다항시행접근 단순 베이지안 문서분류기의 개선

(Improving Multinomial Naive Bayes Text Classifier)

김 상 범 † 임 해 창 ††
(Sang-Bum Kim) (Hae-Chang Rim)

요 약 단순 베이지언 분류모형은 구현이 간단하고 효율적이기 때문에 실용적으로 사용하기에 적합하다. 그러나 이 분류모형은 많은 기계학습 도메인에서 우수한 성능을 보임에도 불구하고 문서분류에 적용되었을 경우에는 그 성능이 매우 낮은 것으로 알려져왔다. 본 논문에서는 단순 베이지언 분류모형 중 가장 성능이 우수한 것으로 알려진 다항시행접근 단순 베이지언 분류모형을 개선하는 세가지 방법을 제안한다. 첫 번째는 범주에 대한 단어의 확률추정방법을 문서모델에 기반하여 개선하는 것이고, 두 번째는 문서의 길이에 따라 범주와의 관련성이 선형적으로 증가하는 것을 억제하기 위해 길이에 대한 정규화를 수행하는 것이며, 마지막으로 범주판정에 중요한 역할을 하는 단어들의 영향력을 높여주기 위하여 상호정보가중 단순 베이지언 분류방법을 사용하는 것이다. 제안하는 방법들은 문서분류기의 성능 평가를 위한 벤치마크 문서집합인 Reuters21578과 20Newsgroup에서 기존의 방법에 비해 상당한 성능향상을 가져옴을 알 수 있었다.

키워드 : 정보검색, 문서분류, 단순 베이지안 학습

Abstract Though naive Bayes text classifiers are widely used because of its simplicity, the techniques for improving performances of these classifiers have been rarely studied.

In this paper, we propose and evaluate some general and effective techniques for improving performance of the naive Bayes text classifier. We suggest document model based parameter estimation and document length normalization to alleviate the problems in the traditional multinomial approach for text classification. In addition, Mutual-Information-weighted naive Bayes text classifier is proposed to increase the effect of highly informative words.

Our techniques are evaluated on the Reuters21578 and 20 Newsgroups collections, and significant improvements are obtained over the existing multinomial naive Bayes approach.

Key words : information retrieval, text classification, naive Bayes learning

1. 서 론

자동문서범주화(automatic text categorization)은 자연어로 이루어진 문서에 미리 정의된 여러개의 범주들 중 하나 혹은 그 이상을 할당하는 작업이다. 이는 디지털형태로 저장된 텍스트 문서들이 방대하게 늘어남에 따라, 이러한 정보들을 효율적으로 관리하고 검색하기 위하여 그 중요성이 크게 부각되었고 활발히 연구되는 분야이다. 자동문서범주화와 유사한 작업으로 문서 필터링(filtering)이나 라우팅(routing)등이 있는데, 이 모든 작업들은 주어진 하나 혹은 다수의 문서들 중 어떠한 문서가 사용자의 요구에 더욱 적합한가, 혹은 주어진 문서가 어떠한 범주에 더욱 적합한가를 판정하는 문제로 귀결된다. 따라서, 이러한 작업에 사용되는 문서분류기는 문서가 주어졌을 때 범주에 대한 순위화와 범주가 주어졌을 때 문서의 순위화를 잘 할 수 있는 능력이 요구된다. 이를 위하여 최근 수년간 기계학습분야에서 연구되어온 최근린법(kNN)[1], 지지벡터기계(SVM)[2], 단순 베이즈 분류기[3], 결정트리[4] 등 여러 가지 통계적인 학습 방법들이 이 문제에 적용되어 왔다.

이중 단순 베이즈 문서분류기는 구현이 용이하고 기

† 학생회원 : 고려대학교 컴퓨터학과
sbkim@nlp.korea.ac.kr
†† 종신회원 : 고려대학교 컴퓨터학과 교수
rim@nlp.korea.ac.kr
논문접수 : 2002년 5월 8일
심사완료 : 2002년 12월 11일

존의 검색엔진에 쉽게 적용될 수 있다는 장점으로 인하여 문서분류시스템을 구축하는데 많이 사용되어 왔으나, 다른 학습방법에 비하여 비교적 낮은 성능을 보인다고 알려져왔다[5]. 그러나 단순 베이즈 분류기법은 문서분류를 제외한 많은 문제영역에서 매우 좋은 성능을 보여주어왔기 때문에[6], 이 방법이 문서분류에서 좋은 성능을 보여주지 못해왔던 이유를 밝히는 것은 대단히 의미있는 일이다. 본 논문에서는 기존의 단순 베이즈 방법, 특히 다항분포기반 단순 베이즈 문서분류 접근방법이 갖고 있던 문제를 지적하고 이를 완화할 수 있는 몇가지 기법들을 제안하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 간단하게 지금까지 사용되어왔던 두가지 단순 베이즈 문서분류 모델에 대해 간단히 알아보고 그 문제점을 지적한다. 3장에서는 지적된 문제들을 완화시킬 수 있는 몇가지 기법들에 대해 설명한다. 4장에서는 실험을 통하여 제안하는 방법이 매우 효과적임을 입증할 것이며 5장에서는 결론 및 향후연구로 끝을 맺고자 한다.

2. 단순 베이즈 문서 분류 모형

단순 베이즈 분류방법은 전통적으로 많이 연구가 되어온 확률기반 분류방법이다. 이 방법은 베이지언 확률이론에 그 이론적 근거를 두고 있으며, 특정 클래스가 주어졌을 때 객체를 표현하는 각 속성은 다른 속성들과 조건부 독립을 이룬다는 가정을 통하여 모델을 단순화함으로써 매우 실용적으로 많은 문제에 적용되어왔다. [6]는 여러 실험 영역에서 이 방법이 복잡도가 높은 다른 많은 학습방법들에 비해 뒤지지 않는 성능을 보인다고 밝힌바 있다.

