비분류표시 데이타를 이용하는 분류 기반 Co-training 방법

(A Co-training Method based on Classification Using Unlabeled Data)

윤 혜 성 † 이 상 호 †† 박 승 수 ††† 용 환 승 †† 김 주 한 ††††
(Hye-Sung Yoon) (Sang-Ho Lee) (Seung-Soo Park) (Hwan-Seung Yong) (Ju-Han Kim)

요 약 생물 정보학 등 많은 응용 분야에서 데이타 분석을 할 때는 적은 수의 분류표시된 데이타(labeled data)와 많은 수의 비분류표시된 데이타(unlabeled data)가 있을 수 있다. 분류표시된 자료는 사람의 노력이 요구되기 때문에 얻기가 어렵고 비용이 많이 들지만, 비분류표시된 자료는 별 어려움 없이 쉽게 얻을 수 있다. 이때 비분류표시된 자료를 이용하여 자료를 분류하고 분석하는데 널리 이용되고 있는 방법이 co-training 알고리즘이다. 이 방법은 적은 수의 분류표시된 자료에서 두 가지 뷰(view)로 각 분류자를 학습한다. 그리고 각 분류자는 분석하고자 하는 모든 비분류표시된 자료에서 가장 만족할만한 예측자들을 만들어 나간다. 이렇게 훈련 데이타 셋에서 실험을 여러 번 반복적으로 하게 되면 각 뷰에서 새로운 분류자가 학습되어 분류표시된 자료의 수가 증가한다.

본 논문에서는 비분류표시된 데이타를 이용하여 새로운 co-training 방법을 제시한다. 이 방법은 두 가지 분류자와 WebKB 및 BIND XML의 2가지 실험 데이타를 가지고 평가하였다. 실험 결과로서, 이 논문에서 제안한 co-training 방법이 분류표시된 자료의 수가 매우 적을 때 분류정확성을 효과적으로 향상시킬 수 있음을 보였다.

키워드 : co-training, 분류 알고리즘, 반교사 학습, 반구조화된 데이타

Abstract In many practical learning problems including bioinformatics area, there is a small amount of labeled data along with a large pool of unlabeled data. Labeled examples are fairly expensive to obtain because they require human efforts. In contrast, unlabeled examples can be inexpensively gathered without an expert. A common method with unlabeled data for data classification and analysis is co-training. This method uses a small set of labeled examples to learn a classifier in two views. Then each classifier is applied to all unlabeled examples, and co-training detects the examples on which each classifier makes the most confident predictions. After some iterations, new classifiers are learned in training data and the number of labeled examples is increased.

In this paper, we propose a new co-training strategy using unlabeled data. And we evaluate our method with two classifiers and two experimental data: WebKB and BIND XML data. Our experimentation shows that the proposed co-training technique effectively improves the classification accuracy when the number of labeled examples are very small.

Key words : co-training, classification algorithm, semi-supervised learning, semi-structured data

† 비 회 원 : 이화여자대학교 컴퓨터학과
comet@ewha.ac.kr
†† 종신회원 : 이화여자대학교 컴퓨터학과 교수
shlee@mm.ewha.ac.kr
hsyong@mm.ewha.ac.kr
††† 정 회 원 : 이화여자대학교 컴퓨터학과 교수
sspark@ewha.ac.kr
†††† 비 회 원 : 서울대학교 의과대학 생명의료정보학 교수
juhan@snu.ac.kr
논문접수 : 2003년 6월 25일
심사완료 : 2004년 5월 17일

1. 서 론

DM(Direct Mail) 반응 여부, 가장 높은 수익을 얻을 수 있는 목표 시장을 찾는 문제 등과 같이, 예측하고자 하는 목적 변수(target value)가 있는 데이타 마이닝 과정이나 학습과정을 교사학습(supervised learning)이라고 하고, 예측 대상이 되는 목적변수가 없는 경우를 비교사학습(unsupervised learning)이라 한다. 이러한 학습방법은 우리가 탐사한 지식, 데이타베이스의 종류, 정

보의 종류, 적용할 탐사 기법에 따라서 다르게 적용된다.

현재 생물 정보학 분야에서 얻을 수 있는 바이오 데이타(bio data)들로부터 유용한 정보를 얻고자하는 연구가 많이 진행되고 있다. 하지만 각기 다른 형식과 내용으로 산재해 있는 데이타들에서 우리가 찾고자 하는 목적에 부합된 관련 정보를 추출하기 위해서는 바이오 데이타의 형식에 대한 연구와 그에 적합한 적용 알고리즘들이 필요하다. 더구나 생물 정보학에서 분석하고자 하는 데이타 형식의 대부분이 반구조화된 데이터(semi-structured data) 구조를 가지고 있다. 하지만 이러한 데이타 구조에 적합한 분석방법을 찾고 이를 적용하여 유용한 결과를 얻어내기 위해서는 연구해야 할 사항들이 많다.

