SVM 분류기를 이용한 문서 범주화 연구

An Experimental Study on Text Categorization using an SVM Classifier

정영미(Young-Mee Chung), 임혜영(Hye-Young Lim)*

목 차

초 록

문서 범주화에 이용되는 학습알고리즘 중에서 이원 패턴인식 문제를 해결하기 위해 제안된 SVM은 다른 분류기 보다 우수한 성능을 보이고 있다. 본 연구에서는 Reuters-21578 (ModApte 분할판)을 대상으로 SVM 분류기를 이용하여 단어빈도, 역문헌빈도, 문헌길이 정규화 공식을 자질에 대한 가중치로 적용하여 성능을 평가하고, 선형 SVM과 비선형 SVM의 분류 성능을 비교하였다. 또한 이원 분류기를 승자독식 방법과 쌍단위 분류 방법에 의해 다원 분류기로 확장하여 실험한 후 이원 분류기와의 성능을 비교 분석하였다.

ABSTRACT

Among several learning algorithms for text categorization, SVM(Support Vector Machines) has been proved to outperform other classifiers. This study evaluates the categorization ability of an SVM classifier using the ModApte split of the Reuters-21578 dataset. First, an experiment is performed to test a few feature weighting schemes that will be used in the categorization tasks. Second, the categorization performances of the linear SVM and the non-linear SVM are compared. Finally, the binary SVM classifier is expanded into a multi-class classifier and their performances are comparatively evaluated.

키워드: 문서 범주화, SVM, SVM 분류기, 자동분류

* 연세대학교 문헌정보학과

■ 논문 접수일 : 2000년 11월 30일

1 서 론

문서 범주화(text categorization)는 문서의 내용을 바탕으로 미리 정의된 범주를 문서에 부여함으로써 문서를 자동 분류하는 기법이다(Yang and Pedersen 1997). 특히 문서의 내용을 바탕으로 컴퓨터가 자동으로 범주를 할당하는 자동 문서 범주화는 수작업으로 문서를 분류하는 데 소요되는 시간과 노력, 비용 등을 감소시킴으로써 효율적인 정보의 조직 및 검색을 가능하게 할 뿐만 아니라 최근에는 문서필터링이나 전자우편 범주화 등의 응용분야로 확대 적용되고 있다.

자동 문서 범주화는 전문가 시스템에서와 유사한 방법으로 지식공학과 범주화 규칙에 대한 지식베이스를 이용하는 규칙기반방법이나 수동으로 구축된 학습집단을 대상으로 귀납 학습에 의해 자동으로 범주시스템을 유도하는 귀납적 학습방법 등을 이용하여 범주화 작업을 수행한다. 특히 미리 주어진 예제들의 유사성을 이용하여 일반화 과정을 수행하고 가설을 생성한 다음 새로운 예제에 대한 범주를 예측하는 귀납적 학습방법에 기반한 자동 문서 범주화의 경우, 분류기의 구축과 갱신, 개개인의 관심분야에 대한 범주 생성이 용이하고 업무에 따른 정확률과 재현율 조정이 가능하다는 점에서 많은 연구가 이루어지고 있다(Dumais et al. 1998).

문서 범주화에서 적용 가능한 통계적 이론과 기계학습방법은 다중회귀모형(multivariate regression models), 최근접이웃 분류기(nearest neighbor classifiers), 확률적 베이지언 모형(probabilistic Bayesian models), 결정트리(decision trees), 신경망(neural networks), SVM(Support Vector Machines; 지지벡터기) 등이 있으며, 특히 최근 발표된 SVM 분류기를 이용한 문서 범주화 실험연구 결과를 보면 다른 학습방법을 적용한 것보다 SVM의 분류 성능이 우수한 것으로 밝혀졌다(Dumais et al. 1998).

SVM은 1995년에 Vladimir Vapnik에 의해 이원 패턴인식 문제를 해결하기 위해 제안된 학습방법으로 부정예제로부터 긍정예제를 분류해 낼 수 있는 결정면(decision surface)을 찾아내는 분류모형이다(Vapnik 2000).

SVM은 이원 패턴 분리를 위한 알고리즘으로 개발되었기 때문에 K개의 범주를 학습시키기 위해서는 여러 개의 SVM 분류기를 조합해서 분류업무를 수행하여 다원 분류기로 확장을 하게 된다. 패턴 인식 분야에서는 SVM을 다원 분류기로 확장하는 연구가 진행되고는 있으나, 아직 문서 범주화 분야에서는 SVM 다원 분류기를 중심으로 고찰하는 연구는 거의 수행되지 않았기 때문에 그 성능을 분석해 볼 필요가 있다.

따라서 본 연구에서는 이원 패턴 분류기인 SVM 학습알고리즘을 이용한 문서 범주화 실험을 수행하고, 다원 분류기로 확장할 수 있는 방안에 대해서 검토해 보고자 한다.

2 SVM 분류기의 원리

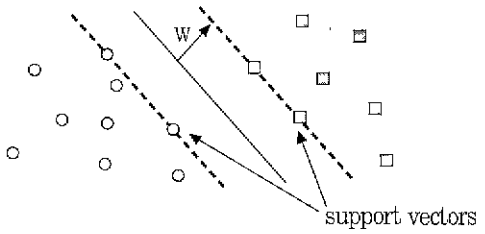
통계적 학습이론(statistical learning theory)에 기반한 SVM은 기존의 통계적 이론에서 이용되는 경험적 리스크 최소화 원칙(empirical risk minimization; ERM)이 아닌 구조적 리스크 최소화 원칙(structural risk minimization; SRM)을 이용하여 일반화 오류를 줄이기 때문에 패턴 인식과 문서 범주화 등

에서 우수한 성능을 보여주고 있다.

SVM은 선형적으로 분리할 수 있는 학습집단에 대해서 최대 마진 분류기를 구축하는 선형 SVM과, 선형적으로 분리할 수 없는 경우에 커널함수에 의해 만들어지는 비선형 결정함수를 이용하여 최적의 초평면을 구축하는 비선형 SVM으로 분류된다.

2.1 선형 SVM

SVM의 가장 간단한 형태는 〈그림 1〉과 같이 최대 마진(margin)을 가지고 부정예제로부터 긍정예제를 분류해 낼 수 있는 결정면(decision surface)을 찾아내는 선형 분류모형이다 (Dumais et al. 1998).



〈그림 1〉 선형 SVM

〈그림 1〉에서 실선은 긍정예제와 부정예제를 분리하는 결정면이고, 실선과 평행인 점선들은 오류를 발생시키지 않으면서 결정면을 이동할 수 있는 공간으로 이것을 마진이라 한다. 즉, SVM은 학습집단에서 마진을 최대화하는 결정면을 찾아내는 알고리즘이라 할 수 있다. SVM에서 마진이 최대화되었을 때, 점선 상의 데이터는 결정면(실선)으로부터 $\frac{1}{||w||}$ 의 거리에 위치하게 되는데 이를 SV(support vectors; 지지벡터)라 하며 학습집단에서 유일하게 유효한 요소가 된다.

