DOI: 10.3745/KIPSTB.2009.16B.6.479

경계범주 자동탐색에 의한 확장된 학습체계 구성방법

최 윤 정[†] · 지 정 규[††] · 박 승 수[†††]

요 약

본 논문은 기존의 목표항목만을 위주로 한 학습체계에서 발생하는 오분류 문제의 해결을 위해 기존의 학습체계에 경계항목을 자동으로 탐색하여 포함시켜 확대시키는 방법을 제안하고 있다. 여러 주제에 걸쳐 다양한 내용을 다루는 복잡한 문서들은 확실히 어느 범주로 분류해야 할지 판가름하기 어려운 성질인 모호성이 강하다. 이러한 경우 모든 경우들을 정확히 구분할 수 있는 최적의 경계를 찾는 일은 더욱 어려운 일이다. 복잡하고 불확실성이 높은 데이터들의 특징은 대부분 분류 경계영역에 위치하므로 이러한 분류경계의 데이터들을 새로운 학습 항목으로 인식시키도록 하는 것이 필요하다. 본 연구에서는 주어진 목표항목 사이의 경계항목을 자동으로 탐색하여 학습체계에 추가하는 학습 체계 확장 알고리즘을 제시하고, 의도적인 학습오류를 발생시킨 후 기존방법과의 비교실험을 수행함으로써 제안방법의 정확성과 안정성을 비교하였다. 실험결과 경계범주를 포함하여 학습 체계를 확장시켰을 때의 예측력은 기존 0.70에서 0.86으로 약 24% 향상 되었고, 오류를 포함시켰을 때의 예측력은 기존 0.52에서 0.79로 약 49% 향상되었다.

키워드 : 기계학습, 학습알고리즘, 능동학습, 계층 분류, 클러스터링

Construction Scheme of Training Data using Automated Exploring of Boundary Categories

Yun-Jeong Choi[†] · Jeong-Gyu Jee[††] · Seung-Soo Park[†††]

ABSTRACT

This paper shows a reinforced construction scheme of training data for improvement of text classification by automatic search of boundary category. The documents laid on boundary area are usually misclassified as they are including multiple topics and features. which is the main factor that we focus on. In this paper, we propose an automated exploring methodology of optimal boundary category based on previous research. We consider the boundary area among target categories to new category to be required training, which are then added to the target category sementically. In experiments, we applied our method to complex documents by intentionally making errors in training process. The experimental results show that our system has high accuracy and reliability in noisy environment.

Keywords : Machine Learning, Learning/Training Algorithms, Active Learning, Hierarchical Classification, Clustering

1. 서 론

지금까지의 학습알고리즘은 분류를 위한 범주(Category)가 미리 결정된 상태에서 학습문서선택과 구성의 문제가 주로 다뤄져 왔다. 본 연구에서는 분류의 목표를 반영하는 분류체계(Classification Scheme)를 수립하는 문제로 확장하기로 한다. 학습문서를 선택하는 문제 이전에 분류체계 설정의 문제로 변환하는 것이다. 분류체계는 분류문제에서 궁극적으로 분류해야 할 대상과 직결된다. 분류의 응용영역과 밀접한 연관이 있으며 효율적인 정보관리 측면에서 매우 중요하게 관리되어야 한다. 예로써 사용자의 활용 목적에 따라 잘 구성된 디렉토리는 문서를 저장하는 역할 뿐 아니라 작업의 계획을 반영하며 문서들을 효율적으로 관리하게 하는 기반이 된다.

본 논문의 목적은 자동분류시스템에서 전문가의 개입비용과 오분류율(Misclassification Rate)을 최소화하는 것이며, 이를 위해 기존의 분류체계를 확장시키는 학습알고리즘을 제안한다.

제안방법의 배경은 다음과 같다.

- 실세계에서의 분류문제들은 점점 더 복잡해지고 불확실성이 강해지므로 최적의 분류경계를 찾기 힘들다.
- 오분류율이 높은 문서들은 대부분 분류경계(Decision Boundary) 영역에서 발생한다.

† 정 회 원 : 서일대학 정보통신과 강의전담교수
†† 종신회원 : 한국연구재단 연구기반조성단장
††† 정 회 원 : 이화여자대학교 컴퓨터공학과 부교수
논문접수 : 2009년 10월 21일
심사완료 : 2009년 10월 29일

분류체계 확장방법의 주요 내용은 다음과 같다.

- 목표범주간의 불확실하고 모호한 경계인접부근을 찾아 별도의 범주로 정의하여 우선적으로 오분류율(False Positive Rate)을 줄인다.
- 각 목표범주와의 상호관련도가 높은 입력 문서들은 본래의 목표범주보다 경계영역상의 놓여져 있을 가능성이 높다.
- 불확실성이 강한 문서들은 경계영역으로 분류될 것이며 후속 작업을 통해 좀 더 근접한 목표항목을 찾아 범주를 결정하도록 하여 정확도(True Positive Rate)를 높인다.

전통적인 기계학습방법의 분류체계는 목표항목으로만 구성되는 것이 일반적이다. 본 논문의 제안방법은 자동으로 탐색된 경계항목을 포함시켜 확장하는 것으로서, 학습 집합 내의 가장 불확실성이 큰 집단을 탐색해내는 알고리즘을 설정하여 매번 수작업으로 경계항목을 지정해야 했던 지난 연구의 내용을 보완하고 있다[1]. 이는 정보력(Information Power)과 구분력(Classification Power)이 큰 문서들을 구분하여 학습데이터로 선택하는 능동학습방법의 개념을 분류체계 설계 과정에 응용한 것이다. 이렇게 확장된 분류체계는 분류기준 및 범주결정방식에 영향을 준다. 또한, 상향식(Bottom-Up) 계층분류를 위해 세분화 시킨 학습데이터 집합은 분류기(Classifier)에게 보다 풍부한 범주의 정보를 제공한다.

본 논문은 다음과 같이 구성된다. 2장에서는 관련연구로서 분류체계의 역할 및 계층구조, 능동학습방법의 개념을 설명한다. 3장에서 학습체계의 확장을 위한 경계범주 자동탐색알고리즘에 대해 기술하고 상향식 계층분류를 위한 확장된 학습데이터 구성방법(Extended Training Set Organization Method, 이하 ETOM) 에 대해 설명한다. 4장에서는 기존의 분류체계과 제안방법으로 확장된 분류체계에 대한 비교실험으로서 자동분류시스템의 동작과정을 보이고, 5장에서 결론을 맺는다.

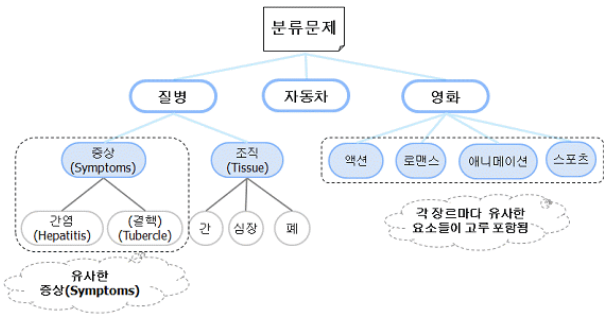
2. 관련 연구

2.1 분류체계와 분류경계

분류체계는 일련의 분류목표와 기준에 따라 설정되며 이 기준이 분류의 경계가 된다. 미리 주어진 범주와 학습 데이터에 분류알고리즘을 적용하여 범주의 고유한 특징을 찾아 범주간 경계를 정하는 일이 학습과정의 역할이다[2-4].