따라서 본 논문에서는 이러한 점들을 고려하여 반구조화된 데이타에서 알려진 정보의 양이 적을 때 데이타를 분석하는 효과적인 방법에 대해서 연구하고 이를 적용하였다. 본 논문의 실험에서 생물학적 데이타 구조인 반구조화된 데이타 구조 형식의 예로 XML 데이타를 적용하였고, 분석 방법으로는 반교사학습(semi-supervised learning) 방법인 co-training 방법에 분류 알고리즘을 적용하였다. 그 결과 기존의 텍스트 분석에서 그 효과가 입증되었던 반교사학습 방법이 바이오 데이타의 분석방법에도 역시 효과가 있음을 알 수 있었다. 또한 제시하는 새로운 co-training 알고리즘이 기존의 알고리즘보다 더 정확한 결과를 보여 주었다.

본 논문의 구성은 다음과 같다. 2절에서는 관련 연구로서 현재까지의 반교사학습에 대한 연구와 그 실험에 이용된 데이타의 특성에 대해 소개하고, 3절에서는 기존 반교사학습 방법의 기반이 되는 co-training 알고리즘과 새롭게 제안한 co-training 알고리즘을 소개한다. 그리고 4절에서는 제안하는 방법에 대한 실험 절차를 보이고, 교사학습 방법과 기존의co-training 방법 각각에 대하여 제안하는 방법과 비교·평가하며, 5절에서 결론을 맺는다.

2. 관련 연구

본 절에서는 기존의 텍스트 분석에서 그 효과가 입증되었으며, 현재도 꾸준히 새로운 분석방법과 적용분야에 대해서 연구되고 있는 반교사학습 방법과 바이오 데이타 구조와 유사한 구조인 반구조화된 데이타에 대해서 설명한다.

2.1 반교사학습

비분류표시된 데이타를 이용한 분석 방법은 텍스트 분류에 있어서 적은 수의 분류표시된 데이타만을 가지고 분석하는 것보다 분석 에러를 줄일 수 있다는 것이 알려져 있다. 교사학습의 대부분의 계산적 모델들은 분류표시된 자료들만을 고려하고 비분류표시된 자료의 역할은 무시한다[1]. 하지만 요즈음 데이타의 양이 점점 많아짐에 따라 비분류표시된 데이타에도 유용한 정보가 있다는 사실이 알려짐에 따라 이로부터 정보를 추출하고자하는 노력이 진행되고 있지만, 기존의 교사학습 방법으로는 많은 문제점들이 있다[2].

반교사학습 방법의 주요 아이디어는 바로 이러한 비분류표시된 자료와 분류표시된 자료를 함께 이용하여 분류자(classifier)를 만들고 이를 활용하는 것이다. 분류표시된 자료가 전혀 알려지지 않다면 이는 클러스터링(clustering) 문제가 되지만, 어느 정도 분류표시된 자료가 알려져 있다면 이는 분류(classification) 문제가 된다. 실제로 의료진단(medical diagnosis), 웹 탐색(web search), 약품 설계(drug design), 데이타베이스 마케팅(database marketing) 등의 여러 분야에서 비분류표시된 자료를 다루지만 이러한 자료를 만드는 데에는 많은 노력과 비용이 들게 된다.

현재까지의 대표적인 반교사학습 알고리즘들의 흐름을 보면 다음과 같다. 우선 1990년대 후반부터 Tom Mitchell의 co-training 알고리즘[1,3]을 기반으로 하는 여러 알고리즘들이 등장하였다. 최근의 흐름을 살펴보면 [4-6]에서는 EM(Expectation Maximization) 방식을 반교사학습에 적용하였다. 이 방법은 소수의 분류표시된 자료에서 한 개의 뷰(view)를 이용하여 분류자를 만든 후, 다수의 비분류표시된 자료에 분류표시된 자료에서 만들어진 분류자와 확률이론을 적용하여 자료의 분포를 만들고 가장 그럴듯한 가정을 찾는 방법으로서, 마치 자료를 클러스터링하는 것처럼 보인다. 그리고 [7]에서 제시하는 co-testing 알고리즘은 비분류표시된 자료에서 가장 높게 분류표시된 자료의 정보를 가진 자료를 찾고 이를 분류표시할 때 사용자에게 분류표시를 할 것인가를 쿼리하여 선택하도록 하는 방식을 제안하였다. 2년 후 같은 팀에서 능동적 학습(active learning)과 반교사학습 알고리즘인 co-testing과 co-EM을 결합하여 분석할 때 도메인에 국한되지 않도록 하기 위하여 co-testing의 쿼리와 비분류표시된 자료를 무조건 분류표시 자료로 분류하지 않고 확률적 이론을 적용하는 co-EM의 장점을 결합한 Co-EMT[8] 알고리즘을 제안하였다. 이밖에도 co-training 방법에 클러스터링을 적용한 방법과[9] SVM(Support Vector Machine)을 적용한 방법[10], 웹 페이지의 모든 종류 데이타를 추출할 때에 랩퍼 유도(wrapper induction) 규칙을 이용하여 분석하는 방법[10,11], 그리고 여러 가지 다른 데이타 셋에 반교사학습 알고리즘을 적용하는 연구가 진행되고 있다.