단순 베이즈 분류방법이 문서분류에 적용될 경우, 목표는 어떤 문서 D가 주어졌을 때 그 문서가 특정 범주 c에 속한 문서일 확률을 구하는 것으로, 그 확률은 베이지언 공식에 따라 다음과 같이 전개된다.

$$\mathrm{Rel}(c, D) = P(c \mid D) = \frac{P(D \mid c)P(c)}{P(D)} \tag{1}$$

한편, 서론에서 밝혔듯이 우리가 관심있는 것은 문서가 주어졌을 때 적합한 범주들을 순위화 하거나 혹은 범주가 주어졌을 때 적합한 문서들을 순위화 하는 것이므로, 실제 확률값을 구하지 못하더라도 이러한 순위가 유지될 수 있다면 수식을 변형시켜 계산상의 편의를 도모할 수가 있다. 따라서, $\mathrm{Rel}(c,D)$ 는 다음과 같이 순서가 유지되는 방법으로 전개될 수 있다.

$$\begin{aligned}\mathrm{Rel}(c, D) &= \log \frac{P(c \mid D)}{P(\bar{c} \mid D)} \\ &= \log \frac{P(D \mid c)}{P(D \mid \bar{c})} + \log \frac{P(c)}{P(\bar{c})}\end{aligned} \tag{2}$$

수식 (2)에서 $\bar{c}$ 는 'c범주가 아님'을 의미하며, 두 번째 항의 경우는 클래스의 사전확률을 의미하는 것으로 많은 경우 상수로 취급될 수 있는 값이다. 따라서 중요한 것은 수식(2)의 첫 번째 항(이하 문서범주로그비)을 추정하는 일이다. 문서범주로그비를 계산하는 방법은 문서를 바라보는 관점에 따라 달라지게 되는데, 크게 다변량베르누이시행 접근(multivariate Bernoulli trial approach)와 다항시행접근(multinomial trial approach)으로 나눌 수 있다.

다변량베르누이시행 접근에서는 문서라는 객체가 사전에 등재된 모든 어휘의 문서내 출현에 대한 베르누이 실험결과로 특징지어지는 대상이라 간주한다. 다시 말하면, 문서는 차원이 자질(feature)어휘집합의 크기 |V|인 벡터로 구성되어 있고 벡터의 각 원소는 자질어휘집합의 특정 단어가 문서에 출현했는지의 여부를 값으로 갖는다고 할 수 있다. 이러한 관점은 확률기반 정보검색을 위해 오래전부터 사용되어온 이진독립모형[7]과 동일한 것이라 할 수 있다. 이 방법에서 문서범주로그비는 독립가정에 의하여 다음과 같이 구할 수 있다. 즉, 문서 D가 자질어휘집합에 속해있는 단어 w_i로 구성되어 있다면,

$$\begin{aligned}&\log \frac{P(D \mid c)}{P(D \mid \bar{c})} \\ &= \log \frac{\prod_{i=1}^{|V|} P(w_i \mid c)^{g(w_i, D)} P(\overline{w_i} \mid c)^{1-g(w_i, D)}}{\prod_{i=1}^{|V|} P(w_i \mid \bar{c})^{g(w_i, D)} P(\overline{w_i} \mid \bar{c})^{1-g(w_i, D)}} \\ &\simeq \log \frac{\prod_{i=1}^{|V|} P(w_i \mid c)^{g(w_i, D)} P(\overline{w_i} \mid c)^{1-g(w_i, D)}}{\prod_{i=1}^{|V|} P(w_i \mid \bar{c})^{g(w_i, D)} P(\overline{w_i} \mid \bar{c})^{1-g(w_i, D)}} - \log \frac{\prod_{i=1}^{|V|} P(\overline{w_i} \mid c)}{\prod_{i=1}^{|V|} P(\overline{w_i} \mid \bar{c})} \\ &= \sum_{i=1, w_i \in D}^{|V|} \log \frac{P(w_i \mid c)(1 - P(w_i \mid \bar{c}))}{P(w_i \mid \bar{c})(1 - P(w_i \mid c))}\end{aligned} \tag{3}$$

where, $g(w_i, D) = 1$ if $w_i \in d_k$, otherwise $g(w_i, D) = 0$

한편, $P(w_i|c)$ 는 확률이 0이 되는 것을 방지하는 가산적 평탄화(additive smoothing) 요소 θ를 사용하여 다음과 같이 추정하며 일반적으로 θ 값은 0.5를 사용한다.

$$P(w_i|c) = \frac{\sum_{k=1}^{|D|} f(d_k, c) g(w_i, d_k) + \theta}{\sum_{k=1}^{|D|} f(d_k, c) + 2\theta}$$

where, $f(d_k, c) = 1$ if c is a category of d_k, otherwise $f(d_k, c) = 0$

수식 (3)의 두 번째 라인과 같이 식을 변형할 수 있는 것은 빼주는 부분이 문서에 상관없이 모두 동일한 상수 값이므로 가능하며, 결국 문서에 존재하는 단어들만을 고려할 수 있게 수식이 변형된다. 즉, 이렇게 해주어도 클래스와 문서사이의 관련성에 대한 순위화에는

영향을 미치지 않는다. [7]에서는 이를 순서유지변형(order-preserving transformation)이라 하여 실제로 확률모델을 정보검색에 사용할 수 있도록 적절히 이용하였다.

다변량 베르누이시행 접근방법의 문제점은 문서내 단어빈도를 고려할 수 없다는데 있다. 즉, 어휘의 문서내 출현여부만을 고려함으로서 어휘들이 문서내에서 차지하는 중요도 정보를 사용할 수 없다는 것이다.

