전자메일 자동관리 시스템을 위한 전자메일 분류기의 성능 비교

김국표*, 권영식*, 백찬영**

*동국대학교 산업공학과, **(주)연성정보기술

Comparison of e-Mail Classifiers for e-Mail Response Management Systems

Kim, Kuk Pyo*, Kwon, Young S*, Baek, Chan Young.**

*Dongguk University, **Yeonsung Infotech Co.

E-mail : molanos@dongguk.edu, yskwon@dongguk.edu, baekcy@yeonsung.co.kr

요 약

인터넷의 발전과 더불어 전자메일 사용자가 증가하게 되고, 기업의 고객접촉채널로서 전자메일에 대한 중요성 또한 증가되고 있다. 고객의 요구에 대해 적시에 적절하게 응답하지 못하면 고객의 불만족이 증가하게 되고, 충성도를 감소시켜 결국 장기적 매출 및 수익성 악화를 초래하게 된다. 따라서 고객의 전자메일에 신속, 정확하게 응답할 수 있는 전자메일 자동관리 시스템의 필요성이 증가되고 있다. 본 연구에서는 나이브 베이지안 학습과 중심점 기반 분류 방법을 이용하여 전자메일 자동관리 시스템에서 전자메일 분류를 수행하는 분류기를 구현한다. 구현된 분류기를 이용하여 실제 기업의 고객 전자메일을 분류하는 실험을 수행하고 두 분류기의 성능을 비교하였다. 실험결과 두 분류기 모두 전자메일 분류에 비교적 우수한 성능을 보였다. 그러나, 클래스 수가 적은 경우 중심점 기반 분류기가 좋은 성능을 보였으나, 학습집합이 작아지면서 두 분류기의 성능 차이는 없었으며, 클래스의 수가 많아지면서 나이브 베이지안 분류기가 더 우수한 성능을 보였다.

1. 서론

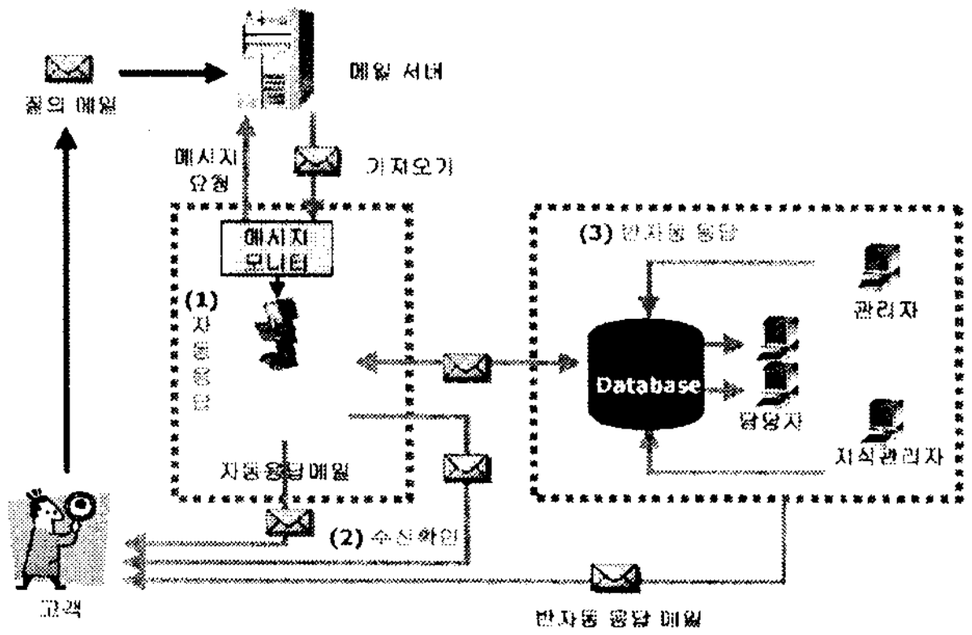
e-비지니스 환경에서 전자메일과 웹을 통한 고객의 질의나 불만사항에 대해 적시에 적절한 응답을 하는 것은 대 고객 서비스 증진뿐 아니라 기업의 생존까지 직결되게 되었다[2]. 시간이 지나감에 따라 축적되어 가는 방대한 전자메일을 사람이 일일이 읽고 응답하는 것은 매우 고비용이며 비효율적이다.