2.2 반구조화된 데이타

최근 데이타베이스를 관리하는 연구는 불규칙하고 자

주 변화하는 데이타 구조를 가지면서 반구조화된 데이타까지 확장되고 있다. 반구조화된 데이타를 직관적으로 정의하면 처리하지 않은 데이타(raw data)도 아니면서, 정형적인 구조도 가지고 있지 않은 데이타를 말한다. 즉, 데이타의 구조가 부분적으로 알려져 있으며 사전예고 없이 변화 가능한 데이타이다. 반구조화된 데이타의 주요 특징을 요약하면 다음과 같다[12].

① 비정규적(irregular)인 구조를 가진다.
② 함축적(implicit)인 구조를 가진다.
③ 부분적(partial)인 구조를 가진다.
④ 스키마가 매우 크고 빠르게 변한다.
⑤ 다양한 데이타 요소를 가진다.
⑥ 스키마와 데이타 사이의 구별이 모호하다.

우리 주변에서 반구조화된 데이타들의 예로는 웹, XML, 지놈(genome), CAD 및 과학 계산 데이타들을 들 수 있다. 본 논문에서는 바이오 데이타의 여러 가지 구조와 비슷하면서도 우리가 쉽게 이해할 수 있도록 생체정보경로 데이타베이스의 XML 데이타를 이용하여 실험하였다. XML 데이타 역시 반구조화된 데이타로서 마크업이 있는 요소의 내용을 시작 태그와 종료 태그로 둘러쌓은 형식을 가진다[13].

본 논문에서 적용한 방법은 반교사학습의 기반이 되는 기존의 co-training과는 다르게 분류자를 만드는데 나이브 베이지(Naive Bayes) 방법을 이용하지 않고 분류표시된 자료와 비분류표시된 자료에 분류 알고리즘을 적용하여 규칙을 생성하고 이를 XML 데이타에 적용하여 실험하였다. 이 접근방법은 풀어야 할 문제의 크기가 커지더라도 변화하는 규칙을 발견하여 적용하기 때문에 데이타의 크기에 관계없이 변화하는 규칙만으로 문제를 해결할 수 있는 장점이 있다. 다음 장에서는 본 논문에서 제시하는 분류 알고리즘을 적용한 새로운 co-training 알고리즘과 기존의 co-training 알고리즘에 대해 살펴본다.

3. 분류 알고리즘을 적용한 co-training 알고리즘

본 논문에서의 실험 목적은 반구조화된 성격의 바이오 데이타 내에 존재하는 여러 가지 규칙을 분류 알고리즘을 적용하여 밝히고, 비분류표시된 데이타에 이 규칙을 적용하여 분류의 정확성을 높이는 것이다. 이를 위한 반교사학습 방법으로 반교사학습의 기반이 되고 있는 co-training 알고리즘을 적용하였다.

3.1 Co-training 알고리즘

아래 그림 1은 [1][3]에서 제시하는 co-training 알고리즘이다.

Co-training 알고리즘은 비분류표시된 데이타를 이용하여 텍스트 분석에 기반이 되는 알고리즘으로 여러 적

전제 :
- 훈련(training) 데이타 셋에서 분류표시된 데이타를 L
- 비분류표시된 데이타를 U
- U에서 랜덤하게 u개의 데이타 샘플을 선택하여 U' 생성

k번 반복 실행 :
- 전체 데이타 x를 2가지 뷰로 나누어 x_1, x_2로 분류
- 분류표시된 데이타 x_1에서 분류자 h_1을 훈련
- 분류표시된 데이타 x_2에서 분류자 h_2를 훈련
- 비분류표시된 U' 에서 p positive와 n negative 데이타 셋을 h_1을 이용하여 분류표시
- 비분류표시된 U' 에서 p positive와 n negative 데이타 셋을 h_2를 이용하여 분류표시
- 분류표시된 데이타 셋 L로 위의 것을 더함
- U에서 U' 를 만들기 위하여 랜덤하게 $2p+2n$ 데이타 셋을 선택

그림 1 Co-training 알고리즘

용분야에서 연구되고 있는 방법이다. [1][3]의 실험에서는 이를 웹페이지 분류에 적용하였는데, 여러 가지 고려해야 할 사항들이 있음에도 불구하고 적은 수의 분류표시된 웹페이지 자료와 많은 수의 비분류표시된 웹페이지 자료를 활용하여 분석한 결과 높은 정확성을 가지는 분석 기법이라는 것이 증명되었다. 기존의 co-training 알고리즘과 본 논문에서 제안하는 분류 알고리즘을 적용한 새로운 co-training 방법을 비교한 실험 결과는 4장에 기술한다.