분류경계를 정하는 문제에 있어서 직관적으로도 뚜렷하게 분리할 수 있는 서로 다른 영역의 범주들의 경계는 선형적 형태로 쉽게 드러난다.

(그림 1)의 증상이 비슷한 질병을 가려내는 것처럼 높은 상관도를 갖는 유사한 영역의 범주를 구분하는 경계는 비선형적 형태로 나타나며, 이때 최적의 결정 경계선을 찾는 일은 거의 불가능하다. 각 범주를 대표하는 특이한 자질들이 대부분 공유되어 있기 때문이며 이렇게 공유된 자질들은 오



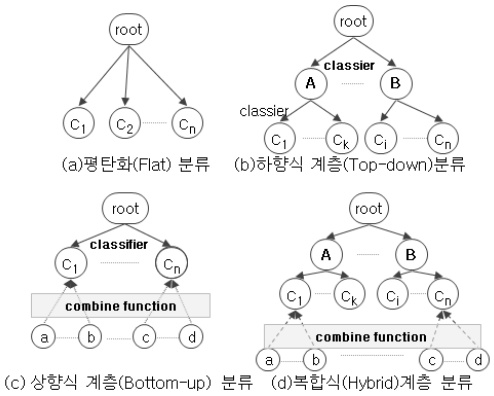
(그림 1) 분류체계 상관도에 따른 분류문제

분류를 일으키는 주 원인이 된다. 즉, 범주를 구성하는 특이하고 고유한 자질들이 일반화 된 자질로 인식됨으로써 그 구분력을 상실하게 되는 것이다. 대표적인 예로는 여러 내용을 함축적으로 담고 있는 과학문서와 평범한 문서로 보이기 위해 작성된 스팸성 문서들이 이에 해당한다. 스팸성 문서들의 분류를 위해 여러 강력한 필터링 규칙이 사용되지만 문서 자체가 필터링 규칙을 벗어나도록 자질들이 교란되도록 작성되어있기 때문에 구분이 매우 어렵다.

2.2 분류체계와 계층구조

분류체계는 분류분석이 필요한 응용영역에서 분류목표와 의도가 반영되어야 하며, 데이터의 분포와 영역 그리고 특성을 고려하여 구성해야 한다. 문서의 범주 집합을 C 라 하고 각각의 목표범주들을 c_1, c_2, …, c_n이라고 하자(|C|=n). 문서 분류를 위한 범주 구성방법으로 (그림 2)와 같이 모든 범주를 루트 노드 아래에 평탄화 된 구조로 배치할 수 있고, 범주를 적당한 크기의 부분 집합으로 묶고 그 부분집합들을 계층적 구조로 배치할 수 있다. 문서분류방법에 있어 전자의 경우를 평탄화 된 분류(Flat Classification), 후자를 계층적 분류(Hierarchical Classification)라고 부른다[5, 6].

계층적 분류방식은 분석데이터의 특성과 영역을 고려한 것으로써, 분류가 이루어지는 방향에 따라 하향식(Top-Down)의 분할(Partition)방법과 상향식의 병합(Agglomeration) 방법으로 나뉜다. 하향식 방법은 입력문서가 상위 범주로부터 시작하여 하위 클래스까지 분류하도록 구성된 분류기에 따



(그림 2) 평탄화 분류와 계층적 분류

라 단계적인 분류가 이루어지는 방식이다. 상향식은 입력문서가 가장 하위범주를 구분하는 분류기와 그 결과를 통합하여 범주를 결정하도록 하는 병합함수(Combine Function)에 의해 상향식으로 올라가면서 분류가 이루어진다. 공통적으로 광범위한 범주를 적당한 크기의 부분 집합으로 나누어 계층적 구조로 배치함으로써 분석데이터에 내포된 내용에 대해 직관적인 이해와 예측이 가능하도록 구성하고 있으며, 평탄화 방법보다 좋은 성능을 나타낸다. 보통 하향식 방법이 상향식 방법에 비해 정확도가 높아 정밀한 분류가 필요한 영역에서 많이 적용되고 있으나 각 계층을 이루는 노드마다 분류기를 구성해야 하므로 학습비용이 매우 높다는 단점이 있다.

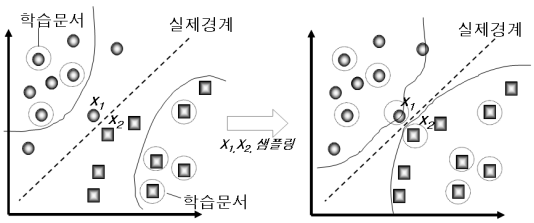
이러한 계층적 분류방법의 장점은 보다 유연하고 정교한 분류전략을 선택할 수 있다는 점이며, 범주의 계층구조가 사람의 직관적인 이해와도 가깝기 때문에 실제 환경에 적용하기에 적합하다. 각 범주의 계층구조를 생성하기 위해서는 분류체계에 대해 사전지식이 있는 전문가가 직접 구성할 수 있고, 보다 객관적인 개요를 파악하거나 정밀한 결과를 얻고자 할 때는 클러스터링 기법을 활용하기도 한다. 계층적 분류와 평탄화된 범주집합의 구성 모두 공통적으로 정보량이 큰 데이터를 훈련데이터로 선택하는 것이 관건이며 이러한 문제를 다루는 방법으로 능동적 학습알고리즘이 대표적이다.

2.3 능동적 학습(Active Learning) 알고리즘

능동적 학습알고리즘은 정보량이 큰 데이터를 학습데이터로 선택하는 것으로써, 여기서 정보량이 큰 데이터란 어느 한 범주에 확실히 지정될 수 있는 구분력을 갖추었거나 현재의 기준으로는 판단력이 부족한 데이터를 의미한다[7, 8]. 능동적 학습알고리즘에서 정보량이 큰 데이터란 판단하는 근거는 기준에 따라 다양한데, 그 중 대표적인 것이 불확실성(Uncertainty) 개념을 이용하는 것이다. 이때의 정보력이 큰 데이터는 불확실성이 높은 것으로서 어느 범주로 구분되어야 할지 애매한 경우, 즉 분류함수가 분류하기 어려운 데이터들로 범주를 판단할 때 확신이 적은 것들을 의미한다. 이러한 데이터들은 대부분 범주를 구분하는 경계선상에 있거나 경계영역에 근접하여 확실히 어떤 범주로 구분되어야 하는지 판단하기 어려운 성질을 지닌다.

(그림 3)은 능동적 학습알고리즘의 기본적 개념을 설명한다. 그림에서 ●와 ■는 각각의 범주로 구분된 개체라고 볼 때, 새로운 데이터들도 잘 분류할 수 있도록 학습시키기 위해 이들을 학습집합에 포함시켜 구성할 수 있다. 어떤 개체가 학습데이터로 적합한가 선택하는 문제에 있어서, 점선으로 표시된 실제경계선과 이 경계선이 움직일 수 있는 범위로 확장시킨 경계영역으로도 뚜렷이 분리되는 지점들의 데이터가 적당하다. 이 경우 능동학습의 의미는 전문가가 범주의 특징을 잘 말해주는 개체 몇 개를 학습데이터들로 선택해 주는 것으로 이해할 수 있다.

