DOI: 10.3745/KIPSTB.2010.17B.2.139

상향식 계층분류의 최적화 된 병합을 위한 후처리분석과 피드백 알고리즘

최 윤 정[†] · 박 승 수[††]

요 약

본 논문은 자동화된 분류시스템의 성능향상을 위한 것으로 오분류율이 높은 불확실성이 강한 문서들의 범주결정방식을 개선하기 위한 후처리분석 방법과 피드백 알고리즘을 제안한다. 전통적인 분류시스템에서 분류의 정확성을 결정하는 요인으로 학습방법과 분류모델, 그리고 데이터의 특성을 들 수 있다. 특성들이 일부 공유되어 있거나 다의적인 특성들이 풍부한 문서들의 분류문제는 정형화된 데이터들에서 보다 심화된 분석과정이 요구된다. 특히 단순히 최상위 항목으로 지정하는 기존의 결정방법이 분류의 정확도를 저하시키는 직접적인 요인이 되므로 학습방법의 개선과 함께 분류모델을 적용한 이후의 결과 값인 순위정보 리스트의 관계를 분석하는 작업이 필요하다. 본 연구에서는 경계범주의 자동탐색기법으로 확장된 학습체계를 제안한 이전 연구의 후속작업으로써, 최종 범주를 결정하기까지의 후처리분석 방법과 이전의 학습단계로 피드백하여 신뢰성을 높일 수 있는 알고리즘을 제안하고 있다. 실험결과에서는 제안된 범주결정방식을 적용한 후 1회의 피드백을 수행하였을 때의 결과들을 단계적이고 종합적으로 분석함으로써 본 연구의 타당성과 정확성을 보인다.

키워드 : 특성추출, 메타 분류, 계층분류, 강화학습, 후처리분석, 피드백

Reinforcement Post-Processing and Feedback Algorithm for Optimal Combination in Bottom-Up Hierarchical Classification

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

ABSTRACT

This paper shows a reinforcement post-processing method and feedback algorithm for improvement of assigning method in classification. Especially, we focused on complex documents that are generally considered to be hard to classify. A basis factors in traditional classification system are training methodology, classification models and features of documents. The classification problem of the documents containing shared features and multiple meanings, should be deeply mined or analyzed than general formatted data. To address the problems of these document, we proposed a method to expand classification scheme using decision boundary detected automatically in our previous studies. The assigning method that a document simply decides to the top ranked category, is a main factor that we focus on. In this paper, we propose a post-processing method and feedback algorithm to analyze the relevance of ranked list. In experiments, we applied our post-processing method and one time feedback algorithm to complex documents. The experimental results show that our system does not need to change the classification algorithm itself to improve the accuracy and flexibility.

Keywords : Feature Extraction, Meta-Classification, Hierarchical Classification, Reinforcement Learning, Post-Processing, Feedback

1. 서 론 [1)]

자동 문서분류란 문서의 내용에 기반하여 미리 정의된 범주에 문서를 자동으로 할당하는 기법과 관련된 연구 분야로서, 여러 사례에 학습시킨 후 새로운 질의에 대해서도 적절한 답을 찾아낼 수 있도록 하는 기계학습과 정보검색(information retrieval)기법을 포함한다. 기계학습 관점에서 보면 문서분류시스템은 주어진 문서와 대응하는 분류 값의 쌍인 {D_i, category value$_i$}들의 집합을 가지고 새로운 문서에 대한 범주를 찾아내는 감독학습(supervised learning)의 문제이다[1-3]. 실제로 문서 D_i 는 0 이상의 범주 값을 가질 수 있으나 분류시스템에서는 어떠한 범주와도 관련이 없더

† 정 회 원 : 서일대학 정보통신과 강의전담 교수
†† 정 회 원 : 이화여자대학교 컴퓨터공학과 부교수
논문접수 : 2009년 12월 16일
수 정 일 : 1차 2010년 1월 20일
심사완료 : 2010년 1월 20일

라도 유사도나 확률 값에 따라 최소한 한 개 이상의 범주로 할당되게 한다.

전문가는 지식 창고(knowledge source)의 역할을 하며 학습과정은 전문가의 지식 축적 과정을 모방한다. 현재 활성화 된 지식에 오류가 있으면 분류과정 전체에 영향을 미치고 오분류로 직결된다. 불완전(incomplete)하고 부정확(incorrect)한 정보에 의한 결정을 신뢰하기 힘든 이유에서 이다. 불완전성에 의한 오류는 정보가 상실되었거나 근거가 부족하여 정확한 결론을 내릴 수 없을 때 발생하며 조건(condition)과 실행(action)에 관한 상호관계가 확실히 설정되어 있지 않아서 A라는 조건이 참일 때 B가 성립된다는 규칙이 확실히 보장되지 않는다. 부정확성에 의한 오류는 지식의 일부가 틀린 경우에 발생하며 논리적인 오류나 전문지식의 부족으로 말미암아 잘못된 가설을 세워 추론하는 경우가 이에 해당한다. 이와 같은 경우 인간은 자신의 지식의 한계를 인지하여 모른다고 생각되는 부분에서는 답을 유보한다. 자신의 결론이 경계선상에 있다면 어떤 불확실성(uncertainty)을 부여하는 현명함을 보일 수 있다. 그러나 전문가 시스템이나 분류시스템에서는 불확실성을 받아들이도록 계획되어있지 않으면 추론의 근거(inference evidence)가 약하더라도 계속해서 답을 출력해 낸다.

이러한 관점에서 볼 때 자동분류시스템에서 오분류를 일으키는 주요 원인은 학습과정과 범주지정방식에서 찾을 수 있다.

첫째, 학습과정에서 예상되는 오류는 불완전하고 부정확한 학습데이터로 인해 명확하지 않는 분류기준이 생성되는 것이 기인한다[4, 6]. 실제로 전문가가 개입하는 학습과정에서도 오류문서가 학습문서집합에 포함되어 잘못 학습되는 경우가 발생하며, 부적절한 학습문서는 분류기준을 왜곡시키게 되므로 오분류가 발생한다. 따라서 학습과정의 문제를 인지하여 올바른 정보 즉 적합한 학습데이터를 수집하여 재학습하는 과정이 필요하다.

둘째, 궁극적인 분류목표에 따라 범주를 지정하는 방식에서 오분류의 원인을 찾을 수 있다[7, 8]. 예를 들어, 한 문서와 후보 범주간의 관련된 차이가 극명하지 않은 경우에는 기존의 단순한 지정방식에 의해 무조건적으로 할당하기보다는 따로 가려내어 재검토의 여지를 주는 것이 타당하다. 그러나 대부분의 기계학습방법을 이용한 자동분류시스템에서는 범주지정방식이 매우 단순하다는 문제점을 갖고 있으며, 이는 직접적으로 분류의 정확도를 저하시키는 요인이 되고 있다. 또한 오분류 된 문서는 전문가가 개별적으로 처리하도록 하는 소극적인 해결방법을 취함으로써 현실적으로 자동문서분류시스템의 목표인 정확성과 신뢰성, 대량의 문서에 대한 최소화된 인간개입의 효과를 기대하기 어렵다.

본 논문은 경계범주를 자동 탐색하여 학습체계의 확장방법(ETOM) 을 제안했던 이전 연구의 후속 연구로서, 최종 범주를 결정하기까지의 결정조건, 재학습 관계를 서로 연결시켜줄 수 있는 보다 강화된 후처리 분석방법과 피드백 알고리즘을 제안하고 있다[15]. 본 논문은 다음과 같이 구성된다. 2장에서는 관련연구로서 분류알고리즘의 특성, 분류기의 결합 및 상호보완방법, 범주결정과정의 대해 설명한다. 3장에서는 본 논문에서 제안한 후처리분석 방법과 피드백 알고리즘(reinforcement post-processing algorithm,이하 Rpost)에 대해 설명한다. 4장에서는 기존의 분류체계과 제안방법의 비교실험으로서 자동분류시스템의 동작과정을 보이고, 5장에서 결론을 맺는다.