실제 기업에서 고객의 전자메일을 분류하고 자동 응답할 수 있는 시스템을 도입할 경우 기업의 생산성을 향상시킬 수 있을 뿐만 아니라, 고객요구를 만족시킴으로써 고객관계를 향상시켜 기업의 수익증대를 꾀할 수 있다.

이러한 프로세스를 자동화하기 위한 전자메일 자동관리시스템(e-mail response management system)에 대한 관심과 활용이 증대되고 있는데 기존의 외국제품으로는 eMailUnlimited, Kana Solutions, ReplyMate 등과 국내의 경우 (주)연성정보기술, (주)스펙트라, (주)네오케스트, (주)다음소프트 등에서 전자메일 자동관리 시스템을 개발하고 있다.

[그림 1]은 전자메일 자동관리 시스템의 한가지 예를 나타낸 것으로 고객의 요구사항을 담은 전자메일이 메일 서버로 실시간으로 들어오게 된다. 이렇게 메일 서버로 들어온 메일을 이 시스템이 불러 들여서 우선 자동응답이 가능한지를 판단하게 된다. 만약 자동응답이 가능한 전자메일일 경우 바로 고객에게 응답 메일을 보내게 된다. 그러나 자동응답이 가능한 메일이 아닌 경우 고객에게 메일

이 처리중이라는 수신확인 메일을 보내어 고객의 요구사항이 현재 처리 진행중이라는 것을 알리고, 다음 단계로 넘어가게 된다. 본 연구에서 구현된 부분인 이 단계는 고객의 문의메일을 업무영역에 따라 분류하여 문의내용을 처리할 수 있는 담당자에게 보내주는 역할을 한다. 각 담당자는 본인이 맡은 업무와 관련 있는 문의 메일에 대해서 신속한 응답을 함으로써 모든 시스템의 처리과정이 끝나게 된다. 이러한 전자메일 자동분류 시스템은 고객의 요구사항에 보다 적합한 담당자가 아주 신속하게 응답함으로써 고객과 기업의 관계 유지에 많은 도움이 된다.



출처:(주)다우기술

[그림 1] 전자메일 자동관리 시스템

본 연구에서는 나이브 베이지안 학습과 중심점 기반 분류 방법을 이용하여 전자메일 자동관리 시스템에서 전자메일 분류를 수행하는 분류기를 구현한다. 구현된 분류기를 이용하여 실제 기업의 고객 전자메일을 분류하는 실험을 수행하고 두 분류기의 성능을 비교함으로써 현업 적용에 적합한 알고리즘과 향후 분류기 개발 방향을 제시한다.

본 연구는 전체 5개의 장으로 구성되어 있으며, 1장에서는 연구의 배경과 목적, 2장은 선행연구로써 나이브 베이지안 학습과 중심점 기반 분류 방법을 간단하게 소개하였다.

3장에서는 본 연구에서 구현된 나이브 베이지안 분류기와 중심점 기반 분류기의 구성을 간략히 소개하고, 4장에서는 실험자료 및 실험결과, 그리고 5장은 결론으로 구성되어 있다.

2. 선행연구

전자메일 분류는 기존 전자메일의 특성을 학습하여 새로운 전자메일을 사전에 정해진 클래스로 분류하는 것이다. 전자메일 분류에 좋은 성능을 보이는 알고리즘은 나이브 베이지안 학습이다. 그러나 이것과 달리 일반 문서 분류에는 중심점 기반 분류 방법이 나이브 베이지안 학습보다 더 우수한 성능을 보인다는 연구 결과를 볼 수 있었다[1, 2, 3, 5, 7, 9, 10, 11].