이러한 개체들이 적절히 분류될 수 있도록 위해 학습집합



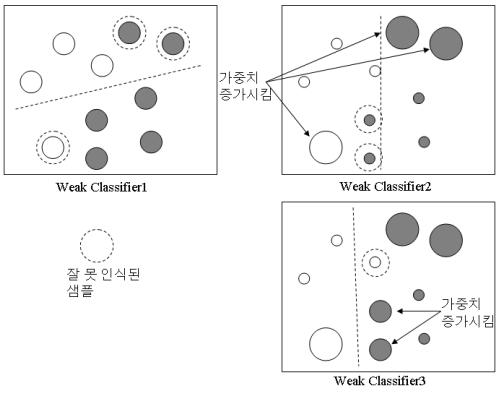
(그림 3) 불확실성 기반의 샘플링 알고리즘의 개념

으로 선택되게 하는 것이 능동적 학습의 핵심이며, 학습집합 선택에 이 개념을 이용하는 것을 불확실성 기반 샘플링 알고리즘(Uncertainty Based Sampling Algorithm)이라고 한다[9]. 자연스럽게 일반적인 감독학습(Supervised Learning) 알고리즘에 비해 전문가의 역할이 좀 더 큰 비중을 차지하게 되며, 불확실성이 높은 데이터들을 학습데이터로 삼기 위해 각각의 내용을 개별 확인하여 지정해 주어야 하는 일도 필요하다. 경계선상의 모든 개체들을 학습데이터로 선택하면 해당영역들의 자질들이 일반화되어 구분력을 잃게 되는 경우도 생기게 된다.

2.4 에이다부스트(Adaboost) 알고리즘

기계학습 기법에서의 학습문서의 구성방법은 좋은 훈련사례가 될 학습문서들을 선택하는 샘플링(Sampling) 방법과 선택한 학습문서를 재구성하는 부스팅(Boosting)방법에 의해 접근된다[10]. AdaBoost알고리즘은 학습문서집합을 재구성하여 더욱 정확도를 크게 하는 학습문서집합을 만들어 내기 위한 알고리즘으로 수많은 연구에서 수학적으로 분석되고 정확도가 향상된다는 사실이 증명되었다.

이 알고리즘은 부스팅 알고리즘을 응용한 것으로 에러율이 0.5보다 작기만 하면 된다는 조건을 만족시키는 약한(Weak) 분류함수들을 부스팅 효과를 통해 정확한 분류함수를 만들고 있다. Adaboost는 기본적으로 확률분포에 의한 가중치를 가지고 학습이 진행된다. (그림 4)와 같이 학습문서의 비중을 조정(Re-Weighting)할 때, 분류결과가 틀린 문서의 비중을 높이는 방식으로 알고리즘을 진행시켜 나가는 것이 특징이다. 예를 들면, 단순한(Weak) 분류기로 주어진 샘플들을 인식하고, 정확히 인식된 샘플에 대해서는 가중치



(그림 4) AdaBoost 알고리즘

를 감소시키고 오인식된 샘플에 대해서는 가중치를 증가시켜서 다음의 단순한 분류기에 이 결과를 반영시켜 분리하도록 한다.

2.5 클러스터링 알고리즘

클러스터링은 데이터에 내재된 특성을 자동으로 추출하여 그룹핑하는 방법으로서 기계학습 기법상 비교사 학습(Unsupervised Learning)에 해당한다. 문서 클러스터링이란 대용량의 문서 집합을 군집화하는 것으로 정보 추출을 위한 중요한 도구로 오래 전부터 다루어져 왔으며, 분류나 정보검색의 전처리(Preprocessing) 단계에 많이 사용된다.

클러스터링의 분석기법은 분류방법과 마찬가지로 평탄식과 계층식으로 나뉘며, 계층식은 다시 분할식과 병합식으로 나뉜다. 분할식에서는 전체 입력문서집합을 하나의 클러스터로 간주하고 반복적으로 더 작은 클러스터로 분할해 나간다. 병합식에서는 각각의 입력문서들을 개별적인 클러스터로 간주하고 유사한 문서들을 반복적으로 그룹핑 한다. 최근에는 전체 입력문서에 대해 문서-문서간의 유사행렬을 미리 생성한 후, 다음단계에서 병합되거나 분할될 때 발생하는 정보의 이익과 손실을 미리 예측해가며 최적의 클러스터를 구성하는 연구가 활발하다.

3. 분류체계의 확장을 위한 경계범주 자동탐색 알고리즘과 상향식 계층분류알고리즘

본 논문에서 제안하는 확장된 분류체계는 먼저 분류문제를 위해 미리 주어진 목표범주집합과 각 목표범주집합 간의 경계범주집합으로 구성된다. 분류문제에 있어서 각 목표범주들을 구분할 수 있는 개념상의 결정 공간(Decision Boundary)이 있다고 가정하면 결정 공간 부근의 영역이 경계범주가 되는 것이다. 또한, 평탄화 된 분류방법이 아니라 비교적 정확성이 높은 상향식 계층분류를 수행하기 위해 각 목표범주들을 하위범주로 나누어 구성한다. 이전 연구에서는 경계범주집합과 하위범주집합을 수작업으로 정의했으나 본 연구에서는 클러스터링 분석을 통해 최적의 영역을 찾아낼 수 있는 방법을 제안한다.

3.1 범주의 정의 : 목표범주와 경계범주

3.1.1 목표범주집합(Target Category Set)

분류문제를 위해 미리 주어진 범주들의 집합으로써, 목표범주집합 C = {c_1, c_2, … , c_n}는 궁극적으로 분류하고자 하는 목표범주 c_i들의 집합을 의미한다. 이때, 각 목표범주들을 위한 학습문서집합인 Tr{C}은 각 목표범주 c_i를 가장 잘 나타내는 문서들로 이루어진 훈련 집합(Training Set) 을 나타낸다.

3.1.2 경계범주집합(Boundary Category Set)

경계범주집합 X = {x_1, x_2, … , x_m}는 각 목표범주들을 구분짓는 불확실한 경계영역을 나타내는 개념적인 공간이다. 임의의 경계범주 x_k 는 목표범주 c_i 와 c_j 의 불확실하고 모호한 경계영역을 의미한다. 만약 c_i 와 c_j 의 상관관계가 적어 서로 공유되는 자질들이 없이 극명한 분리가 가능하다며 이 영역은 존재의 의미가 없다. 반면, 각 목표항목들의 학습문서들의 상호연관도가 높을수록 이 공간의 크기는 작아지며 선형분리가 불가능한 형태를 띄게 된다. 본 논문에서는 목표범주들을 위한 학습문서들의 자질벡터(Feature Vector)를 이용하여 클러스터링 분석을 통해 경계범주를 자동으로 구성하고 있다.