2. 관련 연구

2.1 분류알고리즘

자동분류시스템의 성능에 영향을 미치는 요인으로 분류모델 즉 분류기를 설계하는 분류알고리즘을 들 수 있다. 기계학습분야에 기반을 두고 있는 분류알고리즘은 크게 규칙기반 모델과 연역적 학습모델, 검색모델로 나뉘며, 주어진 학습문서집합을 이용하여 최종적으로 분류함수 또는 분류규칙을 만들어내는 것이 이들의 역할이다[9, 10].

분류에 쓰이는 대표적인 알고리즘으로는 의사결정트리(decision tree), K-최단인접기법(k-nearest neighbor, 이하 KNN), 지지벡터기계(support vector machine,이하 SVM), 나이브베이지안(naive bayesian 이하 NB) 등이 있다. 각 모델에 기인하는 이들 알고리즘들은 각각 특성을 지닌다. 의사결정알고리즘은 SVM과 함께 학습 표본상의 학습변이(variance)가 큰 알고리즘이라고 알려져 있다. 학습 변이란 학습문서집합의 작은 변화가 분류함수에 크게 영향을 주어 분류결과가 달라질 수 있다는 의미이다. 반면 확률모델에 기반한 NB는 오류에 상당히 안정적인 알고리즘으로 알려져 있는데, 안정적이라는 의미는 학습표본의 변화에 민감하게 반응하지 않기 때문에 소수의 오류문서에 의해 결과가 변경될 가능성이 적다는 의미이다.

대체적으로 일반적인 분류문제에서는 정보검색에 기반을 둔KNN과 선형/비선형 최적화 기법이 가능한 SVM이 가장 좋은 성능을 보이고 있다고 알려져 있다[5, 11]. 그러나 이론상으로는 가장 높은 성능을 발휘하지만 현실세계에서는 시공간상의 높은 복잡도로 인해 근사화 된 알고리즘으로 구현되기 때문에 이론상의 성능에 미치지 못한다고 평가되고 있다[1, 12]

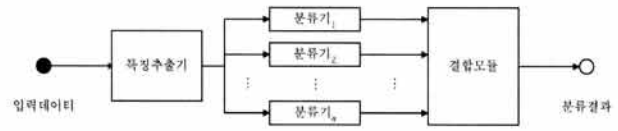
기존의 자동 문서분류의 성능향상을 위한 연구에서는 이러한 분류알고리즘들이 가진 특성을 중심으로 분류모델의 개선방법을 위주로 다루고 있는데, 분류알고리즘 자체를 개선하는 방법과 여러 분류기들의 장점을 취하여 상호 보완방법, 해당 도메인에 적합하도록 특성 가중치 함수를 새로 설계하는 등의 최적화시키는 방법이 제안되었다[4, 7, 12].

2.2 분류기의 결합 및 상호 보완 방법

분류기들을 상호 보완하는 방법으로는 여러 알고리즘을 결합하는 형태를 취하고 있는데, (그림 1)은 분류기들의 결합을 이용한 분류방법의 예를 도식화하여 보이고 있다.

다중 분류기를 이용하는 방법으로는 여러 개의 분류기

(a) 단일 분류기를 이용한 분류방법



(b) 다중 분류기의 결합에 의한 분류방법

(그림 1) 단일/다중분류기를 이용한 분류방법

(weak classifier)들을 같은 환경에서 경쟁시켜 그 중 더 나은 것을 선택하는 투표(competition & voting) 방식과, 앙상블(ensemble) 구조로 구성하여 조합 모델을 만들거나 협동 모델을 생성하는 방법이 제시되었다[7, 13, 14].

이러한 결합형 분류기를 크게 정적구조(static structure)와 동적구조(dynamic structure)로 분류된다. 정적구조 분류기는 입력벡터가 분류기의 결합모듈에 다시 사용되지 않는 경우로, 약 학습(weak learning)을 사용하여 정확도를 높이는 부스팅과 앙상블이 이에 해당한다. 동적구조 분류기는 입력벡터가 분류기 결합모듈에 사용되는 경우이다. 여러 전문가들의 신경망이 네트워크를 통해 결합된 모델(mixture- of-experts)등이 여기에 해당되며, 여러 개의 신경망들 중 어느 것으로 최종 분류결과를 도출하도록 할 것인지 결정한다[13].

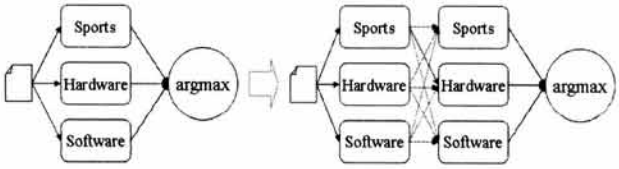
2.3 범주결정방법

단순한 순위정보에 의한 범주의 결정방식을 개선한 연구로서 범주들의 의미상 연관성을 고려하는 방법이 제안되었다. (그림 2)는 범주간의 연관성을 계산하여 분류를 결정하는 의미를 도식화하여 설명하고 있다. 이 방법에서 학습은 기존의 방법들과 동일하게 수행되지만 최종적으로 범주를 할당하는 과정에서 범주들 사이의 관계에 따라 재계산이 이루어진다.

식(1)은 (그림 2)의 방법론을 표현한 것으로 순위정보의 순위값(score)을 계산할 때 범주간의 연관도를 가중치로 하여 순위를 재조정하여 계산하고 있다.

$$Rescore(d_i,c_j) = score(d_i,c_j) + \sum_{\forall i,j\neq j} score(d_i,c_j)\cdot\frac{Sim(c_i,c_j)}{N_j} \quad (1)$$

여기서 score(d_k c_j) 는 분류알고리즘이 계산한 문서 k와 범주 j 간의 초기 연관도이며, Sim(c_i,c_j)는 범주 c_i 와 c_j



(그림 2) 범주간의 연관성(relevance)를 고려한 순위정보

의 유사도를 계산한다. N_j는 범주 C_j에 대한 정규화인수(normalization factor)이다. 이 개념을 여러 가지 분류알고리즘의 순위화 계산에 적용한 예가 있으나 범주결정방식에 있어서는 여전히 최우선 순위에 의해 범주를 결정하고 있다. 이는 범주간의 연관성을 계산하지 않았을 때 보다는 우수하지만 범주의 분류체계가 잘 갖춰져 있는 경우에는 크게 달라진 결과를 얻기 힘들 수 있다. 분류되어야 할 범주가 많고 그 범주들이 종속관계나 하위관계로 이루어져 연관성이 높은 경우에는 재계산된 순위정보에서의 최우선순위 범주가 지정하는 범주의 결정력이 얼마나 정확한지에 대해 실험해 볼 필요가 있다. 따라서 범주결정방식이 갖는 문제점에 대해 단순히 최대값으로 지정할 것이 아니라 순위정보들이 가진 순위와 순위 값들의 분포 값과 같은 메타 정보들을 사용하여 결과의 신뢰도를 측정하는 방법이 요구된다.