SVM은 선형적으로 분리할 수 있는 문제에서 출발한다. 선형 분리가 가능하다는 것은 학습데이터를 두 집합 즉, 긍정예제와 부정예제를 분리시킬 수 있는 결정면이 존재한다는 것이며, 이 결정면은 다음 수식과 같이 나타낼 수 있다 (Dumais et al. 1998).

$$w \cdot x - b = 0$$

여기서 w는 가중치벡터, x는 입력벡터, b는 기준치로, w와 b는 학습데이터로부터 학습된다. 학습문서 집합을 $D = \{(x_i, y_i)\}$ 라고 할 때, 입력데이터 x_i가 범주(class)에 속하면 y_i는 $+1$의 값을 갖고, 속하지 않으면 -1의 값을 갖는다. 결국 SVM은 최적의 w와 b를 찾는 문제이다.

$$w \cdot x_i - b \geq +1 \ (y_i = +1 \text{ 인 경우})$$
$$w \cdot x_i - b \leq -1 \ (y_i = -1 \text{ 인 경우})$$

위의 두 식은 부호함수를 이용하여 다음의 결정함수로 표현할 수 있다.

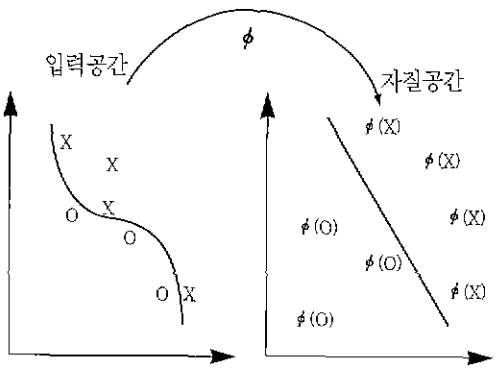
$$f(x) = sign((w \cdot x_i) + b)$$

2.2 비선형 SVM

SVM은 기본적으로 선형 분리가 가능한 문제에서 출발하지만 모든 문제가 선형적으로 분리될 수는 없다. 이처럼 입력데이터의 선형 분리가 불가능할 경우 입력공간을 분리하는 비선형 결정면(nonlinear surface)을 이용하게 되는데, 비선형 결정면의 식을 분석적으로 계산해낸다는 것은 매우 어려운 일이다. 이런 경우 SVM에서는 〈그림 2〉와 같이 고차원의 자질공간을

효율적으로 처리하기 위해서 커널함수 $k(x, x_i) = \phi(x), \phi(x_i)$를 이용하여 입력벡터 x를 고차원 자질공간에서의 벡터로 변형한 후, 선형의 경계선을 찾는 문제로 전환하게 된다.

일반적으로 사용되는 커널함수는 다항식(polynomial) 커널함수, RBF(Radial Basis Function) 커널함수, 다층 퍼셉트론(multi-layer perceptron) 커널함수 등이다.



〈그림 2〉 비선형 SVM

2.3 SVM 다원 분류기

SVM은 원래 이원 패턴 분리를 위한 알고리즘이기 때문에 K-범주의 패턴 분리 문제를 위해서는 입력데이터 x_i의 범주 정보 $y_i = \{1,\ldots,k\}$로 구성되는 학습집단$(x_1, y_1),\ldots,(x_l, y_l)$을 이용하여 여러 개의 SVM 이원 분류기를 조합하여 함수 $f(x) = y$를 추정해야 한다. SVM 다원 분류기 학습을 위한 조합방법으로 승자독식 방법(winner-takes-all)과 쌍단위 분류 방법(pairwise classification) 등이 사용되고 있다(Chin 1998).

(1) 승자독식 방법

승자독식 방법은 가장 간단하며 효율적인 조합방법으로, 두 범주의 결정함수를 이용하여 각각의 k 범주에 대해 이원 결정함수를 구축함으로써 K개의 범주로 확장하는 방법이다(Kressel 1999).

$$f_k : R^N \rightarrow \{\pm 1\} \quad +1 \text{ for all samples in class } k \quad -1 \text{ else}$$

각 범주당 하나씩 총 K개의 SVM 분류기를 학습시키며, K-범주 분류문제에 대한 최적의 범주를 결정하기 위해서 각각의 분류기의 결정함수가 조합된다(Hastie, and Tibshirani 1996).

$$f(x) = \arg_k \max \sum_{i=1}^{l} y_i \alpha_i^k k^k(x, x_i) + b^k$$

(2) 쌍단위 분류 방법

승자독식 방법은 간단하고 효율적인 방법이지만, 경계면을 동시에 만족시킬 수 있는 최적의 다원 범주 결정면(optimal multicategory decision boundaries)을 구축하지는 못한다. 이러한 단점을 보완하기 위해 두 범주간의 경계선을 직접적으로 처리하는 쌍단위 분류 방법이 제안되었다(Weston, and Watkins 1998).

쌍단위 분류 방법은 각 쌍마다 분류기를 만드는 방법으로, K(K-1)/2개의 모든 조합에 대하여 SVM 분류기를 학습시키는 것이다.

$$f_{kl} : R^N \rightarrow \{\pm 1\} \quad +1 \text{ for all samples in class } k \quad -1 \text{ for all samples in class } l$$

〈표 1〉 실험문서집단

학습집단 범주	해당 본문이 있는 기사수	해당 본문이 없는 기사수	총 기사수	검증집단 범주	해당 본문이 있는 기사수	해당 본문이 없는 기사수	총 기사수
earn	2705	68	2773	earn	1044	24	1068
acq	1472	21	1493	acq	635	5	640
money-fx	304	18	322	money-fx	99	3	102
grain	364	2	366	grain	121	0	121
crude	285	0	285	crude	117	1	118
trade	292	5	297	trade	98	0	98
interest	160	11	171	interest	64	1	65
ship	121	0	121	ship	42	0	42
	5703	125	5828		2220	34	2254

학습과정을 거친 후 구축된 K(K-1)/2개의 쌍단위 결정함수(pairwise decision function) $f_k = \Sigma_i f_{ki}$ 중에서 최대값을 갖는 범주에 할당하게 된다.

$$f(x) = \arg\max_k \Sigma_i f_{ki}$$

3 SVM 분류기를 이용한 문서 범주화 실험

3.1 실험설계

3.1.1 실험문서집단

본 연구에서는 Reuters-21578[1] 문서집단을 사용하여 SVM 분류기를 이용한 문서 범주화 실험을 수행하였다. Reuter 문서집단은 문서 범주화 연구에서 가장 널리 사용되는 평가용 문서집합으로, 1987년에서 1991년까지 보도된 2만여 개의 로이터 통신 신문기사들로 구성되어 있다. Reuters-21578 문서집단에서 제안한 ModLewis 분할 방법, ModApte 분할 방법, ModHayes 분할 방법 중에서 ModApte 분할 방법을 이용하여 본 연구의 학습집단과 검증집단을 구성하였다.

본 연구는 문서 범주화 실험을 SVM 다원 분류기의 확장을 중심으로 진행하는 것을 목적으로 하고 있기 때문에, 가장 많은 기사가 할당된 10개 범주 중에서 하나의 범주만이 할당된 기사를 실험집단으로 선정하였다. 그러나 범주 Wheat 와 Corn의 경우 거의 모든 문서들이 그 상위개념에 해당하는 Grain 범주와 동시에 할당되어있기 때문에 복수범주를 허용하지 않을 경우 문서가 추출되지 않게 된다. 그러므로 위의 두 범주를 제외한 8개의 범주에 대해 실험을 수행하기로 하였고, 그 결과 〈표 1〉과 같이 5703개의 학습문서와 2220개의 검증문서로 구성되는 실험문서집단이 선정되었다.

1) http://www.research.att.com/~lewis에 공개되어 있다.