3.2 경계범주 자동탐색 알고리즘

목표범주간의 모호한 경계영역 탐색 및 구성을 위해 각 목표범주들의 학습문서집합을 입력으로 하여 클러스터링 분석을 수행한다. 최적화 된 클러스터들을 구성하기 위해 클러스터의 내부 유사도(Internal Similarity)와 외부 유사도(External Similarity)를 측정하는 기준함수(Criterion Function)를 이용하고, 생성된 클러스터의 품질을 평가하여 경계범주를 자동으로 탐색하여 경계범주집합 X를 구성한다.

・최적화된 클러스터 형성

입력문서들을 S = { S_1,S_2, … , S_k }의 개별적인 클러스터로 구성하기 위해 각 문서들을 일정 기준에 따라 서로 그룹핑 한다. 그림 5는 미네소타대학에서 개발한 CLUTO-클러스터링 알고리즘의 최적화된 클러스터 구성을 위한 기준함수들의 일부이며 식 (1)과 식 (2)는 유사도 검사를 위한 함수이다[11]. 여기서 S_r 은 r 번째 클러스터, k는 생성된 클러스터의 전체 수를 나타낸다. n_r 은 r 번째 클러스터를 이루는 문서의 수로 클러스터의 크기를 의미한다.

$$\text{maximize } I\_Sim = \sum_{r=1}^{k} n_r \left(\frac{1}{n_r^2} \sum_{d_i, d_j \in S_r} sim(d_i, d_j)\right) \quad (1)$$

식 (2)에서 Cr 은 전체문서집합의 중심 값(Centroid Vector), Cr_r는 r 번째 클러스터의 중심 값을 의미한다. n_r 은 r 번째 클러스터를 이루는 문서의 수이다. 즉, 식 (2)에서는 클러스터간의 유사도를 최소화시키는 것이 관건이다.

$$\text{minimize } E\_Sim = \sum_{r=1}^{k} n_r sim(Cr_r, Cr) \quad (2)$$

내부 유사도는 각 클러스터들에 속하는 문서들의 유사도를 나타내며 외부 유사도는 서로 다른 클러스터간의 유사도를 나타낸다. 따라서 이상적인 클러스터는 내부 유사도는 최대가 되고 각 클러스터간의 외부 유사도는 최소가 되도록 형성된다. <표 1>에서 sim(v, u)은 문서 v와 u의 자질벡터를 입력으로 하여 계산되는 유사도 측정값이고, 이러한 함수를 이용하여 최적화 된 k 개의 클러스터를 구성하게 된다. 생성되는 클러스터의 수가 목표범주의 개수보다 작다면 범주간 “is-a”의 관계가 내포되어 있음을 나타내며 분류체계에 대한 재설정이 요구된다.

〈표 1〉 CLUTO알고리즘의 최적화 된 클러스터 생성을 위한 기준함수

기준함수 notation	수식
$\mathcal{I}_1$	maximize $\sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{v,u\in S_i}\mathrm{sim}(v,u)\right)$
$\mathcal{I}_2$	maximize $\sum_{i=1}^{k}\sqrt{\sum_{v,u\in S_i}\mathrm{sim}(v,u)}$
$\mathcal{E}_1$	minimize $\sum_{i=1}^{k}n_i\frac{\sum_{v\in S_i,u\in S}\mathrm{sim}(v,u)}{\sqrt{\sum_{v,u\in S_i}\mathrm{sim}(v,u)}}$
$\mathcal{G}_1$	minimize $\sum_{i=1}^{k}\frac{\sum_{v\in S_i,u\in S}\mathrm{sim}(v,u)}{\sum_{v,u\in S_i}\mathrm{sim}(v,u)}$
$\mathcal{G}_1'$	minimize $\sum_{i=1}^{k}n_i^2\frac{\sum_{v\in S_i,u\in S}\mathrm{sim}(v,u)}{\sum_{v,u\in S_i}\mathrm{sim}(v,u)}$
$\mathcal{H}_1$	maximize $\frac{\mathcal{I}_1}{\mathcal{E}_1}$
$\mathcal{H}_2$	maximize $\frac{\mathcal{I}_2}{\mathcal{E}_1}$

· 경계범주 탐색을 위한 클러스터의 품질 평가

위 기준함수의 조건에 따라 구성된 최적의 클러스터에서 경계범주가 될 공간을 탐색하기 위해 클러스터의 품질을 평가한다. 클러스터의 품질을 평가하기 위한 측정요소로써 입력문서로 사용된 학습문서들의 라벨을 이용하여 클러스터의 엔트로피(Entropy)와 순정도(Purity)를 계산한다. 클러스터의 엔트로피는 한 클러스터에 얼마나 다양한 범주의 개념이 포함되었는가를 가늠할 수 있으며, 이와 반대로 순정도는 한 클러스터가 단일 범주의 문서들로 잘 그룹핑 되었는지 추측할 수 있다.

엔트로피는 원래 물질계의 열적상태를 나타내는 물리량으로서 무질서의 정도를 나타내는 지표로 사용되었다. 내부 복잡도(Internal Complexity)의 개념으로 쓰이는 엔트로피는 일반적으로 그 값이 낮을수록 내부 자질들의 유사도가 높은 좋은 클러스터임을 나타내기 때문에 클러스터 품질평가에 유용하게 적용된다. 클러스터의 크기가 n_r 인 임의의 클러스터 S_r 에 대한 일반적인 엔트로피는 식 (3)과 같이 계산될 수 있다. 여기서 q는 범주의 개수를 나타내며, n_{ri} 는 r번째 클러스터로 그룹핑 된 문서들 중 범주 i 에 해당하는 문서들의 개수를 의미한다. 이 때, 전체 클러스터 S에 대한 엔트로피는 개별 클러스터들의 크기에 가중치를 두어 식 (4)와 같이 합산된다.

$$Entropy(S_r)=-\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r} \tag{3}$$

$$Entropy(S)=\sum_{r=1}^{k}\frac{n_r}{n}Entropy(S_r) \tag{4}$$

엔트로피 개념과 함께 클러스터의 품질을 평가하는 척도로서 클러스터의 순정도를 이용한다. 본래의 범주에 속하는 문서들이 서로 그룹핑 되어 한 클러스터를 이루었는지 관찰하는 의미로서 식 (5)로 표현된다. 클러스터 S_r 가 다양한 범주의 문서들로 구성된 경우, 전체 문서개수인 n_r 대해 가장 많은 문서가 포함됨 특정범주 범주의 문서 개수에 대한 비율을 나타내며, 전체 클러스터에 대해 식 (6) 과 같이 합산된다.

$$Purity(S_r)=\frac{1}{n_r}\max_i(n_r^i) \tag{5}$$

$$Purity(S)=\sum_{r=1}^{k}\frac{n_r}{n}purity(S_r) \tag{6}$$

위 내용에 의해 엔트로피가 작고 순정도가 높을수록 좋은 클러스터임을 의미한다. 이상적으로 완벽한한 클러스터는 단일 범주의 문서들이 단일 클러스터를 생성하는 형태를 띄며 이때의 엔트로피는 0 이고 순정도는 1 이다. 반면, 임의의 클러스터 S_r 이 전체 q 개의 범주들에 대한 학습문서들로 고르게 구성되었다면 1에 가까운 높은 엔트로피와 1/q에 근사하는 낮은 순정도를 갖게 된다. 그러므로 경계범주의 역할을 할 수 있는 공간 탐색을 위해 클러스터의 복잡도를 계산할 때 다음과 같은 간단한 식을 이용할 수 있다.

$$U(S_r)=\frac{Entropy(S_r)}{Purity(S_r)} \tag{7}$$

(그림 5)는 TREC데이터의 스포츠 관련문서들을 CLUTO-클러스터링 시스템의 최적화 된 기준함수에 따라 구성한 최적의 클러스터의 행렬을 나타낸다. 0번 클러스터에서 9번 클러스터에 이르기까지 각 범주들의 학습문서가 한 개 이상 그룹핑 되어 있다. 그러나 8번과 9번을 제외한 클러스터들은 모두 높은 순정도를 갖게 되므로 경계범주영역에서 제외된다.