3. 강화된 후처리분석과 피드백 알고리즘

불확실하고 복잡한 문제들의 분류에 있어서 단순하게 이루어지는 범주지정방식은 시스템의 정확도와 신뢰도를 향상시킬 수 있는 요인으로 지적된다.

$$범주_i = \max_{i=1}^{n} P(c_i \mid D_i) \quad (2)$$

식(2)는 확률모델에서의 일반적인 범주결정방식을 나타낸다. 범주집합 C= {c_1,c_2, … , c_n } 에 대해 문서 D_i 를 분류할 때 최종 범주를 결정하는 과정에서 확률 값이 가장 높은 범주로 지정하고 있다. 이처럼 순위정보(ranking order)의 분석과정 없이 단순히 최상위항목으로 지정하는 것은 직접적으로 분류의 정확도를 저하시키는 요인이 된다. 이로 인해, 유사한 자질들이 서로 비슷한 빈도로 포함된 문서들은 정확히 분류되기 어렵다. 후보 범주들의 차이가 극명하지 않은 경우에는 기존방법으로 할당하기보다 따로 분류해 내어 재검토를 위한 여지를 주어 최종분류항목을 지정하는 것이 더 효과적이지만, 대부분의 기계학습방법을 이용한 대부분의 자동분류시스템에서는 분류기의 계산결과에서 최상위 수치를 갖는 범주로 지정하고 있다.

본 논문은 이전 연구에서 확장된 분류체계로 구성한 학습문서 집합에 계층 분류알고리즘을 적용한 이후의 작업으로, 분류 알고리즘을 적용한 이후의 문서의 분류결과에 대해 범주결정방식을 강화한 후처리분석과 피드백 알고리즘에 대해 설명한다.

3.1 강화된 후처리분석과 피드백 방법의 개요

임의의 분류기에서 문서 D_i가 범주 c_j로 할당되는 경우의 신뢰도는 식 (3)과 같이 정의된다. 아래 식에서 $P(c_j|D_i)$는 문서 D_i가 범주 c_j 에 속할 확률을 의미한다.

$$P(c_1|D_i) + P(c_2|D_i) + \cdots + P(c_j|D_i)$$

$$= 1\ (0 \le P(C_j|D_i) \le 1) \quad (3)$$

[정의 1] 문서 D_i의 후보항목리스트(the List of Candidate Category) 후보항목리스트 $L_i = [l_{i1}, l_{i2}, \cdots, l_{ik}]$는 임의의 입력문서 D_i의 순위화 된 분류결과를 나타낸다, l_{ij}는 각 범주에 대해 주어진 학습문서집합과 입력문서 D_i의 유사도 값인 <category : normalized similarity score>로 이루어진다. 이 때, 문서 D_i 에 대해 j 번째로 유사한 순위의 범주와 해당 유사도 값은 l_{ij}.category, l_{ij}.score로 표기할 수 있다. 후보항목리스트 L_i 에서 최상위후보는 간략히 L_i^1으로 표기한다.

본 논문에서 정의한 문서 D_i 의 후보항목리스트는 분류기로부터 얻어진 범주간의 유사도를 정규화 한 값으로 관련도가 높은 순서로 정렬한 순위화 된 리스트를 나타낸다. 즉, $l_{i1.category}$ 는 문서와 첫번째로 가깝고 $l_{i2.category}$는 두번째로 가까운 범주이다. 순위정보의 수치값은 0과 1사이의 실수로 계산하여, 다음과 같다.

$$l_{i1.score} + l_{i2.score} + \cdots + l_{in.score} = 1 \quad (4)$$

본 논문에서 제안하는 후처리분석 과정은 위에서 얻은 분류결과를 입력으로 하며 다음의 네 단계로서 진행된다.

- -범주결정과정(단계 1): 후보항목리스트의 수치분석에 의해 분류된 결과에 대해 신뢰도가 높은 문서들의 범주를 결정한다.
- -범주결정과정(단계 2): 단계 1에서 결정되지 않은 문서들에 대해 경계범주와 각 목표항목의 관련도를 분석하여 좀 더 가까운 쪽으로 결정한다.
- -범주결정과정(단계 3): 학습문서의 분류결과인 후보항목리스트를 입력으로 한 분류결과를 얻는다. 이는 피드백과정의 입력으로 사용된다.
- -피드백과정(단계 4): 각 단계에서 결정된 결과의 변이를 분석하여 피드포워드제어와 피드백제어를 통해 범주결정과정과 재학습-재분류의 과정을 반복한다.

3.2 범주결정과정

3.2.1 단계 1 : 분류결과의 신뢰도가 높은 문서들의 범주 결정

문서 D_i를 후보항목리스트(L_i)의 최상위항목으로 결정하기 위하여 후보항목리스트의 수치정보를 이용하여 신뢰도를 측정한다. 이 과정에서는 분류결과 값으로서 일률적으로 최상위항목으로 지정했던 기존의 범주결정방식을 개선하여 적정임계값 이상을 만족하는 경우에만 최상위항목으로 결정한다.

[정의 2] 최상위항목 수치의 임계값:*min_confidence*
최상위항목으로서 가져야 할 수치값의 하한값을 의미한다.

[정의 3] 1순위-2순위간 수치값의 임계값: *diff_confidence*
1순위와 2순위 격차값의 하한값을 의미한다. 이 값은 사용자가 조절할 수 있는 파라미터로서, 최상위항목의 수치값이 *min_confidence*보다 작지만, 순위값들의 격차가 큰 문서들도 최상위범주로 범주를 결정해주기 위한 것이다.

· 단계 1에서 사용하는 신뢰도 측정요소와 역할 :
min_confidence , diff_confidence

단계 1에서는 후보항목리스트의 최상위범주로 결정하기 위한 측정요소로서 위에서 정의한 *min_confidence* 와 *diff_confidence*를 이용한다.

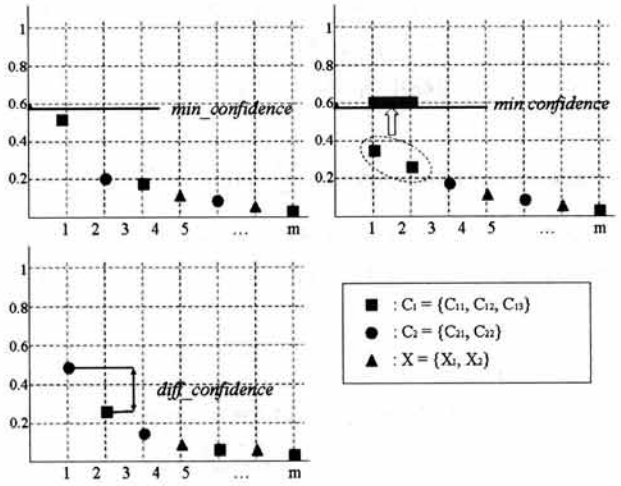
(그림 3)은 이들의 역할을 설명하고 있다. *min_confidence*는 후보항목리스트의 1위의 수치 값과 비교하게 되는 임계값이며, *diff_confidence*는 후보항목리스트의 1순위와 2순위의 수치 값의 신뢰도 설정을 위한 값으로 순위간 격차와 비교하게 되는 임계값이다. 이들은 분류결과에서 최상위범주로 지정할 때 강한 확신을 가진 문서들을 우선적으로 분류해내기 위한 것이다. 최상위항목으로 결정하기 위한 *min_confidence* 값은 전체 문서의 개수 N에 대해 $l_{i1\cdot score}$ 의 평균값을 적용하며 식 (5)와 같다. 따라서 *min_confidence* 수치를 높게 할수록 결과에 강한 확신을 가지는 소수의 문서에 대해서만 범주를 지정하게 되므로 분류결과의 신뢰도는 높아진다.

$$\mu(L^1) = \frac{1}{N}\sum_{i=1}^{N} l_{i1.score} \quad (5)$$

위의 식 (4)에서 ($l_{i1.score}$ - $l_{i2.score}$) 값 즉 순위별 편차가 클수록 분류결과를 확신할 수 있다. *diff_confidence*는 1순위와 2순위의 수치 값들이 그들의 평균과 얼마나 떨어져 있는지를 나타내는 것으로써, 다음의 식 (6)와 같이 표현된다.

$$\sigma = \sqrt{\sum_{j=1}^{2} \left| l_{ij}.\text{score} - \mu(L^1 - L^2) \right|^2} + \alpha \quad (6)$$

α : control parameter



(그림 3) 최상위후보항목으로 결정하기 위한 수치 임계값의 역할

$$\mu(L^1 - L^2) = \frac{1}{N}\sum_{i=1}^{N}(l_{i1.score} - l_{i2.score}) \quad (7)$$

위에서 정의한 *min_confidence*와 *diff_confidence*는 후보 항목리스트의 최상위항목에 대한 관련 수치값이 적어도 *min_confidence* 이상이거나, 1순위와 2순위의 편차가 적어도 *diff_confidence* 이상일 경우에만 범주를 지정하도록 하는 신뢰도 측정값으로 사용된 것이다. 여기서 사용된 α 는 조절상수로서 1순위와 2순위의 격차구간을 넓이거나 좁혀 사용할 수 있다. 신뢰도가 높다는 것은 문서분류함수가 문서를 정확하게 분류할 가능성이 크다는 것을 의미한다. 또한 이 값을 낮추어 정할 경우에는 보다 많은 문서들에 대해 범주를 지정할 수 있으나 결정된 범주에 대해서 비교적 낮은 신뢰도를 갖게 된다.