그에 비해 다항시행 접근방법은 문서를 어휘가 출현하는 사건의 연속적인 발생열로 간주하는데, 이때 어휘가 출현하는 사건은 어휘사전과 각 어휘의 발생확률로 정의된 모델에 의해 일어난다. 이 모델에 의하면, 발생확률이 $p_1, p_2, \cdots, p_{|V|}$ 인 어휘 $w_1, w_2, \cdots, w_{|V|}$가 tf_1, tf_2, ... , $tf_{|V|}$ 회 출현한 문서 D가 존재할 확률은 다음과 같이 구해지므로,

$$P(D) = |D|! \prod_{i=1}^{|V|} \frac{p_i^{tf_i}}{tf_i!}$$

다항시행 접근방법에 의한 문서범주로그비는 아래와 같이 계산할 수 있다.

$$\log \frac{P(D|c)}{P(D|\bar{c})} = \log \frac{|D|! \prod_{i=1}^{|V|} \frac{P(w_i|c)^{tf_i}}{tf_i!}}{|D|! \prod_{i=1}^{|V|} \frac{P(w_i|\bar{c})^{tf_i}}{tf_i!}} = \sum_{i:w_i \in D} tf_i \cdot \log \frac{P(w_i|c)}{P(w_i|\bar{c})} \quad (4)$$

한편, $P(w_i|c)$는 유니그램 언어모델에서의 파라미터 추정과 유사하게 다음과 같이 계산한다.

$$P(w_i|c) = \frac{\theta + \sum_{k=1}^{|D|} g(d_k, c) \cdot tf_{ki}}{\theta \cdot |V| + \sum_{k=1}^{|D|} g(d_k, c) \cdot dl_k} \quad (5)$$

수식 (5)에서 tf_{ki}는 문서 k에서 단어 i가 나타난 회수를 의미하고, dl_k는 문서 k에 있는 총 토큰수를 의미한다. [3]는 4개의 실험집합에서 단순 베이즈 문서분류를 위한 다변량 베르누이시행 접근과 다항시행 접근방법을 적용하였는데, 문서내 단어빈도를 고려하는 다항시행 접근방법이 다변량 베르누이시행 접근방법에 비하여 좋은 성능을 보여주었다.

그러나 다항시행 접근방법은 두 가지 문제점을 갖고 있다. 첫 번째 문제점은 파라미터 추정에 있어서 문서의 경계를 고려하지 않는다는 점이다. 이러한 파라미터 추정은 그림 1과 같은 문제를 야기한다. 그림 1에서 두 개의 학습문서에 나타난 세 어휘에 대해 기존의 방법은 범주가 주어졌을 때 어휘의 출현확률을 모두 동일하게 추정하고 있다. 그러나 두 개의 문서가 동일한 정도로 범주 c에 대한 내용을 다루고 있다고 가정한다면, 길이가 긴 문서에서 1회 출현한 것과 길이가 짧은 문서에서 1회 출현한 것에 같은 비중을 두는 것은 직관적으로 옳지 않다. 왜냐하면 길이가 짧은 문서에 사용된 어휘가 긴 문서에서 사용된 어휘보다 범주 c의 내용을 잘 나타내고 있을 가능성이 크기 때문이다.

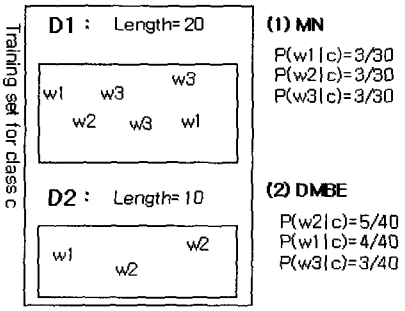

그림 1 기존의 파라미터 추정방법(MN)과 제안하는 방법(DMBE)의 실례

두 번째 문제점은 수식 (4)에서 나타나듯 어떤 문서와 범주와의 관련성이 문서내 단어 빈도 tf에 선형적으로 비례하고 있다는 것이다. 이는 문서가 고정되어 있는 상황에서 범주들을 비교하는 태스크의 경우에는 문제가 되지 않지만, 범주가 고정되어 있을 때 여러 문서들이 그 범주와 갖는 관련성을 비교하는 태스크의 경우 길이가 긴 문서를 지나치게 선호할 수 있다. 즉, 문서 길이에 대한 정규화가 이루어지지 않는다는 것이다.

본 논문에서는 위에서 언급한 두 가지 문제를 완화할 수 있는 두 가지 방법을 제시하고 이를 다항시행 단순 베이즈 문서분류기에 적용한 뒤, 문서범주화에서 널리 사용되는 두 개의 벤치마크 실험집합에 대해 평가를 수행하도록 한다.

3. 제안하는 방법

3.1 문서모델에 기반한 파라미터 추정

기존의 다항시행접근 단순 베이즈 문서분류기에서는 어떤 범주에서 어휘가 출현할 확률을 구할 때 그 범주에 속한 학습문서집합을 하나의 거대한 문서로 보고 확률을 추정한다. 이는 유니그램 언어모델(unigram language model)에서 어휘의 확률을 구하는 일반적인 방법과 동일하다. 따라서 여러 개의 학습 문서로 구성된 말뭉치에서 확률을 추정할 때 길이가 긴 문서가 짧은 문서에 비해 확률 추정에 상대적으로 많이 관여하는 결과를 가져온다.

본 논문에서 제안할 첫 번째 가설은 학습문서에서 범주모델의 파라미터를 추정함에 있어서 각 문서는 동일한 정도로 범주모델의 파라미터를 추정하는 일에 참여해야한다는 것이다. 즉, 문서분류기의 학습을 위해 제공

된 학습문서들은 그 길이에 상관없이 모두 동일한 정도로 중요하다고 가정하기 때문이다.