본 연구에서 찾아내고자 하는 경계영역은 높은 엔트로피와 낮은 순정도를 보이는 8번과 9번과 같은 클러스터이다. 경계영역은 두 개 이상의 목표범주에 대한 학습문서가 그룹핑 되어 목표항목 사이에 새로운 공간을 형성하는 클러스터들로 구성한다. 이 영역은 목표범주집합 내에 존재하는 매우 불확실한 공간이다. 이 때 경계영역을 위한 학습문서들은 해당 클러스터에 그룹핑 된 문서들로 구성한다.

CLUTO 알고리즘의 모든 내부 유사도와 외부 유사도의 조합으로 클러스터를 수행하고 엔트로피와 순정도 를 평가하여 본 결과, 내부 유사도의 기준함수로 I2를 적용한 H2 에서 가장 좋은 클러스터를 생성하는 것으로 보고되어 있다[11].

본 논문에서는 CLUTO-클러스터링 알고리즘의 기준함수를 이용하여 최적화된 클러스터를 생성하고, 엔트로피와 순정도값을 이용하여 목표범주간의 모호한 경계를 탐색한다.

<표 2>는 경계범주집합의 자동구성을 위한 알고리즘의 의사코드를 나타낸다. CLUTO 클러스터링 알고리즘을 이용한 경계범주 구성방식은 다음과 같다.

· 목표범주들에 대한 학습문서 Tr{C}를 입력으로 한다.

· 모든 문서를 별도의 클러스터로 초기화하여, 클러스터 내의 내부 유사도를 최대화 하고 외부 유사도는 최소화

clustering: [I2=2.29e+03] [8580 of 8580], Entropy: 0.155, Purity: 0.885

cid	Size	ISim	ISdev	ESim	ESdev	Entpy	Purty	base	bask	foot	hock	boxi	bicy	golf
0	359	+0.168	+0.050	+0.020	+0.005	0.010	0.997	0	358	1	0	0	0	0
1	629	+0.106	+0.041	+0.022	+0.007	0.006	0.998	628	0	1	0	0	0	0
2	795	+0.102	+0.036	+0.018	+0.006	0.020	0.995	1	1	1	791	0	0	1
3	762	+0.099	+0.034	+0.021	+0.006	0.010	0.997	0	1	760	0	0	0	1
4	482	+0.098	+0.045	+0.022	+0.009	0.015	0.996	0	480	1	1	0	0	0
5	844	+0.095	+0.035	+0.023	+0.007	0.023	0.993	838	0	5	0	1	0	0
6	1724	+0.059	+0.026	+0.022	+0.007	0.016	0.996	1717	3	3	1	0	0	0
7	1175	+0.051	+0.015	+0.021	+0.006	0.024	0.992	8	1	1166	0	0	0	0
8	853	+0.043	+0.015	+0.019	+0.006	0.461	0.619	46	528	265	8	0	0	6
9	957	+0.032	+0.012	+0.015	+0.006	0.862	0.343	174	38	143	8	121	145	328

(그림 5) 엔트로피(entropy)와 순정도(purity)를 이용한 경계범주 설정 예

하도록 구성된 최적의 K개의 클러스터 $S=\{ S_1,\cdots,S_k \}$를 얻는다.

- 최적화 된 클러스터 $S=\{ S_1,\cdots,S_k \}$에 대하여 각각의 클러스터 S_r 내의 내부 복잡도 $U(S_r)$을 계산하여, 임계값 $u = 0.5$ 이상인 클러스터 S_r 을 경계범주 x_j 로 구성한다.

서로 다른 목표범주를 위한 학습문서들의 자질과 자질 값이 유사하여 서로 균등한 사이즈로 그룹핑 된 경우 높은 엔트로피와 낮은 순정도를 갖게 되며 내부 복잡도 값은 대체로 1 이상의 값을 갖는다. 반면, 식 (6)의 순정도 계산에 있어서 가장 많은 문서가 할당된 단일 범주에 대한 문서개수로 계산되므로, (그림 5)의 3번 클러스터와 같이 범주별 문서가 비균등하게 포함된 경우에는 0.5 이상의 상대적으로 높은 순정도를 갖는다. 전체 7개의 범주 중 5개의 범주에 대한 문서가 포함되어있으나 대부분 특정 범주의 문서로 이루어져 있기 때문에 낮은 엔트로피와 높은 순정도를 갖게 되는 것이다. 이러한 경우 대부분 내부복잡도 $U(S_r)$은 0.3 미만의 값을 갖게 된다. 따라서 본 연구에서는 내부 복잡도를 위한 임계값(Threshold)의 초기값은 오차값을 포함하여 0.5로 설정하였다.

<표 2>의 과정을 통해 기존의 목표범주와 각 범주에 대한 학습 문서간의 상관도 분석으로 복잡도가 높은 모호한 클러스터를 탐색한 후 이들을 경계범주로 구성할 수 있으며, 확장된 분류체계는 다음과 같이 정의된다.

[정의 1] 확장된 분류체계(Expanded Target Category Set)

확장된 분류체계는 기존의 목표범주에 경계범주를 포함시킨 것으로써 확장된 분류체계에 의한 목표범주 집합 $EC= C \cup X$ 로 구성된다. 이 때, 전체 학습문서집합 $Tr\{EC\} = Tr\{C\} \cup Tr\{X\}$로 구성한다.

본 연구에서 제안한 확장된 분류체계의 주요내용을 정리하면 다음과 같다. 미리 주어진 목표범주집합 C는 분류목표와 의도를 반영하는 궁극적인 분류체계로서의 의미를 갖는다. 그러나 목표범주의 개념적인 상관도가 높은 경우 경계공간이 뚜렷하지 않으며 각 목표범주들과의 연관성이 높은 문서들은 오류가능성이 높다. 경계범주집합 X는 목표범주들을 대표하는 자질들이 겹쳐져 있는 공간을 의미하며, 확실히 어느 범주로 구분해야 할지 불확실한 문서들은 각 목표항목보다 이 공간에 분류될 가능성이 높다. 확장된 분류체계는 불확실한 문서들을 별도로 처리 할 수 있도록 하는 분석체계의 기틀을 설계하는 것이며, 경계항목집합 X를 정의한 이유가 바로 여기에 있다.

