 대부분의 방법은 미 분류(unlabeled) 데이타가 할 수 있는 역할들을 무시한 채 분류 데이타에만 의존해왔다. 이 경우 분류(classification)작업을 하고자 할 경우 높은 정확도의 분류값을 유지하기 위하여 충분한 분류 데이타가 필요하게 되는데 많은 경우에 기계학습 환경에서 분류 데이타는 미 분류 데이타보다 생성하기가 상당히 어려운 경우가 많이 있다[2]. 예를 들어서 웹 문서를 분류하는 응용분야에 있어서 분류가 된 문서를 얻는 것은 사람이 각 문서를 직접 읽고 분류해야하는 과정이 필요하므로 많은 노력과 시간이 필요한 일이다. 하지만 분류가 되지 않은 문서들은 비용을 들이지 않고도 인터넷 등에서 거의 무한대의 양을 생성할 수 있다. 따라서 이와 같이 상대적으로 쉽게 얻을 수 있는 많은 양의 미 분류 데이타와 작은 양의 분류 데이타를 이용하여 분류 정확도를 높이는 방법들에 대한 연구는 큰 효용성을 가지고 있으며 최근 들어서 활발히 연구가 진행되는 분야중의 하나이다[3-7].

† 종신회원 : 동국대학교 정보통신학과 교수
chlee@dgu.ac.kr
†† 비 회 원 : 한국외환은행 정보시스템부
sominlee@empal.com
논문접수 : 2004년 3월 8일
심사완료 : 2004년 8월 3일

미 분류 데이타를 활용하는 대표적인 방법 중의 하나는 협력학습(co-training) 알고리즘이다[3,4,7]. 이는 카네기 멜론 대학의 Tom Mitchell 교수를 주축으로 개발된 알고리즘으로서, 분류 데이타를 두 개의 독립적인 속성그룹으로 나누어(feature split) 두 개의 분류자(classifiers)로 학습한 후 미분류 데이타를 분류해보고 분류 결과 중 가장 신뢰성이 높은 데이타를 선택하여 분류 데이타에 포함한다. 이와 같은 과정을 반복하여 학습 모델을 만든 후 이 두 모델 결과를 결합하여 최종 결과를 산출한다. 카네기 멜론 대학의 협력학습알고리즘은 특히 웹 문서를 분류하는 분야에서 연구되어 왔으며[7], 웹 페이지 내용 및 그 페이지를 가리키고 있는 하이퍼링크의 두 가지 속성 그룹으로 자연스럽게 나눌 수 있다.

본 연구는 이와 같은 기존의 방법과는 달리 협력학습 방법을 일반 데이타베이스 데이타에 적용시킬 수 있는 시도를 하고자한다. 즉, 두 개의 독립적인 속성 그룹으로 나누는 가정을 따르지 않고 전체 속성을 사용할 수 있는 새로운 협력학습방법을 제안하며, 다수의 실험 결과 제안한 방법이 카네기 멜론대학의 협력학습방법보다 성능을 향상시킬 수 있음을 보이고자한다.

2. 관련 연구

2.1 미분류 데이타를 이용한 감독자 학습방법

분류 데이타를 수집하는 것은 미 분류 데이타를 모으는 것보다 많은 시간과 노력이 필요하다. 예를 들어 이미지 패턴 분류의 경우에 분류된 데이타를 만드는 일은 일일이 인간의 판독 작업에 의해 분류를 해야 하기 때문에 많은 비용과 시간을 필요로 한다. 하지만 상대적으로 미 분류 이미지 데이타는 거의 비용을 들이지 않고 손쉽게 얻을 수 있다. 따라서 분류 정확도를 높이기 위하여 미 분류 데이타를 활용할 수 있다면 아주 효과적일 것이다.

최근 텍스트 마이닝에 대한 관심이 증대되고, 더불어 웹 문서 분류 분야에서 '정확도를 높이기 위하여 작은 양의 분류데이타와 상대적으로 얻기 쉬운 미분류 데이타를 어떻게 이용할 수 있을까'하는 아이디어로 출발한 연구들이 활발하게 진행 중에 있다. 순수하게 분류 데이타만을 이용하여 학습하는 감독자학습(supervised learning)과 미분류 데이타만을 이용하여 학습하는 비감독자학습(unsupervised learning)과 구분하여 미분류 데이타를 이용하여 감독자학습의 성능을 향상시키는 학습방법을 준 감독자학습(semi-supervised learning)방법이라고한다[5].