3.2 분류 알고리즘으로 특징 뷰(feature view) 생성

Co-training의 아이디어와 분류 알고리즘을 적용하여 반구조화된 데이타 즉, XML에서 특징 뷰를 선택하는 문제에 대하여 설명한다. 현재까지 co-training을 응용하고 적용한 알고리즘들은 클러스터링 방법이나 통계적인 방법 혹은 다른 도메인의 데이타들을 결합하여 패턴을 발견하는 방법을 적용하였다. 하지만 실제로 반구조화된 데이타 안에는 일정한 여러 가지 패턴과 규칙이 있기 때문에, 이를 찾기 위한 분류 알고리즘을 co-training 방법에 적용하였다.

실험 XML 데이타로는 생체정보경로(biopathway) 데이타베이스인 [1]*BIND*(*The Biomolecular Interaction* Network Database)를 사용하였는데, 이는 화학적 요소들(chemical compounds)과 반응(reaction)으로 구성된 복잡한 네트웍 모델링으로 구성되어 있다. XML 데이타 내에서 특징 뷰를 갖는 분류자를 만들기 위해서 기존의 co-training 방법과 같이 두 가지 다른 뷰로 데이타를 나누었다. 즉, XML 문서 구조는 태그와 요소내용으로 구성되어 있으므로 각각 태그와 요소내용으로 나누어 각각의 뷰에서 분류 알고리즘인 C5.0을 적용하였다. 기

1) http://www.bind.ca

존의 co-training 방법에서 적용한 웹페이지 분류에서는 문서를 나누는 2가지 뷰로서 요소내용과 웹페이지를 가리키는 하이퍼링크(hyperlink)로 나누었다. 왜냐하면 웹페이지는 HTML 문서로 되어 있기 때문이고 태그는 사용되는 규칙이 정해져 있지만 요소내용을 보충할만한 특징은 가지고 있지 않다. 즉, HTML의 태그는 문서의 구조를 규정하는 것과 태그의 요소내용의 표현을 지정하는 역할만을 하고 요소명의 대소문자 구분을 하지 않는다. 하지만 XML 문서에서의 태그는 요소내용을 설명할 수 있는 핵심 단어로 구성하는 경우가 많고 변화할 수 있는 규칙이 있을 뿐만 아니라 대소문자도 구분한다.

분류표시된 데이타와 테스트 데이타에 대해 위에서 선택된 각 뷰를 이용하여 분류 알고리즘을 적용하고 패턴을 발견하기 위한 과정은 다음과 같다.

(1) BIND XML 문서를 태그 데이타와 요소내용 데이타로 나눈다.
(2) 태그 데이타는 그 자체가 특징을 갖는 데이타이므로 "<"와 "/>"만 제거한 후 태그 데이타의 빈도수만을 가지고 분류 알고리즘인 C5.0을 적용한다.
(3) 요소내용 데이타는 텍스트 데이타이므로 전처리 과정을 수행한다. 우선 텍스트내의 관 사, 동사, 전치사 등과 같은 정지-리스트(stop-list)를 적용하여 제거하고, 단어에서 어근만을 추출하는 스테밍(porter stemming) 알고리즘을 적용한 후 분류 알고리즘인 C5.0을 적용한다.
(4) (2)와 (3)에서 각각의 규칙을 발견하고 이 규칙이 바로 특징 뷰로 비분류표시된 데이타를 구분할 수 있는 기반이 되는 처음 분류자로 만들어지며 이것을 훈련시킨다.

다음 절에서는 위의 이러한 과정으로 만들어지는 분류자를 가지고 비분류표시 데이타에서 분류표시된 데이타로 만들어 나가는 새로운 co-training 알고리즘에 대해 설명한다.

3.3 분류방법을 적용한 co-training 알고리즘

제안하는 co-training 방법은 2가지 뷰에 의해 분류된 것을 각각 태그에 의한 분류자와 요소내용에 의한 분류자로 구분하고, co-training 알고리즘을 다음과 같이 진행시킨다.

(1) 분류표시된 데이타에서 2가지 뷰로 훈련데이타 셋을 나누고 이를 각각 태그 분류자와 요소내용 분류자로 나눈다.
(2) 비분류표시된 데이타 집합에서 분류표시된 데이타를 만들기 위해서 각각의 뷰에서 분류 알고리즘인 C5.0을 적용한다. 그리고 분류표시된 데이타에서 만들어진 규칙과 태그 분 류자를 사용하여 처음의 루트(root) 분류규칙으로 분류와 비분류 데이타로 구분하고, 분 류표시된 데이타로 선택 가능한 데이타 셋과 아닌 데이타 셋으로 비분류 표시된 데이타 셋을 구성한다.
(3) 단계(2)와 유사하게 구분되어진 데이타 셋에 처음의 분류표시된 데이타에서 만들어진 각각의 태그 분류자와 요소내용 분류자를 적용하여 데이타 셋을 분류표시한다.
(4) 위의 단계 (2)-(3)의 분류표시된 데이타 셋에서 새롭게 각각의 분류자들을 생성한다.
(5) 비분류표시된 데이타가 안정적으로 분류표시된 데이타로 만들어질 때까지 단계 (2)에서 (4)를 반복한다.

알고리즘에서 사용된 표기는 그림 2에서 설명하고, 그림 3은 위의 단계를 알고리즘으로 표현한 것이다.