이 가설에 따라 제안하는 방법에서는 문서모델에 기반한 파라미터 추정(DMBE:Document Model Based Estimation)을 사용한다. 이 방법은 해당 범주에 속해 있는 각 학습문서에서의 단어 출현확률을 구한 뒤 평균을 계산함으로써 그 범주에서 단어가 출현할 확률을 구하는 방식을 사용한다. 그림 1은 제안하는 방법에 의한 확률의 추정 예를 기존의 방법과 함께 보여주고 있는데, 결과적으로 전체 문서집합에서의 출현빈도는 다른 단어들과 동일하지만 길이가 짧은 문서에서 상대적으로 많이 출현한 w2가 높은 확률값을 갖게 됨을 알 수 있다. 이러한 결과는 직관적으로도 납득하기 쉽다. 왜냐하면 길이가 짧은 문서에는 어떤 내용을 표현하기 위해 불필요한 단어들 없이 중요한 단어들만을 주로 사용했다고 생각할 수 있고, 이러한 단어들이 특정 범주에서 출현할 확률을 높여주는 것은 타당하기 때문이다.

다른 관점에서 생각하면, 우선 어떠한 범주 모델(topic model)이 존재하고 각 학습문서들은 이 모델에 의해 생성되었으며, 범주 모델의 파라미터를 추정하기 위해 그 모델에서 생성된 학습문서들 각각에서 단어 출현 확률을 구하고 이의 평균값을 범주 모델의 파라미터로 사용한다는 것이다. 즉, 기존의 수식(5)는 다음 수식(6)과 같이 수정된다.

$$P(w_i|c)=\frac{\sum_{k=1}^{|D|} g(d_k,c)\cdot P(w_i|d_k)}{\sum_{k=1}^{|D|} g(d_k,c)} \quad (6)$$

수식(6)에서 $P(w_i|d_k)$ 는 다음과 같이 계산한다.

$$P(w_i|d_k)=\begin{cases}\dfrac{\theta+tf_{ik}}{\theta\cdot|V|+dl_k} & \text{if } tf_{ik}>0\\ \dfrac{\theta}{\theta\cdot|V|+\text{avdl}} & \text{if } tf_{ik}=0\end{cases} \quad (7)$$

수식(7)은 제안하는 방법이 단순한 평탄화기법을 사용하여 실제로 특정 단어가 나타나지 않은 문서에도 작은 양의 단어빈도를 제공하고 있다는 것을 의미한다. 다만, tf_{ik}가 0인 경우 문서길이 dl_k가 아닌 문서집합에서의 평균 문서길이 avdl(average document length) 값을 사용하는 이유는 실제 구현상의 효율성을 위함이다.

제안하는 추정방법의 특징은 길이가 짧은 문서에서 나온 단어를 강조한다는 것 이외에도 다항시행접근 문서분류기에서 일종의 "가중치"로 작용하는 단어별 로그오즈값 $\left(\log\frac{P(w_i|c)}{P(w_i|\overline{c})}\right)$이 그림 2와 같이 안정된 분포를 보인다는 것이다. 그림 2는 Reuters21578 문서집합에서의 alum 범주에 대하여 각 로그오즈값 범위에 속하는 단어들의 개수를 나타내고 있는데, 기존의 방법으로 추정을 수행할 경우에 비해 제안하는 방법으로 추정을 할 때 평균 로그오즈값을 중심으로 높은 값과 낮은 값을 갖는 단어들의 개수가 균형을 이루는 것을 알 수 있다.

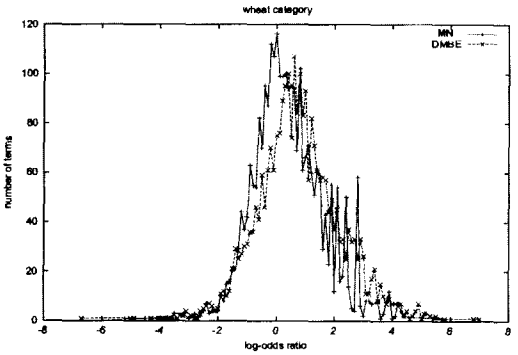

그림 2 Reuters21578 문서집합의 "wheat" 범주에 대한 기존의 파라미터 추정방법(MN)과 제안하는 파라미터 추정방법(DMBE)에 따른 로그오즈비값의 분포

3.2 문서 길이 정규화

다항시행접근 단순 베이즈 분류기에서는 문서범주로 그비를 사용할 경우 문서길이에 대한 정규화가 전혀 이루어지지 않아 길이가 긴 문서에 대해 지나치게 높은 값을 할당해주는 결과를 초래한다[8]. 예를 들어 어떤 문서 d와 범주 c와의 관련성이 Rel(c,d) 라 할때 만일 동일한 문서 d를 n개 이어놓은 새로운 문서 d'을 생각한다면, d'과 범주 c와의 관련성은 n×Rel(c,d)가 되어 문서의 길이에 선형적으로 비례하게 됨을 알 수 있다.

본 논문의 두 번째 가설은 문서와 범주의 관련성이 문서의 길이에 선형적으로 비례해서는 안된다는 것이다. 이는 길이가 긴 문서의 경우 동일한 내용을 말하면서 좀더 많은 단어를 사용한다는 용장가설(verbosity hypothesis)[10]이 실제 문서에 어느정도는 적용된다는 것을 의미하는 것으로, 길이가 길다고 해서 더 많은 내용을 포함하고 있다고 단언할 수는 없다는 것이다.