이러한 준 감독자학습 방법 중 가장 널리 알려져 있는 것은 협력학습(co-training)알고리즘([3,4,7])과 EM (Expectation-Maximization)알고리즘([8])이다. 협력학습(co-training)방법은 전체 속성을 독립적인 두개의 속성그룹으로 분리하여 두 개의 분류자가 학습/분류/결합하여 정확도를 높이는 방법이다. 이는 카네기 멜론 대학에서 제안되어 현재 여러 분야에서 조사, 연구되고 있는 대표적인 준 감독자학습 방법이다. 이 방법은 웹 문서의 분류에 적용되었는데 우선 웹 문서를 분류하기 위한 데이타를 목표 값에 대해서 두 개의 독립적인 속성 그룹으로 나눈다. 즉, 웹 문서를 분류하기 위한 데이타를 크게 웹 페이지와 그 웹 페이지를 가리키는 하이퍼링크(hyperlink) 데이타로 구분한다. 웹 문서 분류의 경우는 전체 데이타를 굳이 독립적인 속성그룹으로 나누는 노력을 하지 않아도 자연적으로 두 개의 속성 그룹으로 나뉘는 장점이 있다. 이는 웹 문서 분류 작업이기 때문에 가능한 일일 것이다.

또한 협력학습방법에서는 한 개가 아닌 두 개의 독립적인 분류자(classifier)를 사용하여 학습한다. 첫번째 분류자는 웹페이지 내용의 단어를 이용하고, 또 다른 분류자는 하이퍼링크의 단어만을 이용하여 다음과 같이 학습한다. (1) 각 분류자는 미 분류 데이타의 목표 값을 예측한다. (2) 이렇게 예측한 미 분류 데이타 중 가장 신뢰도 높은 데이타를 선택한 다음 분류 데이타 셋에 추가한다. 즉, 각 분류자는 미 분류 데이타를 분류 데이타화하면서 학습한다. 이러한 프로세스는 미분류 데이타가 모두 분류 데이타가 될 때까지 반복한다. (3) 결국, 모든 데이타는 분류 데이타가 되고 다시 두 분류자는 전체 데이타를 재학습한 후, 테스트 데이타에 대해 결과를 산출한다. 그림 1은 미분류 데이타를 이용한 협력학습알고리즘의 기본적인 내용을 설명하고 있다.

Given:
- Set L of labeled training examples
- Set U of unlabeled training examples

Loop:
1. Learn classifier H from L
2. Learn classifier F from L ($H \neq F$)
3. Allow H to label p positive and n negative examples from U
4. Allow F to label p positive and n negative examples from U
5. Add these self-labeled examples to L

그림 1 미분류 데이타와 두 분류자를 이용한 협력학습 방법

카네기 멜론 대학의 협력학습방법 실험에서는 웹 페이지를 분류하기 위해 naive Bayesian을 사용하는 감독자 학습 알고리즘보다 협력학습 알고리즘이 분류 정확도를 향상시킨다는 것을 보였다[7]. 대학 교과 홈페이지의 여부를 분류하기 위해 협력학습 알고리즘을 이용한 실험을 통해 다음과 같은 결과를 얻었다. 웹 페이지가 교수 홈페이지인지, 아닌지를 분류하는 데이타로써 16

개의 분류 데이타와 약 800개의 미분류 데이타를 컴퓨터학과의 웹 사이트에서 구했고, 협력학습 각 반복에서 각 분류자는 1 개의 참인 데이타와 3 개의 거짓인 미분류 데이타를 분류 데이타에 추가하였다. 협력학습의 30번의 반복이후, 오직 분류 데이타만 사용한 감독자학습방법경우의 89% 정확도에 비하여 결합된 협력학습방법은 95% 의 정확도를 보였다.

