나이브베이즈 문서분류시스템을 위한 선택적샘플링 기반 EM 가속 알고리즘

장 재 영[†] · 김 한 준[††]

요 약

본 논문은 온라인 전자문서환경에서 전통적 베이지안 통계기반 문서분류시스템의 분류성능을 개선하기 위해 EM(Expectation Maximization) 가속 알고리즘을 접목한 방법을 제안한다. 기계학습 기반의 문서분류시스템의 중요한 문제 중의 하나는 양질의 학습문서를 확보하는 것이다. EM 알고리즘은 소량의 학습문서집합으로 베이지안 문서분류 알고리즘의 성능을 높이는데 활용된다. 그러나 EM 알고리즘은 최적화 과정에서 느린 수렴성과 성능 저하 현상을 나타내는데, EM 알고리즘의 기본 가정을 따르지 않는 온라인 전자문서환경에서 특히 그러하다. 제안 기법의 주요 아이디어는 전통적 EM 알고리즘을 개선하기 위해 불확정성도 기반 선택적 샘플링 기법을 활용한 것이다. 성능평가를 위해 Reuter-21578 문서집합을 사용하여, 제안 알고리즘의 빠른 수렴성을 보이고 전통적 베이지안 알고리즘의 분류 정확성을 향상시켰음을 보인다.

키워드 : 자동문서분류, 기계학습, EM 알고리즘, 나이브베이즈, 불확정성, 선택적샘플링

Accelerating the EM Algorithm through Selective Sampling for Naïve Bayes Text Classifier

Jae-young Chang[†] · Han-joon Kim[††]

ABSTRACT

This paper presents a new method of significantly improving conventional Bayesian statistical text classifier by incorporating accelerated EM(Expectation Maximization) algorithm. EM algorithm experiences a slow convergence and performance degrade in its iterative process, especially when real online-textual documents do not follow EM's assumptions. In this study, we propose a new accelerated EM algorithm with uncertainty-based selective sampling, which is simple yet has a fast convergence speed and allow to estimate a more accurate classification model on Naive Bayesian text classifier. Experiments using the popular Reuters-21578 document collection showed that the proposed algorithm effectively improves classification accuracy.

Key Words : Text Classification, Machine Learning, EM Algorithm, Naïve Bayes, Uncertainty, Selective Sampling

1. 서 론

온라인 전자문서의 규모가 커짐에 따라 이를 관리하는 정보시스템은 자동문서분류(automated document categorization) 기술에 큰 중요성을 두게 되었으며, 최근 기계학습(machine learning) 이론의 도입으로 문서분류에 관한 연구가 활발하게 진행되고 있다. 자동문서분류란 유입되는 문서를 그 내용에 적합한 카테고리로 자동 할당하는 것을 의미한다. 기계학습을 이용한 방법은 준비된 학습문서집합으로부터 각 카테고리의 특성을 학습하는 귀납적 프로세스(inductive process)가 있어 문서를 분류할 수 있는 모델을 생성한다. 대표적인 기계학습 방법으로는 Naïve Bayes[9], K-nearest neighbors[5], SVM(support vector machine)[6] 등이 있다. 이 중에서 Naïve Bayes 알고리즘은 통계 기반 학습기법 중 하나로 모델 파라미터의 구성이 간단하면서도 복잡한 모델 구성을 가지는 다른 기법들과 견주어 성능이 우수하여 많은 문서분류 프로젝트에서 채택되고 있다.

본 논문은 Naïve Bayes 학습 알고리즘을 사용하여 실세계 온라인 문서환경에서 문서분류의 성능을 높이는데 초점을 맞춘다. 실세계 온라인 문서환경에서는 문서집합을 학습문서와 분류할 대상문서(테스트문서)를 미리 구분할 수 없다. 그래서 온라인 전자문서의 자동분류시스템은 학습문서집합을 미리 확보하고 있지 않은 상태에서 지속적으로 유입되는 문서를 분류해야 한다. 온라인상태의 문서분류에서는 학습문서를 처음부터 충분히 확보하는 것이 불가능하기 때문에 학습결과가 유입되는 문서의 순서에 따라 영향을 받게

※ 이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었으며 (KRF-2004-003-D00355), 또한 정보통신부 및 정보통신연구진흥원의 대학IT연구센터 육성지원사업 (IITA-2005-C1090-0502-0016)의 연구결과로 수행 되었음.
† 종신회원 : 한성대학교 컴퓨터공학과 조교수
†† 정 회 원 : 서울시립대학교 전자전기컴퓨터공학부 조교수(교신저자)
논문접수 : 2006년 2월 28일, 심사완료 : 2006년 4월 19일

된다. 결과적으로 문서분류시스템에서는 학습문서집합이 지속적으로 보충 또는 보강되는 기법이 요구된다.

기계학습에 기반을 둔 문서분류 알고리즘에서 어려운 문제 중의 하나는 정확한 분류모델을 도출하기 위해서는 충분히 많은 학습문제가 필요하다는 것이다. 그러나 실세계 온라인 문서환경에서는 초기에 충분한 학습 문서를 확보하기 힘들고 학습 문서의 질에 일관성이 없다[12]. 따라서 이러한 환경에서 최소의 노력으로 양질의 학습문서를 확보하는 문제가 중요한 이슈가 된다. 이는 문서분류시스템을 상용화하고자 할 때 해결해야 할 중요한 문제이다[10]. 이에 대한 하나의 해결 방안은 A.P. Dempster에 의해서 고안된 반복적(iterative) 알고리즘인 EM(Expectation Maximization) 알고리즘을 이용하는 것이다[4]. 이는 카테고리 레이블이 부여된 학습 문서가 부족한 상황에서 레이블이 없는 미분류 상태인 문서를 이용하여 레이블이 있는 학습 문서를 보강하는 기법으로서, 문서분류시스템에서 학습문서가 충분하지 못한 경우에 실용적이면서도 성능 및 효과 면에서도 우수한 방법으로 알려져 있다.