두번째 가설에 따라 제안하는 방법에서는 문서 길이 정규화 방법을 사용한다. 문서의 길이를 정규화 함은 실제 문서내 단어빈도(tf)를 정규화 요소로 나누어줌을 의미하는데, 해당 문서의 길이와 전체 문서들의 평균 길이를 결합한 방식을 취하는 것이 일반적이다. 이때 결합하

는 방식에는 크게 두 가지 방법이 있을 수 있다. 첫 번째 방법은 해당 문서의 길이와 전체 문서들의 평균길이를 결합파라미터 α를 사용하여 선형적으로 결합하는 방법으로, 문서 k의 정규화 요소는 다음과 같이 계산된다.

$$NF_k = \alpha \cdot dl_k + (1-\alpha) \cdot avdl \quad (8)$$

다른 하나의 방법은 이 두 요소를 지수적으로 결합하는 방법으로, 문서 k의 정규화 요소는 다음과 같이 계산된다.

$$NF_k = dl_k{}^{\alpha} \cdot avdl^{1-\alpha} \quad (9)$$

선형적인 결합방법은 기존의 정보검색연구에서 많이 사용되었던 방법이나[9,10], 지수적인 결합방법은 본 논문에서 제안하는 방법이다. 두 방법 모두 α의 값에 따라 정규화의 정도를 조정할 수 있고 그 정도도 비슷하긴 하지만, 지수적 결합방법의 경우 곱으로 결합되어 있기 때문에 범주에 대해 문서들의 순위화를 수행하는 경우 $avdl^{1-\alpha}$항이 상수로 취급되어, $NF_k = dl^{\alpha}$로 단순화할 수 있다는 장점이 있다. 또한 지수적인 결합에 의한 정규화를 수행할 경우 앞에서 언급했던 d'(=문서 d를 n개 붙여놓은 것) 예제에 있어 관련성이 $n^{1-\alpha} \cdot \mathrm{Rel}(d, c)$가 됨을 알 수 있는데, 이는 정규화 요소와 파라미터 α에 의한 분류기의 동작변화를 직관적으로 이해할 수 있도록 해 준다는 장점이 있다. 실험 결과, α가 약 0.8일 때 성능이 가장 좋았는데, 이는 만일 동일한 문서 2개를 이어놓았다면 그 문서와 어떤 범주 c와의 관련성은 원래 문서와 C와의 관련성에 비해서 약 $2^{0.2} \approx 1.15$ 배정도가 된다는 것을 알 수 있다.

3.3 상호정보 가중 단순 베이즈 분류기

문서분류시스템을 구축하기 위하여 많은 실험에서 자질추출을 수행해왔다[11]. 자질추출을 하는 이유는 문서 벡터공간을 축소하여 학습의 시간적 공간적 효율성을 도모할 수 있을 뿐 아니라 유용한 자질만을 사용하고 불필요한 단어들은 배제함으로서 정확한 분류기 학습이 이루어 질 수 있을 것이라는 기대 때문이다. 그러나 아직 전체 단어들 중 자질단어집합만을 사용하여 학습을 수행하였을 때 성능향상이 보장될 수 있다는 연구결과는 없다. 오히려 모든 단어를 사용할 경우 좋은 성능을 보인다는 실험결과가 나온바 있다[2]. 또한 미리 자질추출을 수행하는 접근 자체가 매우 비효율적인 경우가 있는데, 지속적으로 새로운 자질들이 추가되어야 하는 시스템이 그 예이다. 이러한 시스템에서 미리 자질추출을 수행하는 것이 비효율적인 이유는 기존의 자질집합을 사용하여 구축된 분류기가 다시 재구축되어야 하기 때문이다.

본 논문에서는 미리 자질추출을 하는 것이 아니라 분류할 시점에서의 말뭉치 통계정보를 이용하여 자질단어의 중요성을 반영할 수 있도록 다항시행접근 단순 베이즈 분류기를 수정한 가중 단순 베이즈(Weighted Naive Bayes) 분류기를 제안한다. 즉, 수식(4)를 다음과 같이 수정한다.

$$\log \frac{P(D \mid c)}{P(D \mid \bar{c})} = \log \frac{|D|! \prod_{i=1}^{|V|} \frac{(P(w_i \mid c)^{fw_{ic}})^{tf_i}}{tf_i!}}{|D|! \prod_{i=1}^{|V|} \frac{(P(w_i \mid \bar{c})^{fw_{ic}})^{tf_i}}{tf_i!}} = \sum_{i=1, i \in D}^{|V|} fw_{ic} \cdot tf_i \cdot \log \frac{P(w_i \mid c)}{P(w_i \mid \bar{c})} \quad (8)$$

위 수식에서 fw_i는 자질단어 w_i의 가중치로 정의된다.

가중 단순 베이즈 분류의 입장에서, 기존의 자질추출을 미리 수행하는 단순 베이즈 분류기는 이진가중(binary weighted) 단순 베이즈 분류기의 사용으로 해석될 수 있다. 즉, 자질단어일 경우 $fw_i = 1$ 그렇지 않으면 $fw_i = 0$ 인 가중 단순 베이즈 분류기이기 때문이다.

