SVM을 이용한 한글문서 범주화 실험

Categorization of Korean documents using Support Vector Machines

최성환, 임혜영, 정영미(연세대학교 대학원 문헌정보학과)

Choi Sung-Hwan, Lim Hye-Young, Chung Young-Mee
Dept. of Library and Information Science, Graduate School of Yonsei University

자동 문서 범주화에 이용되는 학습분류기 중에서 SVM은 자질 차원을 축소하지 않고도 좋은 성능을 보이고 있다. 본 실험에서는 KTSET 텍스트 컬렉션을 대상으로 두 개의 SVM 분류기를 이용하여 자질축소 및 자질표현에 따른 성능비교 실험을 하였다. 자질축소를 위하여 χ^2 통계량을 자질선정기준으로 사용하였으며, 자질값으로는 단어빈도 및 문헌빈도의 두 요소로 구성되는 다양한 가중치를 사용하였다. 실험결과 SVM은 자질축소에 큰 영향을 받지 않고 가중치 유형에 따라 성능의 차이를 보였다.

1. 서 론

오늘날 온라인 정보의 급속한 증가는 모든 주제에 대해 손쉬운 접근을 가능하게 한다. 그러나 대량의 정보를 적절하게 활용하기는 쉽지 않으며, 앞으로도 이런 현상은 계속될 것으로 보인다. 대량의 정보를 효과적으로 조직하기 위한 한 가지 접근방법은 주제에 따라 문서들을 자동으로 범주화시키는 것이다.

문서 범주화의 특징은 고차원의 입력공간을 가진다는 것이다. 고차원의 입력공간을 피하기 위한 방법은 대부분의 자질이 부적합하다고 가정하고 문서를 가장 잘 표현할 수 있는 자질들만을 선정하는 것이다. 그러나 문서 범주화에서 대부분의 자질이 실제로 상당한 정보를 포함하기 때문에 자질 선정으로 인한 정보의 손실을 가져올 수도 있다. 또한 문서 범주화에서 문서는 용어들의 벡터로 표현되며, 이러한 문서벡터는 수많은 용어 중에서 실제로 출현한 용어들만이 값을 갖는 밀도가 희박한 벡터(sparse vector)가 되기 쉽다. 본 연구에서는 이런 특징을 갖는 문서들의 범주화에 적합한 SVM 분류기를 이용하여 문서 범주화 실험을 수행하였다.

2. SVM 학습알고리즘

SVM은 Vapnik에 의해 1979년에 제안되었지만 최근에 이르러서야 문서 범주화 등 여러 분야에서 많은 연구가 진행되고 있다. SVM의 가장 간단한 형태는 <그림 1>과 같이 최대 마진(margin)을 가지고 부정 예제로부터 긍정 예제를 분류해 낼 수 있는 결정면(decision surface)을 찾아내는 선형분류모형이다(Dumais 1998).

<그림 1>에서 실선은 긍정예제와 부정예제를 분리하는 결정면이고, 실선과 평행인 점선

들은 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.

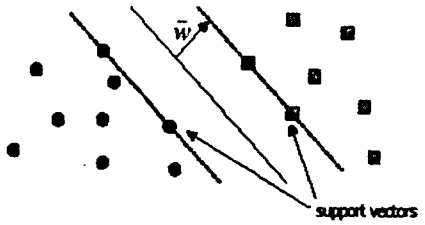

그림 1. 선형 SVM

SVM의 기본 원리는 선형 분리 가능한 문제에서 출발한다. 선형 분리가 가능하다는 것은 학습데이터를 두 집합 즉, 긍정예제와 부정예제로 구분짓는 결정면이 존재한다는 것이다. 이 결정면은 다음 수식과 같이 나타낼 수 있다.

$$\vec{w} \cdot \vec{x} - b = 0$$

여기서 $\vec{w}$는 가중치벡터, $\vec{x}$는 입력벡터, b는 기준치로, $\vec{w}$와 b는 학습 데이터로부터 학습된다. 학습문서 집합을 $D = \{(\vec{x}_i, y_i)\}$ 라고 할 때, 입력벡터 $\vec{x}_i$가 범주(class)에 속하면 y_i는 +1의 값을 갖고, 속하지 않으면 -1의 값을 갖는다. 결국 SVM은 최적의 $\vec{w}$와 b를 찾는 문제이다.

$$\vec{w} \cdot \vec{x}_i - b \geq +1 \quad for \quad y_i = +1$$
$$\vec{w} \cdot \vec{x}_i - b \leq -1 \quad for \quad y_i = -1$$

3. 범주화 실험 및 결과

3.1. 실험집단 및 문서표현

본 실험에서는 KTSET 테스트 컬렉션을 이용하여 각각 50건 이상의 문서를 포함하는 6개 범주를 대상으로 실험하였다. 전체 488건의 문서를 7:3 비율로 학습집단 342건, 검증집단 146건으로 나누었으며 구체적인 실험집단 구성은 <표 1>과 같다.