전자메일과 일반문서의 차이점을 살펴보면, 전자메일은 일반 문서와 달리 형식이 없고, 주로 개인적인 내용을 다루며, 문서 길이가 짧다는 특징을 가지고 있다. 또한 표준어가 아닌 신조어, 인터넷 용어를 많이 사용하며 맞춤법이 틀린 경우가 많다는 것이다.

2.1 나이브 베이지안 학습

나이브 베이지안 학습은 문서 분류에 가장 널리 이용되고, 그 성능 또한 우수하다고 증명되었다[3, 4, 7, 9]. 나이브 베이지안 학습은 문서들이 클래스에 속할 사후 확률을 계산하여, 그 사후 확률이 가장 큰 클래스로 문서를 분류하는 것이다. 클래스에 대한 사후 확률은 각 속성(단어)들이 서로 독립이라는 가정을 가지고 베이즈 규칙에 의해서 계산된다.

또한, 이 알고리즘은 단어의 존재 유무만을 고려하는 multi-variate Bernoulli event 모델과 단어의 빈도를 고려하는 multinomial 모델의 두가지 모델이 있다. 일반적으로 multinomial 모델이 더 우수한 성능을 보이므로 본 연구에서도 이 모델을 이용하였다[3, 11].

2.2 중심점 기반 분류

중심점 기반 분류기는 K-근접점 분류와 같은 패러다임을 가진다[5]. 중심점 기반 분류 알고리즘은 전자메일을 단어 빈도(term frequency)와 역문서 빈도(inverse document frequency)를 이용하여 표현할 수 있다. 이와 같은 방법으로 표현된 여러 개의 전자메일은 사전에 정해진 임의의 K개의 클래스에 속하게 된다.

학습 단계에서 K개 클래스의 중심점(centroid)을 계산하고, 코사인 함수를 이용하여 전자메일과 중심점간의 유사도(similarity)를 계산한 후, 가장 큰 유사도 값을 가지는 클래스로 메일을 분류하게 된다.

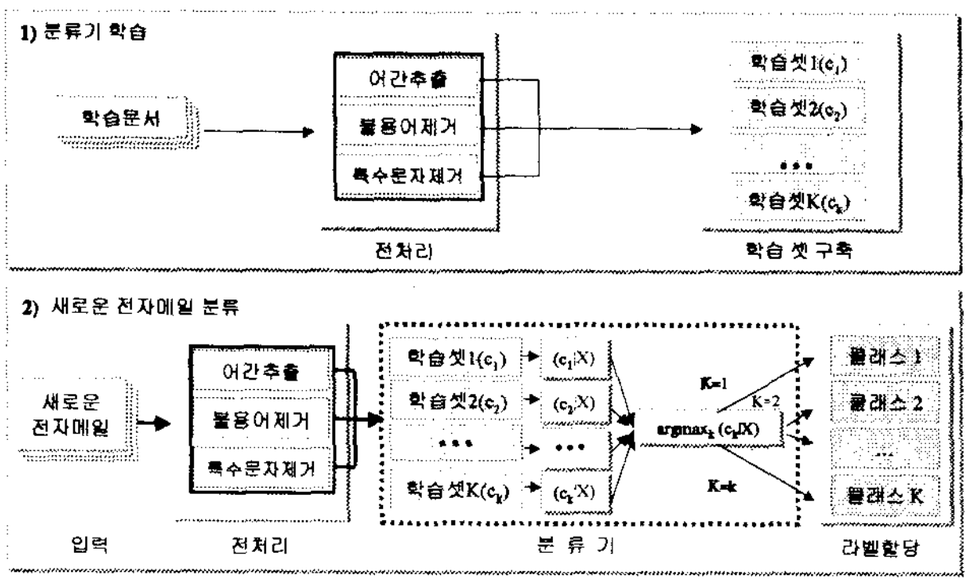
3. 구현된 두 가지 분류기

3.1 나이브 베이지안 분류기

본 연구에서 구현된 나이브 베이지안 분류기는 [그림 2]와 같이 학습단계와 분류단계로 구성되어 있다.