일반적으로 임계값을 사용하는 시스템에서는 최적의 결과를 얻기 위해 사용자로 하여금 반복 작업을 통해 적절한 값을 정하도록 유도한다. 사용자는 최적의 값을 찾기 위해 여러 시행착오(trial and error)를 거쳐 휴리스틱(heuristic)하게 찾아가는 것이다. 구현시스템에서는 설정된 임계값으로 범주를 결정할 때 마다 보고를 작성하여 사용자에게 제공하며, 사용자는 시스템이 제시하는 정보를 기준으로 최적의 임계값을 찾아 적용해 갈 수 있도록 하였다.

위의 임계값을 이용하여 조건에 해당하면 문서 D_i의 범주는 $L_i^1.category$로 결정된다. 그렇지 않으면 분류지정 보류항목인 'U' 로 구분한다. 단계 1에서 U로 구분된 문서의 후보리스트는 범주를 결정할 만한 확신이 매우 작은 문서로써 이는 후속 조치가 필요한 재분석대상이 되는 문서임을 나타낸 것이다. 이 문서들은 단계 1에서 경계범주로 분류된 문서와 함께 단계 2에서 분석된다.

3.2.2 단계 2 : 경계항목을 포함한 문서의 범주 결정

단계 2는 단계 1에서 결정이 보류된 문서(U)와 경계항목(X)으로 지정된 문서를 입력으로 한다. 단계 1에서 궁극적인 목표범주로 결정되지 못한 문서들을 분류해 내기 위한 과정으로 후보항목리스트 L_i의 항목 패턴의 분석이 이루어진다. 경계항목 x_k 가 목표범주 c_i 와 c_j 를 구분하는 경계선 인접영역에 위치되는 문서들로 구성된 범주일 때, 문서 D_i 가 경계항목인 x_k로 분류되었다면 목표범주 c_i 와 c_j 에 걸쳐 관련된 정도가 유사하다는 의미로 이해할 수 있다. 즉 경계항목과 좀 더 가까운 쪽의 목표항목을 찾는 과정이다.

분석을 위해 필요한 용어로서 피보트(pivot) 항목을 다음과 같이 정의한다. 피보트항목 P는 문서 D_i의 후보항목리스트 L_i내에서 순위가 가장 높은 경계항목(x_j)를 의미한다. 여기서 P는 각 목표항목 간의 거리를 측정하여 가까운 곳을 찾기 위한 기준점의 역할을 한다.

· 단계 2에서 사용하는 신뢰도 측정요소와 역할 : 피보트항목 P와 목표항목간의 관련도

피보트항목 P와 각 목표항목들 간의 관련도를 계산하여 보다 가까운 목표항목으로 범주를 지정한다. 식 (8)은 후보항목리스트 내에서 P와 목표항목 c_i의 관련도를 계산하는 함수이다.

$$Dist(P, c_n) = \sum_{j=1(l_{ij}.\text{category} \in c_n)}^{m}(RD(P, l_{ij.category}) * w_i) \quad (8)$$

$$RD(P, l_{ij.category}) = |P.rank - j| \quad (9)$$

$$w_r = \log(\sqrt{(r+\alpha)}) \quad (10)$$

α : control parameter

$$Cn \leftarrow Min\{Dist(P, Cn)\} \quad (11)$$

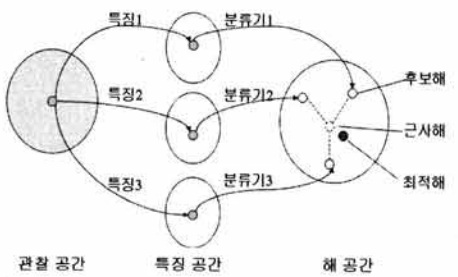
식 (9)의 $RD(P, l_{ij.category})$ 함수는 피보트항목 P의 순위와 $l_{ij.category}$ 의 순위격차를 의미하며 인접해 있을 경우는 1이다. 후보항목내에 포함된 목표항목 c_i의 세부항목(subclass) c_{ik} 들과 P와의 관련도는 목표항목 c_i로 합산시켜서 궁극적인 목표항목 c_i로 지정될 수 있도록 한다. 식 (10)에서 r은 순위를 나타내고 w_r 은 그 순위에 가중치를 부여하기 위한 함수이다. 이는 P와 인접한 목표항목의 순위에 따라 차이를 주기 위해 사용되었다. 상위후보항목은 의미상으로 가까운 관계를, 하위순위는 보다 먼 관계임을 나타낼 수 있도록 조절상수를 사용하여 로그함수를 취하였고 식 (8)에 적용된다.

3.2.3 단계 3 : 후보항목리스트의 순위 정보를 자질로 한 학습과 분류

단계 1에서는 분류결과의 확신이 높은 문서들과 낮은 문서들이 구분되었으며, 단계 2를 진행에 오는 동안 분류보류로 지정된 문서나 경계항목으로 지정된 문서의 분류가 이루어졌다. 여기서는 후보항목리스트들의 항목 수치값과 항목들의 패턴이 분석대상이 되었다. 따라서 문서의 분류결과인 후보항목리스트를 문서와 범주간의 관계를 대표하는 요약된 문서표현방식으로 활용하는 방법을 고려한다.

· 다중특징을 사용한 최적해 근사방법

입력데이터의 다중특징을 사용하는 것은 분류시스템에 풍부한 정보를 제공하여 데이터의 다양한 양상을 학습하게 함으로써 분류성능향상에 도움을 준다. (그림 4)는 다중특징을 사용하여 최적해를 구하는 기본 아이디어를 표현하고 있다. 특징추출 및 선택은 관찰공간으로부터 내재되어 있는 범주



(그림 4) 다중 특징을 사용한 최적해 근사방법

의 정보를 찾아 최적해로 향하는 최적의 경로를 얻기 위한 작업이다. 그러나 어떤 문제의 최적의 특징은 정의하기 어렵고, 정답조차 알지 못하는 상황이라면 최적해인지 여부를 가늠할 수 있는 객관적인 근거를 찾아야할 것이다. 따라서 여러 가지 특징을 동시에 사용하는 방법을 고려할 수 있다. 복수개의 특징을 고려하는 것은 다양한 후보해(candidate)를 얻게 함으로써 최적해와 가까운 근사해를 구할 수 있도록 도와준다. 여러 개의 분류알고리즘과 여러 개의 분류기의 결합모델을 설계하는 이유가 바로 여기에 있다. 본 연구에서는 최적의 해를 찾기 위한 다중특징 추출을 위해 관찰공간을 확장시키기로 한다.