범 주	분류번호	총 문서수	학습집단	검증집단
1	C2	93	65	28
2	D1	67	47	20
3	D2	74	52	22
4	D4	65	46	19
5	H2	76	53	23
6	I2	113	79	34

표 1. 실험집단

문서의 색인어 추출은 한성대학교에서 개발한 형태소 분석기 HAM을 이용하였다. 전처리 단계를 거친 후 SVM^{light}은 학습집단과 검증집단의 문서를 다음과 같이 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 위해 고유번호를 부여하였다.

3.2. 실험환경

Visual Foxpro6.0 프로그램을 이용하여 사전구축 및 자질표현을 한 후, UNIX(r) System에서 Joachims(1998)의 SVM^{light} 소프트웨어를 사용하여 실험하였다. 일반적으로 문서범주화에서는 다양한 자질선정기법을 통해 자질공간을 축소시키는 동시에 성능을 향상시키는 실험

이 많이 이루어져왔다. 그러나 SVM은 자질공간을 축소하지 않고도 다른 학습분류기에 비해 성능이 우수하다는 연구결과들이 나와있다 (Joachims 1998; Dumais 1988). 본 연구에서는 SVM 분류기를 이용하여 자질축소 및 용어가중치 유형에 따른 SVM의 범주화 성능을 실험하였다.

3.3. 실험 결과

실험 결과에 대한 평가방법으로는 정확도(accuracy)를 사용하였다. 자질축소 실험에서는 자질선정기준으로 χ^2 통계량을 사용하였으며 200개씩 자질수를 축소하여 실험하였다. 문헌벡터의 자질값으로는 binary, 전통적인 TFIDF, 스파크 존스가 제시한 역문헌빈도에 TF를 곱한 TFsparck, 정보이론에 근거한 신호량에 TF를 곱한 IT를 용어가중치로 사용하였다. <그림 2>를 보면 자질수를 29% 축소했을 때 가장 좋은 성능을 보였으며, binary로 자질표현을 했을 때 가장 좋은 성능을 보이고 있다. 전체적으로 SVM이 자질축소보다는 가중치 유형에 의해 영향을 많이 받고 있다는 것을 알 수 있다. 이것은 SVM이 자질공간의 차원에 독립적이며 자질의 수가 아니라 데이터를 분리해내는 마진에 기반하기 때문이다. 즉 데이터가 자질공간으로부터 함수를 이용하여 최대 마진으로 분리해낼 수 있다면 방대한 자질공간에서조차 우수한 성능을 보일 수 있다는 것을 의미한다.

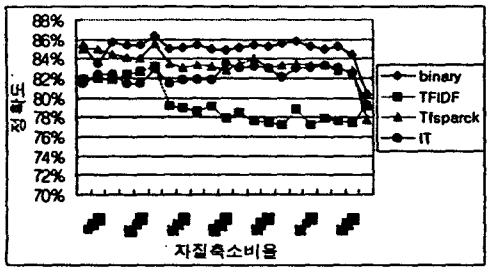

그림 2. 자질축소에 따른 성능변화

SVM은 기본적으로 선형 분리가 가능한 문제에서 출발하지만, 모든 문제가 선형적으로 분리될 수는 없다. 이런 경우 SVM에서는 고차원의 자질공간을 효율적으로 처리하기 위해서 커널함수 ($k(x,y):=(\Phi(x),\Phi(y))$을 이용하며, 이 커널함수는 <그림 3>과 같이 입력공간을 자질공간으로 사상시키는 역할을 한다.

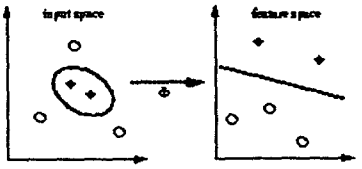

그림 3. SVM의 기본개념

커널함수별 성능비교 실험에서는 문헌빈도가 2 이하인 용어를 제거한 후 binary, TF, 스파크존스의 역문헌빈도, TFsparck, IT로 문서를 표현하였으며, 실험에 사용한 커널함수는 다음과 같다.