EM 알고리즘은 많은 장점을 가진 기법이지만 반면에 몇 가지 문제점을 안고 있다. 첫째는 레이블이 있는 문서의 수가 최적에 가까운 분류모델을 생성할 만큼 충분히 많을 경우 레이블이 없는 문서는 더 이상 모델을 개선하는데 도움이 안 될 수 있으며, 심지어는 이미 추정된 최적에 근사된 모델을 오히려 왜곡시킬 가능성도 갖고 있다. 둘째, EM 알고리즘은 그 수렴속도가 매우 느리며 그 속도는 레이블이 없는 문서의 크기에 따라 민감하게 달라진다. 마지막으로, EM 알고리즘은 패턴이 hill-climbing 알고리즘과 유사하여 모델 파라미터(parameter)의 해(solution)가 지역 최적치(local optimum)에 빠질 수 있다[10]. 따라서 초기에 모델 파라미터 값이 어떻게 설정되느냐에 따라 도달하게 되는 최적치가 민감하게 달라진다.

본 논문에서는 이러한 전통적 EM 알고리즘의 문제점을 해결하기 위해 Naïve Bayes 프레임워크를 기반으로 학습문서를 자동적으로 보강하는 수정된 EM 알고리즘을 제안한다. 수정된 EM 알고리즘의 기본 아이디어는 다음과 같다. 우선 주어진 학습문서를 이용하여 한번의 EM 과정을 적용한 후에, 추정된 분류모델을 기반으로 레이블이 없는 문서집합으로부터 학습 문서에 보강될 문서를 추출한다. 이 문서는 학습 문서에 추가된 후 다시 EM 과정을 적용하고 이러한 과정들이 특정 조건이 만족될 때까지 반복적으로 수행된다. 여기서 학습 문서에 보강될 문서를 결정하는 과정은 기존의 선택적샘플링기법(selective sampling)[8]에서 제안된 문서 평가 기준중의 하나인 분류애매도(categorization uncertainty)를 이용하게 된다. 그러나 선택된 문서에 레이블을 부여하는 과정은 기존의 수작업과는 달리 자동적으로 수행된다.

본 논문에서는 Reuters-21578 문서집합[11]을 이용한 실험을 통해 제안 알고리즘의 우수성을 평가하였다. 우선 레이블이 있는 문서집합의 크기에 따라 전통적 EM 알고리즘과 본 논문이 제안한 수정된 EM 알고리즘의 분류성능을 비교하고, 분류애매도의 임계값 변화에 따른 분류성능의 변화를 관찰한 실험 결과를 제시한다. 본 논문의 구성은 다음과 같다. 우선 2장에서는 본 논문이 제안하는 알고리즘에 대한 이론적 배경을 기술하고, 3장에서는 수정된 EM 알고리즘을 제안한다. 4장에서는 실험결과를 제시하고, 마지막으로 5장에서는 결론과 향후 연구 과제를 제시한다.

2. 문서분류를 위한 이론적 배경

2.1 Naïve Bayes 알고리즘

본 절에서는 통계기반 학습기법 중 하나인 Naïve Bayes (간단히 NB라고 표기함) 알고리즘을 소개한다. 이 알고리즘은 본 논문에서 제안하는 EM 알고리즘뿐만 아니라 전통적인 EM 알고리즘의 기본 바탕이 되고 있다. NB는 서론에서 언급한 바와 같이 단순하면서도 정확한 추정능력을 발휘한다고 알려져 있어 많은 문서분류 프로젝트에서 채택되어 왔다[2, 13, 15].

NB 알고리즘에 의한 문서분류는 주어진 문서를 분류하기 위해 베이즈(Bayes) 정리에 의해 아래 식과 같이 주어진 문서에 대한 카테고리의 사후확률값(posterior probability)을 추정함으로써 이루어진다.

$$Pr(c_j|d_i) = \frac{Pr(c_j)Pr(d_i|c_j)}{Pr(d_i)} \quad (1)$$

여기서 $Pr(c_j)$는 전체문서집합에서 임의 추출한 문서가 카테고리 c_j에 속할 사전확률값(prior probability), $Pr(d_i|c_j)$는 카테고리 c_j에 속하는 문서집합에서 임의로 추출된 문서가 d_i일 확률값, $Pr(d_i)$는 전체문서집합에서 임의 추출한 문서가 d_i일 확률값을 의미한다.

식(1)의 사후확률값을 추정할 수 있도록 모델 θ_{NB}를 추정하고 나면 문서분류자 $f_{\theta_{NB}}$는 입력된 문서에 대해 모든 카테고리에 대한 사후확률값을 계산한 후에 가장 큰 사후확률값을 가지는 카테고리를 리턴한다. 이를 다음 식[2]과 같이 표현할 수 있다.

$$f_{\theta_{NB}}(d_i) = argmax_{c_j \in C} Pr(c_j|d_i) = argmax_{c_j \in C} \frac{Pr(c_j)Pr(d_i|c_j)}{Pr(d_i)} \quad (2)$$

여기서 사후확률값을 계산하기 위해서는 $Pr(c_j)$, $Pr(d_i|c_j)$, $Pr(d_i)$를 계산해야 한다. 즉 각 항이 분류모델을 구성하는 파라미터가 된다. 그런데 $Pr(d_i)$는 모든 카테고리

2) $argmax_{x \in X} f(x)$는 $f(x)$의 값을 최대로 하는 x 값을 의미한다.

에 대한 $Pr(c_j)$와 $Pr(d_i|c_j)$의 값을 모두 더한 값과 같기 때문에 따로 계산할 필요는 없다. $Pr(c_j)$와 $Pr(d_i|c_j)$의 계산은 주어진 학습문서에 대하여 각 모델 파라미터의 추정 방법은 MAP 가설(Maximum A Posteriori hypothesis)[3]에 따른다.