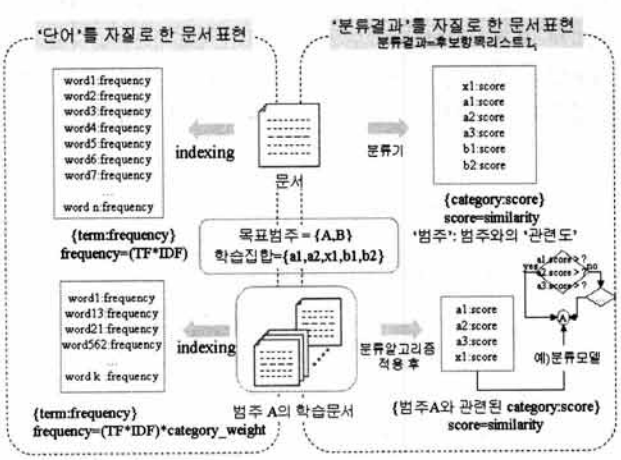
· 문서표현(document representation)을 위한 관찰공간 확장

일반적으로 문서들은 문서에 포함된 주요 단어들의 집합(bag-of-word)으로 취급되며 {term : frequency}의 쌍으로 표현된다. 일반적인 분류시스템에서의 학습이란 학습데이터에서 추출된 {feature : weight}의 특징 집합과 분류알고리즘을 이용하여 해당범주를 가장 잘 나타낼 수 있는 고유의 특징을{feature : category}들의 집합으로 모델링하는 것이다. 문서 분류시스템에서의 관찰공간은 단어를 자질로 한 특성벡터 즉, {term : category}로 구성된 문서모델과 {term : frequency}로 표현된 문서이다. 한편, 문서의 분류결과는 {category : similarity}의 집합으로써 문서와 각 범주간의 유사도 측정값들이 표현된다.

(그림 5)는 문서의 분류결과를 특성으로 갖는 축소된 자질공간에서의 문서표현방법을 나타낸다. ETOM에 의한 계층 분류결과인 후보항목리스트의 패턴은 {목표항목 또는 경계항목의 subclass : similarity}들이 순위화 된 형태로 문서의 분류결과를 예측할 수 있는 함축된 자질들의 공간으로 활용될 수 있다.

· 후보항목리스트의 순위정보를 자질로 한 다중학습(multi-training)

(그림 5)에서 나타낸 것과 같이 문서의 분류결과인 후보항목리스트는 목표범주를 구분할 수 있는 정보력을 지니므로, 이를 새로운 관찰공간으로 인식하여 문서와 목표범주에 대한 관계 표현 방법에 이용한다. 확장된 학습체계로 구성된 전체학습문서집합을 이용하여 후보항목리스트의 항목을 자질로 하고 수치정보를 자질 값으로 하는 새로운 분류기준을 만든다.



(그림 5) 다중 특징을 이용한 문서표현방법

학습패턴의 형태는 <표 1>과 같다. 문서 D_i의 후보항목리스트 L_i 에서 순위정보를 입력값, 실제 분류값을 목표값으로 한 {항목: 항목과의 관련도}의 집합으로 학습패턴을 만든다. 즉, 문서를 다른 각도에서 관찰한 특징을 이용하여 다중학습의 효과를 주는 것이다.

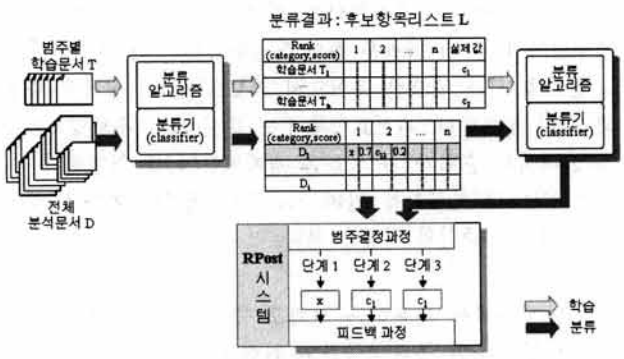
위의 학습패턴에 의한 다중학습 결과로서 후보항목의 패턴을 조건항으로 갖고 목표항목을 출력값으로 갖는 새로운 분류 규칙이 생성된다. 예를 들면, 의사결정트리 알고리즘을 적용한 경우 후보항목리스트의 각 항목과 수치값들의 패턴을 조합으로 한 규칙들이 생성된다. 이 분류규칙에는 일련의 항목별 수치정보들이 자질 값으로 사용되며 범주결정과정에서의 임계값들과 같은 기준들이 일반화되어 나타난다.

(그림 6)은 후보항목리스트의 순위정보를 자질로 한 다중학습의 개념과 과정을 보인다. 화살표는 단계별 입력 및 출력내용을 나타낸다. 결과값을 알고 있는 학습문서들의 분류결과로 분류규칙을 만들고, 이를 입력문서들에 적용하여 단계 3의 결과를 얻는다. 이는 후보항목리스트의 순위정보와 실제값으로 만들어진 분류규칙에 대한 결과로서, 이전의 단계 1과 단계 2에서 얻은 예측값과 비교되어 최적해를 찾아가기 위한 새로운 경로의 역할을 한다.

문서들을 다른 시각으로 관찰하여 분류한 단계 3의 결과는 단계 1의 결과와 함께 분류시스템의 성능을 측정할 수 있는 기반이 된다. 이 결과들은 이후 작업인 피드백 분석을 위한 입력으로 사용되며 이 값들의 변이들로 시스템의 성능을 평가한다.

〈표 1〉 후보항목리스트의 순위정보(항목과 수치값)와 실제값으로 구성한 학습 예

L_i / D_i	1		2		3		4		5		Actual Class
	category	score	category	score	category	score	category	score	category	score	
D_1	C_{21}	.98	C_{11}	.01	X_1	.01	X_2	.01	C_{12}	.00	C_2
D_2	C_{21}	.39	C_{12}	.20	X_2	.17	C_{11}	.13	X_1	.10	C_2
D_3	X_1	.29	C_{11}	.28	C_{12}	.17	C_{21}	.15	X_1	.01	C_1
D_4	C_{22}	.28	C_{21}	.23	X_2	.17	C_{11}	.16	C_{12}	.15	C_2



(그림 6) 문서와 후보항목리스트를 이용한 다중학습과 분류과정

3.3 피드백(feedback) 과정

3.3.1 피드백 방법의 개요

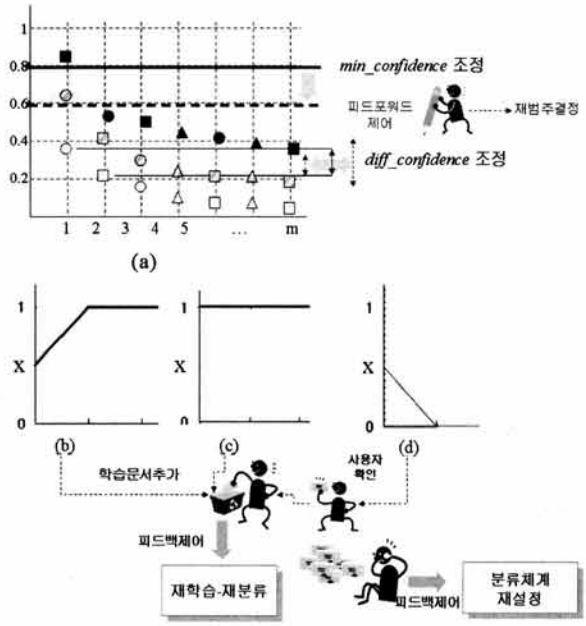
본 연구의 피드백 과정은 경계범주가 포함된 확장된 분류체계 하에서 계층분류를 수행하고, 이후 후처리 분석 알고리즘으로 범주가 결정된 문서들에서 오류의 위치와 원인을 찾아 적절한 피드백 제어방법을 제시하는 데 목적이 있다. 범주결정과정의 단계 1에서는 분류결과에 대해 적정임계값을 이용하여 이를 만족하는 경우에만 범주를 결정하였다. 해당 신뢰도를 만족하지 않는 경우에는 이를 보류하고 다음 단계에서 처리되도록 함으로써 오류율을 줄이고자 한 것이다. 단계 2에서는 지정이 보류된 문서들에 대해 경계항목과 목표항목간의 의미상의 거리차를 계산하는 방법으로 목표범주를 결정하였다. 단계 3에서는 문서의 새로운 관찰공간으로써 후보항목리스트를 학습대상으로 분류값을 얻었다.