두 번째의 준 감독자 학습방법은 EM 알고리즘으로서 이 방법은 각 속성마다 데이타 값의 분포가 정해져있다고 가정하고, 그 분포를 결정하는 파라미터를 알게 되면 결측치(missing value)를 예측해 낼 수 있다는 방법이다[8]. 미 분류 데이타의 목표 값을 하나의 결측치로 가정하고 그 값을 알아내기 위해 EM 알고리즘을 적용시킨다. EM은 원래 결측치를 추정하기 위해 사용되는 알고리즘으로 많이 사용되었으나, 근래에 미 분류 데이타의 목표 값을 마치 결측치로 간주하여 목표 값을 찾아 분류 데이타에 포함시켜 감독자학습 알고리즘을 사용하여 성능을 향상시키는데 사용된다. 즉 각 속성의 값에는 임의의 확률분포가 존재한다고 가정하고 그 분포를 결정하는 파라미터를 알게 되면 결측치를 예측할 수 있는 방법이다. 현재 알려진 데이타를 이용하여 파라미터의 초기치를 결정하고, 결정된 파라미터에 따른 확률분포를 이용하여 결측치를 예측한 후, 예측한 데이타를 포함한 전체 데이타에 의해 파라미터를 조정하며, 이와 같은 프로세스를 파라미터가 더 이상 변하지 않을 때까지 반복하는 방법이다.

앞에서 설명한 준 감독자 방법들 중에서 본 논문에서는 협력학습 방법에 대하여 기존의 문제점과 이들을 보완하는 내용을 연구하였다.

1.2 점증적 학습방법과 반복적 학습방법

협력학습 방법은 미 분류 데이타를 분류하는 방식에 따라 점증적 학습방법(incremental method)과 반복적 학습방법(iterative method)의 두 가지로 분류할 수 있다[3]. 첫 번째는 점증적 학습방법(incremental method)으로서 이 방법은 목표 값에 대해 독립적인 두 개의 데이타 셋인 분류데이타를 이용해 학습한 후 미 분류 데이타를 분류, 그중 가장 신뢰도가 높은 데이타 하나를 선택하여 분류 데이타에 포함시켜 위의 프로세스를 미 분류 데이타가 모두 분류될 때까지 반복하는 방법이다. 즉, 분류 데이타의 크기는 분류자에 의해 분류된 미 분류 데이타에 의해 점점 늘어난다. 두 개의 학습 모델이 만들어지면 이 두 모델을 결합해서 답을 찾게 된다. 1.1 절에서 설명한 협력학습방법은 점증적 학습방법이고 개략적인 내용은 그림 2와 같다.

두 번째 협력학습방법은 반복적 학습방법(iterative method)으로서 이 방법은 점증적 학습방법과 마찬가지로, 독립적으로 속성이 나뉜 두 개의 분류 데이타를 학습하는 방법은 동일하다. 그러나 학습한 후, 미 분류 데이타를 모두 분류하는 프로세스를 사용자가 정한 조건(대부분은 unlabeled data 값의 변동이 없는 조건)이 만족할 때까지 반복하는 것이 첫 번째 방법과 다른 점이다. 다음으로 두 모델을 결합하여 답을 찾아낸다. 기본적인 반복적 학습방법의 내용은 그림 3과 같다.

Incremental Co-training Method

1. **Input** : a sample $S = LAB \cup UNL$ (LAB : labeled data, UNL : unlabeled data)
2. **Loop while there exist unlabeled data**
3. with input LAB_1, W_1 outputs h_1 (W_1 : classifier 1)
4. with input LAB_2, W_2 outputs h_2 (W_2 : classifier 2)
5. pick the a unlabeled examples classified as positive by h_1 and about which h_1 is most confident
6. pick the a unlabeled examples classified as positive by h_2 and about which h_2 is most confident
7. add the labeled examples in steps 5 and 6 to set LAB
8. **Output** : combine(h_1, h_2)

그림 2 점증적 협력학습방법

Iterative Co-training Method

1. **Input** : a sample $S = LAB \cup UNL$ (LAB : labeled data, UNL : unlabeled data)
2. **Loop while condition**
3. with input LAB_1, W_1 outputs h_1 (W_1 : classifier 1)
4. label UNL with h1
5. with input LAB_2, W_2 outputs h_2 (W_2 : classifier 2)
6. label UNL with h2
7. **Output** : combine(h_1, h_2)

그림 3 반복적 협력학습방법

1.3 기존 협력학습방법의 문제점

앞 절에서 설명한 것과 같이 협력학습방법은 정확도를 높이기 위해 두 개의 독립적인 속성 그룹으로 나누고 두개의 분류자를 이용하여(주로 naive Bayesian 알고리즘 사용) 학습한 후 미 분류 데이타의 목표 값을 예측하여 가장 신뢰도가 높은 데이타를 분류 데이타에 포함시키고 위의 과정을 반복하는 방법이다. 여기서 데이타를 두 개의 독립적인 속성 그룹으로 나누는 작업이 중요하게 여겨지게 되는데 웹 페이지 데이타와 그 웹 페이지를 가리키는 하이퍼링크 데이타로 자연스럽게 나눌 수 있는 웹 문서 분류분야에서는 효과적으로 사용할 수 있겠지만, 다른 분야에서는 이러한 가정이 별로 적합하지 않다.