3.1.2 실험내용 및 방법

문서 범주화에서 모든 단어를 문서 표현에 사용할 경우 방대한 문서벡터가 형성되어 학습의 효율성이 떨어지고 많은 학습시간이 소요되게 된다. 그러므로 일차적으로 어간 추출과정(stemming)과 불용어 목록(stopword list)을 이용하여 차원을 축소하게 된다. 본 연구에서는 Reuters-21578 문서집단 중에서 본문 내용을 가진 학습집단 5703개의 기사에 대해 〈BODY〉 태그 안에 있는 기사 내용에 포함된 모든 단어를 추출하였다. 추출된 단어들을 대상으로 Porter의 스태밍 알고리즘[2]을 이용하여 어간을 추출하고, SMART 시스템의 불용어 사전을 이용하여 524개의 불용어를 제거한 다음 총 15086의 단어를 자질로 추출하였다.

대부분의 문서 범주화 실험의 경우 위와 같은 전처리 과정을 거친 후에 학습문서에 나타나는 여러 용어(단어 혹은 구)들 중에서 범주화 학습에 유용하게 사용할 만한 용어를 선정하는 자질 선정 과정을 거치게 된다. 그러나 문서 범주화에서 대부분의 자질이 실제로 상당한 정보를 포함하기 때문에 자질 선정으로 인한 정보의 손실을 가져올 수도 있다. 그러므로 적합한 용량으로 학습하도록 고차원 자질 공간을 통제하는 SVM 학습알고리즘은 문서 범주화에 적당한 방법이다(Joaichms 1998). 본 연구에서는 별도의 자질 축소 실험을 수행하지 않고 문헌빈도가 2 이하인 단어들을 제외한 5614개 단어들을 자질로 선정하였다.

본 연구는 Joachims(1998)의 SVM^{light} 소프트웨어[3]를 이용하여 실험을 수행하며, SVM^{light}에서는 학습집단과 검증집단의 문서를 다음과 같은 벡터로 표현한다.

〈line〉.=.〈class〉 〈feature〉:〈value〉...
〈feature〉:〈value〉 ... 〈feature〉:〈value〉

각 〈line〉에서 〈class〉는 문서가 해당 범주에 속하면 +1, 아니면 -1로 표현되며 〈feature〉는 용어, 〈value〉는 용어의 자질값에 해당한다. 각 용어는 실험을 용이하게 하기 위해서 고유번호를 부여하였다.

본 연구에서는 다음의 4 단계로 실험을 수행하였으며, 실험과정의 전체 구성도는 〈그림 3〉과 같다.

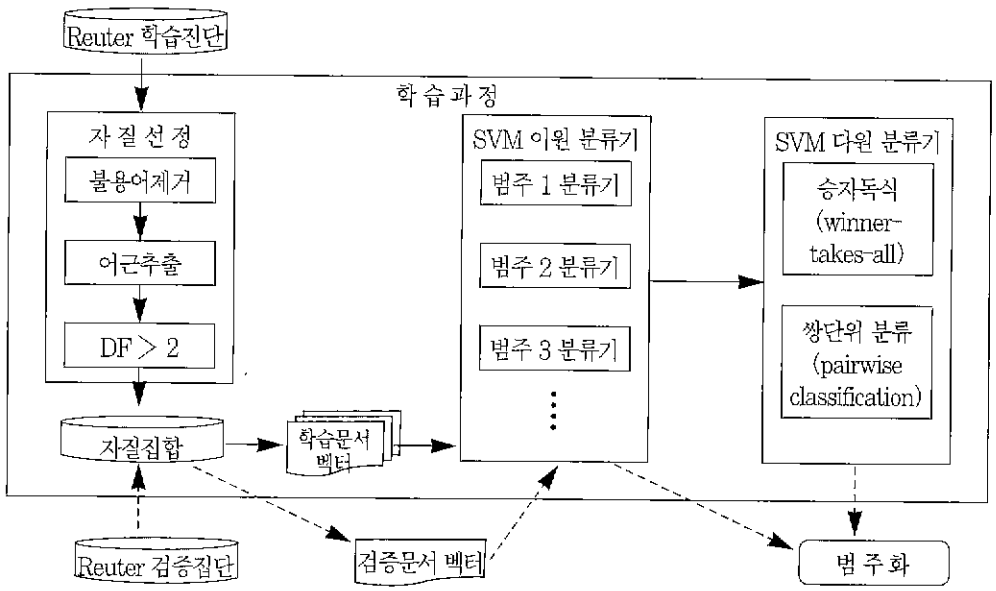
첫째, 다양한 단어빈도 가중치 공식과 SMART 시스템에서 사용되는 역문헌빈도 공식과 코싸인 문헌길이 정규화 공식을 선정된 자질들에 대한 가중치 공식으로 이용하여 그 성능을 비교 분석하였다.

둘째, SVM에서는 학습과정이 이루어지기 전에 사용자가 직접 파라미터 값을 결정해야 하며 파라미터 값에 따라 SVM의 성능은 달라지게 된다. 이렇게 미리 결정해야 하는 파라미터는 마진폭과 분류 오류사이의 타협점을 찾아주는 오류 패널티 변수 C 값과 커널함수의 파라미터이다. 따라서 본 연구에 적합한 C 값과 커널함수의 파라미터 값을 먼저 결정하고, 이를 바탕으로 선형 SVM과 다항식 커널함수 및 RBF 커널함수를 이용한 비선형 SVM의 분류 성능을 비교 분석하였다.

셋째, 이원 패턴 분리 알고리즘인 SVM을 K-범주의 패턴 분리 문제에 적용하기 위해서는 여러 개의 SVM 분류기를 조합하여 함수 $f(x)=y$

2) http://open.muscat.com/stemming에 공개되어 있다.

3) http://www-ai.cs.uni-dortmnd.de/SOFTWARE/SVM_LIGHT/svm_light.eng.html에 공개되어 있다



〈그림 3〉 SVM 분류기를 이용한 문서 범주화 실험 설계도

를 추정해야 한다. 본 연구에서는 SVM 다원 분류기로 확장하기 위해 승자독식 방법과 쌍단위 분류 방법을 이용하여 실험을 수행한 후, SVM 이원 분류기와 다원 분류기의 성능을 비교 분석하였다.

마지막으로 SVM 분류기의 성능을 다른 분류기 성능과 비교 평가하기 위하여 간단하면서도 초기 문서 범주화 연구부터 지금까지 꾸준히 사용되고 있는 나이브 베이즈 분류기를 이용하여 문서 범주화 실험을 수행하였다.

3.1.3 실험결과 평가방법

범주화 성능 평가척도로는 정확률(precision)과 재현율(recall), 정확도(accuracy)를 사용하였다. 각 범주에 대한 분류결과를 표현하는 다음의 2×2 분할표를 이용하면 정확률, 재현율, 정확도 공식은 다음과 같다.

	적합문헌	부적합문헌
범주에 할당	a	b
범주에 할당되지 못함	c	d

$$\text{정확률} = \frac{a}{a+b}$$

$$\text{재현율} = \frac{a}{a+c}$$

$$\text{정확도} = \frac{a+d}{a+b+c+d}$$

특히 반비례 관계에 있는 정확률과 재현율을 하나의 값으로 나타내기 위하여 다음의 F_1척도를 사용하였다(van Rijsbergen 1979).

$$F_1(P,R) = \frac{2P \cdot R}{P+R} \text{ (P : 정확률, R : 재현율)}$$

위의 척도들을 이용하여 각 범주에 대한 분류 성능을 측정하더라도 하나의 범주에 대한 성능을 바탕으로 그 분류기를 평가할 수는 없다. 그러므로 범주별로 측정한 분류 성능을 바탕으로 모든 범주의 성능을 통합하여 평균치를 구하는 것이 바람직하며, 이러한 평균치를 내는 방법으로는 매크로 평균(macro-averaging) 기법과 마이크로 평균(micro-averaging) 기법이 있다. 매크로 평균 기법은 각 범주의 출현 빈도에 관계없이 모든 범주에 같은 가중치를 주고 각 범주별 평균을 구하기 때문에 저빈도 범주에 영향을 많이 받는 질문 지향적 방법으로 정보검색 시스템의 평가를 위해 자주 사용된다. 마이크로 평균 기법은 각 문서에 같은 가중치를 주고 문서별 평균을 구하는 방법으로, 고빈도 범주에 영향을 많이 받는 문서 지향적 방법으로 문서 범주화 방법의 성능 평가 척도로 주로 사용되는 방법이다(Yang, and Liu 1999). 본 연구에서는 고빈도 범주와 저빈도 범주에 대한 분류기의 성능을 모두 평가해야 하므로 두 기법을 함께 사용하였다.