본 과정을 위해 검증문서를 이용하여 각 단계에서 결정한 범주의 정확성을 평가하고 각 단계별 진행상태의 출력 값으로 부터 오류패턴들을 감지하여 분류기와 환경변수들의 특성으로부터 일관된 결과를 보이는가에 대해 안정성을 평가한다. 본 논문에서 구현한 피드백 시스템에서는 오류를 사전에 방지하는 피드포워드 제어방식과 오류탐지 후의 결과를 학습과정에 반영되도록 재학습을 지시하는 피드백 제어방식을 사용한다. 각 제어방식의 기능은 다음과 같이 요약된다.

- 피드포워드 제어방식은 실행해 옮기기 전에 미리 결함을 예측하여 미연에 방지하도록 움직여주는 피드백 방식이다. 사용자는 RPost의 범주결정단계로의 피드포워드 제어를 함으로서 파라미터로 주어진 임계값을 조정하여 범주를 재결정하도록 한다. 경계항목을 포함하지 않으면서 범주별 수치들의 편차가 매우 작은 문서들은 사용자가 직접 분석할 수 있도록 제시한다.
- 피드백 제어방법으로 해당범주에 새로운 학습문서들을 보강하여 재학습-재분류하는 소극적 제어와 전문가가 개입하여 전반적인 분류체계를 재구성하는 적극적 제어방법을 제시한다.

(그림 7)은 위에서 설명한 피드백 방법을 도식화하여 표현하고 있다. 여기서 분류값이 실제값과 다르면 0, 같으면 1로 나타내었고, 각 단계에서 경계항목으로 지정되거나 지정이 보류된 것은 X로 표기하였다.

(그림 7)의 (a)는 앞에서 설명한 피드포워드 제어를 나타낸다. 적극적인 개입이 필요한 경우는 다음과 같다. 우선, (그림 7)의 (d)와 같이 단계 2에서 지정된 결과가 실제값과 다른 경우는 전반적으로 분류체계를 재정의 하도록 지시한다. 범주결정방법으로 분류되지 않는 문서들, 특히 경계범주를 포함하지 않는 불확실성이 높은 문서들이 많다는 것은 ETOM의 학습과정이 지향하는 것에 반하여 학습문서집합을 구성하는 분류체계나 학습문서가 적합하지 않다는 의미이기 때문이다. 또한, (그림 7)의 (b)와 (c)의 경우와 같이 RPost의 범주결정과정에서 높은 신뢰도를 가지고 지정된 문서는 소극적인 개입을 통해 학습문서로 새로이 추가하여 분류력



(그림 7) RPost시스템의 피드백 방법의 개요: 피드포워드 제어, 피드백 제어

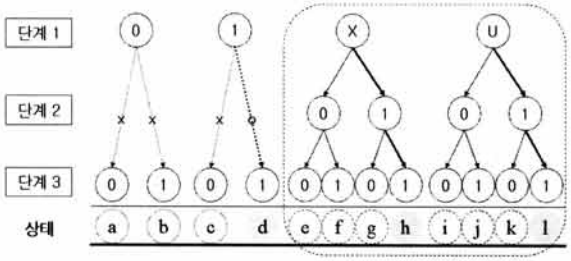
을 향상시킬 수 있도록 한다.

· 각 단계별 결과의 변이(variance) 해석 및 평가

위에서 살펴본 단계별 결과가 의미하는 바를 근거로 하여 문서 D_i에 대한 범주가 결정되어져가는 패턴들의 변이를 통해 오류의 위치와 주요 원인을 예측할 수 있게 된다. 이는 피드백 위치와 동작을 설정하는 데 매우 유용하게 응용된다.

(그림 8)은 이론상 각 단계에서 결정될 수 있는 범주들의 변이 패턴을 나타내며 이들로부터 시스템의 정확성과 안정성을 예측할 수 있다. 각 단계의 출력값과 목표항목이 같으면 1(true), 다르면 0(false), 출력값이 경계항목 X이거나 분류보류항목 U 이면 그대로 표시하였다. (그림 8)에서 단순한 문서들일 경우 초기 분류값은 1, 목표항목 사이에서의 분류가 애매한 문서들의 초기값은 X와 U로 나타난다.

이상적인 학습집합과 확장된 분류체계로 학습된 경우 RPost 의 범주결정과정에서 a, b, c의 상태는 거의 일어날 가능성이 없을 것이며, RPost에서 지향하는 결과의 패턴은 d, h, l 의 상태가 된다. 반대로 a, b, c의 상태가 빈번히 일어난다면 학습집합과 분류체계를 초기화시켜야 할 정도의 문제가 있다는 의미로 해석할 수 있다.



(그림 8) 각 단계에서 발생가능 한 결과값과 변이

이 값들의 변이는 시스템이 얼마나 정확한 결과를 예측하고 있는가에 대한 정보와 얼마나 안정적으로 동작하고 있는가에 대한 정보를 가진다. 정확성과 안정성을 평가하는 개요는 다음과 같다.

-정확성 : 시스템이 결정한 결과와 실제값이 같은가 (precision)
-안정성 : 시스템이 결정한 결과의 값들이 일관적인가 (consistency)

· 유효성 평가함수에 의한 자동화된 피드백 시스템의 설계

성능측정요소로써 현재의 학습체계에 대한 결과의 정확성과 안정성을 평가하며, 이들을 계산하는 유효성 평가함수를 정의한다. 구현 시스템에서는 평가결과에 따라 전문가가 적극적 개입을 제시할 수 있도록 하였고, 소극적인 제어에 의해 새로운 학습문서를 설정하여 재학습하는 동작을 자동화시켰다.

▪ 정확성 평가함수

제안 시스템이 가장 잘 동작하고 분류결과의 정확성을 높게 평가할 수 있는 경우는 (그림 8)의 (b)이다. 첫 회의 단계 1에서 미분류 되었으나 다음 단계에서 정확히 분류되는 형태로 [X|U 1 1] 또는 [X|U - 1]…[1 - -]로 진행하는 경우이다. 이는 식 (12) 과 같이 나타낸다.

$$Good(D_i) = ((step1 = X \mid U) AND (step2 = TRUE)) \\ OR((step1 = X \mid U) AND (FinalRound.step1 = TRUE)) \quad (12)$$

둘째, (그림 8)의 (c)와 같이 비교적 단순한 문서들의 [1 - 1]…[1 - 1]로의 진행은 제안시스템의 보편적으로 동작한다고 간주한다.

$$Fair(D_i) = (step1 = TRUE) AND (FinalRound.step1 = TRUE) \quad (13)$$

셋째, 제안시스템이 가장 이례적으로 동작하고 분류결과의 정확성이 가장 낮다고 측정할 수 있는 경우는 (그림 8)의 (a)이다. 단순한 문서들이 반복을 거듭하여도 제대로 분류되지 않는 경우로써 예를 들어 [0 - -] … [0 - -]로 진행하는 형태에 해당한다.

$$Poor(D_i) = (step1 = False) AND (FinalRound.step1 = False) \quad (14)$$

▪ 안정성 평가함수

입력문서들에 대하여 시스템에서 단계별로 예측한 분류결과들이 일치하는지를 검사하여 현재의 학습체계의 안정성(stability)을 평가한다. 현재의 학습체계에 의해 분류기준이 잘 동작하고 있다고 판단할 수 있는 예는 한 회(round)내에서의 분류 결과들이 일치하는 경우이다. 둘째, 현재의 학습체계에 의한 분류기준에 문제가 있는 경우는 위의 예와 반대인 각 단계의 결과 값이 상이한 경우이다. 따라서 본 논문에서는 앞에서 기술한 바 있는 단계별 결과의 의미를 이용하여 식 (15)와 같이 간단한 규칙을 이용한다.

$$E(stability) = 1 - \frac{1}{N} \sum XOR(step1 \mid step2, step3) \quad (15)$$

ETOM+RPost에 의한 분류결과의 정확성과 학습체계의 안정성을 검사하는 유효성평가 함수의 의의는 다음과 같다.