먼저 학습단계에서는 어간추출, 불용어 제거, 특수문자 제거와 같은 전처리 과정을 거친 다음, 각각의 클래스별로 학습집합을 구성한 후 유일한(unique) 단어들의 문서에서 발생빈도를 계산함으로써, 각 단어들과 클래스의 사전확률을 계산하게 된다.

분류단계에서는 학습단계와 같은 전처리 과정을 거친 후, 학습단계에서 계산된 확률값을 이용하여 전자메일이 클래스에 속할 사후확률을 계산하고, 그 사후확률 값이 가장 큰 클래스로 분류하게 된다.



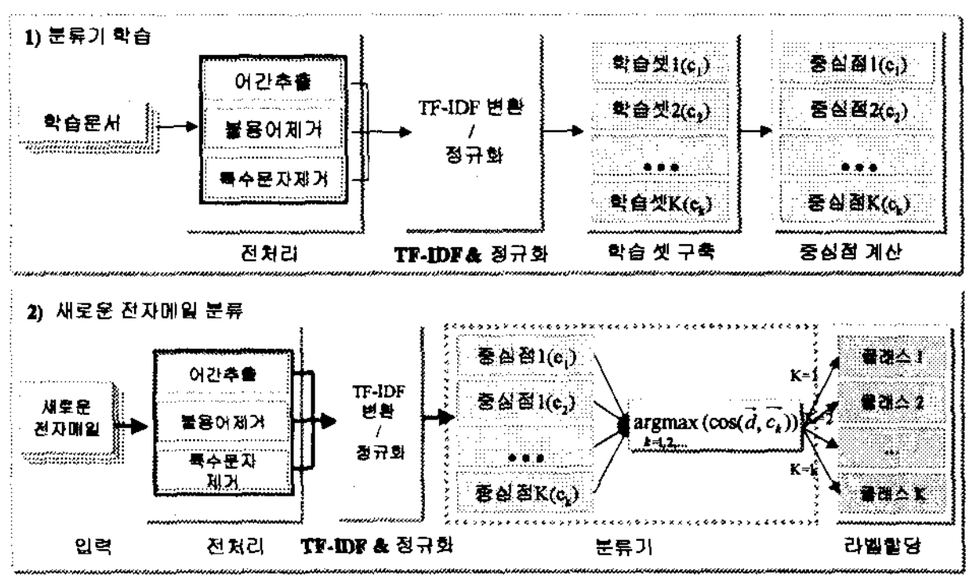
[그림 2] 나이브 베이지안 분류기

4.2.2 중심점 기반 분류기

본 연구에서 구현된 중심점 기반 분류기는 [그림 3]을 통해서 알 수 있듯이 나이브 베이지안 분류기와 같은 학습단계와 분류단계로 구성되어 있다.

학습단계에서는 나이브 베이지안 분류기와 동일한 전처리 과정을 거친다. 이 단계에서 얻은 유일한(unique) 단어들의 빈도(tf)와 역빈도(idf)를 이용하여 TFIDF형태의 벡터로 표현한 후, 벡터의 크기를 1로 정규화 시킨다. 이렇게 정규화된 벡터로 표현된 학습집합을 이용하여 각 클래스의 중심점을 계산한다.

분류단계에서는 학습단계와 같이 전처리 과정, TFIDF 변환과 정규화를 거친 후, 코사인 함수를 이용하여 테스트 전자메일과 각 클래스의 중심점과의 유사도를 계산하게 된다. 최종적으로 가장 큰 유사도를 가지는 클래스로 메일을 분류하게 된다.



[그림 3] 중심점 기반 분류기

4. 실험

4.1 실험자료

본 연구에서는 온라인 쇼핑몰 전자메일과 카드회사 전자메일로 나이브 베이지안 분류기와 중심점 기반 분류기의 성능 비교 실험을 하였다.

먼저 온라인 쇼핑몰 전자메일의 내용은 [표 1]과 같다. 온라인 쇼핑몰 전자메일을 이용한 실험 목적은 학습집합의 구성과 크기에 따른 분류기의 성능 변화를 관찰하는 것이다. 따라서 학습집합의 비율을 70%, 50%, 30%로 변화시키면서 각각 10번씩 반복 실험을 하였다.