예를 들어 이메일(e-mail) 데이타의 경우나, 일반 문서 데이타인 경우는 웹 페이지처럼 자연스럽게 두 개의 속성그룹으로 나눌 수 있는 명확한 기준이 존재하지 않는다[9]. 또한 속성끼리의 연관성이 강한 기존의 관계형 데이타베이스의 데이타인 경우 이를 두개의 독립적인 속성그룹으로 분할하는 것은 현실적으로 거의 불가능하다. 이러한 이유로 협력학습방법을 제안한 카네기 멜론의 연구자들도 "데이타 셋의 과반수이상이 명확하게 또

는 자연스럽게 두 개의 속성그룹으로 나뉠 수 있는 기준이 없는 데이타이기 때문에 협력학습방법은 강력한 파라다임에도 불구하고 널리 적용할 수 없다"라고 지적하고 있다[7]. 따라서 본 논문에서는 데이타 속성을 두 개의 독립된 집합으로 구분할 필요가 없으며 따라서 훨씬 다양한 분야에서 협력학습을 응용할 수 있는 새로운 협력학습 모델을 제시하고자한다.

3. 새로운 협력학습방법

기존의 협력학습방법은 독립적인 두 개의 속성 그룹으로 나누는 기본적인 가정아래 두 개의 분류자로 학습, 결합하여 높은 정확도를 얻기 위해 제안되었다. 하지만 기존의 협력학습방법은 독립적인 두 개의 속성그룹으로 나누는 가정 때문에 일반 데이타베이스에 적용하기가 쉽지 않다. 왜냐하면, 일반 데이타베이스 데이타는 속성간의 관계를 자연적으로 나눌 수 없으며 두 개의 독립적인 속성들의 그룹으로 나눌 수 있는 데이타를 찾기가 쉬운 일이 아니며, 속성끼리 관계가 밀접한 관계가 있기 때문에 기존의 방법을 사용하기에는 부적합하다. 또한 웹 페이지가 아닌 일반 텍스트문서 데이타나 이메일 데이타인 경우도 마찬가지로 기존방법의 가정을 따르는데 문제가 있다.

본 논문은 독립적인 속성그룹으로 데이타를 나눠야 하는 기존의 협력학습방법 대신에, 전체 속성을 사용하여 학습하는 새로운 협력학습방법을 제안한다. 본 연구에서 제안하는 새로운 협력학습방법은 우선 전체속성을 이용하여 두개 이상 다수의 분류자가 학습을 한 다음, 각 분류자는 독립적으로 미 분류 데이타의 목표 값을 예측하고, 그 중 가장 신뢰도가 높은 데이타를 선택하여 분류 데이타에 포함시킨 후, 재학습을 한다. 이러한 프로세스를 미 분류 데이타가 모두 분류될 때까지 반복한다. 위의 과정을 통해 다수의 학습 모델이 생성된다. 이때 마지막으로 생성된 학습 모델들을 통해 테스트 데이타를 분류하고자 할 경우 각 데이타마다 다수의 답이 나오는데, 투표(voting)를 통한 방법이나, 답들 중에서 가장 확률 값이 높은 답을 선택하여 최종 결과를 산출한다.

제안한 협력학습방법은 전체 데이타 중 두 개의 데이타를 나누어서 각기 다른 데이타를 학습한 후, 결과를 산출하는 기존의 방법보다, 전체 데이타를 전문가 여러 명이 학습한 후 각각 답을 내준 후, 그 결과들을 통합하여 산출하는 결과가 더 나을 것이라는 것은 직관적으로 생각해 보았을 때에도 예상이 될 것이다. 본 논문에서 제안한 방법이 기존의 협력학습 방법에 비하여 보완된 사항을 요약하면 다음과 같다.