V_1, V_2	: 각각의 두가지 뷰
A	: 분류 알고리즘
h_1, h_2	: 뷰에서 분류 알고리즘을 이용하여 만들어지는 분류자
L	: 분류표시된 데이타
U	: 비분류표시된 데이타
k	: 반복 실행수
C_i	: 비분류표시된 데이타의 부분집합
E_1, E_2	: 비분류표시된 데이타에서 분류표시될 가능성이 있는 데이타 집합

그림 2 표기

다음을 k 번 반복 실행
1. A, $V_1(L)$, $V_2(L)$ 이용하여 분류자 h_1과 h_2를 만든다.
2. U의 클래스 C_i ($C_i \subset U$)에 대해서,
   2.1 h_1과 h_2를 이용하여 C_i에서 가장 만족할만한 예측값들을 만들고 이를 각각 E_1, E_2라 한다.
   2.2 U에서 E_1과 E_2를 제거하고 각각 h_1과 h_2로 분류표시하여 L에 더한다.
   2.3 위의 결과 $U = U - C_i$, $L = L + C_i$로 U와 L을 수정한다.
3. h_1과 h_2의 예측자들을 결정하고, 이를 결합한다.

그림 3 분류 알고리즘을 적용한 co-training

4. 실험 데이타 및 결과

새로운 co-training 방법은 두 가지 실험 데이타를 사용하여 결과를 비교하였다. 첫 번째 데이타는 기존의 co-training 방법에서 실험한 데이타인 WebKB 데이타로써 본 논문의 분류 알고리즘을 적용한 co-training 방법을 이용하여 기존의 실험 결과와 비교하였다. 그리고 두 번째 데이타로는 3절에서 설명하였듯이 생체정보경로 BIND XML 문서를 가지고 실험하였다.

이 모든 실험의 프로그램 실행 환경은 마이크로소프트 윈도우 2000에서 비쥬얼 베이직 프로그래밍 언어를 사용하여 구현하였으며, 구현된 프로그램은 마이크로소

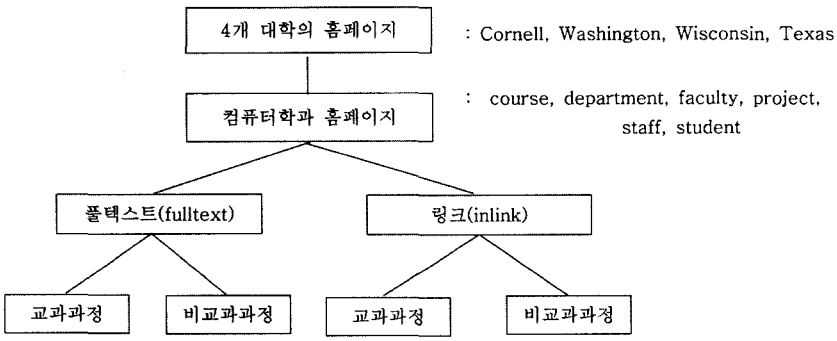


그림 4 WebKB 전체 데이타 구성도

프트 닷넷(.NET) 도구를 이용하였다. 그리고 분류 알고리즘은 SPSS사의 클레멘타인 5.2.1 버전에서 C5.0 알고리즘을 이용하여 그 규칙을 적용하였다. 다음 절들에서 각 데이타에 대한 설명과 함께 실험 결과를 보인다.

4.1 WebKB 데이타

이 데이타 셋은 1997년 CMU(Carnegie Mellon University)의 WebKB 프로젝트에 이용된 데이타로서 다양한 대학의 컴퓨터학과의 웹페이지를 포함한다. 전체 1,051개의 웹페이지를 수작업으로 분류하여 교과과정(course: 230개) 페이지와 비교과과정(non-course: 821개) 페이지로 구분하였다. 데이타를 다운받으면[2] 두개의 디렉토리로 구분되어 있는데 웹페이지의 텍스트를 포함하고 있는 텍스트 파일인 풀텍스트(fulltext) 디렉토리와 그 페이지를 가리키는 하이퍼링크 텍스트 파일인 링크(inlink) 디렉토리로 나누어져 있다. 그리고 각 두개의 디렉토리에는 4개 대학을 5개 카테고리로 구분하여 놓았다. 데이타의 전체 구성도는 그림 4와 같다.

[1], [13]의 실험과 동일한 방법으로 전체 1,051개의 데이타 중에서 약 25%인 263개 데이타 셋을 테스트 데이타로 하였고, 처음 12개 데이타 셋을 분류표시된 셋으로 하여 실험하였다.