[표 1] 온라인 쇼핑몰 전자메일

구 분	개수	내 용
시스템 문의	189	패스워드 분실이나 메일에서 글씨가 깨지는 문제 등
상품 문의	149	상품의 재고여부, 색깔, 사이즈, 주문 확인, 주문 취소 등
사업 문의	61	배너광고 교환, 제휴, 업체 입주 등

[표 2]의 카드회사 전자메일은 클래스 수가 증가함에 따라 정확도가 어떻게 변화하는가를 실험하기 위한 자료로써 학습집합과 테스트 집합의 비율을 70:30으로 나누어서 10회 반복 실험을 하였다.

[표 2] 카드회사 전자메일

구 분	개수	내 용
클래스1	227	결제, 결제일 관련 문의
클래스2	189	승인내역, 거래내역, 승인취소 관련 문의
클래스3	242	연체, 할부, 대출, 현금서비스, 수수료 관련문의
클래스4	212	제휴카드, 제휴서비스, 부가서비스 관련 문의
클래스5	135	홈페이지 이용 관련 문의
클래스6	205	카드종류 변경 / 해지 관련 문의
클래스7	205	카드 신규발급, 분실 / 훼손 재발급 관련 문의

4.2 성능 평가 방법

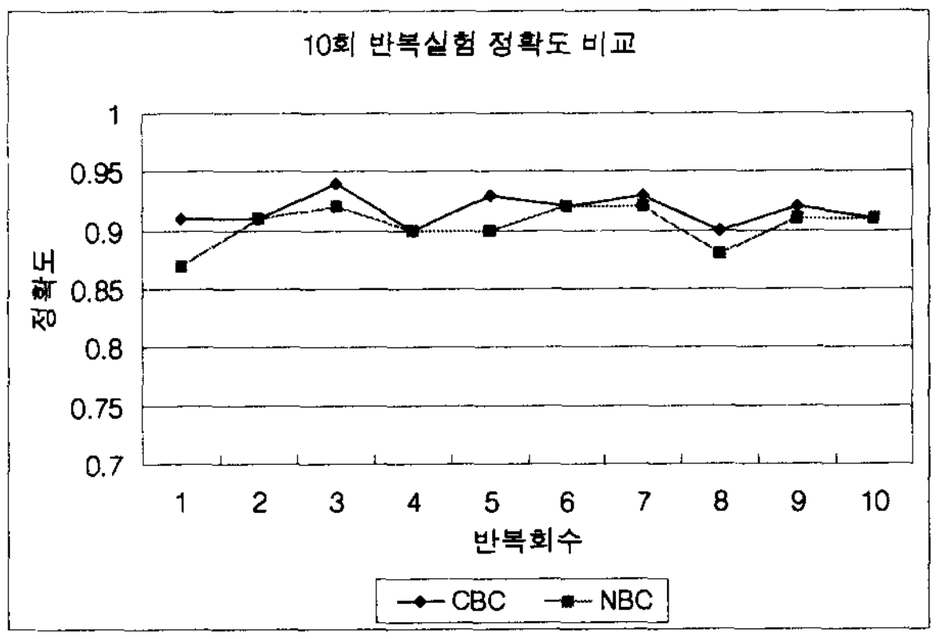
문서 분류 시스템의 성능 평가 방법에는 정확도(accuracy), 에러(error), 정확율(precision), 재현율(recall) 등 다양한 방법들이 있다[2,11]. 그러나 본 연구에서는 정확도 관점에서 두 시스템의 성능을 평가하였다. 그 이유는 실제 현업에서 여러 종류의 전자메일 분류기를 구현하여 적용하고 있지만, 이러한 분류기가 시간이 지남에 따라 정확도가 떨어진다는 것이 문제가 되고 있기 때문이다.