우선 $Pr(c_j)$의 계산은 아래 식과 같이 학습문서 D^t 중에서 c_j에 존재하는 문서들의 수를 카운트함으로써 쉽게 얻을 수 있다.

$$Pr(c_j) = \frac{c_j\text{에존재하는문서의수}}{|D^t|} \tag{3}$$

다음으로 $Pr(d_i|c_j)$의 계산을 위해서 NB 학습 알고리즘에서는 그 조건부 확률의 계산을 간단히 하기 위해 카테고리 변수에 대해서 문서 내에 존재하는 각 단어들이 서로 독립(independent)이고, 문서내의 단어 위치와 그 단어의 출현 확률사이에도 독립성이 존재한다고 가정한다. 따라서 $Pr(w_{ik}|c_j)$를 카테고리 c_j에 존재하는 학습문서 내에 용어 w_{ik}가 출현할 확률이라 할 때 $Pr(d_i|c_j)$는 $\prod_{k=1}^{|d|} Pr(w_{ik}|c_j)$로 표현될 수 있다. 결과적으로 식 (2)는 다음과 같이 변환된다.

$$\begin{aligned} f_{\theta_{NB}}(d_i) &= argmax_{c_j \in C} Pr(c_j|d_i) \\ &= argmax_{c_j \in C} Pr(c_j) \cdot \prod_{k=1}^{|d|} Pr(w_{ik}|c_j) \end{aligned} \tag{4}$$

이제 $Pr(w_{ik}|c_j)$의 값을 추정하는 문제만 남게 되는데, $TF(w_{ik}, c_j)$를 카테고리 c_j에서 단어 w_{ik}가 출현하는 빈도수라 하고 V를 학습문서의 전체 용어집합이라 할 때, $Pr(w_{ik}|c_j)$의 MAP 추정값은 $\frac{TF(w_{ik}, c_j)}{\Sigma_{w_{ik} \in V} TF(w_{ik}, c_j)}$와 같다.

그런데 이 식은 특정 카테고리에 존재하는 않는 단어에 대해서는 0값을 가지게 되므로, 이를 보정하기 위해 Laplace smoothing[9]을 적용한 식은 다음과 같다.

$$Pr(w_{ik}, c_j) = \frac{1 + TF(w_{ik}, c_j)}{|V| + \Sigma_{w_{ik} \subset V} TF(w_{ik}, c_j)} \tag{5}$$

식 (3)과 (5)에 의하여 모델 파라미터들이 추정되면 이를 사용하여 문서분류자 함수를 완성할 수 있다. NB 학습기법은 단어 간의 독립성, 단어위치와 카테고리간의 독립성을 가정하고 있다. 하지만 실제 문서 데이터는 문법의 사용 및 주제 표현 등의 이유로 이러한 독립성이 맞지 않는다. 그럼에도 불구하고 NB 학습에 의한 문서분류 정확도가 복잡한 모델 파라미터를 가지는 다른 기법의 성능과 비슷하다고 알려져 있다[16].

지금까지 설명한 NB 학습기법은 문서분류 시스템의 관점에서 다음과 같은 장점을 갖는다. 첫째, 학습문서집합을 한번만 스캔하여 모델 파라미터를 구축하기 때문에 학습 과정의 속도가 다른 기법에 비해 빠르다. 둘째, 학습문서집합의 크기 또는 유형에 따라 분류모델이 크게 달라지지 않는다. 이는 온라인 학습 측면에서 볼 때 매우 바람직한 성질이다. 셋째, 주어진 카테고리 집합에 대한 모델 파라미터를 수정하기가 용이하다. 이는 재학습이 매우 간단하게 수행될 수 있음을 의미한다. 식 (3)과 (5)에서 보는 바와 같이 학습문서가 추가되거나 삭제되었을 때 새로운 $Pr(c_j)$와 $Pr(w_{ik}|c_j)$의 값을 구하기 위해서는 기존의 값에 추가된 또는 삭제된 부분에 값만을 더 보태거나 또는 빼기만 하면 된다[9]. 마지막으로, NB 학습기법은 HTML 형식의 반구조적(semi-structured) 전자문서에 포함되어 있는 태그(tag) 정보의 중요도를 쉽게 반영할 수 있다. 다시 말해서, 중요 태그에 포함되어 있는 용어의 출현횟수에 적당한 비중값을 곱해주기만 하면 된다.

NB 학습기법을 포함한 대부분의 기계학습 기법들은 학습에 적합한 문서들이 이미 확보되었다는 것을 가정한다. 그러나 실세계에서는 고품질의 학습문서를 초기부터 확보하기 어려우며 학습문서집합을 지속적으로 갱신하는 것이 바람직하다. 이러한 맥락에서 EM 알고리즘은 학습문서 확보에 좋은 해결방법이라 할 수 있다. 특히 본 절에서 소개한 NB기법을 EM 알고리즘에 적용할 경우 적은 학습문서 만으로도 문서분류의 성능을 획기적으로 향상 시킬 수 있다[10, 13].