3.2 실험결과 분석

3.2.1 1차 범주화 실험결과

본 연구는 상위 8개 범주들을 대상으로 복수주제를 허용하지 않는 범위에서 실험집단을 선정하였기 때문에 선행연구의 분류 성능과 비교 분석하는 것에는 한계가 있다고 판단된다. 그러므로 어간 추출과정과 불용어 제거 등의 전처리 과정을 거친 모든 자질에 대해서 출현여부를 나타내는 이진값과 단순 TF(term frequency)값을 자질값으로 부여한 후, C = 1000인 선형 SVM 분류기를 이용하여 1차 범주화 실험을 수행하였다. 〈표 2〉에 나타난 것과 같이 정규화를 적용하지 않고 단순 TF만을 자질값으로 부여한 경우가 마이크로 평균 F1 값이 94.83%로 성능이 가장 좋게 나타났다. 기존의 SVM 분류기를 이용한 문서 범주화 실험결과보다 본 연구의 1차 범주화 실험결과가 더 좋게 나타난 것은 복수 주제를 허용하지 않는 범위에서 실험집단을 선정하여 주제간의 관련성을 최소화시킴으로써, 결과적으로 주제 집중도를 높이게 되어 분류 능력을 향상시켰기 때문으로 파

〈표 2〉 1차 범주화 실험결과

자질값		정확도	정확률	재현율	F1값
문헌길이 정규화 미적용					
이진값	macro-avg	98.52	88.55	81 23	84.73
	micro-avg	98.52	95 16	92 93	94.03
단순 TF	macro-avg	98.69	89.90	85.64	87.72
	micro-avg	98.69	95.61	94.06	94.83
문헌길이 정규화 적용					
이진값	macro-avg	98.47	89.75	75.36	81.93
	micro-avg	98 47	95.14	92 48	93.79
단순 TF	macro-avg	98.29	91.61	71.92	80.58
	micro-avg	98.29	95 71	90 41	92.98

악된다. 그러므로 본 연구에서 수행한 다양한 자질값과 선형 SVM 및 비선형 SVM의 분류 성능은 1차 범주화 실험결과를 바탕으로 비교 분석한다.

3.2.2 자질값에 따른 분류 실험결과

자질값을 표현하는 용어가중치는 단어빈도(TF), 역문헌빈도(IDF), 문헌길이 정규화(DL)의 세 가지 요소로 구성된다. 특히 전문검색시스템에서는 길이가 긴 문헌일수록 각 단어의 출현빈도가 높고 출현하는 단어의 종류가 많다는 두 가지 원인 때문에 짧은 문헌에 비해서 검색될 확률이 높아지는 문제가 발생하며, 이는 문서 범주화에서도 마찬가지여서 길이가 긴 문헌일수록 다른 문헌과의 유사도가 상대적으로 높게 될 여지가 있으므로 문헌길이 정규화가 필요하게 된다(이재윤, 최보영, 정영미 2000).

본 연구에서는 다양한 단어빈도, 역문헌빈도, 코사인 문헌길이 정규화의 세 요소를 적용한 용어가중치를 자질값으로 부여하여 분류 실험을 수행하였다.

단어빈도 가중치는 이진값, 단순 TF, 로그 TF, 루트 TF, 보정(augmented) TF, Okapi TF을 사용하였으며, 각 단어빈도 가중치 공식은 〈표 3〉과 같다.

〈표 3〉 단어빈도 가중치 공식

	공식
이진값	$1\ (if\ tf > 0), 0$
단순 TF	$TF = tf$
로그 TF	$TF = 1 + log(tf)$
루트 TF	$TF = \sqrt{tf}$
보정 TF	$TF = (1-w) + w \times \frac{tf}{\max tf}$
Okapi TF	$TF = \frac{tf}{2 + tf}$

역문헌빈도로는 다음의 공식을 이용하였다.

$$IDF(t) = \log(N/df(t))$$

- N : 장서내 전체 문서수
- $df(t)$: 용어 t를 포함하는 문서수

코싸인 정규화 공식은 다음과 같으며, 이 공식에서 w는 정규화시키기 이전의 용어가중치를 나타낸다.

$$\text{정규화 가중치} = \frac{w}{\sqrt{w_1^2}}$$

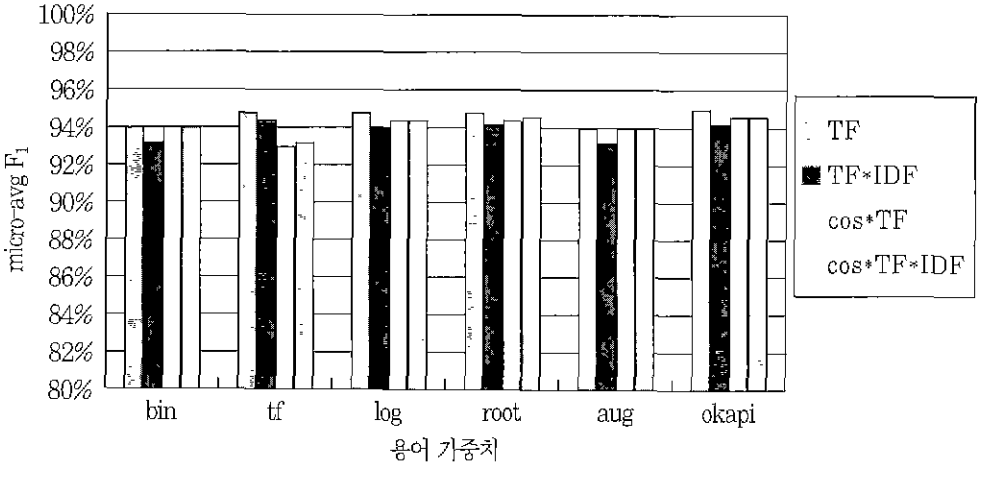
〈표 4〉와 〈그림 4〉는 자질값에 따른 분류 성능을 매크로 평균치와 마이크로 평균치를 이용하여 나타낸 것이다. 문헌길이 정규화 공식의 경우 역문헌빈도를 적용하였을 때 성능이 향상되었지만, 문헌길이 정규화 공식을 적용하지 않았을 때에는 TF만을 사용한 것이 더 좋은 성능을 보였다. 다양한 단어빈도 가중치 중에서 루트 TF를 사용하였을 때 가장 좋은 분류 성능을 보였으며, 그 외에도 Okapi TF와 로그 TF도 우수한 성능을 보이는 것을 알 수 있다. 그러나 각 용어가중치에 따른 성능의 차이는 그다지 크지 않으며, TF만 사용한 경우가 역문헌빈도와 정규화를 적용한 경우보다 우수한 것으로 나타났다. 결과적으로 이진값이나 단순 TF를 자질값으로 이용하더라도 SVM에서는 충분히 좋은 성능을 보이는 것을 알 수 있다.