표 1은 교과과정 홈페이지를 기존의 co-training 방법과 분류 알고리즘을 적용한 새로운 co-training 방법으로 웹페이지를 분류하여 에러율을 비교한 표이다. 웹페이지 기반 분류자와 하이퍼링크 분류자 모두 기존의 교사학습과 co-training 방법보다 분류 알고리즘을 적용한 새로운 co-training 방법이 작은 에러율을 보인다. 또한 이 둘을 결합한 분류자를 적용한 에러율이 훨씬 작은 것을 알 수 있다. 이러한 결과를 자세히 분석해보면 기존의 co-training 알고리즘에 적용한 WebKB 데이타는 정지 리스트와 스태밍 알고리즘을 적용하지 않았다. 하지만 본 논문에서 제안하는 알고리즘은 이 두 가지 방법을 적용하였고 분류 알고리즘 자체가 규칙을 발견해주기 때문에 보다 정확한 결과를 가져왔다고 볼 수 있다.

표 1 Co-training과 제안된 co-training의 웹페이지 분류 에러율 비교(%)

	교사학습	co-training	New co-training
웹페이지 분류자	12.9	6.2	5.3
하이퍼링크 분류자	12.4	11.6	9.3
결합 분류자	11.1	5.0	1.0

또한 기존의 co-training 방법의 에러율을 살펴보면 하이퍼링크기반 분류자 에러율(11.6%)이 웹페이지기반 분류자 에러율(6.2%)보다 높다. 그 이유는 WebKB 데이타에서 하이퍼링크기반 분류자가 웹페이지기반 분류자보다 그 양과 질에 있어서 현저히 떨어지는 데이타 값들을 가지고 있었다. 물론 두 가지 뷰로 데이타 셋을 나누어서 이 두가지 뷰로 인한 분류자가 데이타 분석에 있어서 교사학습보다 좋은 결과를 가지고 온다는 것을 알 수 있었지만, 만약 뷰를 둘로 나누었을 때 현저히 낮은 데이타 값을 가질 경우 동등한 다중 뷰(multi-view)를 선택하는 방법도 고려해 보아야 한다고 판단된다.

4.2 XML 데이타

새로운 co-training 방법을 테스트하기 위한 다른 하나의 실험 데이타로서, 바이오 데이타의 여러 가지 구조와 비슷한 반구조화된 데이타로 생체정보경로 BIND XML 문서 중에서 총 1,000개를 가지고 실험하였다. BIND XML 문서는 3가지 다른 XML 문서로 구성되어 있는데 이중에서 50개만을 분류표시 데이타로, 나머지를 비분류표시 데이타로 구성하였다. BIND는 상호작용(interactions), 복합체(complexes), 경로(pathways)의 세 가지 XML 문서로 구성되어 있다. 이 세 가지 문서에서 분류의 정확성을 살펴보기 위하여 복합체 XML

2) http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-51/www/co-training/data/

문서를 목적함수(target function)로 하여 복합구성 XML 문서를 긍정(positive) 값으로 하고, 나머지 상호작용과 경로 XML 문서는 부정(negative) 값으로 하여 긍정과 부정 값을 정확히 찾아내는지 조사하였다.

각 실험에서는 우선 전체 1,000개의 문서 중에서 21.4%인 214개의 문서를 테스트 셋으로 두었다. 그리고 나머지 데이타 셋은 분류 알고리즘을 적용한 co-training 알고리즘으로 실험하였다.

표 2 5번 반복 후 분류 알고리즘을 적용한 co-training과 교사학습의 에러율(%)

	요소내용기반 분류자	태그기반 분류자	결합 분류자
교사학습	7.5	11.7	6.0
New co-training	3.7	4.7	0.0

표 3 상위 10% 제거후 변화 에러율(%)

	요소내용기반 분류자	태그기반 분류자	결합 분류자
New co-training	4.2	8.4	2.8

표 2는 각 행별 분류자를 교사학습 알고리즘을 이용한 분류표시된 데이타 셋에 적용한 다음 이를 다시 테스트 데이타에 적용한 경우의 에러율과 본 논문에서 제안한 새로운 co-training 실험결과의 비교표이다. 이 표에서 보면 교사학습보다 제안한 co-training 알고리즘이 요소내용기반 분류자와 태그기반 분류자의 두 가지 모두에서 더 좋은 결과가 나왔고, 결합 분류자에 의해서는 100%의 정확도를 보인다. 이는 분류 알고리즘의 정확성에 그 이유가 있을 수도 있지만 우연히 단어의 빈도수가 높아서 그럴 수도 있으므로, 여러 번 분류 알고리즘을 적용하여 데이타 셋의 상위 10%에서 발견되는 규칙을 제거한 다음에 제거된 데이타 내에서 분류 알고리즘을 적용하여 co-training 방법을 수행하였다. 그 결과가 표 3인데 이 또한 교사학습보다 에러율이 작음을 알 수 있으며, 이는 자주 나타나는 단어의 빈도수에 의하여 우연히 만들어지는 규칙이 아니라는 사실을 말해준다.

따라서, 본 실험결과에서 예측해 볼 수 있는 것은 생물정보학 연구에서 자주 나타나는 서열(sequence) 데이타와 마이크로어레이(microarray) 데이타 등과 같은 값들이 반구조적인 형태로 구성되어 있고, 또한 분류표시된 데이타의 정보도 많지 않을 때 이에 대한 적절한 분류 및 분석 방법으로서 이 논문에서 제안하는 방법을 적용하면 데이타를 좀 더 정확히 분류할 수 있고 분석 시간 및 비용을 절감하는 등 여러 가지 이점이 있다고 생각된다.