2.3 EM 알고리즘

EM 알고리즘은 A.P. Dempster에 의해서 고안된 반복적 알고리즘으로서, 수동적으로 카테고리 레이블을 부여한 학습 문서가 부족한 상황에서 카테고리 레이블이 없는(미분류인) 학습문서(unlabeled training documents)를 이용하여 MAP추정을 통해 레이블이 있는 학습문서를 보강하는 기법이다[4]. 학습문서의 수가 작을 때는 모델 파라미터를 추정하는데 있어 그 정확도가 떨어지고 변이(variance) 또한 커진다. 하지만 레이블이 없는 문서에 대하여 확률적으로 카테고리 레이블 정보를 얻어내어 레이블이 있는 학습문서집합을 보강함으로써 추정된 모델을 개선할 수 있다. (그림 1)은 NB 학습기법을 사용하여 EM 알고리즘에 의해 모델 파라미터를 개선하는 알고리즘을 보여준다. 앞서 언급한 바와 같이 NB 문서분류자에서의 학습은 주어진 학습문서집합 D^{tl}의 MAP 추정을 계산한다. 즉, $l_{NB}(D^{tl}) = argmax_{\hat{\theta}_{NB}} Pr(D^{tl}|\hat{\theta}_{NB}) \cdot Pr(\hat{\theta}_{NB})$이 된다. 여기서 $l_{NB}(D^{tl})$는 D^{tl}을 입력받아 분류모델을 생성하는 함수를 의미한다. 초기에는 레이블이 있는 학습문서집합 D^{tl}만을 이용하여 모델 파라미터인 $\hat{\theta}_{w|c}$와 $\hat{\theta}_c$를 추정한다. 그리고 문서분류자는 레이블이 없는 학습문서집합 D^{tu}에 존재하는 각 문서 d_i에 대

3) MAP가설은 학습문서 D가 주어질 때, 가능한 가설집합 H에서 'maximally probable hypothesis' h_{MAP} ($\in H$)를 선택하는 것이다.

$$h_{MAP} \equiv argmax_{h \in H} = Pr(h|D) = argmax_{h \in H} \frac{Pr(D|h)Pr(h)}{Pr(D)}$$

입력: 학습문서집합 $D^{l} = D^{tl} \cup D^{tu}$

출력: 추정된 분류모델 $\hat{\theta}_{NB} = \{\hat{\theta}_{wc}, \hat{\theta}_{c}\}$

BEGIN

$l_{NB}(D^{tl}) = argmax_{\hat{\theta}_{NB}} Pr(D^{tl}|\hat{\theta}_{NB}) \cdot Pr(\hat{\theta}_{NB})$ /* MAP 계산 */

$\hat{\theta}_{NB} = l_{NB}(D^{tl})$; $D^{temp} = D^{tl}$ /* 초기 추정모델 계산 */

WHILE (추정모델 $\hat{\theta}_{NB}$가 더 이상 향상되지 않을 때까지)

(E-단계) $D^{temp} = D^{temp} \cup \{\langle d_i, f_{\hat{\theta}_{NB}}(d_i)\rangle\}$ for each $d_i \in D^{tu}$

(M-단계) $\hat{\theta}_{NB} = l_{NB}(D^{temp})$; $D^{temp} = D^{tl}$

END

(그림 1) 전통적 EM 알고리즘

해 확률적으로 얻어진 카테고리 레이블(probabilistic category label) 즉, $Pr(c_j|d_i)$를 계산한다 (E-단계). 다음 단계로 학습문서와 분류된 문서를 모두 사용해서 새로운 모델 파라미터를 추정한다 (M-단계). 이 두 작업은 추정되는 모델 파라미터가 더 이상 개선되지 않을 때까지 진행한다. 이 EM 알고리즘은 제안했던 Dempster에 의해 그 수렴성이 증명되었다. 다시 말해서, EM과정의 각 루프에서 $Pr(D^{tl}|\hat{\theta}_{NB}) \cdot Pr(\hat{\theta}_{NB})$ 값이 그 이전루프와 비교했을 때 같거나 항상 증가한다.

이 EM 알고리즘은 적은 양의 학습문서를 가지고 문서분류의 정확도를 획기적으로 높이는데 기여하였다. [10]에서는 70%의 분류정확도를 얻기 위해 기본적인 NB 알고리즘은 2,000개의 레이블이 있는 학습문서가 필요하였지만, 여기에 EM 알고리즘을 적용하였을 때는 동일한 수준의 분류정확도를 성취하는데 단지 600개의 학습문서가 필요하였다. (그림 1)의 전통적 EM 알고리즘은 그 수렴 속도가 매우 느리며, 그 속도는D^{tl}의 크기에 따라 민감하게 달라지며, 이는 성격상 hill-climbing 알고리즘과 유사해서 모델 파라미터가 지역최적치(local optimum)에 빠질 수 있다[10]. 따라서 초기에 모델 파라미터 값이 어떻게 설정되느냐에 따라서 도달하게 되는 최적치가 민감하게 달라진다. 본 논문에서는 이러한 EM 알고리즘의 두 가지 단점을 분류애매도를 이용한 학습문서집합의 확장을 통해 극복하고자 한다.

3. EM 가속(acceleration) 알고리즘

본 장에서는 앞서 기술한 전통적인 EM 알고리즘의 한계를 극복하기 위해 분류애매도를 이용한 EM 가속알고리즘을 소개한다.

3.1 분류애매도를 이용한 학습문서의 선택적 샘플링

우선 분류애매도를 정의하기 위해 능동적학습의 한 가지 유형인 선택적 샘플링(selective sampling) 기법[8]을 소개한다. EM 알고리즘에서는 현재 학습된 모델 파라미터를 토대로 레이블이 없는 학습 문서에 대해서 카테고리 레이블을 추정한다. 이와는 달리 선택적 샘플링은 레이블이 없는 문서 중에서 가장 최적의 학습문서를 선택하는 모듈인 '학습문서선택자(sampler)'가 있어 이것이 선택하는 문서에 대해 문서분류 전문가가 카테고리 레이블을 부여하게 된다. 결국 이 문서를 학습문서에 병합되어 문서분류모델을 최적화하는데 다시 활용되게 된다. 이때 가장 중요하게 고려해야할 점은 어떠한 방법으로 최적의 문서를 선택하느냐이다. 문서선택의 가장 큰 기준은 현재 추정된 분류모델에 대해서 그 문서의 정보량이다. 문서의 정보량이 많다는 것은 그것의 분류 결과가 애매하다(ambiguous)는 것을 의미하며 그것이 학습문서집합에 반영되었을 때 현재 추정된 분류모델에 큰 영향을 줄 수 있다는 것을 의미한다. 이는 마치 의사결정트리(decision trees)에서 분기(split)의 기준을 정하는 단계에서 가장 정보량이 많은 필드를 선택하는 것과 유사하다. 이 정보량의 크기가 결국 분류애매도가 된다.