3.2.3 파라미터 결정을 위한 사전실험

SVM에서는 학습과정이 이루어지기 전에 사용자가 직접 파라미터를 결정해야 하며 파라미터 값에 따라 SVM의 성능은 달라지게 된다. 이렇게 미리 결정해야 하는 파라미터는 학습과정

〈표 4〉 자질값에 따른 분류 성능

			TF				TF*IDF			
			정확도	정확률	재현율	F_1값	정확도	정확률	재현율	F_1값
문헌길이정규화미적용	이진	macro	98.52	88.12	82.37	85.15	98.33	87.13	80.81	83.85
		micro	98.52	94.70	93.37	94.03	98.33	94.20	92.29	93.24
	단순	macro	98.68	89.59	85.52	87.51	98.68	88.20	85.27	86 71
		micro	98 68	95.55	93.82	94.68	98.68	95.32	93.44	94.37
	로그	macro	98.68	89.68	84.53	87.03	98.52	88.94	83.63	86.20
		micro	98.68	95.51	93.87	94.68	98 52	94.95	93.10	94.02
	루트	macro	98 71	89.97	84.72	87.26	98.61	88.96	84.54	86.70
		micro	98.71	95.73	93.87	94.79	98 61	95.11	93.69	94.40
	보정	macro	98.50	88.51	81.18	84.68	98 30	87.12	79.58	83.18
		micro	98.50	94.98	92.92	93.94	98.30	94.31	91.93	93.11
	Okapi	macro	98.73	90 10	85.12	87.54	98.56	88.80	84.23	86.45
		micro	98.73	95.73	94.01	94.86	98.56	94.92	93 51	94.21
문헌길이정규화적용	이진	macro	98.49	90.01	76.55	82.73	98.51	90.58	75.70	82.47
		micro	98.49	95.06	92 74	93 89	98.51	95.95	91.94	93.90
	단순	macro	98.30	91.66	72.62	81.03	98.34	92.03	73.48	81.72
		micro	98.30	95.62	90.54	93.01	98.34	96.34	90.13	93.13
	로그	macro	98.62	90.82	79.67	84.88	98.63	91.31	79.32	84.89
		micro	98.62	95.48	93.38	94.42	98.63	96.08	92 84	94.43
	루트	macro	98.63	90.68	79 97	84.99	98 65	91.72	80.12	85.53
		micro	98 63	95.49	93.46	94.47	98.65	96.13	92 97	94.52
	보정	macro	98.51	90.20	76.74	82.93	98 51	90 77	75.72	82.56
		micro	98.51	95.95	91.98	93.92	98.51	95 95	91.98	93.92
	Okapi	macro	98.63	90.45	79.61	84.68	98.65	91 41	79.79	85.21
		micro	98.63	95.53	93.42	94.46	98.65	96.04	93.02	94.50



〈그림 4〉 자질값에 따른 성능 비교

〈표 5〉 파리미터 값 결정을 위한 사전실험 - C값

		정확도	정확률	재현율	F_1값
			Okapi TF		
C = 1000	macro	98.73	90.11	85.14	87.44
	micro	98.73	95.74	94.01	94.87
C = 100	macro	98.73	90.11	85.14	87.44
	micro	98.73	95.74	94.01	94.87
C = 10	macro	98.72	90.42	84.55	87.27
	micro	98.72	95.86	93.83	94.84
C = 1	macro	98.69	91.19	83.85	87.23
	micro	98.69	95.94	93.52	94.71
			정규화(루트 TF*IDF)		
C = 1000	macro	98.65	91 72	80.12	85.53
	micro	98.65	96 13	92.97	94.52
C = 100	macro	93.95	74 65	20.25	26.42
	micro	93.95	100.00	52.41	68.78
C = 10	macro	87 50	0.00	0.00	0.00
	micro	87.50	-	0 00	-
C = 1	macro	87.50	0.00	0.00	0.00
	micro	87.50	-	0.00	-

에서 마진폭과 분류 오류 사이의 타협점을 찾아주는 오류 패널티 변수 C 값과 비선형 SVM에 적용하는 커널함수의 파라미터이며, 본 연구에서는 SVM 분류기를 이용한 문서 범주화 실험에 앞서 이에 적합한 C 값과 커널함수의 파라미터 값 결정을 위해 다음의 사전실험을 수행하였다.

(1) C 값 결정을 위한 사전실험

C 값은 마진폭과 분류 오류 사이의 타협점(trade-off)을 찾아주는 역할을 담당하며, 분리할 수 없는 데이터에 대한 오류 패널티 값이다. 일반적으로 C 값이 0으로 수렴할수록 학습데이터를 정확하게 분류하는 용어보다는 마진을 최대화하는 용어에 더 중점을 둠으로써 아주 넓은 마진폭을 갖는 간단한 모형을 구축하게 되며, C 값이 커질수록 최적의 초평면을 구축하여 학습집단의 모든 데이터를 정확하게 분류하려는 경향이 있다(Gunn 1998). 그러나 C 값을 아주 크게 정의하게 되면 입력데이터에 대해서 정확하게 분류할 수 있다고 하더라도 오류가 포함된 선형 분리가 가능하지 않은 데이터에 대해 분류 성능은 보장할 수 없게 된다. 결국 적당한 C 값을 선정하는 것은 모형 복잡도를 통제하여 일반화 성능을 향상시키는 효과를 지니게 된다

그러므로 본 연구에 적합한 C 값 결정을 위해 자질값에 따른 분류 실험에서 가장 좋은 성능을 보인 Okapi TF와 정규화한 루트 TF*IDF 가중치를 자질값으로 이용하여 선형 SVM으로 실험하였다. 실험결과 〈표 5〉에서 나타난 것과 같이 C 값이 1000일 때 가장 우수한 성능을 보였으며, 이는 최적의 초평면을 구축하여 학습집단의 모든 데이터를 정확하게 분류하고 모형 복

〈표 6〉 파리미터 값 결정을 위한 사전실험 - 다항식 커널함수

		정확도	정확률	재현율	F_1값
		Okapi TF			
d = 2	macro	98.72	91.44	81.30	86.07
	micro	98.72	96.37	93.24	94.78
d = 3	macro	98.38	89.29	73.57	80.67
	micro	98.38	96.00	90.85	93.35
d = 4	macro	98.03	88.11	68.18	76.88
	micro	98.03	95.25	88.68	91.85
d = 5	macro	97.55	87.76	61.67	72.44
	micro	97.55	93.76	85.45	89.41
		정규화(루트 TF*IDF)			
d = 2	macro	98.78	89.76	86.97	88.09
	micro	98.78	95 17	95 00	95.08
d = 3	macro	98.86	89.59	89.98	89.63
	micro	98.86	95.08	95.85	95 46
d = 4	macro	98.84	88.41	90.45	89.29
	micro	98.84	94.75	96 03	95.39
d = 5	macro	98.79	87.11	90.51	88 67
	micro	98 79	94.33	96.08	95.20

잡도를 통제하여 일반화 성능을 향상시켰기 때문으로 해석된다.

(2) 커널함수의 파라미터 결정을 위한 사전 실험

SVM은 기본적으로 선형 분리가 가능한 문제에서 출발하지만, 모든 문제가 선형적으로 분리될 수 없기 때문에 SVM에서는 고차원의 자질공간을 효율적으로 처리하기 위해서 커널함수 $k(x, x_i = \phi(x), \phi(x_i))$를 이용하게 된다 (Vapnik 2000). 비선형 SVM의 범주화 성능 평가를 위한 실험에서는 다항식 커널함수와 RBF 커널함수를 이용하며, 커널함수의 파라미터 값 결정을 위해 Okapi TF와 정규화한 루트 TF*IDF 가중치를 자질값으로 부여한 후 C 값은 1000으로 하여 사전실험을 수행하였다.