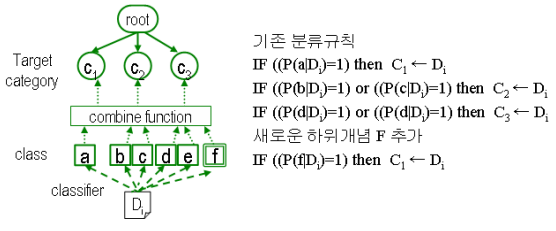
3.3 상향식(Bottom-Up) 계층분류를 위한 학습 집합

3.3.1 상향식 계층분류 방법의 개요

앞 절에서 설명한 바 있듯이 계층분류방법은 분류방향에 따라 하향식과 상향식으로 나뉘며 평탄화 된 분류방법보다는 많은 비용이 들지만 분류의 정확성은 높은 편이다. 정확도 측면에서는 하향식 방법이 좀 더 좋은 결과를 보이고 있으나 각 노드마다 분류기를 구성해야 하므로 매우 높은 비용이 든다는 단점이 있다. 반면, 상향식 방법은 하향식 방법에서와 같이 높은 비용이 들지 않으면서 일반적인 평탄화된 분류방법보다 좋은 결과를 보인다.

상향식 분류방식에서 학습문서들은 원래 범주보다는 작은 의미의 항목(Class)들로 구분해야 한다. 항목들은 분류 문제에 따라 '학년-반' 같은 계층처럼 동등 레벨)일 수도 있고 '영화-장르' 같은 하위 레벨이 될 수도 있다. 이러한 범주의 계층적 의미를 이용하여 상향식 계층분류방식으로 항목들의 앙상블(Ensemble)로 범주를 구성하는 것은 분류기에 보다 풍부한 클래스 정보를 제공하여 학습문서의 다양한 양상을 학습하게 하며 보다 유연하고 정교한 분류전략을 선택할 수 있게 함으로써 분류 성능향상에 도움을 준다.

(그림 6)은 상향식 분류를 수행하는 가장 단순한 방법으로써 상향식 분류방법의 유연성을 나타낸다. 기존의 목표범주 c_i 에 새로운 항목 f가 추가될 경우, 학습문서들로 F에 대한 분류모델을 생성해야 한다. 이 때, 하향식 방법처럼 분



(그림 6) 상향식 분류방식

〈표 2〉 경계범주집합의 자동구성을 위한 알고리즘의 의사코드

```
Input : 주어진 목표범주집합에 대한 학습문서 Tr{C},  C = {c1, c2, … , cn}
Output : 경계범주집합 X= {x1, x2,… , xm}
//Method :
//A:  Cluto-클러스터링 알고리즘을 이용한 입력문서들간의 연관도 분석
//A-1. 목표범주집합에 대한 모든 입력문서 Tr{C}에 대한 자질벡터 V={v1, …,v N} 생성,  N= |C|
//A-2. 각 문서의 개별 자질벡터로 클러스터 초기화
//A-3. 기준함수(criterion function) maximize  I_SIM(Internal Similarity) = I2 / E_SIM(External Similarity) = E1 를
       만족하는 최적의 K개의 클러스터 S={S1,…,Sk} 생성

//Expanded Target Category Strategy
B:  최적의 클러스터 S={S1, …,Sk}에 대하여 내부 복잡도가 높은 클러스터 탐색
B-1. 초기화 :  X= {x1, x2,.., xm} 구성을 위한 인덱스  j = 1 ,내부 복잡도를 위한 임계값 u = 0.5
B-2.  FOR  r  = 1  to  k,  j=1  {          // S={S1,…,Sk }
        IF ( U(Sr) >  u) THEN {
              Tr{xj} ← documents in Sr        // Construction xj  using  Sr
              j++
        }
      }
```

류기 전체를 학습할 필요 없이 F에 대한 개별 분류모델을 만들어 기존의 분류모델에 추가하고 병합함수의 분류규칙, 즉, 범주별 할당규칙을 갱신한다.

3.3.2 상향식 계층분류를 위한 범주와 학습문서집합의 구성

앞 절에서는 경계영역에서 발생하는 오분류 문제를 다루기 위해 경계범주를 포함시킨 형태로 분류체계를 확장시켜 제안하였다. 이에 보다 정확한 분류결과를 얻기 위한 방법으로서 각 범주들의 하위개념을 구성하여 상향식 계층분류를 수행하고 하위범주를 목표범주로 통합하기 위해 병합함수를 설계한다. 한 개의 범주는 여러 개의 더 작은 항목으로 분할 될 수 있다. 주어진 분류 문제와 목표범주가 나타내는 개념의 크기에 따라서 목표범주들은 한 개 이상의 세부항목(Subclass)으로 분할한다.

[정의 2] 세부항목집합(Subcategory Set)

임의의 목표범주 c_i 에 대한 세부범주는 $sc_i = \{c_{i1}, \cdots, c_{it}\}$로 구성될 수 있으며, 전체 목표범주집합에 대한 세부범주집합은 $SC = sc_1 \cup sc2 \cup \cdots \cup sc_n$ 로 나타낸다. 목표범주가 포함하는 개념의 영역에 따라 각 세부범주의 크기는 달라질 수 있고, 목표범주의 하위개념이 존재하지 않는다면 세부범주의 크기는 1 이다.

(그림 7)은 (그림 5)의 클러스터들의 계층 트리를 나타낸다. 이러한 계층구조는 각각의 클러스터를 병합할 때 발생하는 기준함수, 즉 내부 유사도와 외부 유사도의 변화량을 이용하여 생성할 수 있다. 자식 클러스터를 병합해 가는 과정에서 정보의 이익이나 손실이 발생하게 되는데, 이들의 적정한 값을 만족하며 클러스터를 생성해 간다.

<표 3>은 목표범주집합을 세부범주들로 구성하기 위한 알고리즘의 의사코드를 나타낸다. 계층분류를 위한 학습집합은 목표범주 c_i 의 세부항목집합 $Sci = \{c_{i1}, \cdots, c_{in}\}$와 c_j의

```
-------------------------------------------------------------------
Hierarchical Tree that optimizes the I2 criterion function...
-------------------------------------------------------------------
                              base bask foot hock boxi bicy golf
-------------------------------------------------------------------
14
|-----11
|     |-------------6     1717    3    3    1    0    0    0
|     |---10
|          |---------1      628    0    1    0    0    0    0
|          |---------5      838    0    5    0    1    0    0
|-13
  |--------- 9
  |          |-------7        8    1 1166    0    0    0    0
  |          |-------3        0    1  760    0    0    0    1
  |-12
     |--------8
     |         |---0           0  358    1    0    0    0    0
     |         |---4           0  480    1    1    0    0    0
     |
     |---------------2        1    1    1  791    0    0    1
-------------------------------------------------------------------
```

(그림 7) 클러스터들의 계층 트리

세부항목집합 $Sc_j = \{c_{j1}, \cdots, c_{jk}\}$들로 구성된다. 세부항목집합 개수인 n과 k는 각 범주가 나타내는 문제의 영역 및 크기에 따라 클러스터링 분석에 의해 결정된다.