NB 학습기법은 간단히 각 카테고리 c에 대하여 그것의 개념을 설명할 수 있는 단어들 $W(\subset V)$의 확률분포 $Pr(w \in W|c)$를 추정한 것이라 할 수 있다. 이런 관점에서 현재 학습된 분류모델을 통해 주어진 문서의 분류가 애매하다면 아직 카테고리에 대한 올바른 단어 분포가 형성되지 않았음을 의미한다. 다시 말해서, 주어진 문서에 출현하는 단어들에 대해서 그 문서의 진리 카테고리(true category)와 다른 카테고리간의 확률분포 $Pr(W|c)$의 차별성이 크지 않다는 것이다. 반대로 그 문서의 분류가 확실하다는 것은 그 문서가 출현하는 단어에 대해서 진리 카테고리의 단어 분포가 다른 카테고리와의 단어 분포와는 매우 다름을 의미한다. 여기서 확률분포가 비슷하고 다른 정도를 확률분포간 거리의 개념으로 정량화할 수 있다. 일반적으로 확률 분포간의 거리는 Kullback-Leibler(KL) 거리함수를 활용할 수 있다[1]. KL 거리는 두 분포간의 차이를 정량화한 정보 이론적 수치로서 어떤 분포로부터 샘플링된 메시지를 다른 분포에 대해서 최적의 코드를 사용해서 보내는데 필요한 정보의 비트량을 의미한다. 이 KL 거리함수를 이용하여 현재 학습된 분류모델 상에서 단어 분포간의 거리를 측정함으로써 분류애매도를 결정할 수 있다. 문서 d와 두 카테고리 c_i, c_j에 대해서 나타나는 단어분포사이의 KL 거리함수는 다음과 같다[4].

4) 본래 KL 거리함수는 비대칭성을 가지므로, 대칭성을 가지도록 식 (6)과 같은 형태를 가진다. 이는 참고문헌 [7]에서 제시한 KL 거리함수의 대칭화를 위한 여러 방안 중에서 간단한 형태를 채택한 것이다.

$$KLdist_d(Pr(W|c_i), Pr(W|c_j)) = \frac{1}{2}\left\{\sum_{w_k \in d} Pr(w_k|c_i) \cdot \log\frac{Pr(w_k|c_i)}{Pr(w_k|c_j)} + \sum_{w_k \in d} Pr(w_k|c_j) \cdot \log\frac{Pr(w_k|c_j)}{Pr(w_k|c_i)}\right\} \quad (6)$$

여기서 $|C|$는 현재 카테고리 집합에 존재하는 카테고리의 수를 의미한다. 주목할 것은 분포 $Pr(w_k|c_i)$와 $Pr(w_k|c_j)$ 간의 거리값이 모든 단어에 대해서가 아니라 주어진 문서 d에 출현하는 단어에 대해서만 계산한다는 것이다. 전체 카테고리에 대한 KL 거리값의 평균값이 작다(또는 크다)는 것은 현재 학습된 단어분포가 문서 d에 출현하는 단어에 대해서는 분별력이 약하다(또는 강하다)는 것을 의미한다. 문서 d에 대해서 카테고리에 대한 분별력이 큰 단어분포가 형성되기 위해서는 카테고리에 대한 단어분포의 KL 거리값이 커야한다.

이 거리함수를 사용하여 문서 d의 분류애매도 $CU_{NB}(d)$는 다음과 같이 정의할 수 있다.

$$CU_{NB}(d) = 1 - \frac{\sum_{c_i, c_j \in C} KLdist_d(Pr(W|c_i), Pr(W|c_j))}{|C| \cdot (|C|-1)} \quad (7)$$

분류애매도 CU_{NB}는 모든 카테고리에 대해서 $KLdist_d$의 값을 산술평균한 후 분류애매도의 의미를 갖기 위해 1에서 그 평균값을 뺀다.

3.2 EM 가속 알고리즘을 이용한 NB 문서 분류

일반적으로 분류모델의 추정 오차 $|\theta - \hat{\theta}|$는 $O(\frac{1}{\sqrt{|D^{tl}|}})$의 속도를 가지고 0으로 수렴한다고 알려져 있다[3]. 이는 레이블이 있는 문서가 증가할수록 진리모델(true model)을 찾는 수렴 속도가 지수적으로 증가함을 의미한다. 이를 EM 알고리즘 측면에서 보면 만약 EM 과정의 E-단계에서 옳다고 판단되는 문서를 알 수 있어 이를 D^{tl}에 병합시킨다면 분류모델의 추정 속도가 한층 높아지는 것을 기대할 수 있다.

EM과정이 어느 정도 진행된 후에는 E-단계에서 D^{tu} 집합의 문서들을 분류한 결과, 각 문서의 분류애매도 값에 따라서 분류가 정확한지 여부를 간접적으로 추정할 수 있다. 다시 말해서, 분류애매도가 매우 낮다면 분류가 정확하게 된 것으로 예상할 수 있다는 것이다. 이를 확인하기 위해 분류애매도와 분류결과를 비교한 실험을 수행하였다.

(그림 2)는 분류애매도와 분류결과를 비교한 그래프이다. 이때 사용된 분류애매도는 3.1절에서 제안한 CU_{NB}값을 이용하였다. 이 그림에서 가로축은 문서 d에 대한 분류애매도 $CU_{NB}(d)$의 값을 오름차순으로 정렬하였을 때 구간 값을 나타내고 세로축은 옳게 분류된 문서와 그렇지 않은 분류된 문서의 상대적인 비율을 나타낸다. 이 그림을 통해 대체적으로 분류애매도 값이 커질수록 잘못 분류된 문서의 비율보다 많아지고 있음을 관찰할 수 있다. 결과적으로 이 실험은 각 문서의 분류기준이 애매하지 않을 경우에만 올바르게 분류될 수 있음을 보여준다.