$$\text{정확도(accuracy)} = \frac{\text{정분류 메일 수}}{\text{전체 메일 수}}$$

$$\text{에러(error)} = \frac{\text{오분류 수}}{\text{분류된 메시지 수}} = 1 - \text{정확도}$$

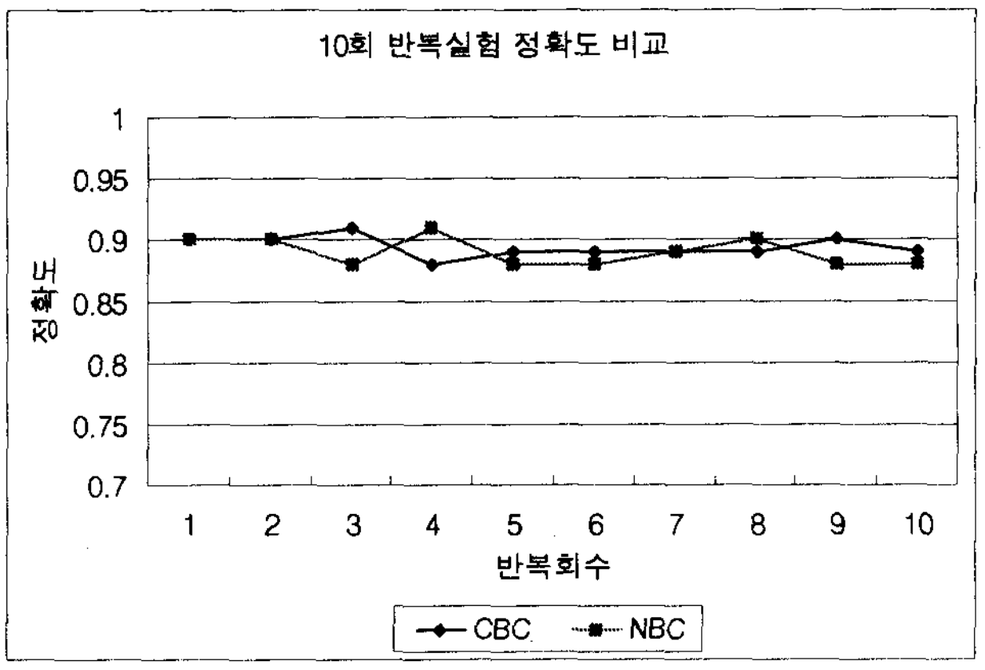
4.3 실험결과

아래 [그림 4]에서 [그림 6]은 온라인 쇼핑몰 전자메일에 대한 실험결과이다. [그림 4]와 [표 3]을 통해서 중심점 기반 분류기(CBC)가 더 좋은 정확도를 보인 것을 알 수 있다.