- Linear kernel
$$k(x,y)=x\cdot y$$
- Polynomial kernel
$$k(x,y)=((x\cdot y)+1)^d$$
- Radial Basis Function kernel
$$\exp(-\gamma\|x-y\|^2)$$

실험결과 <표 2>에서 나타난 것처럼 binary로 자질값을 표현했을 때 RBF 함수의 분류정확도가 86.76%(γ=0.01)로 선형함수보다 좋은 성능을 보였다. 그러나 Polynomial과 RBF함수는 에러를 최소로 하는 파라미터들이 국부적 해(local minimum)를 찾을 가능성이 크며 모델의 복잡도가 선형함수보다 크다. 그러므로 분류성능과 모델의 복잡도를 고려했을 때 선형 SVM이 문서범주화에 간단하고 효율적인 방법이라고 할 수 있다.

	linear	polynomial (degree)		RBF (γ)		
		(2)	(3)	(0.01)	(0.1)	(1.0)
binary	86.07	85.16	84.02	86.76	83.45	83.33
TF	83.79	82.19	81.85	73.63	83.33	83.33
sparck	85.16	82.19	83.33	83.33	83.33	83.33
TFsparck	84.71	83.11	82.99	83.33	83.33	83.33
IT	81.39	81.96	81.51	83.33	83.33	83.33
평균	83.95	82.92	82.74	82.08	83.35	83.33

표 2. 가중치에 따른 커널 실험 결과

분류기 비교 실험에서는 가장 좋은 성능을 보였던 binary를 자질값으로 하여 Joachims의 SVM^{light}와 Royal Holloway University of London에서 개발한 SVM 분류기의 성능을 비교하였다. <표 3>을 보면 SVM^{light}가 RHUL의 SVM 보다 약 10% 이상의 높은 분류정확도를 보이는 것을 알 수 있다. 분류기의 또 다른 평가척도인 학습시간과 새로운 문서의 분류시간은 SVM^{light}의 경우 CPU 시간으로 평균 0.46초, RHUL SVM은 평균 101초 소요되었다. 즉, 분류의 정확도와 학습시간/분류속도의 측면에서도 SVM^{light}가 RHUL SVM 보다 문서범주화에서 더 우수한 성능을 보여주었다.

범주	SVM^{light}			RHUL SVM		
	linear	poly d=2	RBF γ=0.01	linear	poly d=2	RBF γ=0.01
1	91.78	86.30	92.47	73.29	80.14	82.19
2	80.82	82.19	81.51	82.88	84.25	86.30
3	84.25	84.93	86.30	76.71	78.77	84.93
4	88.36	89.04	89.73	82.88	85.62	86.30
5	87.67	85.62	87.67	75.34	80.82	84.25
6	83.56	82.88	82.88	60.27	68.49	76.03
평균	86.07	85.16	86.76	75.23	79.68	83.33

표 3. SVM 분류기 비교

4. 결 론

SVM에서는 자질축소보다는 가중치에 따라 범주화 성능에서 큰 차이를 보였다. 실험결과 binary로 자질값을 주었을 때 가장 높은 성능을 보였으며 분류기 비교에서는 문서범주화를 위해 개발된 SVM^{light}가 더 좋은 성능을 보였다. 정보검색시스템에서는 검색결과를 향상시키기 위해 색인어나 질의어에 가중치를 부여하여 용어의 특정성을 반영한다. 본 연구에서는 SVM분류기를 이용하여 실험한 결과 정보검색의 하위개념인 문서 범주화에서도 용어에 대한 가중치 유형에 따라 성능의 차이를 보인다는 것을 확인하였다.

참 고 문 헌

오장민, 장병탁, 김영택. 1999. SVM 학습을 이용한 다중 클래스 뉴스그룹 문서 분류. 한국정보과학회 가을 학술발표 논문집 (II), 제26권 2호, 60-62.

Dumais, S. T., Platt, J., Heckerman, D, and Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. In CIKM-98: Proceedings of the 7th International Conference on Information and Knowledge Management, 148-155.

Joachims, Thorsten. 1998. Text categorization with Support Vector Machines: learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, 137--142.

Royal Holloway Univ. of London Homepage, <http://svm.cs.rhbnc.ac.uk>

Vapnik, Vladimir N. 1995. Nature of statistical learning theory.

Yang, Y., Pedersen, J.O. 1997. A comparative study on feature selection in text categorization. In Machine Learning: Proceedings of the 14th International Conference (ICML97), 412-420.