본 논문에서는 이 가중치를 상호정보의 절대값을 사용한다. 절대값을 사용하는 이유는 음의 상호정보를 갖는 단어도 그 절대값이 크다면 문서분류에 적절한 어휘가 될 수 있기 때문이다. 한편, 일반적으로 상호정보는 자질추출과정에서 사용되는데 단어와 범주에 대한 문서의 개수를 상호정보계산에 이용하나[11], 본 논문은 문서에 대한 관점이 다항시행접근 방식이므로 상호정보도 단어의 출현 회수를 사용하여 [3]과 유사한 방법으로 다음과 같이 fw_i를 계산한다.

$$fw_{ic} = \left| \log \frac{P(w_i, c)}{P(w_i)P(c)} \right| = \left| \log \frac{\left(\sum_{i=1}^{|V|} \sum_{k=1}^{|D|} tf_{ik} \right) \left(\sum_{k=1}^{|D|} g(d_k, c) \cdot tf_{ik} \right)}{\left(\sum_{k=1}^{|D|} tf_{ik} \right) \left(\sum_{i=1}^{|V|} \sum_{k=1}^{|D|} g(d_k, c) \cdot tf_{ik} \right)} \right|$$

4. 실험 및 결과

제안하는 문서분류기의 성능을 평가하기 위하여 본 논문에서는 문서분류기 성능평가에 널리 사용되는 Reuters21578과 20Newsgroup 벤치마크 문서집합을 사용하였다. Reuters21578은 1987년에서 1991년까지 보도된 2만여개의 로이터 신문 기사들로 구성되어 있으며 여러 형태의 버전으로 나누어 사용할 수 있는데, 본 논문에서는 "ModApte" 분리 버전을 사용하다. 이 버전은 학습문서 9,603개와, 테스트문서는 3,299개로 구성되어 있다. 한편 총 135개의 범주중 적어도 한 개 이상의 학습문서와 테스트문서에 할당되어 있는 90개의 범주에 대해서 분류기의 성능을 평가하였다. Reuters21578은 범주할당

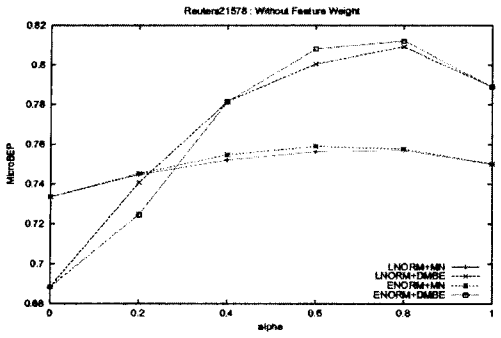

그림 3 자질가중치를 적용하지 않았을 때 정규화파라미터값에 따른 분류기의 성능변화 비교

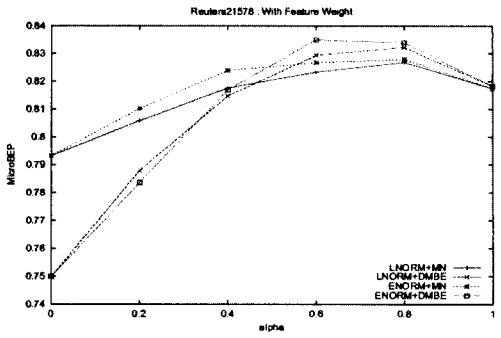

그림 4 자질가중치를 적용했을때 정규화파라미터값에 따른 분류기의 성능변화 비교

이 매우 불균형하여 고빈도 상위 10개 범주가 전체 문서집합의 71.2%에 할당되어 있다. 따라서 Reuters21578에 대한 문서분류기의 성능평가 실험에는 일반적으로 마이크로 평균 손익분기점(micro-average break even point)를 사용한다[5]. 즉, 각 범주별로 정확률(precision)과 재현율(recall)이 같아지는 손익분기점을 계산한 뒤 이를 각 범주에 속한 문서들의 개수에 대하여 가중치를 두어 평균을 내는 것이다.

20Newsgroup은 19,997개의 유즈넷 기사들로 구성되어 있으며 이 기사들은 20개의 서로 다른 뉴스그룹에서 모은 것이다. 따라서 각 기사에는 단 한 개의 범주가 할당되어 있으며 범주별 기사들의 개수도 대략 1000개씩으로 거의 일정하다. 따라서 이 문서집합에 대한 성능평가는 기계학습 태스크에서 일반적으로 사용하는 정확도(Accuracy)로 수행하였다. 즉, 각 문서에 대하여 범주들을 랭킹한 뒤 가장 높은 관련성을 갖는 범주를 할당하고 이를 평가하는 것이다.

그림 3과 그림 4는 자질가중치를 적용하지 않은 경우와 적용했을 경우, 정규화 파라미터값에 따른 문서분류기들의 성능을 보여주고 있다. MN과 DMBE는 기존의 파라미터 추정방법과 본 논문에서 제안하는 문서모델기반 파라미터 추정방법을 의미하고 LNORM과 ENORM은 문서정규화를 위한 선형적 결합방법과 지수적인 결합방법을 의미한다.

α값이 작은 경우, 즉 정규화를 거의 수행하지 않을 경우에는 기존의 추정방법이 좋은 성능을 보여주나 0.6에서 0.8정도의 값에서는 제안하는 DMBE가 좋은 성능을 보이며 이때 분류기가 최상의 성능을 보여줌을 알 수 있다. 한편, ENORM이 LNORM에 비해서 다소 좋은 성능을 보여주나 그다지 의미있는 차이는 없었다.

위 실험에서 주목할만한 것은 정규화를 수행하지 않았을 경우 DMBE 방법의 성능이 매우 저하가 된다는 사실이다. 그 이유는 3장에서 언급했던 각 단어들의 로그오즈값 분포에서 찾을 수 있다. 그림 2를 보면 제안하는 방법에 비해 기존의 추정방법에서 음의 값을 갖는 단어들이 많음을 알 수 있는데, 이러한 특징으로 인하여 기존의 추정방법에 의한 분류기는 문서의 길이가 길어짐에 따라 범주와 문서의 관련성이 지나치게 증가하는 현상을 억제할 수 있게 된다. 따라서 문서길이 정규화를 하지 않더라도 ($\alpha=0$인 경우) 그다지 성능의 저하가 일어나지 않는 것이다.