- 전체 속성을 두개의 독립된 속성 집합으로 구분할 필요가 없다.
- 2 개 이상 임의 개수의 분류자를 결합할 수 있게 하여 더욱 정확도를 향상시킬 수 있다.

본 논문에서 사용된 분류자로서는 기계학습방법 중 널리 사용되고 있는 신경망학습 방법 중 역전파(back-propagation)알고리즘([10])을 이용한다. 역전파 알고리즘은 학습 데이타의 에러에 별로 민감하지 않고, 비교적 높은 정확도를 가지는 알고리즘이다. 또한 여러 가지 파라미터를 이용하여 얻고자 하는 정확도를 조정할 수 있는 장점이 있다.

특히 협력학습알고리즘에서 미 분류 데이타를 분류한 후 그 중 가장 신뢰도가 높은 것을 선택하는 것이 중요하게 되는데, 본 논문에서는 역전파 알고리즘의 결과가 각 클래스의 가중치(weight)로 표현되는 것을 이용, 각 분류한 결과의 가중치 값이 가장 큰 것을 신뢰도가 높은 것으로 간주하였다. 즉, 분류데이타를 이용하여 학습한 후, 미 분류 데이타를 분류, 분류한 데이타 중 가장 가중치 값이 큰 것을 분류 데이타에 포함하여 재학습하는 프로세스를 미 분류 데이타가 모두 분류될 때까지 반복한다. 또한 다수의 분류자가 최종 분류 데이타를 이용하여 학습한 후, 최종 결론을 내기 위해 분류자끼리 결합을 해야 하는데 이때에도 마찬가지로 각 분류자가 예측한 값들 중에 가장 가중치 값이 높은 값을 최종 답으로 선택하는 방식을 따랐다. 그림 4는 본 논문에서 제안한 협력학습방법에 대한 내용을 요약한 것이다.

1. **Input** : An initial collection of labeled(LAB) and unlabeled data(UNL)
2. **Loop while there exist Unlabeled Data**
 2.1 Learn n different classifiers($C_1, C_2, \cdots, Cn$) using LAB (without feature split)
 2.2 Pick a data u from UNL, each C_i classifier outputs h_i independently
 2.3 Label u as h_i with the highest confidence value
 2.4 u is added to the labeled data
3. **Output** : n classifiers predict class labels for new data.
 n outputs are combined together

그림 4 제안한 협력학습방법

4. 실험 결과

제안한 협력학습방법에 대한 실험을 위해서 본 논문에서는 UCI machine learning repository [11] 에서 제공하는 벤치마킹데이타를 이용하였다. 이 데이타는 기계학습 연구 분야에서 보편적으로 사용되는 실용성 있는 데이타로써 이 중에서 Animal, Breast-cancer, Voting, Post-operative patient 의 데이타들을 사용하여 실험을 하였다.

본 실험에서는 다양하게 파라미터를 변화시켜 두 개의 신경망 알고리즘을 분류자로 이용하였다. 분류 데이타는 전체 데이타의 20%를 랜덤 샘플링 하여 생성하였

고, 나머지 데이타는 미 분류 데이타로 간주하여 사용하였다. 신경망 알고리즘은 초기 가중치 값에 따라 결과가 다르게 나올 수 있으므로 초기 가중치를 다르게 한 세 번의 실행에서 얻은 결과의 평균을 얻었다. 신경망은 10개의 은닉노드로 구성되었으며 학습률은 0.01, 모멘텀은 0.1 로 실험하였다.

각 실험 데이타마다 (1) 분류 데이타만을 학습하여 분류하는 감독자 학습방법, (2) 분류/미분류 데이타를 이용한 기존의 카네기 멜론 대학의 협력학습방법(CMU 방법), (3) 그리고 본 논문에서 제안한 협력학습방법의 세 가지 방법에 대하여 성능을 실험, 비교하였다. 특히, 기존 CMU 방식은 데이타베이스 속성들을 독립적인 속성그룹으로 나눈다는 것은 불가능하기 때문에, 전체 데이타를 랜덤하게 두 개의 그룹으로 분할하는 방법을 사용하였고, 세 번 랜덤하게 속성을 나누어 만들어진 세 개의 데이타 셋을 이용하여 실험하였다.

본 논문의 실험에서 각 데이타는 감독자학습방법과, 카네기 멜론 대학의 협력학습방법, 본 논문에서 제안한 방법의 에러율을 비교 분석하였다.