· 다항식 커널함수

$$k(x, x_i) = ((x \cdot x_i) + 1)^d$$

· RBF 커널함수

$$k(x, x_i) = \exp(-\gamma \| x - x_i \|^2)$$

그 결과 〈표 6〉과 〈표 7〉에서와 같이 자질값으로 Okapi TF를 가중치를 사용한 경우 다항식 커널함수에서는 degree 가 2일 때 가장 우수한 분류 성능을 보였으며, RBF 커널함수는 γ값이 0.1일 때 가장 우수한 성능을 보였다. 또한 정규화한 루트 TF*IDF 가중치를 자질값으로 이용한 경우 다항식 커널함수에서는 degree 가 3일 때, RBF 커널함수는 γ값이 2일 때 가장

〈표 7〉 파리미터 값 결정을 위한 사전실험 - RBF 커널함수

		정확도	정확률	재현율	F1값
		Okapi TF			
$\gamma = 0.1$	macro	98.67	95.25	76.98	85.15
	micro	98.67	97 78	91.39	94.48
$\gamma = 0.5$	macro	95.93	84 94	30.19	44.54
	micro	95.93	99 32	66.86	79.92
$\gamma = 1$	macro	94.03	85.81	20.02	26.78
	micro	94.03	99.57	52.41	68.67
$\gamma = 2$	macro	91.85	74.98	11.58	20.06
	micro	91.85	99.87	34.85	51.67
$\gamma = 3$	macro	90.70	49.98	8.93	12.21
	micro	90.70	99.84	28.14	43.91
		정규화(루트 TF*IDF)			
$\gamma = 0.1$	macro	96.13	71.46	40.99	52.10
	micro	93.13	98.30	70.25	81.94
$\gamma = 0.5$	macro	98.65	91.62	80.12	84.38
	micro	98.65	96.09	92.97	94.50
$\gamma = 1$	macro	98.78	89.74	86.87	88.01
	micro	98.78	95.21	95 00	95.10
$\gamma = 2$	macro	98.83	88.41	90 44	89.28
	micro	98.83	94.75	95.99	95.36
$\gamma = 3$	macro	98.78	86.46	90.71	88.44
	micro	98 78	94.21	96.17	95.18

우수한 성능을 나타내었다.

3.2.4 SVM 분류기 성능 평가

Okapi TF와 정규화한 루트 TF*IDF 가중치를 이용하여 자질값을 표현한 후, 파라미터 값 결정을 위한 사전실험에서 우수한 성능을 보인 C 값과 비선형 SVM에 적용하는 커널함수의 파라미터 값을 이용하여 SVM 이원 분류기를 평가하였다. 〈표 8〉에서 나타난 것과 같이 Okapi TF를 자질값으로 이용한 경우 선형 SVM이 비선형 SVM 보다 다소 우수한 성능을 보였으나, 정규화한 루트 TF*IDF 가중치를 적용시켜 자질값으로 사용하였을 때에는 다항식 커널함수를 적용한 비선형 SVM이 우수한 분류 성능을 보였다. 그러나 전반적으로 선형 SVM과 비선형 SVM의 정확도는 거의 차이가 없으며, F1 값에서 약간의 차이를 보이고 있다. 그러나 비선형 SVM의 경우 학습과정 전에 미리 커널함수의 파라미터 값을 결정해야 하며, 결정된 파라미터들은 국부적인 해(local minimum)를 찾을 가능성이 크고 모형 복잡도 역시 선형 함수보다 크다. 그러므로 분류 성능과 모형 복잡도를 모두 고려했을 때 선형 SVM이 문서 범주화에 간단하고 효율적인 방법으로 평가된다.

본 연구의 SVM 다원 분류기 성능 평가를 위해서는 이원 분류기에서 모든 커널함수에 대해

〈표 8〉 SVM 이원 분류기의 분류 성능

SVM			정확도	정확률	재현율	F_1값
			Okapi TF			
선형 SVM		macro	98.73	90 10	85.12	87.54
		micro	98.73	95.73	94.01	94 86
비선형	다항식 (d = 2)	macro	98.72	91.44	81.30	86 07
		micro	98 72	96.37	93.24	94.78
	RBF (γ = 0.1)	macro	98.67	95.25	76.98	85.15
		micro	98.67	97.78	91.39	94.48
			정규화(루트 TF*IDF)			
선형 SVM		macro	98.65	91.72	80.12	85.53
		micro	98.65	96.13	92.97	94.52
비선형	다항식 (d = 3)	macro	98.86	89.59	89 98	89.63
		micro	98.86	95.08	95 85	95.46
	RBF (γ = 2)	macro	98.83	88.41	90 44	89.28
		micro	98.83	94.75	95.99	95.36

〈표 9〉 SVM 다원 분류기의 분류 성능

SVM			정확도	정확률	재현율	F_1값
			승자독식 방법			
선형 SVM		macro	98.96	91.44	89.94	90.53
		micro	98.96	95.86	95.86	95 86
비선형	다항식 (d = 3)	macro	99.01	91 08	89.99	90.37
		micro	99.01	96 04	96.04	96.04
	RBF (γ = 2)	macro	99 10	91.70	90.36	90 89
		micro	99.10	96 40	96.40	96.40
			쌍단위 분류 방법			
선형 SVM		macro	98.06	86.28	70.48	77.58
		micro	98.06	92.83	91.58	92.20
비선형	다항식 (d = 3)	macro	98.05	73.12	69.74	71.39
		micro	98.05	92.90	91.36	92.12
	RBF (γ = 2)	macro	98.06	73 69	70.06	71.83
		micro	98.06	92.87	91.54	92.20

우수한 성능을 보였던 정규화한 루트 TF*IDF 가중치를 자질값으로 이용하여 문서 범주화 실험을 수행하였다. 그 결과 〈표 9〉에 나타난 것과 같이 승자독식 방법이 쌍단위 분류 방법보다 우수한 분류 성능을 보였다. 더욱이 쌍단위 분류 방법의 경우 본 연구에서 선정한 8개의 범주

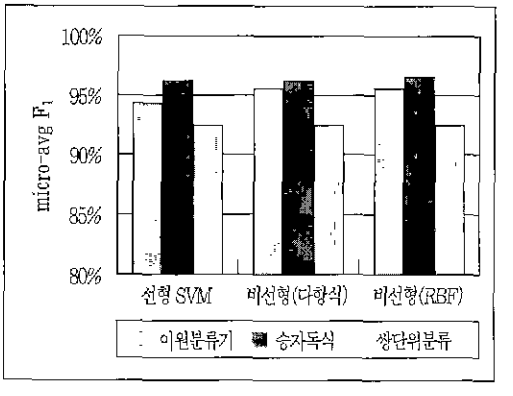
〈표 10〉 SVM 이원 분류기와 다원 분류기의 성능 비교

SVM			정확도	정확률	재현율	F_1값
SVM 이원 분류기						
선형 SVM		macro	98.65	91.02	80.12	85.53
		micro	98.65	96.13	92.97	94.52
비선형	다항식 (d = 3)	macro	98.86	89.59	89.98	89.63
		micro	98.86	95.08	95.85	95.46
	RBF (γ = 2)	macro	98.83	88 41	90.44	89.28
		micro	98.83	94.75	95.99	95.36
SVM 다원 분류기			승자독식 방법			
선형 SVM		macro	98.96	91.44	89.94	90.53
		micro	98 96	95.86	95.86	95.86
비선형	다항식 (d = 3)	macro	99.01	91.08	89.99	90.37
		micro	99 01	96.04	96.04	96.04
	RBF (γ = 2)	macro	99.10	91.70	90.36	90.89
		micro	99.10	96.40	96.40	96.40
SVM 다원 분류기			쌍단위 분류 방법			
선형 SVM		macro	98.06	86 28	70.48	77.58
		micro	98.06	92.83	91.58	92.20
비선형	다항식 (d = 3)	macro	98.05	73 12	69.74	71.39
		micro	98.05	92.90	91.36	92.12
	RBF (γ = 2)	macro	98.06	73.69	70 06	71.83
		micro	98.06	92.87	91.54	92.20