그림 5는 Reuters21578 문서집합의 wheat 범주에 대하여 정규화를 수행하지 않은 MN과 DMBE의 성능을 보여주고 있는데, 문서의 길이가 길어짐에 따라 DMBE의 경우 범주와의 관련성이 급증하는 반면 MN은 그렇

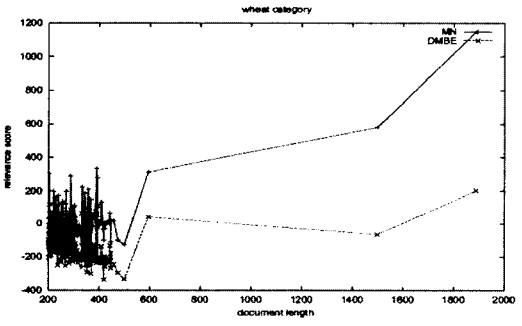

그림 5 Reuters21578 문서집합의 "wheat" 범주에 대한 기존의 파라미터 추정방법(MN)과 제안하는 파라미터 추정방법(DMBE)에 따른 문서길이별 연관성 점수값의 분포

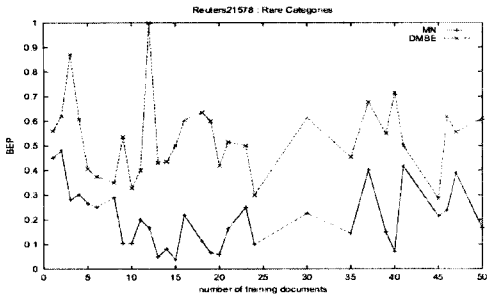

그림 6 학습문서개수가 적은 범주에서의 기존 파라미터추정방법(MN)과 제안하는 방법(DMBE)에 의한 분류기 성능

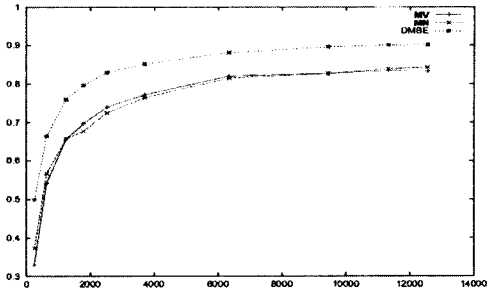

그림 7 20Newsgroup 문서집합에서 학습문서의 개수에 따른 다변량접근방법(MV), 다항시행접근방법(MN), 제안하는 방법(DMB E)에 기반한 각 분류기의 성능변화

지 않은 것을 알 수 있다. 따라서 길이 정규화를 수행하지 않을 경우 DMBE의 성능은 매우 저하되나 적절한 길이 정규화를 수행할 경우 그림 3과 그림 4에서 볼 수 있는바와 같이 문서분류기의 성능을 극대화 할 수 있다는 결론을 내릴 수 있다.

그림 6은 Reuters21578에서 학습문서의 개수가 상대적으로 적은 범주들에 대한 문서분류기의 성능을 나타낸 것인데, 이 실험에서 DMBE는 MN의 성능에 비해 월등히 우수한 것을 알 수 있다. 즉, 학습문서의 개수가 많지 않을 때도 문서모델에 기반한 파라미터 추정방법은 안정적으로 기존 방법에 비해 우수한 성능을 보이게 함을 알 수 있다. 이러한 현상은 20Newsgroup 에 대한 실험에서도 마찬가지였다. 그림 7은 학습문서의 개수를 늘려가면서 20Newsgroup 에 대한 실험결과인데 학습문서의 개수가 매우 적은 경우에도 DMBE의 성능이 MN의 성능이나 2장에서 언급했던 다변량접근 분류기(MV)의 성능보다 훨씬 높은 것을 알 수 있다.

마지막으로 표1 과 표2는 Reuters21578과 20Newsgroup에서의 성능을 요약해 놓은 것이다. 표1에서 10cat.은 주요 10개의 범주에 대한 실험결과를 마이크로 평균낸 것이고 90cat.은 전체 범주에 대한 실험결과를 마이크로 평균낸 것이다. 표2에서는 길이정규화에 대한 실험이 없는데, 이는 20Newsgroup의 경우 문서가 고정되었을 때 가장 높은 관련성을 갖는 범주를 할당하기 때문에 문서간의 비교가 없어 길이정규화가 의미가 없기 때문이다. 대신, 이 실험에서는 2장에서 언급했던 다변량접근 단순 베이지언 분류기의 실험결과(MV)를 함께 실었다. Reuters21578의 경우 제안하는 파라미터 추정방법과 정규화 방법을 모두 사용했을 때 10 cat.의 경우 약 4%, 90 cat.의 경우 10.6% 포인트 성능향상을 보

표 1 Reuters21578에서의 성능평가

	자질 가중치를 적용하지 않은 경우			자질 가중치를 적용한 경우		
	MN	MN +ENORM	DMBE +ENORM	MN	MN +ENORM	DMBE +ENORM
10범주	0.8496	0.8822	0.8833	0.8531	0.8973	0.8952
90범주	0.7336	0.7576	0.8121	0.7932	0.8280	0.8339

표 2 20Newsgroup에서의 성능평가

	자질 가중치를 적용하지 않은 경우			자질 가중치를 적용한 경우		
	MV	MN	DMBE	MV	MN	DMBE
10범주	0.8333	0.8431	0.9016	0.8914	0.8707	0.8972

였음을 알 수 있다. 또한, 상호정보가중 단순 베이지언 분류기를 사용하게 되면 성능은 더욱 증가하였으나 10 cat.의 실험에서 DMBE보다 MN이 높은 성능을 보여 상호정보가중 단순 베이지언 분류방법에 대한 고찰은 더욱 필요할 것으로 보인다. 20Newsgroup에서도 마찬가지로 기존의 방법에 비해 가중치를 적용하지 않은 경우와 상호정보가중 단순 베이지언 모두 제안하는 DMBE를 사용했을 때 MV나 MN에 비해 성능이 향상됨을 알 수 있었다. 표1과 표2가 보여주고 있는 성능은 최고의 성능을 보인다고 알려져 온 SVM이나 kNN을 이용한 다른 실험에서의 성능[2, 5]에 비해 크게 뒤지지 않는 것으로 평탄화 기법등 다른 부분에서의 개선이 이루어진다면 단순 베이지언 문서분류기의 성능은 다른 학습방법에 비해서도 우수한 성능을 나타낼 수 있을것으로 기대된다.