표 1은 Animal를 이용한 실험결과이다. 이 데이타는 16 개의 속성으로 이루어져 있고, 분류데이타 23 개 미분류 데이타 78 개를 이용하여 실험하였다. 제안한 방법의 결과는 1.9%의 에러율을 보이는데 반해 감독자학습방법은 4.9%, 기존 협력학습방법은 16.8%의 에러율을 나타냈다. 물론 기존 방법은 세 번의 랜덤하게 속성을 분리한 가공된 데이타를 가지고 실험한 결과이다. 이때, 기존의 방법과 감독자학습방법의 결과는 경우에 따라서 다르게 나타나는데, 이는 두 개의 데이타가 얼마나 독립적으로 나누어졌느냐에 따라 카네기 멜론의 협력학습방법이 높은 정확율을 나타남을 예측하게 된다.

두 번째 데이타인 Breast-cancer 데이타의 경우 가장 높은 정확도를 갖는 실험으로 비교해 보았을 때 본 연구에서 제안한 방법은 2.5%의 에러율을 보이는 데 반해 감독자 학습방법은 2.7%, 카네기 멜론의 방법은 2.8%의 에러율을 나타냈다. 이 데이타는 CMU방법과 감독자 학습방법, 제안한 방법의 에러율이 거의 비슷하게 결과가 산출되었다. 이는 분리된 속성그룹들이 거의 독립적으로 분리가 되어 기존 CMU방법의 성능이 제안한 방법과 비슷하게 나온 것으로 판단된다.

표 3은 Voting 데이타 셋을 이용한 실험결과이다. 87개의 분류데이타와 348 개의 미 분류 데이타를 이용하여 실험한 결과 본 논문에서 제안한 알고리즘은 11.9%의 에러율을 보인데 반해, CMU방식은 평균 21.41%의 에러율을 나타내었다. 또한 87개의 분류데이타만을 이용하여 감독자학습방법을 실행한 결과는 15.6%의 에러율을 보였다.

표 4는 Post-operative patient 데이타 셋을 이용한 실험 결과이다. 약간 정확도가 떨어지는 데이타이긴 하지만, 이 데이타 셋에서도 마찬가지로 제안한 협력학습방법이 가장 낮은 에러율을 보였다. 18 개의 분류데이타

표 1 각 알고리즘별 분류 에러율 - Animal data set

			분류자1	분류자2	결합
Animal (16/101) #L:23 #U:78	제안한 Co-training	Supervised	7.9	6.9	4.9
		Co-training	5.9	3.9	1.9
	기존의 Co-training 실험 결과 (CMU방식)		데이타1	데이타2	결합
		Supervised	7.9	18.8	8.9
		Co-training	10.8	23.7	17.8
		Supervised	10.8	3.9	4.9
		Co-training	28.7	14.8	16.8
		Supervised	14.8	12.8	9.9
		Co-training	14.8	16.8	18.8

표 2 각 알고리즘별 분류 에러율 - Breast-cancer data set

			분류자1	분류자2	결합
Breast Cancer (10/699) #L:139 #U:560	제안한 Co-training	Supervised	5.8	5.2	5.2
		Co-training	3.7	4.4	2.5
	기존 Co training 실험 결과 (CMU방식)		데이타1	데이타2	결합
		Supervised	5.4	3.8	2.8
		Co training	4.7	3.7	3.4
		Supervised	4.2	5.7	2.7
		Co-training	4.0	5.8	2.8
		Supervised	4.7	3.5	3.4
		Co-training	7.5	4.8	8.7

표 3 각 알고리즘별 분류 에러율 - Voting data set

			분류자1	분류자2	결합
Voting (17/435) #L: 87 #U:348	제안한 Co-training	Supervised	17.25	16.19	15.6
		Co-training	13.5	12.12	11.9
	기존 Co-training 실험 결과 (CMU방식)		데이타1	데이타2	결합
		Supervised	22.98	25.51	21.5
		Co-training	25.05	24.36	22.57
		Supervised	18.16	12.64	13.26
		Co-training	33.33	13.33	17.7
		Supervised	22.06	18.62	17.91
		Co-training	30.8	21.37	23.96