4. 실험 및 결과

제안방법의 효율성과 정확성을 검증하기 위해 실험한 방법 및 결과를 정리한다. 실험 대상은 유즈넷(UseNet)의 뉴스그룹 중 다계층 구조가 내포되어있는 문서들이다. 유즈넷에는 수많은 뉴스그룹들이 다양하게 형성되어 있는데 주제들의 범주 그 자체가 계층구조를 자연스럽게 형성하고 있다. 가장 넓은 개념으로 컴퓨터(comp), 과학(sci), 토의(talk), 여가(rec) 등이 있고 하위로 갈수록 더 구체적인 개념으로 구성된다. 이곳의 문서들은 전체 약 2만 건의 문서 중 540건의 문서가 두개이상의 범주에 교차할당되어 있다. 본 실험에서는 서로 혼돈될 수 있는 범주 내에서의 분류정확도를 검증하기 위하여 컴퓨터관련 문서들(comp이하)을 주 대상으

〈표 3〉 계층분류(bottom-up)을 위한 목표범주집합 분할 알고리즘

```
Input : 주어진 목표범주집합에 대한 학습문서 Tr{C},  C = {c1,c2, …, cn}
Output : 세부범주집합 SC = sc1 ∪ sc2 ∪ … ∪ scn ,  sc1 = {ci1,ci2, … , cit }

//최적의 클러스터 S={S1, …,Sk} 에 대한 계층구조 생성
//Method :  각 클러스터의 병합정보(Merging cost) 로 형성된 계층 트리 분석
//Measure : 두 개의 자식(child) 클러스터 Sa 와 Sb 를 병합할 때 예상되는 기준함수의 변화량
C-1. 최적화 된 K 개의 클러스터에서 경계범주로 구성된 클러스터 제거,  S={S1, …,Sh},  h ≤ k
C-2. 최적의 클러스터 S={S1, …,Sh}에 대한 계층구조 생성
C-3. 목표범주집합 C= {c1, c2, …, cn}에서 목표범주 ci 에 대한 클러스터 탐색 및 세부범주구성
    FOR   i = 1 to n   {
        j=1
       // 목표범주 ci 의 단일범주로 이루어진 클러스터 Sr 의 상위 parent  node 탐색
       Parent_node = Search_ node(ci)          //  Parent_node 제거하여 child  node 분리
       Remove_node(Parent_node)          // child_node들로 세부범주(subclass)구성 및 학습문서설정
       While(child_node){
            Tr(scij) ← documents in child_node with clusters contained documents of ci
            j ++;
        }
    }
```

로 하여 목표범주 C={comp.graphics,comp.sys.ibm.pc.hardware, comp.sys.mac.hardware,comp.windows.x}로 정하였다.

4.1 실험조건 및 계획

기존방법과 제안방법의 비교를 위한 실험 조건을 <표 3>과 같이 정리하였다. 기존방법과 제안방법의 정확도를 비교하기 위하여 자동으로 탐색된 경계범주의 포함여부에 대한 비교와 함께 평탄화 방법과 계층분류에 대한 비교실험을 수행한다. 또한 안정성을 비교하기 위하여 각 목표범주의 자질들을 혼돈시키는 의미에서 서로 다른 범주들에 속하는 문서들을 오류문서로 사용하여 의도적인 교차 학습을 수행하였다.

경계범주를 포함시키지 않은 경우 계층분류를 위한 실험에서는 하향식 계층분류방법을 적용하였고, 분류기 구성이 복잡하지만 성능이 우수한 SVM분류기를 이용하였다. 경계범주를 포함시켰을 때의 계층분류는 기존방법과의 비교를 위해 하향식 분류방법보다 정확도가 낮은 상향식 계층분류를 수행하였다. 분류도구로 학습오류(Noisy)에 강하고 복잡도가 적다고 알려진 Naïve Bayesian(NB) 분류기를 적용하였다.

실험계획은 다음과 같다.

1) 실험 실험문서에 대해 기존 법인 기본 분류체계를 이용하여 학습한 후, 평탄 및 계층 분류를 수행한다.
2) 제안방법 자동 탐색된 경계범주를 포함한 확장된 분류체계 하에서 학습한 후, 평탄 및 계층 분류를 수행한다.
3) 학습과정에 의도적인 오류를 발행시킨 후 1)과 2)를 수행시켜 결과를 비교한다.

〈표 4〉 기존/제안방법의 정확성 비교를 위한 실험조건

		분류체계	분류방법	실험조건	
	실험	경계범주(X)	Flat/Hierarchy	분류 알고리즘	학습 오류
기존 방법	E_00	X	Flat	SVM	X
	E_01	X	Hierarchy	SVM	X
	E_02	X	Flat	SVM	O
	E_03	X	Hierarchy	SVM	O
제안 방법	E_10	O	Flat	NB	X
	E_11	O	Hierarchy	NB	X
	E_12	O	Flat	NB	O
	E_13	O	Hierarchy	NB	O

〈표 5〉 자동 탐색알고리즘에 의해 생성된 확장된 학습체계

확장된 목표범주		계층분류(bottom-up)
목표범주(C)	경계범주(X)	하위항목(subclass)
G(comp.graphics)	X1: GI, X2: IM, X3: GIMW,	G1,G2,G3
I(comp.sys.ibm.pc. hardware)		I1,I2,I3
M(comp.sys.mac. hardware)		M1,M2
W(comp.windows.x)		W1,W2,W3

4.2 실험결과

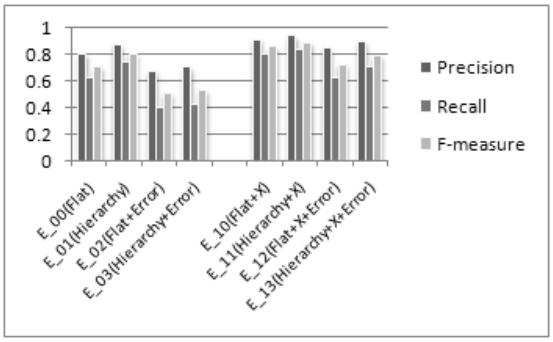
검증을 위해 학습에 사용되지 않은 문서로서 각 목표범주마다 50 개의 문서들로 구성하여 200개의 문서들로 결과를 확인하였다. 이때 정확도(Accuracy)는 F-measure값을 이용하였다.

$$F-measure = \frac{2\times\left(\frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i+\sum_{i=1}^{m} FP_i}\times\frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i+\sum_{i=1}^{m} FN_i}\right)}{\frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i+\sum_{i=1}^{m} FP_i}+\frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i+\sum_{i=1}^{m} FN_i}}$$

<표 6>은 <표 4>에서 정리한 실험조건에 따른 결과로서 제안방법의 우수성과 타당성을 나타내고 있다. 평탄분류를 수행한 E_00과 E_10, 계층분류를 수행한 E_01과 E_11에서의 예측력이 각각 0.16, 0.05만큼 증가하여 모두 경계범주를 구성하여 학습한 제안방법의 결과가 더 우수하였다.