5. 결론 및 향후연구

본 논문에서는 기존의 다항시행접근 단순 베이지언 분류기가 갖고 있는 문제점을 살펴보고 이를 개선할 수 있는 세가지 방법을 제안하였다. 기존의 분류기가 갖고 있는 첫 번째 문제점은 범주에 대한 단어의 출현확률을 계산할 때 길이가 서로 다른 학습문서에서의 출현회수를 모두 동일하게 간주하여 부적절한 경우를 발생시키는 것으로, 이를 개선하기 위하여 문서모델에 기반한 파라미터 추정방법을 제안하였다. 이 방법은 길이가 서로 다른 학습문서들을 동등한 중요도를 갖는 독립된 말뭉치로 간주하여 이 각각의 문서에서 확률을 구한 다음 평균을 냄으로서 결과적으로 길이가 짧은 문서에서 나타난 단어가 더욱 중요한 단어일 가능성이 높다는 직관을 파라미터 추정에 반영할 수 있도록 하였다. 두 번째 문제점은 문서의 길이가 길어질수록 범주와의 관련성이 선형적으로 증가하게 된다는 것으로, 이를 개선하기 위하여 정보검색에서 많은 효과를 보아왔던 길이 정규화 기법을 사용하였다. 특히, 문서의 길이와 전체 문서집합에서의 평균길이를 선형적으로 결합하는 방식이 아닌 지수적으로 결합하는 방식을 제안하여 좀 더 간결하면서도 비슷하거나 우수한 성능을 보일 수 있도록 하였다.

마지막으로 자질추출을 선행하는 기존의 분류기는 일반적인 가중 단순 베이지언 분류기에서 이진가중치를 사용하는 것임을 보였고, 이 가중치를 분류시점에서의 상호정보를 사용하여 계산함으로서 문서의 범주예측을 위해 도움이 많이 되는 단어를 강조하는 분류기를 제안하였다. 제안하는 방법들은 기존의 다항시행접근 단순 베이지언 분류기의 성능을 향상시켰으며 이는 다른 여러 가지 기계학습 기법에 의해 학습된 문서분류기에 비해 뒤지지 않음을 실험적으로 입증하였다.

향후 연구로는 가중 단순 베이지언 분류기에서의 가중치를 구하는 방법으로 상호정보 이외의 방법들을 사용해 볼 것이며, 문서모델에 기반한 파라미터 추정방법이 문서분류뿐 아니라 일반적인 확률기반 정보검색모델에서 어떠한 방법으로 적용될 수 있는지에 대한 연구를 지속적으로 수행할 것이다.

참 고 문 헌

[1] Yang, Y., "Expert network : Effective and efficient learning from human decisions in text categorization and retrieval", In Proceedings of SIGIR-94, 18th ACM International Conference on Research and Development in Information Retrieval, pp. 13- 22, 1994.

[2] Joachims, T., "Text categorization with support vector machines: learning with many relevant features", In Proceedings of ECML-98, 10th European Conference on Machine Learning, pp. 137-142, 1998.

[3] McCallum, A. K., and Nigam, K. , "A comparison of event models for naive bayes text classification", In Proceedings of AAAI-98 Workshop on Learning for Text Categorization, pp. 137-142, 1998.

[4] Lewis, D. D., and Ringuette, M., "A comparison of two learning algorithms for text categorization", In Third Annual Symposium on Document Analysis and Information Retrieval, pp. 81-93, 1994.

[5] Yang, Y. and Liu, X., "A re-examination of text categorization methods.", In Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, pp. 42-49, 1999.

[6] Domingos, P. and Pazzani, M. J., "On the optimality of the simple bayesian classifier under zero-one loss", Machine Learning, Vol. 29, No 2/3, pp. 103-130, 1997.

[7] Sparck Jones, K.., Walker, S. and Robertson, S.E., "A probabilistic model of information retrieval: development and comparative experiments. Information Processing and Management Vol. 36, Part 1 pp. 779-808; Part 2 pp. 809-840, 2000.

[8] Lewis, D. D., "Naive (Bayes) at forty: The independence assumption in information retrieval", In Proceedings of ECML-98, 10th European Conference on Machine Learning, pp. 4-15, 1998.

[9] Singhal, A., Buckley, C. and Mitra, M., "Pivoted

Document Length Normalization", In Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, pp. 21-29, 1996.

[10] Robertson, S.E. and Walker, S., "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval", In Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, pp. 232-241, 1994.

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

김 상 범

1998년 고려대학교 컴퓨터학과 졸업 (B.S.). 2000년 고려대학교 대학원 컴퓨터학과 이학석사 (M.S.). 2000년 3월 ~ 현재 고려대학교 대학원 컴퓨터학과 박사과정. 관심분야는 정보검색, 텍스트마이닝, 자연어처리, 기계학습 등

임 해 창

1991년~현재 고려대학교 컴퓨터학과 교수. 1993년 인지 과학회 이사. 1994년~1998년 한국 정보과학회 편집위원. 1998년 5월~2000년 5월 한국정보과학회 한국어정보처리연구회 운영위원장. 1999년 3월~2000년 8월 고려대학교 컴퓨터과학기술연구소 연구소장. 관심분야는 자연어처리, 구문분석, 정보검색, 기계학습