를 대상으로 28개의 분류기를 이용하여 학습과정을 진행시켜야 하기 때문에 범주 결정과정도 복잡하고 시간도 많이 소요되었다. 그러므로 승자독식 방법이 문서 범주화에 적합한 SVM 다원 분류기 확장 방법인 것으로 평가된다.

3.2.5 SVM 이원 분류기와 다원 분류기의 성능 비교

SVM 이원 분류기와 승자독식 방법 및 쌍단위 분류 방법에 의해 다원 분류기로 확장하여 문서 범주화 실험을 수행한 결과가 〈표 10〉과 〈그림 5〉에 나와있다. 이 실험에서 승자독식 방법을 이용한 SVM 다원 분류기의 성능이 SVM



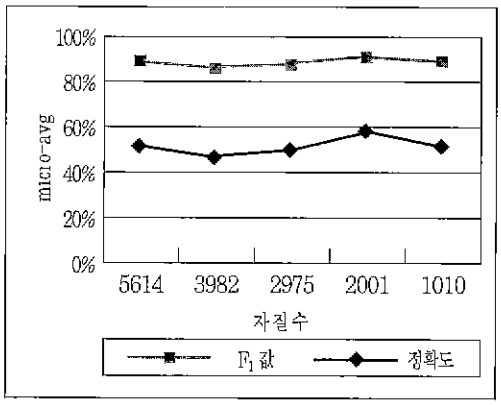
〈그림 5〉 이원 분류기와 다원 분류기의 성능 비교

이원 분류기보다 다소 향상되었다. 이원 분류기의 경우 다항식 커널함수를 적용한 비선형 SVM이 가장 우수한 성능을 보인 반면, 다원 분류기 실험에서는 선형 SVM과 RBF 커널함수를 이용하였을 때 더 좋은 성능을 나타냈다. 이는 방향성과 차수에 의존적인 다항식 커널함수의 특성으로 인해 K개의 분류기를 이용하여 학습에서 얻어진 결과를 하나로 수렴하는 것이 어렵기 때문으로 해석된다.

그러나 SVM 분류기를 이용하여 다원 범주에 대한 결정면을 찾아내기 위해서 8개의 범주를 대상으로 실험을 수행한 결과 승자독식 방법의 경우 8개의 분류기를 학습시켜야 하고, 쌍단위 분류 방법에서는 28개의 분류기를 학습시켜야 하기 때문에 범주 결정과정도 복잡하고 시간도 많이 소요된다. 그러므로 분류 성능과 학습과정의 복잡도를 모두 고려해봤을 때 SVM 이원 분류기가 문서 범주화에 더 적합한 것으로 평가된다.

3.2.6 SVM 분류기와 나이브 베이즈 분류기의 성능 비교

SVM 분류기의 성능을 다른 분류기 성능과 비교 평가하기 위해서 간단하면서도 초기 문서 범주화 연구부터 지금까지 꾸준히 사용되고 있는 나이브 베이즈(Naive Bayes)[4] 분류기를 이용하여 문서 범주화 실험을 수행하였다. 기존의 선행연구에서 SVM 분류기는 자질 축소를 하지 않더라도 우수한 분류 성능을 보였지만, 나이브 베이즈 분류기의 경우 자질 축소에 따라 분류 성능이 영향을 받는 것으로 나타났다 (Joachims 1998). 그러므로 나이브 베이즈 분류기 실험을 위해서 문헌빈도를 자질 선정 기준으로 사용하여 1010개(DF>41), 2001개(DF>14), 2975개(DF>7), 3982개(DF>4), 5614개(DF>2)의 자질을 선정하여 자질 축소 실험을 수행하였으며, <그림 6>에서 나타난 것과 같이 가장 우수한 성능을 보인 2001개(DF>14)개의 자질을 선정하였다. 두 분류기 모두 자질값으로 이진값을 이용하여 문서 범주화 실험을 수행하였다.



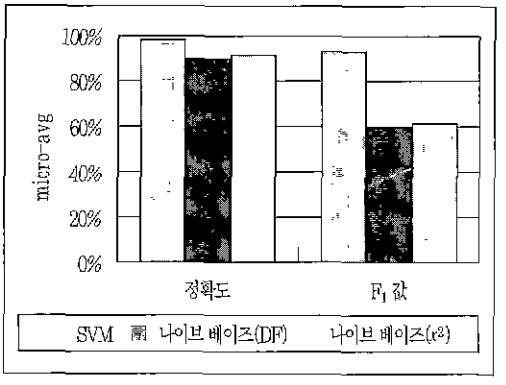
<그림 6> 자질수에 따른 나이브 베이즈의 성능 변화

실험결과 <표 11>과 <그림 7>에서처럼 나이브 베이즈 분류기는 마이크로 평균 F1 값이 59.64%를 보인 반면, SVM 분류기는 94.03%의 성능을 보임으로써 나이브 베이즈 분류기보다 우수한 학습방법임을 증명하였다. 특히 기존의 자질 선정과 관련된 연구(Yang and Perdersen 1997)에서 우수한 성능을 보였던 카이제곱 통계량을 자질 선정 기준으로 사용하여, 2000개의 자질을 선정한 후 나이브 베이즈의 성능을 평가해 본 결과, 마이크로 평균 F1 값이 60.98%를 보임으로써 SVM 분류기를 이용하여 문서 범주화 실험을 수행하였을 때보다 여전히 낮은 성능을 보였다.

4) http //robotics.stanford.edu/users/sahami에 공개되어 있다.

〈표 11〉SVM 분류기와 나이브 베이즈 분류기의 성능 비교

	SVM		나이브 베이즈			
			문헌빈도		카이제곱 통계량	
	정확도	F_1 값	정확도	F_1 값	정확도	F_1 값
earn	98.60	98.53	70.36	76.60	75.63	78.35
acq	97.57	95.66	69.73	49 24	78.96	47.94
money-fx	97.88	75.90	92.52	48.46	94.19	53.76
grain	99.23	92.77	92.16	4.49	92.88	3.66
crude	98.83	88.79	92.93	15.76	94.32	14.86
trade	98.87	87.18	92.21	29.56	92.48	35.02
interest	98.02	62.07	95.99	5.26	96.85	14 63
ship	99.19	74.29	97.16	0.00	97.66	7.14
macro	98.52	84.73	87.88	28.67	90.37	31.92
micro	98.52	94.03	87.88	59.64	90.37	60.98



〈그림 7〉 SVM과 나이브 베이즈 분류기의 성능 비교

4 결 론

본 연구에서는 Reuters-21578 ModApte 분할판을 실험문서집단으로 하여 SVM 분류기를 사용한 문서 범주화 실험을 수행하였으며, 자질값과 선형 SVM 및 커널함수에 따른 비선형 SVM의 성능을 비교하였다. 또한 이원 분류기를 다원 분류기로 확장할 수 있는 방안에 대하여 검토한 후, SVM 이원 분류기와 다원 분류기의 성능을 비교 분석하였다. 본 연구를 통해 밝혀진 사실은 다음과 같다.