첫째, 정확성 평가에 있어서 입력문서 D_i 가 단계 1에서 목표항목으로 한 번에 분류되는 경우보다, 분류가 보류되거나 경계항목으로 지정되었더라도 과정을 진행해가면서 목표값을 가지는 경우에 더 큰 보상점수를 주었다. 정확히 분류하기 어려운 경계영역에 속한 문서가 적절하게 분류된다는 것은 RPost시스템이 잘 동작한다는 의미이기 때문이다. 또한 과정을 거듭하면서 끝내 오분류되는 문서는 가중치 값보다 큰 값으로 벌점을 주었다. 정확성은 검증문서를 이용하여 분류경계 영역에 있는 문서들이 정확하게 분류되었는지를 검사하고 식 (16)와 같이 합산한다.

$$E(accuracy) = \frac{1}{N}\left[\sum Good(d_i) \times benefit + \sum Fair(d_i) - \sum Poor(d_i) \times penalty\right] \quad (16)$$

둘째, 입력문서에 대한 최종 회(Round)의 분류의 결과값을 이용하여 현재 학습체계의 안정성을 평가한다. 이 때, 초기에는 미분류되거나 경계범주로 지정된 문서가 최종 회의 단계 1에서 결정될 가능성이 있으므로 이를 함께 검사하였다. 각 범주간의 안정도는 식 (17)과 같이 합산된다.

$$E(stability) = 1 - \frac{T\_uncertain\_case + T\_simple\_case}{N} \quad (17)$$

본 논문의 기준으로 정의한 정확도 E(accuracy) 값의 범위는 $-1.5 \leq E(accuracy) \leq 1.2$ 이다.

불확실한 문서와 단순한 문서가 같은 비율로 있을 경우, 단순한 문서들만 모두 옳게 분류되면 $E(accuracy)$값은 0.5이다. 이 때, 경계영역의 문서 50%도 옳게 분류되면 0.8이다.

안정도 $E(stability)$ 값의 범위는 $0 \leq E(stability) \leq 1$ 이고, 전체 문서의 결과 값들이 일치하면 1이다.

위의 정확도와 안정도의 두 값을 척도로 하여 다음과 같은 피드백 제어방법을 제시한다. 해석에 따른 피드백 제어방법을 요약하면 다음과 같다.

첫째, $E(accuracy)$와 $E(stability)$ 값이 높으면 분류프로세스가 잘 동작하고 있다고 판단한다. 이 때, 높은 확신을 가지고 분류된 문서들은 새로운 학습문서로 활용한다.

둘째, $E(accuracy)$와 $E(stability)$ 값이 상반될 경우, 특히 낮은 정확도에서 높은 안정성을 보일 경우에는 일차적으로 범주결정과정의 임계값을 상향조율 할 수 있다. 그러나 이후의 결과가 향상되지 않으면, 높은 안정성을 보이는 만큼 학습체계에 이상이 있다는 의미로 해석한다. 따라서 전문가의 적극적 개입을 통해 분류체계를 재설정하도록 한다.

셋째, 높은 정확도에서 낮은 안정성을 보이는 경우에는 시스템이 현재의 학습문서와 검증문서에 과적용(overfit) 되어있음을 의미한다. 분류시스템은 새로운 문서들의 정확한 분류값을 예측할 수 있어야 한다. 일차적으로 문서 분류결과의 변이가 큰 범주들을 파악하여 소극적 개입에 의해 학습문서를 재설정한다. 구현시스템에서는 분류 결과값이 서로 일치하지 문서들에서 변이패턴을 보이는 범주구간을 파악하여 사용자에게 제시함으로써, 해당 범주간 적극적인 재학습이 필요함을 알린다.

4. 실험 및 결과

제안방법의 정확성 검증을 위한 실험으로서, 범주들이 다계층 구조를 가질 때 해당범주들의 하위개념들로 정확히 구분되는지 실험하여 정리한다. 실험대상은 이전연구에서 사용한 유즈넷(UseNet)의 뉴스그룹 컴퓨터(comp.)그룹의 문서들이며, 후처리분석과정과 피드백적용을 추가시켜 이전의 결과와 비교하였다.

4.1 실험조건 및 계획

이전 연구에서의 방법과 본 연구에서 제안한 방법의 비교를 위한 실험 조건을 <표 2>와 같이 정리하였다. 기존방법과 제안방법의 정확도를 비교하기 위하여 자동으로 탐색된 경계범주의 포함여부, 평탄화 방법과 계층분류에 대한 비교, 그리고 후처리 분석의 여부에 따른 비교 실험을 수행한다. 비교를 위해 이전연구의 실험계획을 따른다. 경계범주를 포함시켰을 때의 계층분류는 기존방법과의 비교를 위해 하향식 분류방법보다 정확도가 낮은 상향식 계층분류를 수행하였고, NB 분류기를 적용하였다. 상향식 계층 분류에 있어서 병합함수의 역할은 후처리분석 알고리즘이 수행한다.

〈표 2〉 기존/제안방법의 정확성 비교를 위한 실험조건

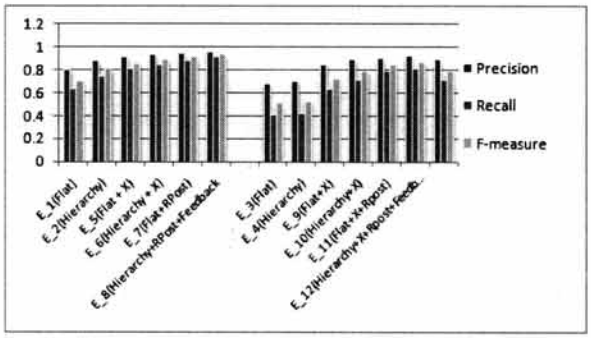
		분류체계	분류방법	실험조건		
	실험	경계범주 (X)	Flat/Hierarchy	분류 알고리즘	학습 오류	후처리& 피드백
기존 방법	E1	X	Flat	SVM	X	X
	E2	X	Hierarchy	SVM	X	X
	E3	X	Flat	SVM	O	X
	E4	X	Hierarchy	SVM	O	X
제안 방법	E5	O	Flat	NB	X	X
	E6	O	Hierarchy	NB	X	X
	E7	O	Flat	NB	X	O
	E8	O	Hierarchy	NB	X	O
	E9	O	Flat	NB	O	X
	E10	O	Hierarchy	NB	O	X
	E10	O	Flat	NB	O	O
	E12	O	Hierarchy	NB	O	O

4.2 실험결과

제안방법의 후처리 분석방법에서 신뢰도 측정을 위해 사용된 설정값은 다음과 같다. 검증문서의 사전 실험결과로 *min_confidence =0.5, diff_confidence =0.3* 으로 설정 후 정확도가 유지됨을 알 수 있었다. 입력문서를 최우선 범주로 결정하기 위한 수치조건은 *min_confidence=0,6 diff_confidence =0.3* 으로 정하였다. 유효성 평가함수의 안정도가 0.7 이하일 때 재학습-재분류하도록 하는 적극적 개입에 의한 피드백을 실시하였다. 검증을 위해 학습에 사용되지 않은 문서로서 각 목표범주마다 50개의 문서들로 구성하여 200개의 문서들로 결과를 확인하였다. 이때 정확도는 F-measure값을 이용하였다.