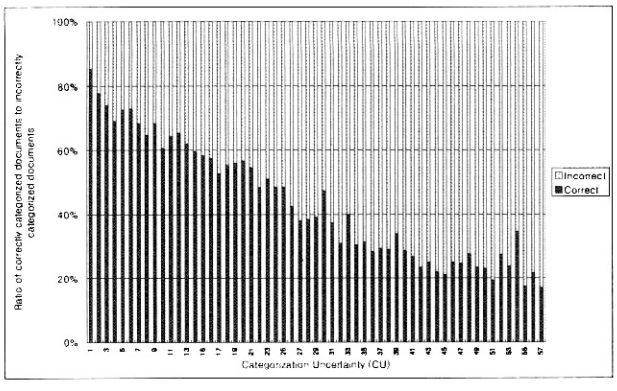
이 실험에 근거하여 전통적인 EM 알고리즘을 수정할 수 있다. 그것의 기본 아이디어는 E-단계에서 현재 분류모델로 분류된 문서집합 D^{tc}의 문서 중에서 분류애매도가 가장 낮은 문서를 레이블이 있는 학습문서집합 D^{tl}에 병합하는 것이다.

(그림 3)은 수정된 EM 알고리즘을 보여준다. 이 알고리즘에서는 충분한 EM 과정의 루프를 수행하여 모델 파라미터의 변화가 없을 때 선택적 샘플링을 통해 레이블이 있는 문서를 추가한다. 따라서 각 EM과정을 진행하기 전에 이전 루프와 비교하여 현재의 분류모델에 변화가 있는지를 검사한다. 이때 모델에 변화가 없으면 레이블이 있는 문서 추가의 필요성을 나타내는 *Augment*변수를 TRUE로 설정하고

```
입력: 학습문서집합 D^t = D^tl ∪ D^tu
출력: 추정된 분류모델 θ̂_NB = {θ̂_wc, θ̂_c}

BEGIN
l_NB(D^tl) = argmax_θNB Pr(D^tl|θ_NB) · Pr(θ̂_NB)  /* MAP 계산 */
θ̂_NB = l_NB(D^tl);  D^temp = D^tl      /* 초기 추정모델 계산 */
WHILE(TRUE)
        if(추정모델 θ̂_NB가 더 이상 향상되지 않으면) Augment = FALSE
        else Augment = TRUE
  (E-STEP)  D^tc = {⟨d_i, f_θ̂NB(d_i)⟩} for each d_i ∈ D^tu
        D^temp = D^temp ∪ D^tc
        if(Augment is TRUE)
                d_best = argmin_{d_j ∈ D^tc} CU_NB(d_j)
                if(CU_NB(d_best) > μ)
                        D^tu = D^tu − {d_best}
                                D^tl = D^tl ∪ {d_best}
                else    EXIT
  (M-STEP)  θ̂_NB = l_NB(D^temp);  D^temp = D^tl
END
```

(그림 2) 수정된 EM 알고리즘



(그림 3) 분류애매도와 분류결과의 상관관계 실험결과

그렇지 않으면 FALSE로 설정한다(줄번호 4-5)[5]. *Augment* 변수를 TRUE로 설정되었으면 D^{tc}의 문서 중에서 분류애매도가 가장 작은 문서들을 선택한다. 이때 선택된 문서들은 그들의 분류애매도가 D^{tl}에 추가할 정도의 기준 값보다 크게 되면 D^{tl}에 추가되어 다음 단계를 수행한다. 이 과정을 반복하여 충분한 문서가 추가되면 분류모델이 안정적인 상태로 진입하게 되고 동시에 D^{tl}에 추가할 정도의 기준 값(μ)보다 큰 문서가 더 이상 존재하지 않을 것이다. 그러면 반복문을 중단하고 최종적인 분류모델을 출력하게 된다.

이와 같이 EM과정에서 D^{tl} 집합을 조심스럽게 증가시키면 최적의 모델로 한층 더 가까이 가게 하는 효과를 가진다. 또한 이는 분류모델을 교란(perturbation)시키기 때문에 이전 모델이 지역 최적치(local optimum)에 빠져 있는 경우라면 이를 벗어나게 하는 효과도 가진다. 결과적으로 더 나은 분류모델을 찾기까지의 수렴속도를 증가시킨 것이다. 실제적으로 제안하는 EM가속 알고리즘의 효과가 크게 하기 위해서는 집합 D^{tl}을 확장하는 시점을 분류모델의 진화가 어느 정도 이루어진 후에 수행하는 것이 바람직하다. 이는 EM과정의 초기 시점에서는 분류모델이 불안정하여 특정 문서의 분류애매도가 상대적으로 낮다고 해서 그 문서가 분류가 옳다고 판단하기가 어렵기 때문이다.

4. 성능 분석

4.1 실험 환경

본 장에서는 제안된 알고리즘의 성능을 평가하기 위해 수행된 실험 결과를 제시한다. 시험된 사용된 데이터는 Reuters-21578로 이 문서집합은 135개의 카테고리로 레이블링 되어있는 21,578개의 문서로 구성되어 있다. 이 문서집합은 다른 문서집합에 비해 카테고리 분류결과의 오류가 적게 발생하기 때문에 문서분류 알고리즘의 성능을 평가하기 위해 자주 사용되지만[11], 카테고리당 문서 비율이 극심하게 한 쪽으로 치우쳐 있기 때문에 문서의 수가 많지 않은 카테고리에 대한 평가를 정확히 하기 어렵다. 따라서 본 실험에서는 문서 개수가 많은 카테고리인 'earn', 'acq', 'money-fx', 'grain', 'crude', 'trade', 'interest', 'ship'을 포함하여 27개의 자주 출현하는 카테고리 문서 중에서 중복된 카테고리를 갖지 않는 문서 8,407개를 선정하였다.