첫째, 자질값에 따른 성능 비교 실험에서는 문헌길이 정규화 공식의 경우 역문헌빈도를 적용하였을 때 성능이 향상되었지만, 문헌길이 정규화 공식을 적용하지 않았을 때에는 TF만을 사용한 것이 더 좋은 성능을 보였다. 다양한 단어빈도 가중치 중에서 루트 TF를 사용하였을 때 가장 좋은 분류 성능을 보였으며, Okapi TF와 로그 TF도 우수한 성능을 나타냈다. 그러나 각 용어가중치에 따른 성능의 차이는 그다지 크지 않으며, 이진값이나 단순 TF를 자질값으로 이용하더라도 SVM에서는 충분히 좋은 성능을 보이는 것을 알 수 있다.

둘째, SVM에서는 학습과정이 이루어지기 전에 사용자가 마진폭과 분류 오류 사이의 타협점을 찾아주는 오류 패널티 변수 C 값과 비선형 분류기에서 적용되는 커널함수의 파라미터를 직접 결정해야 한다. 본 연구에서는 문서 범주

화에 적합한 SVM의 오류 패널티 변수 C 값과 비선형 SVM에 적용되는 커널함수의 파라미터를 결정하기 위해 사전실험을 수행하였으며, 사전실험에서 우수한 성능을 보였던 C 값과 커널함수의 파라미터 값을 이용하여 SVM 이원 분류기의 성능 비교 실험을 수행하였다. 실험결과를 보면 선형 SVM과 비선형 SVM의 정확도는 거의 차이가 없으며, F_1 값에서 약간의 차이를 보이고 있기 때문에, 선형 SVM이 학습 과정 전에 미리 커널함수의 파라미터 값을 결정해야 하는 비선형 SVM에 비해 문서 범주화에 간단하고 효율적인 방법이라고 평가된다.

셋째, SVM 이원 분류기를 다원 분류기로 확장하여 실험을 수행한 결과 승자독식 방법은 95.86%(선형 SVM의 마이크로 평균 F_1 값)을 보임으로써, 92.20%(선형 SVM의 F_1 값)의 쌍단위 분류 방법보다 우수한 분류 성능을 나타냈다. 이것은 SVM 이원 분류기의 분류 성능(선형 SVM의 F_1 값 94.52%)보다도 다소 향상된 결과이다. 그러나 분류 성능과 학습과정의 복잡도를 모두 고려해봤을 때 SVM 이원 분류기가 문서 범주화에 더 적합한 것으로 평가된다.

마지막으로 초기 문서 범주화 연구부터 지금까지 꾸준히 사용되고 있는 나이브 베이즈(Naive Bayes) 분류기를 이용하여 문서 범주화 실험을 수행한 후 SVM 분류기와의 성능을 비교 분석해본 결과, 나이브 베이즈 분류기는 마이크로 평균 F_1 값이 59.64%를 보인 반면, SVM 분류기는 94.03%의 성능을 보임으로써 나이브 베이즈 분류기보다 우수한 학습방법임을 증명하였다.

SVM 학습알고리즘에서는 이용자가 직접 파라미터 값을 결정하여 실험을 수행해야 하며 파라미터 값에 따라 분류기의 일반화 성능이 달라지게 된다. 가장 적합한 파라미터 값을 결정하기 위해서는 파라미터에 대한 이론적인 이해와 실험과정에서 시행 착오를 거치면서 파라미터 값을 조정해야 하기 때문에 SVM 이론에 익숙하지 않은 이용자가 직접 파라미터를 결정한다는 것은 어려운 일이다. 그러므로 이러한 파라미터 값들을 SVM의 VC 차원과 관련하여 기계학습과정에서 결정될 수 있는 연구도 진행되어야 할 것이다.

참고문헌

오장민, 장병탁, 김영택. 1999. "SVM 학습을 이용한 다중 클래스 뉴스그룹 문서 분류." 『한국정보과학회 가을 학술발표 논문집 (II)』, 26(2) : 60-62.

이재윤, 최보영, 정영미. 2000 "문헌 자동분류에서 용어가중치 기법에 대한 연구." 『제7회 한국정보관리학회 학술대회 논문집』. 41-44.

최성환, 임혜영, 정영미. 2000. "SVM을 이용한 한글문서 범주화 실험." 『제7회 한국정보관리학회 학술대회 논문집』, 29-32.

Chin, K. K. 1998. Support Vector Machines *applied to speech pattern classification.* Ph.D.diss., Darwin College. [online]. [cited 2000.10.03]. <http ://svr - www. eng. cam. ac. uk/ ~kkc21 / thesis_main /thesis_main. html>.

Cristianini, Nello and John Shawe-Taylor. 2000. An introduction to Support Vector Machines and other kernel-based learning methods. Cambridge : Cambridge University Press.

Dumais, Susan et al. 1998. "Inductive learning algorithms and representations for text categorization." Proceedings of ACM-CIKM 98, 148-155. [online]. [cited 2000.09.18]. <http://research.microsoft.com/~sdumais/cikm98.doc>.

Gunn, Steve. 1998. "Support Vector Machines for classification and regression." University of Southampton. [online]. [cited 2000.09.18]. <http://www.isis.ecs.soton.ac.uk/resources/svminfo>.

Hastie, T. and R. Tibshirani. 1996. "Classification by pairwise coupling." Stanford University and University of Toronto. [online]. [cited 2000.09.30]. <http://www-stat.stanford.edu/~hastie>.

Joachims, Thorsten. 1997. "SVM light: Implementation of the decomposition *training algorithm.*" [online]. [cited 2000.09.01]. <http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT>.

Joachims, Thorsten. 1998. "Text categorization with Support Vector Machines: Learning with many relevant features." Proceedings 10th European Conference on Machine Learning (ECML). [online]. [cited 2000.09.18]. <http://www-ai.cs.uni-dortmundde/SOFTWARE/SVM_LIGHT>.

Kressel, Ulrich H. G. 1999. "Pairwise classification and Support Vector Machines." In Advanced in kernel methods: Support Vector Machines. Cambridge: MIT Press.

Salton, G., and C. Buckley. 1988. "Term-

weighting approaches in automatic text retrieval." *Information Processing & Management*, 24(5): 513-523.

Scholkopf, Bernhard, Chris Burges, and Alex J. Smola. 1999. Advances in kernel methods : Support Vector Machines. Cambridge : MIT Press.

van Rijsbergen, C. J. 1979. Information Retrieval. London : Butterworths.

Vapnik, V. 2000. The nature of Statistical Learning Theory. 2nd ed. New York : Springer-Verlag.

Weston, C., and C. Watkins. 1998. "Multi-class Support Vector Machines." Royal Holloway University of London. [online]. [cited 2000. 10. 05] <http://www.clrc.rhbnc.ac.uk>

Yang, Yiming and J. O. Pedersen 1997. "A comparative study on feature selection in text categorization." Machine Learning: Proceedings of the Fourteenth International Conference (ICML97), 412-420.

Yang, Yiming and Xin Liu. 1999. "A re-examination of text categorization methods." Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 42-49. [online]. [cited 2000.09 03].<http://www.cs.cmu.edu/~yiming> .