〈표 3〉 기존방법과 제안방법의 실험결과

	실험	Precision	Recall	F-measure	New Training Documents	
					min_confidence =0.6	*diff_confidence =0.3*
기존 방법	E1	0.806	0.632	0.708		
	E2	0.881	0.750	0.810		
	E3	0.679	0.407	0.508		
	E4	0.706	0.423	0.529		
제안 방법 (ETOM+ RPost)	E5	0.916	0.811	0.860	(81)	(44)
	E6	0.941	0.846	0.890	(95)	(31)
	E7	0.948	0.885	0.915	+81	+44
	E8	0.963	0.912	0.936	+95	+31
	E9	0.852	0.632	0.725	(63)	(36)
	E10	0.897	0.714	0.795	(69)	(41)
	E11	0.903	0.795	0.845	+63	+36
	E12	0.931	0.810	0.866	+69	+41



(그림 9) 기존방법과 제안방법의 성능 비교

<표 3>과 (그림 9)는 실험결과를 나타낸다. 전반적으로 기본 분류체계보다 경계범주를 포함한 확장된 분류체계를 적용하였을 때 성능의 향상을 보였고, 이에 후처리 분석과 피드백분석과정을 모두 진행한 E7과 E8에서 정확도가 더 향상되었다. 이는 오류가 포함된 실험인 E3과 E4, E9~E12의 실험에서 더 잘 나타난다. 추천된 학습문서를 포함하여 후처리분석을 수행한 E12의 결과와 비교하면 F-measure값이 0.34 포인트 상승되어 오류가 포함되기 이전과 비교하여 볼 때 높은 안정도를 나타내고 있다.

5. 결 론

최근의 자동화된 시스템은 전문가의 개입을 최소화하면서도 높은 수준의 정확도와 성능을 보장할 수 있도록 지능적으로 설계되고 있다. 오류율이 높은 복잡하고 불확실한 데이터들의 분류성능의 향상을 위해 지난 연구에서는 결정 경

계면이 가진 문제점을 위해 경계범주를 자동 탐색할 수 있는 알고리즘과 기존 분류체계의 확장 방법인 ETOM을 제안하였다. 본 논문은 단순하게 이루어지는 기존의 범주지정방식을 개선하고 문서의 분류결과의 신뢰도를 측정할 수 있는 모델인 RPost를 제안하였다.

본 논문의 결과는 다음과 같이 요약할 수 있다. 첫째, 강화된 후처리 분석으로 보다 적합한 범주로 할당 될 수 있도록 하였다. 신뢰할 만한 결과는 우선 분류해내고 신뢰도가 결여되는 것은 후속처리가 되게 하여 오류율을 감소시킬 수 있었다. 둘째, 각 단계별 진행 상태에 대한 결과를 분석함으로써 오류가 발생하는 원인과 위치를 예상할 수 있으므로 적절한 피드백 위치와 지침을 마련할 수 있었다. 셋째. 전체적인 분류시스템을 평가할 수 있는 유효성함수를 설계하였다. 본 제안방법의 취지에 따라 시스템의 성능을 평가하는 척도를 정의하고 분류가 정확히 진행되는지, 현재의 학습체계가 안정적으로 수행되는지를 평가하였다.

본 논문은 다음과 같은 의의를 갖는다. 분류알고리즘이나 분류기의 성능에 영향을 미치는 자질값과 환경변수 등 여러 가지 세부적인 요인들에 비교적 둔감하고 효율적으로 동작하도록 하였다. 이는 학습문서의 구성방식과 분류 알고리즘의 성능에 전적으로 의존하지 않는 안정적인 분류 프로세스가 되게 하는 역할을 한다. 본 논문에서 제안한 방법들은 정확한 분류가 필요한 여러 분야에 손쉽게 적용될 수 있도록 설계되었다. 확장된 분류체계에 의한 학습방법은 그 대상의 형태에 의존하지 않고 적용할 수 있으며, 분류결과의 범주를 결정하는 후처리방법은 알고리즘의 성능에 의존하지 않도록 고안되어 있다. 스팸문서와 같이 침입탐지 시스템에서의 일반적인 패턴으로 위장된 공격패턴에 대한 이상탐지 및 오용탐지 분석에도 적용될 수 있으며, 기계학습으로 분석하기에 복잡한 이미지분류에도 활용될 수 있다.

참 고 문 헌

[1] T.,Joachims,"Text Categorization with Support Vector Machines: Learning with Many Relevant Features," In Proc. of ECML-98 pp.137-142, 1998.

[2] 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.

[3] D.,Koller,S.,Tong,"Active Learning for Parameter Estimation in Bayesian Networks," In Neural Information Processing Systems, 2001.

[4] D.,David, J., Catlett, "Heterogeneous Uncertainty Sampling for Supervised Learning," In Proc. of the 11th ICML, pp.148-156, 1994.

[5] D.,R.,Wilson, et al "Reduction Techniques for Exemplar-based Learning Algorithms," Machine Learning, Vol.38. No.3, pp.257-286, 2002.

[6] T., Zhang, and F.Oles, "A Probability Analysis on the Value of Unlabeled Data for Classification Problems," In Proc. of 17th Machine Learning (ICML), 2000.

[7] K.,A.,Kofahi, and A., Tyrrell, et.al, "Combining Multiple Classifiers for Text Categorization," In Proc. of ACM CIKM, pp.97-104, 2001.

[8] S.B.,Kim and H.C.,Rim, "Recomputation of Class Relevance Score for Improving Text Classification," In Proc. Conference of Computational Linguistics and Intelligent Text Processing (CICLing), LNCS, Vol.2945, pp.580-583, Feb., 2004.

[9] Shanfeng Zhu and Ichigaku Takigawa et. al, "Field Independent Probabilistic Model for Clustering Multi-field Documents," Information Processing and Management, Vol.45, pp.555-570,2009.

[10] Qinrong Feng, Duoqian Miao and Yi Cheng, "Hierarchical decision rules mining," Expert Systems with Application, Vol.37, pp.2081-2091, 2010

[11] Nicolás García-Pedrajas and Domingo Ortiz-Boyer, "Boosting k-nearest neighbor classifier by means of input space projection," Expert Systems with Applications,Vol. 36, pp.10570-10582, 2009.

[12] David A. Bell, J. W. Guan, Yaxin B, "On Combining Classifier Mass Functions for Text Categorization", IEEE Trans. Knowl. Data Eng. Vol.17, No.10, pp.1307-1319,2005.

[13] G.P. Zhang, "A Neural Network Ensemble Method with Jittered Training Data for Time Series Forecasting," Information Sciences. Vol.177, pp.5329-5346.2007.

[14] S. B. Cho, "Ensemble of Structure Adaptive Self-Organizing Maps for High Performance Classification," Information Science, Vol. 123, No.1-2, pp.103-114, 2000.

[15] 최윤정,지정규,박승수, "경계범주 자동탐색에 의한 확장된 학습체계 구성방법", 정보처리학회논문지(B), Vol. No. pp.~ pp. 2009. 12.

최 윤 정

e-mail : cris@seoil.ac.kr

1997년 서원대학교 전자계산학과(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

박 승 수

e-mail : cris@seoil.ac.kr

1970년~1974년 서울대학교 수학과(공학사)

1974년~1976년 한국과학기술원 수학과(석사)

1976년~1988년 미국 텍사스 오스틴대학 전산학 박사

1988년~1991년 미국 켄사스대학 컴퓨터학과 조교수

1991년~현 재 이화여자대학교 컴퓨터공학과 부교수

관심분야 : 인공지능, 데이터마이닝, 상황인식, 유비쿼터스, 바이오인포매틱스