그리고 알고리즘이 효과적으로 수행되기 위해서는 문서분류 성능에 도움을 줄 수 있는 속성(용어)을 선정하는 것이 중요하다. 본 실험에서는 문서분류 알고리즘에서 자주 사용하는 문서빈도(document frequency : DF) 기반 방식의 속성 선정 기법을 사용하였다. 이 기법은 주어진 단어의 문서빈도 값만을 계산하여 그 값이 임계값 이상인 용어를 선택하는 것으로 작은 빈도를 갖는 용어들은 문서 분류에 큰 도움을 주지 못한다[14, 16]. 따라서 본 실험에서는 빈도수가 가장 높은 3,000개의 속성들을 선정하였다.[6]

문서분류성능은 정보검색에 보편적으로 사용하는 재현율(recall)과 정확도(precision)를 병합한 F1-측정치를 사용한다. 먼저 재현율과 정확도를 설명하기 위해 각 카테고리에 할당된 문서들을 다음과 같이 네 가지 집합으로 나눈다.

a : 해당 카테고리에 정확하게 분류된 문서의 개수
b : 해당 카테고리에 틀리게 분류된 문서의 개수
c : 해당 카테고리에 속하지만 이 카테고리에 할당되지 않은 문서의 개수
d : 해당 카테고리에 속하지 않고, 이 카테고리에 할당되지 않은 문서의 개수`

위 네 가지 집합의 크기를 사용하여 재현율과 정확도는 다음과 같이 정의한다.

$$recall = \frac{a}{a+c} \qquad precision = \frac{a}{a+b}$$

F1-측정치는 재현율과 정확도의 조화평균값으로서 아래 식과 같이 정의된다.

$$F = \frac{2 \times recall \times precision}{recall + precision}$$

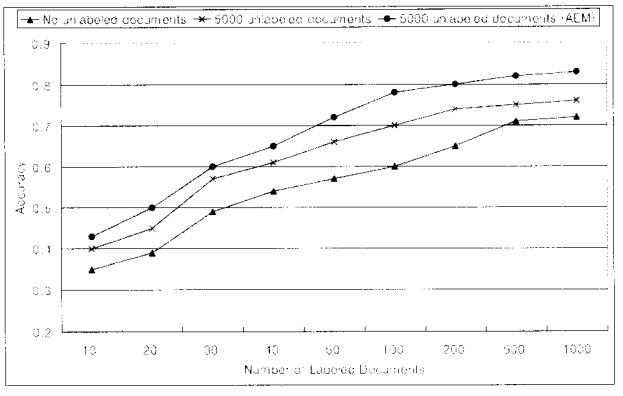
그리고 전체 카테고리에 대한 F-측정치를 얻기 위해서 카테고리 별로 F-측정치를 계산한 후 이 값들을 평균하는 Macro-averaging방식을 채택하였다.

4.2 실험 결과

본 절에서는 EM가속 알고리즘의 효과를 증명하기 위한 실험 결과를 제시한다. (그림 4)는 레이블이 있는 문서집합 $|D^{tl}|$의 크기를 변화시킬 때 전통적인 EM 알고리즘과 제안하는 EM 가속 알고리즘의 분류정확도의 변화를 보여준다.

5) 분류모델이 행상되었는지의 여부를 결정하기 위한 수식은 [10]에서 제안되었다.

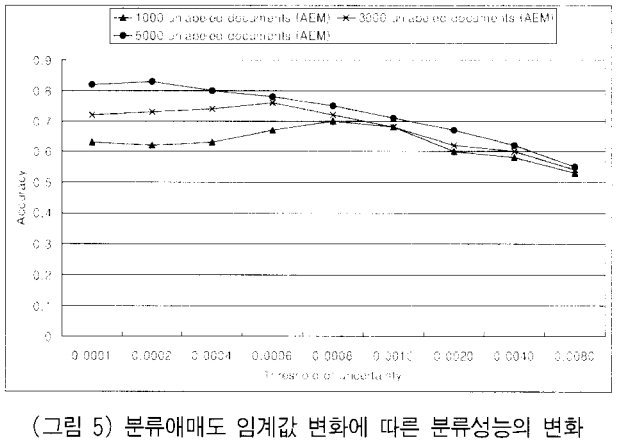
6) 이 실험에서 선정된 속성의 수(3000개)는 [16]에서 수행된 실험의 결과에 따른 것이다.



(그림 4) 학습문서집합의 크기 변화에 따른 분류성능의 변화

이미 [10]에서 알려진 바와 같이 전통 EM 알고리즘은 레이블이 없는 문서가 주어지지 않을 때보다 분류성능이 나아졌다. 그런데 전통 EM 알고리즘은 레이블이 있는 문서집합의 크기가 적정치 이상이 되면 성능향상의 정도가 감소하는 성질을 가진다. 그림에서 보는 바와 같이 $|D^{tl}|$의 크기가 500에 접근함에 따라 성능향상의 폭이 감소하고 있다. 하지만 EM가속 알고리즘의 경우 학습문서의 개수가 같은 적정치 이상이 되었음에도 불구하고 지속적으로 성능이 향상되는 것을 확인할 수 있다. 또한 초기 학습문서의 개수에 상관없이 EM가속 알고리즘의 성능이 전통 EM 알고리즘보다 대체적으로 높다. 이는 EM가속 알고리즘의 EM과정에서 분류애매도가 가장 작은 문서가 올바른 카테고리로 할당되고 있음을 말해준다.