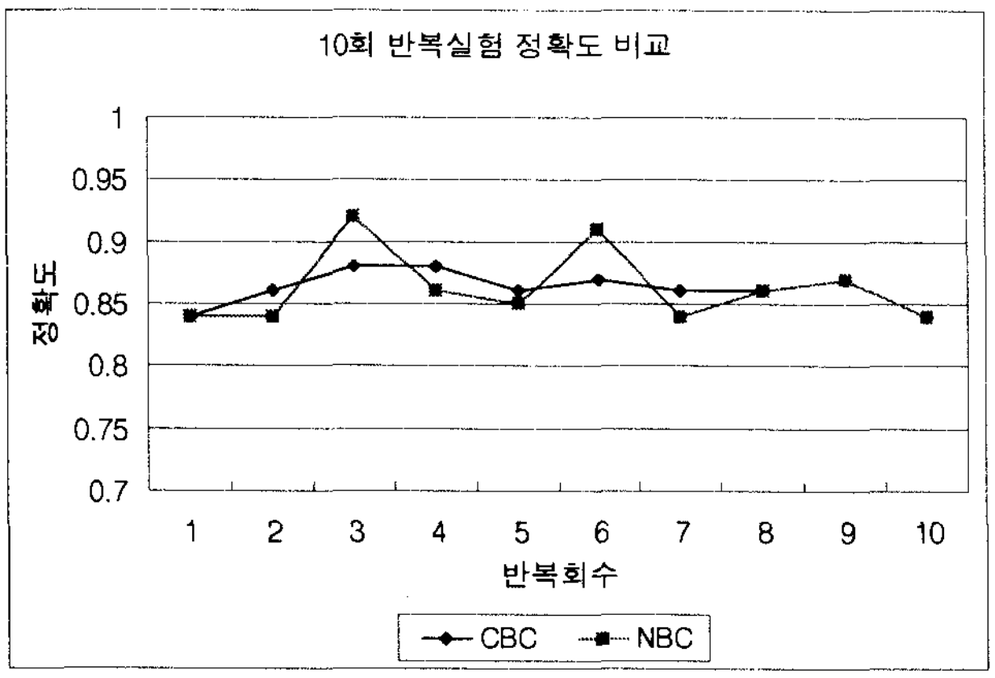


[그림 4] 학습집합이 70인 경우 정확도 비교

그러나 [그림 5], [그림 6], [표 3]으로부터 학습집합의 비율이 50, 30으로 작아지면서 두 분류기의 정확도 차이는 없다는 것을 알 수 있다.



[그림 5] 학습집합이 50인 경우 정확도 비교



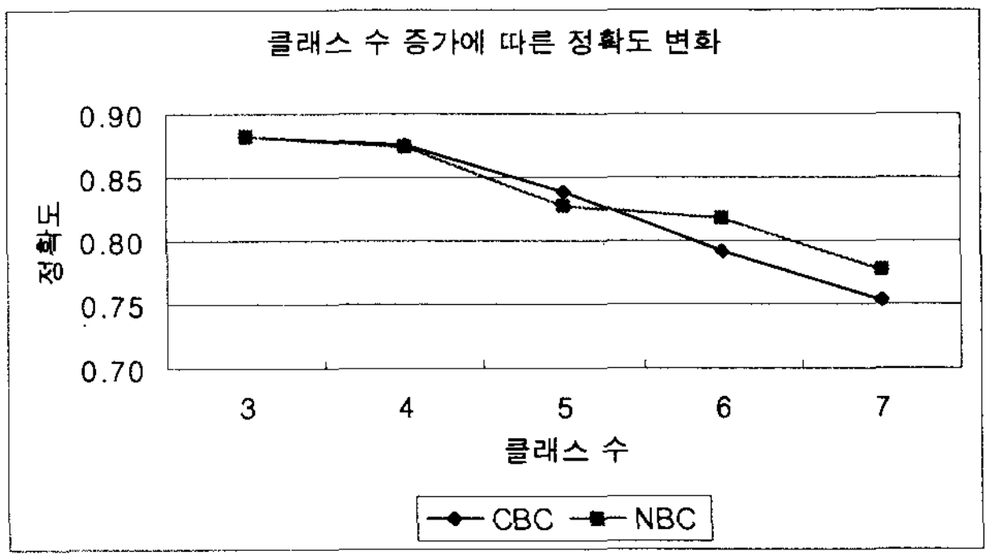
[그림 6] 학습집합이 30인 경우 정확도 비교

본 연구에서는 두 분류기의 정확도 차이가 통계적으로 유의한지 T-검정을 실시하였다[8]. 에러율은 전체 테스트 문서와 두 분류기의 에러 차이의 비로 정의하였다. 그리고 귀무가설은 두 분류기의 에러율의 차이는 없다는 것이며, 검정통계량은 $t=\frac{\bar{p}}{\sqrt{S^2/n}}$ 이며 이 값은 자유도가 $n-1$인 t-분포를 따른다. 여기서 $\bar{p}$는 에러율의 평균을 나타내고, S^2은 에러율의 분산, 그리고 n은 실험 반복 회수를 나타낸다. 검정결과는 [표 3]과 같다.