5. 결론 및 향후 연구과제

생물정보학 등 많은 응용 분야에서 데이타 분석을 할 때, 데이타는 적은 수의 분류표시된 데이타와 많은 수의 비분류표시된 데이타로 구성될 수 있다. 분류표시된 자료들은 사람이 직접 분류하기 때문에 시간과 비용이 많이 들며, 비분류표시된 자료들은 그 반대로 인터넷 등을 통해서 값싸고 쉽게 얻을 수 있지만 자료가 많아도 그 자체로는 학습을 하지 못한다는 단점이 있다. 따라서 정규 교사학습 알고리즘의 수행 결과를 향상시키기 위해서 소수의 분류표시된 자료와 다수의 비분류표시된 자료를 어떻게 적절히 처리할 것인가 하는 기법들이 많이 연구되고 있다.

본 논문에서는 이러한 문제점을 해결하기 위하여 기존의 co-training 알고리즘의 아이디어에 분류 알고리즘을 적용하여 기존의 결과보다 더 나은 결과를 보였고, 또한 이 방법을 바이오 데이타의 구조와 비슷한 반구조화된 데이타 분석에도 적용해 보았다. 논문에서 제안한 알고리즘과 기존의 co-training 알고리즘을 비교하기 위하여 웹데이타인 WebKB와 그 실험값을 비교하였고, 또한 반구조화된 XML 데이타에서도 새로운 co-training 알고리즘이 좋은 결과 도출에 영향을 미친다는 것을 알 수 있었다. 그리고 아직 표준화가 되어 있지 않은 생체정보경로 XML 문서를 가지고 실험한 것이기 때문에 이에 따른 문서 간 비교 분류를 하기 위해 적절한 방법으로 변수를 조절하고 선택하여 실험하였다.

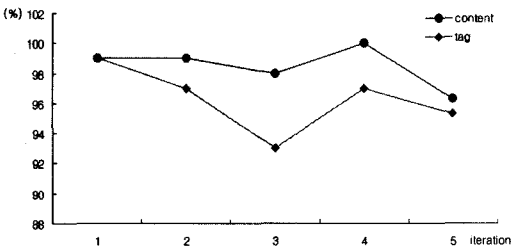


그림 5 5번의 분류정확도 반복실험

그림 5의 y축 값은 5번의 반복실험 결과 생체정보경로 XML 문서를 찾는 정확도 비율을 나타낸다. 생체정보경로 XML 문서를 분류하는 데는 분류자 요소내용이 태그보다 co-training 수행에 더 도움을 준다는 것을 살펴볼 수 있다. 이것은 요소내용이 태그보다 문서마다 너무나 많은 서로 다른 단어가 사용되고 생체정보경로 XML 태그 특성상 목적 함수에 적절한 정확성을 표현할 수 있는 분류자를 찾는 것이 제한적이기 때문이다. 하지만 교사학습에 비해서 더 나은 정확성을 가지기 때문에 분류 알고리즘을 적용한 co-training 알고리즘의

반복실행으로 XML 문서 즉, 반구조화된 성격의 데이타 값을 갖는 경우에서도 효과적임을 알 수 있었다. 그리고 생체정보경로 XML 문서 실험을 통하여 앞으로 바이오 데이타 분석을 위한 여러 응용 문제에서 반교사학습 방법이 바이오 데이타 분석에 적합한 알고리즘이라는 것을 알 수 있었다.

이 실험은 제한된 XML 문서 즉, 한 가지 데이타 셋과 한 개의 목적 함수를 가지고 실험하였다. 그리고 실험 데이타 값을 XML 문서로 하였기 때문에 분류자를 선택하는 데에 있어서 한정된 선택밖에 할 수 없었다. 하지만 현재 생체정보경로의 XML 문서구조와 문서 안에서 자주 나타나는 단어의 패턴을 살펴볼 수 있었다. 또한 제안한 알고리즘이 두 가지 뷰로 데이타를 나누어 분류하고 데이타 내에서 규칙을 발견하여 반구조화된 데이타를 분석하였기 때문에 바이오 데이타를 분석할 때에도 효과적일 것이라고 판단된다.

본 실험에서 중요한 사항은 분류 알고리즘과 반교사학습 방법의 결합으로서 분류표시된 데이타가 많지 않을 경우에 분류 알고리즘을 적용하여 분류자를 만들고, 데이타 내에서 규칙을 생성한 다음에 그 결과를 비분류표시된 데이타에 적용한 것이다. 또한 바이오 데이타를 분석할 때 분류표시된 데이타의 수가 많지 않다면 비분류표시된 문서를 추가적으로 사용했을 경우에 사용하지 않았을 때보다 문서 분류의 정확도가 증가하였으므로 반교사학습 알고리즘 적용은 유용한 방법이라 할 수 있다. 앞으로의 계획은 분류자를 선택할 때 데이타에 따른 다양한 뷰를 자동적으로 선택하는 방법과 새로운 적용 알고리즘을 설계하여 다양한 바이오 데이타에 적용해 보는 것이다.