EM가속 알고리즘의 성능을 결정할 수 있는 중요한 인자는 레이블이 없는 문서 중에서 레이블이 있는 문서집합으로 전이시키는 기준인 분류애매도의 임계값이다. (그림 5)는 분류애매도의 임계값을 증가시킬 때 EM가속 알고리즘의 분류정확도의 변화를 보여준다. 그림에서 보는 바와 같이 분류애매도 임계값이 너무 작거나 너무 큰 경우에는 분류정확도가 떨어지는 현상을 보인다. 그 임계값이 클수록 레이블이 없는 문서 중에서 학습문서집합으로 이동하는 문서의 개수가 커질 것이다. 이때 임계값이 커진다는 것은 자동 분류결과 분류애매도 값이 작지 않은 문서들이 학습문서 집합으로 옮아갈 가능성이 커짐을 의미한다. 학습문서집합으로 이동하는 문서 중에서 카테고리에 잘못 할당된 문서의 개수가 많아져 추정된 분류모델의 정확도를 떨어뜨리게 된다. 분류애매도 임계값이 너무 작은 경우에 분류성능이 떨어지는 것은 학습문서집합으로 이동하는 문서의 수가 작아지게 되어 충분히 학습문서가 확보되지 못하기 때문이다.



(그림 5) 분류애매도 임계값 변화에 따른 분류성능의 변화

5. 결 론

본 논문에서는 문서분류 시스템에서 양질의 학습 문서를 확보하는 문제를 해결하기 위해 NB 알고리즘의 프레임워크를 기반으로 수정된 EM 알고리즘을 제안하였다. 이 알고리즘은 전통적인 EM 알고리즘과는 달리 학습된 분류모델에 의해 분류된 유입문서들에 대해 선택적 샘플링 기법을 통해 현재 상태에서 최적의 학습 문서를 선택할 수 있으며, 결과적으로 초기 구축된 분류모델이 짧은 시간 내에 정확도가 높은 분류모델로 진화가 가능하다. 또한 EM 알고리즘의 M-단계에서 분류애매도를 이용하여 학습문서를 자동적으로 결정함으로써 EM 알고리즘의 수렴 속도와 성능이 개선되었다.

현재 자동분류모델은 카테고리의 계층적 구조를 고려하고 있지 않다. 그러나 NB 학습 기법은 계층적 구조를 사전 지식으로 사용할 경우에 좀 더 정확하게 문서를 분류할 수 있는 가능성을 지닌다. 따라서 향후에는 계층적 NB 학습 기법에 중점을 두고 연구를 진행할 예정이다.

참 고 문 헌

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[2] R. Aggrawal, R.J. Bayardo, and R. Srikant, "Athena: Mining-based Interactive Management of Text Databases," *Proceedings of the 7th International Conference on Extending Database Technology (EDBT 2000)*, pp.365-379, 2000.

[3] V. Castelli, and T.M. Cover, "On the Exponential Value of Labeled Samples," *Pattern Recognition Letters*, Vol.16, No.1, pp.105-111, 1995.

[4] A. P. Dempster, N. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Vol.B39, pp.1-38, 1977.

[5] E. Han, G. Karypis, and V. Kumar, "Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification," *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '91)*, pp.53-65, 1991.

[6] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," *Proceedings of the 10th European Conference on Machine Learning*

(ECML '98), pp.137-142, 1998.

[7] D. Johnson and S. Sinanovic, "Symmetrizing the Kullback-Leibler Distance", *submitted to IEEE Transactions on Information Theory*, 2001.

[8] M. Lindenbaum, S. Markovitch, and D. Rusakov, "Selective Sampling for Nearest Neighbor Classifiers," *Proceedings of the 16th National Conference on Artificial* Intelligence (AAAI '99), pp.366-371, 1999.

[9] T. M. Mitchell, "Bayesian Learning," *Machine Learning*, McGraw-Hill, New York, pp.154-200, 1997.

[10] K. Nigam. Using Unlabeled Data to Improve Text Classification. PhD thesis, Carnegie Mellon University, 2001.

[11] D. D Lewis, "Reuters-21578 text categorization test collection," *http://www.daviddlewis.com/resources/testcollections/reuters21578/*, 1997.

[12] L. Ralaivola and F. d'Alché-Buc, "Incremental Support Vector Machine Learning: A Local Approach", *Lecture Notes in Computer Science*, Vol.2130, pp.322-328, 2001.

[13] J. Rennie, L. Shih, J. Teevan and D. Karger, "Tackling the poor assumptions of Naive Bayes text classifiers", *Proceedings of the 20th International Conference on Machine Learning (ICML-2003)*, pp.616-623, 2003.

[14] M. Sahami, S. Yusufali, and M.Q. Baldonado, "SONIA: A Service for Organizing Networked Information Autonomously," *Proceedings of ACM Conference on Digital Library (ADL '98)*, pp.200-209, 1998.

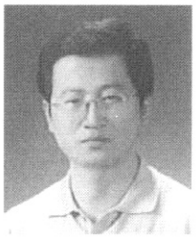
[15] K.M. Schneider, "Techniques for Improving the Performance of Naive Bayes for Text Classification", *Proceedings of the 6th International Conference on Intelligent Text Processingand Computational Linguistics (CICLing-2005)*, pp.682-693, 2005.

[16] Y. Yang, and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *Proceedings of the 14th International Conference of Machine Learning (ICML '97)*, pp.412-420, 1997.

장 재 영

e-mail : jychang@hansung.ac.kr
1992년 서울대학교 계산통계학과(학사)
1994년 서울대학교 계산통계학과(석사)
1999년 서울대학교 계산통계학과(박사)
2000년 3월~현재 한성대학교 컴퓨터공학과 조교수
관심분야 : 데이터베이스, 데이터마이닝

김 한 준



e-mail : khj@uos.ac.kr
1994년 서울대학교 계산통계학과(학사)
1996년 서울대학교 전산과학과(석사)
2002년 서울대학교 컴퓨터공학부(박사)
2002년 2월~2002년 12월 서울대학교 공과대학 Post-Doctor
2002년 12월~현재 서울시립대학교 전자전기컴퓨터공학부 조교수
관심분야 : 데이터베이스, 데이터마이닝, 정보검색