[표 3] 학습집합의 비율에 따른 성능 검정

구 분	학습집합의 비율(%)		
	70	50	30
|T|-값	3.4716	0.7099	0.2210
기각역	$t_{9,0.975}$ = 2.262 (유의수준 0.05)		

[그림 7]은 카드회사 전자메일로 클래스 수가 증가하면서 정확도가 어떻게 변화하는가를 실험하였다. 그래프를 통해서 두 분류기 모두 클래스 수가 증가하면서 정확도가 떨어지는 것을 볼 수 있다. 클래스 수가 작은 경우 두 분류기의 정확도에는 차이가 없으나, 클래스 수가 5개인 경우 중심점 기반 분류기가 다소 우수한 정확도를 보였다. 그러나 클래스 수가 6, 7개로 증가할 경우 나이브 베이지안 분류기(NBC)가 훨씬 우수한 정확도를 보인다는 것을 알 수 있다.



[그림 7] 클래스 수 증가에 따른 정확도 비교

클래스 수에 따른 두 분류기의 정확도 차이 검정에 대한 결과는 아래 [표 9]와 같다.

[표 9] 클래스 수에 따른 성능 차이 검정

구 분	클 래 스 수				
	3개	4개	5개	6개	7개
|T|-값	0	0.62691	6.12372	15.44494	23.35946
기각역	$t_{9,0.975}$ = 2.262 (유의수준 0.05)				

5. 결론

본 논문에서는 나이브 베이지안 분류기와 중심점 기반 분류기를 구현하여 한글 전자메일 분류 성능을 비교 연구하였다. 연구 결과 두 분류기 모두 비교적 우수한 전자메일 분류 성능을 보여 주었다.

그러나, 클래스 수가 적은 경우 중심점 기반 분류기가 좋은 성능을 보였으나, 학습집합이 작아지면서 두 분류기의 성능 차이는 없었다. 또한 클래스의 수가 많아지면서 두 분류기 모두 정확도가 떨어졌지만, 정확도의 감소 폭이 나이브 베이지안 분류기가 중심점 기반 분류기에 비해 작았다.

웹 문서 분류에서 중심점 기반 분류기가 나이브 베이지안 분류기보다 더 좋은 성능을 가진다는 연구결과와 상반되는 실험결과가 나온 이유는 전자메일이 일반문서와는 다른 성격을 가졌기 때문이라고 해석될 수 있다.

현업에서 전자메일의 분류 클래스는 본 연구의

클래스 수보다 더 많다. 그리고 시간과 비용을 고려한다면 학습집합을 크게 한다는 것 또한 불가능할 것이다. 따라서 본 연구 결과로 미루어 현업의 전자메일 분류 시스템에는 나이브 베이지안 분류기를 적용하는 것이 적합하다고 판단된다.

향후에는 중심점 기반 분류기처럼 학습집합의 구성이나 크기에 관계없이 일정한 범위내의 정확도를 유지하면서 나이브 베이지안 분류기처럼 클래스 수가 증가하면서도 정확도의 하락 폭이 크지 않은 전자메일 분류기가 개발되어야 할 것이다.

[참고문헌]

[1] 윤종식, "배깅과 부스팅을 이용한 나이브 베이지안 이메일 분류기의 성능향상", 동국대학교 석사학위논문, 2001.

[2] 황호순, "프론트 앤드 e-CRM을 위한 전자메일 분류기 개발", 동국대학교 석사학위논문, 2001.

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

[4] D. Lewis and M. Ringuette, "Comparison of two learning algorithms for text categorization" , In Tenth European Conference on Machine Learning, 1998.

[5] Eui-Hong (Sam) Han and George Karypis, "Centroid-Based Document Classification: Analysis & Experimental Results", PAKDD 2000.

[6] G. Salton, "*Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*", Addison Wesley, 1989.

[7] Sahami, S. Dumais, D. Heckerman & E. Horvitz, "A Bayesian Approach to Filtering Junk e-mail", AAAI Technical Report WS-98-05, 1998.

[8] T.G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms" . Neural Computation, 10(7), 1998.

[9] Tom M. Mitchell, "*Machine Learning*", The McGraw-Hill Company, 1997.

[10] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization", ICML, 1997.

[11] Yanlei Diao, Hongjun Lu & Dekai Wu, "A Comparative Study of Classification Based Personal E-mail Filtering", PAKDD 2000.