참 고 문 헌

[1] T. Mitchell, "The Role of Unlabeled Data in Supervised Learning," *Proceedings of the 6th International Colloquium on Cognitive Science (ICCS)*, pp.254-278, 1999.

[2] S. Goldman and Y. Zhou, "Enhancing Supervised Learning with Unlabeled Data," *Proceedings of the 7th International Conference on Machine Learning (ICML)*, pp.327-334, 2000.

[3] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," *Proceedings of the 11th Annual Conference on Computational Learning Theory* (COLT), pp.92-100, 1998.

[4] K. Nigam and R. Ghani, "Analyzing the Effectiveness and Applicability of Co-training," *Proceedings of Information and Knowledge Management*, pp.86-93, 2000.

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

[6] K. Nigam, A. K. Mccallum, S.Thrun, and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning*, 39(2/3), pp.103-134, 2000.

[7] I. Muslea, S. Minston and C. Knoblock, "Selective Sampling with Redundant Views," *Proceedings of National Conference on Artificial Intelligence*, pp.621-626, 2000.

[8] I. Muslea, S. Minston and C. Knoblock, "Active+ Semi-Supervised Learning=Robust Multi-view Learning," *Proceedings of International Conference on Machine Learning (ICML)*, pp.435-442, 2002.

[9] B. Raskutti, H. Ferra and A. Kowalczyk, "Combining Clustering and Co-training to Enhance Text Classification Using Unlabelled Data," *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD)*, pp.620-625, 2002.

[10] M. Figueiredo, A. K. Jain and M. H. Law, " A Feature Selection Wrapper for Mixtures," *Proceedings of the First Iberian Conference on Pattern Recognition and Image Analysis, Puerto de Andratx, Spain, June*, 2003.

[11] I. Muslea, S. Minston and C. Knoblock, "Active Learning with Strong and Weak Views: A Case study on Wrapper Induction," *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2003.

[12] P. Buneman, "Tutorial: Semistructured Data," *Proceedings of ACM Symposium on Principles of Database Systems*, pp.117-121, 1997.

[13] D. Suciu, "Semistructured Data and XML," *Proceedings of International Conference on Foundations of Data Organization (FODO)*, 1998.

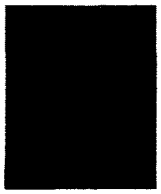
[14] O. Chapelle, J. Weston and B. Scholkopf, "Cluster Kernels for Semi-Supervised Learning," *Advances in Neural Information Processing Systems (NIPS 2002), MIT Press. Cambridge, MA*, 2003.

윤 혜 성

1999년 2월 동덕여자대학교 전산통계학과 학사. 2001년 2월 이화여자대학교 컴퓨터학과 석사. 2001년 3월~현재 이화여자대학교 컴퓨터학과 박사과정. 관심분야는 데이타마이닝, Bioinformatics

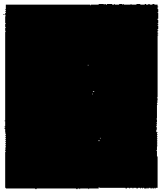
이 상 호

1979년 서울대학교 계산통계학과 이학사. 1981년 한국과학기술원 전산학과 이학석사. 1987년 한국과학기술원 전산학과 공학박사. 1990년 미국 일리노이대학교 전산학과 방문교수. 현재 이화여자대학교 컴퓨터학과 교수



박 승 수

1974년 2월 서울대학교 수학과 학사 1976년 2월 한국과학기술원 전산학과 석사. 1988년 5월 (미)텍사스 대학 컴퓨터학 박사. 1988년~1991년 (미)켄사스 대학 컴퓨터학과 조교수. 1991년~현재 이화여자대학교 컴퓨터학과 교수. 관심분야는 인공지능, 데이타마이닝, Bioinformatics



용 환 승

1983년 서울대학교 컴퓨터공학과 학사 1985년 서울대 대학원 컴퓨터공학과 공학석사. 1985년~1989년 한국전자통신연구소 연구원. 1994년 서울대 대학원 컴퓨터공학과 공학박사. 2002년 8월~2003년 2월 IBM T.J. Watson 연구소 객원연구원. 1995년~현재 이화여자대학교 컴퓨터학과 부교수 관심분야는 객체-관계 데이타베이스 시스템, 멀티미디어 데이타베이스, OLAP 및 데이타 마이닝, 바이오정보학, 유비쿼터스 컴퓨팅

김 주 한

1988년 2월 서울대학교 의과대학. 1995년 2월 서울대학교 의과대학 석사. 1998년 2월 서울대학교 의과대학 박사. 2000년 5월~2001년 8월 (미)하버드의대 생명의료정보학 교수. 2001년 2월 (미)MIT 공학석사. 2001년~현재 서울대학교 의과대학 생명의료정보학 교수. 관심분야는 Bioinformatics, Medical Informatics, Pattern Recognition