표 4 각 알고리즘별 분류 에러율 - Postoperative patient data set

			분류자1	분류자2	결합
Post-operative Patient (9/90) #L: 18 #U: 72	제안한 Co-training	Supervised	38.88	37.77	37.01
		Co-training	33.33	34.54	31.07
	기존 Co-training 실험 결과 (CMU방식)		데이타1	데이타2	결합
		Supervised	28.88	38.88	36.45
		Co-training	30	32.22	33.01
		Supervised	35.55	34.44	36.45
		Co-training	30	33.33	34.57
		Supervised	35.55	32.22	34.89
		Co-training	31.11	28.89	32.77

표 5 실험결과요약

	Supervised	Co-training (CMU)	제안한 Co-training
Animal	4.9	16.8	1.9
Breast-Cancer	2.7	2.8	2.5
Voting	15.6	21.41	11.9
Postoperative Patient	37.01	33.45	31.07

와 72 개의 미분류 데이타를 이용하였고, 순수하게 분류 데이타만 이용하여 학습한 결과는 37.01%, CMU방법의 평균에러율은 33.45%를 나타내었다.

표 5는 위의 실험들의 결과를 제안한 협력학습방법과, CMU방법, 감독자학습방법의 결과를 비교하기 위하여 요약해 놓은 표이다.

그림 5는 본 연구에서 제안한 방법과 다른 방법들의 에러율을 그래프로 표시한 내용이며 벤치마킹데이타를 이용한 실험결과는 본 논문에서 제안한 방법의 실험 결과가 가장 높은 정확도를 갖는다는 것을 보여준다. 각 데이타별로 실험한 알고리즘의 에러율을 나타내는 그래프를 보면 CMU방식과 감독자학습방법과의 순위는 경우에 따라서 다르지만 본 논문에서 제안한 협력학습방법은 이 두 가지 방법에 비해 대부분 높은 결과를 갖는다는 것을 이 실험을 통해서 알 수 있다.

결과적으로 본 논문에서 제안한 협력학습방법이 실험한 방법 중에서 에러율이 가장 낮음을 알 수 있었다. 또

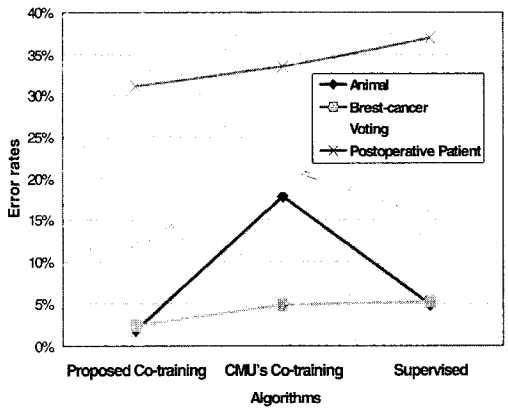

그림 5 각 알고리즘별 평균 에러율

한, 기존 CMU 방법에서는 어떻게 속성 그룹을 나누느냐에 따라서 감독자학습방법과 에러율의 차이를 보였다. 즉, 기존 카네기 멜론대학의 협력학습방법은 데이타를 얼마나 독립적인 속성그룹으로 나누는가가 높은 정확도를 내기 위한 중요한 변수로 작용된다는 것을 알 수 있다.

5. 결 론

본 논문에서는 모든 속성을 이용하여 정확도를 높이는 새로운 협력학습방법을 제안하였다. 두 개의 독립적인 속성그룹으로 나누어 학습하는 기존 협력학습 방법

은 일반 관계형 데이타베이스 데이타에 적용할 때 속성을 독립적인 속성그룹으로 나눈다는 것이 거의 불가능한 일이기 때문에 적용하기가 쉽지 않다. 따라서 본 논문에서는 기존협력학습방법 대신 속성을 나누지 않고 전체 속성을 사용하여 다수의 분류자로 학습하는 새로운 형태의 협력학습방법을 제안하였다. 또한 제안한 방법의 타당성을 입증하기 위하여 분류 데이타만을 학습하는 감독자학습방법, CMU의 협력학습방법, 그리고 제안한 방법을 다수의 데이타를 이용하여 실험하였다. 실험 결과, 본 논문에서 제안한 방법이 기존 방법보다 성능을 향상시킬 수 있음을 알 수 있었다.