〈표 6〉 기존방법과 제안방법의 실험결과

	실험	Precision	Recall	F-measure
기존 방법	E_00	0.806	0.632	0.708
	E_01	0.881	0.750	0.810
	E_02	0.679	0.407	0.508
	E_03	0.706	0.423	0.529
제안 방법	E_10	0.916	0.811	0.860
	E_11	0.941	0.846	0.890
	E_12	0.852	0.632	0.725
	E_13	0.897	0.714	0.795



(그림 8) 기존방법과 제안방법의 성능 비교

전체 10%의 오류를 추가한 상황인 E_02와 E_12, E_03와 E_13의 예측력은 최고 약 60%로 떨어졌으나 제안방법에서는 약 10%정도의 차이를 보여 의도적인 학습오류를 발생시켜 실험한 결과에서 상대적으로 안정된 경과를 보인다. NB 알고리즘은 SVM과 비교하여 학습표본오류에 덜 영향을 받는 편이고 SVM은 오류에 민감한 것으로 알려져 있다[2, 12]. 이는 확률모델과 벡터모델에 의한 문서표현방법의 한계를 나타내고 있는 결과이기도 하다. 전반적으로 분류작업에 소요되는 시간적 비용이 매우 크지만 다른 분류기보다 우수하다고 알려진 SVM 보다 단순하고 용이한 NB을 제안방법과 함께 이용했을 때의 정확율이 더 향상되었다.

5. 결 론

본 논문은 자동화 된 분류시스템의 성능향상을 위한 것이며, 향상시키려는 성능은 대상의 정확성(Accuracy)뿐 만 아니라 학습과정에서 일어날 수 있는 여러 상황으로부터의 안정성(Stability)이다.

본 연구의 결과는 다음과 같이 요약할 수 있다. 확장된 분류체계를 위하여, 경계범주를 자동탐색 할 수 있는 방법을 제안하였다. 복잡하고 불확실성이 높은 문서들의 자질벡터들은 목표항목을 구분하는 경계와 가까이 위치하기 때문에 정확히 분류되지 못하는 특성이 있었다. 따라서 분류경계에 인접한 부근을 새로운 항목으로 인식시켜서 학습체계를 확장시키는 방법이었다. 기존연구에서는 수작업으로 매뉴얼하게 구성하였으나 본 연구에서는 클러스터링 알고리즘과 엔트로피를 이용하여 자동으로 탐색하고 있다. 또한 제안방법은 상향식 계층분류를 위한 구성에도 쉽게 적용될 수 있다. 임의의 목표범주는 다양한 주제를 포괄하여 여러 계층으로 구성될 수 있으므로 범주의 영역과 범위를 분할하도록 한다. 이는 목표 범주들 간의 개념이 유사하여 오분류율이 높다면, 해당 범주를 구성하는 계층적인 개념을 더 구체적으로 나누어서 적절한 범주를 찾도록 하는 것이다. 본 논문의 제안방법은 자동문서분류시스템의 성능향상을 위해서 학습과정과 분류모델에 집중되어있는 시선을 기계학습기법을 이용한 전통적인 문서분류시스템의 전체 프로세스로 옮겨놓았다는 점에서 의의가 있다. 기존의 분류성능을 향상을 위한 연구들은 대부분 분류알고리즘을 개선시키는데 주력해왔으며, 그 범위는 전통적인 분류절차 하에서 통계적인 방법을 응용하는 것에 제한되어 있는 편이었다. 본 논문에서는 오분류율이 높은 문서들에 대한 궁극적인 오분류의 원인을 찾아 해결방법을 모색하면서 제안방법으로 접근해 나가는 방식을 취하고 있다. 또한 정보검색 시스템에서 뿐 아니라 정확한 분류가 필요한 여러 분야에 쉽게 적용될 수 있도록 설계되었다. 확장된 분류체계에 의한 학습방법은 그 대상의 형태에 의존하지 않고 적용할 수 있으며, 스팸문서와 같이 일반적인 패턴으로 위장된 공격패턴에 대한 이상탐지 및 오용탐지분석 문제와 기계학습으로 분석하기에 복잡한 이미지 분류에도 활용할 수 있다.

향후 연구로는, 분류가 응용되는 문제에 따라 분류체계와 결정조건의 관계를 서로 연결시켜줄 수 있는 보다 강화된 후속 처리에 대한 연구가 필요할 것이다.

참 고 문 헌

[1] 최윤정, 박승수, "학습방법 개선과 후처리분석을 이용한 자동문서분류의 성능향상 방법," 한국정보처리학회논문지, Vol.12, No.7, pp.811-822, 2005.

[2] T., Joachims, "Text categorization with support vector machines: learning with many relevant features," In Proc. of ECML-98 pp.137-142, 1998.

[3] Y., Yang, "Expert Network:Effective and Efficient Learning form Human Decisions in Text Categorization and Retrieval," in Proc. of 17th ACM, pp.13-22, 1994.

[4] Y., Yang, "An Evaluation of Statistical Approaches to Text Categorization," Journal of Information Retrieval, Vol.1, No.1, pp.67-88, 1999.

[5] M., Ruiz, P.Srinivasan, "Hierarchical text categorization using neural networks," Information Retrieval, Vol.5, No.1, pp.87-

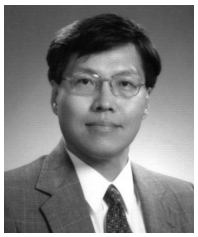
118, 2002.
[6] O., Dekel, J., Keshet, “Large margin hierarchical classification.,” In Proc. of the ICML’04, pp.209- 216, 2004.
[7] D. Koller, S., Tong, “Active learning for parameter estimation in Bayesian networks,” In Neural Information Processing Systems, 2001.
[8] D., Cohn, “Less is more: Active learning with support vector machines,” In Proc.17th International Conference on Machine Learning, pp.839-846, 2000.
[9] D., David, J., Catlett, “Heterogeneous Uncertainty Sampling for Supervised Learning,” In Proc. of the 11th ICML, pp. 148-156, 1994.
[10] D., Raj,et.al, “Boosting for document routing,” In Proc. of the AGM CIKM, pp.70-77, 2000.
[11] CLUTO-Clustering Algorithms, http://glaros.dtc.umn.edu/gkhome/views/cluto
[12] C., Cortes, V., Vapnik, “Supprot Vector Network,” Machine Learning, Vol.20, pp.273-297, 1995.

최 윤 정

e-mail : cris@seoil.ac.kr
1997년 서원대학교 전자계산학과 졸업(학사)
2001년 이화여자대학교 컴퓨터학과(공학석사)
2007년 이화여자대학교 컴퓨터학과(공학박사)
2007년～2008년 서강대학교 컴퓨터학과 Post. Doc
2009년～현 재 서일대학 정보통신과 강의전담교수
관심분야 : 인공지능, 기계학습, 온톨로지, 상황정보인식, 유비쿼터스 센서네트워크

지 정 규

e-mail: jgjee@nrf.go.kr
1987년 서울산업대학교 전자계산학과(학사)
1989년 숭실대학교 전자계산학과(공학박사)
1998년 숭실대학교 전자계산학과(공학박사)
1981년～1996년 (주)삼호, 서울특별시 시설관리공단 전산실
1996년～현 재 한국연구재단 연구기반조성단장
관심분야 : 멀티미디어, 데이터베이스, 영상처리



박 승 수

e-mail : sspark@ewha.ac.kr
1974년 서울대학교 수학과(공학사)
1976년 한국과학기술원 수학과 석사
1988년 미국 텍사스 오스틴 대학 전산학(박사)
1988～1991년 미국 켄사스대학 컴퓨터학과 조교수
1991～현 재 이화여자대학교 컴퓨터공학과 부교수
관심분야 : 인공지능, 온톨로지, 시멘틱웹, 상황인식, 유비쿼터스 컴퓨팅