추후 연구과제로서는 협력학습방법에서 미 분류 데이타를 분류 데이타화하는 과정 중 분류자에 의해 분류된 미 분류 데이타에서 가장 신뢰도 높은 데이타를 선택하는 작업이 중요하다. 또한, 미 분류 데이타가 많을 경우 신뢰도 높은 데이타를 몇 개를 선택할 것인가의 문제도 정확도에 영향을 미칠 수 있다. 따라서 예측된 미 분류 데이타 중 몇 개를 분류 데이타로 만드는 것이 학습시간을 단축시킬 것인지에 대해 연구해 볼 필요가 있다.

또한, 본 논문에서는 카네기 멜론 대학의 방법에 대한 실험을 하기 위해 학습 데이타의 속성을 랜덤하게 분할하는 방법을 사용하였는데, 서로 독립적인 속성으로 분할하는 방법을 이용하여 제안한 실험결과와 기존방법과의 좀 더 정확한 비교가 될 수 있게 할 것이다.

아울러 본 논문에서 사용한 데이타 외에 텍스트 데이타를 이용한 실험을 추가하여 카네기멜론 대학 방법과의 직접적인 비교를 하는 것도 필요할 것으로 생각된다.

마지막으로 본 연구의 실험에서는 신경망 알고리즘을 분류자로 사용하였지만, 이외에도 다양한 감독자 학습 알고리즘(예를 들면 의사결정나무, 기억기반학습, 나이브 베이지안등)을 이용하여 제안한 협력학습방법을 적용한 실험이 필요할 것이다.

참 고 문 헌

[1] T. G. Dietterich, "Machine Learning Research Four Current Directions," *AI Magazine*, 1997.

[2] T. Mitchell, "The Role of Unlabeled Data in Supervised Learning," In *Proc. of the Sixth International Colloquium on Cognitive Science*, 1999.

[3] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-training," In *Proceedings of COLT '98*, 1998.

[4] K. Nigam & R. Ghani, "Understanding the Behavior of Co-training," In *Proceedings of KDD-2000 Workshop on Text Mining*, 2000.

[5] S. Goldman & Y. Zhou, "Enhancing Superviased Learning with Unlabeled Data," In *Proceedings of ICML2000*, 2000.

[6] K. Nigam & R. Ghani, "Analyzing the Effectiveness and Applicability of Co-training," In *Proceedings of the 9th International Conference on Information Knowledge Management*, 2000.

[7] K. Nigam & A. McCallum & S. Thrun & T. Mitchell, "Learning to Classify Text from Labeled and Unlabeled Documents," In *Proceedings of the 15th National Conference on Artificial Intelligence AAAI-98*, 1998.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of Royal Statistical Society*, Vol 39, pp. 1-38, 1977.

[9] S. Kiritchenko & S. Matwin, "Email Classification with Co-Training."

[10] S. Haykin *"Neural Networks: A Comprehensive Foundation,"* Prentice Hall, 1999.

[11] P. Murphy & D. Aha, UCI Repository of Machine Learning Databases, 1995. (http://www.ics.uci.edu/~mlearn/ MLRepository.html)

[12] K. P. Beneett & A. Demiriz & R. Maclin, "Exploiting Unlabeled Data in Ensemble Methods," In *Proceedings of the SIGKDD'02*, 2002.

[13] Virginia R. de Sa, "Learning Classification with Unlabeled Data," *Advances in Neural Information Processing Systems 6*, pp. 112-119.

[14] F. Cozman & I. Cohen, "Unlabeled Data can Degrade Classification Performance of Generative Classifiers," *Technical Report*, HP labs, 2002.

이 창 환

1982년 2월 서울대학교 계산통계학과 졸업(학사). 1988년 8월 서울대학교 계산통계학과 졸업(석사). 1994년 8월 University of Connecticut, Dept. of Computer Science(박사). 1982년 3월~1987년 2월 한국기계연구소. 1994년 12월~1996년 2월 AT&T Bell Laboratories, Middletown, USA. 1996년 3월~현재 동국대학교 정보통신학과 부교수. 관심분야는 기계학습, 마이닝, 생물정보학 등

이 소 민

1996년 3월~1999년 8월 동국대학교 정보관리학과 졸업(학사). 2000년 9월~2002년 8월 동국대학교 정보통신공학과 졸업(석사). 2002년 8월~현재 한국외환은행 정보시스템부. 관심분야는 기계학습, 데이타 마이닝