SVM 학습을 이용한 다중 클래스 뉴스그룹 문서 분류

오 장 민° 장 병 탁 김 영 택
서울대학교 컴퓨터공학과 서울대학교 컴퓨터공학과 서울대학교 컴퓨터공학과
jangmin@nova.snu.ac.kr btzhang@comp.snu.ac.kr ytkim@comp.snu.ac.kr

Classification of Multiclass Newsgroup Documents Using SVM Learning

Jang Min Oh Byoung-Tak Zhang Yung Taek Kim
Dept. of Computer Engineering Dept. of Computer Engineering Dept. of Computer Engineering
Seoul National University Seoul National University Seoul National University

요 약

다중 클래스 문서 분류는 주어진 여러 개의 관심사별로 문서를 선별해 주는 문제이다. 문서 분류 문제의 특징은 문서가 매우 높은 차원으로 표현된다는 것이다. 다른 학습 알고리즘에 비해 SVM 알고리즘은 차원을 전혀 줄이지 않고 문제를 해결한다. 본 논문에서는 SVM 학습 알고리즘을 이용하여 대규모의 뉴스그룹 문서 분류 문제를 다룬다. 다중 클래스 문서 분류를 위해서 각 클래스에 대한 SVM 학습 결과를 효과적으로 결합하였으며 실험을 통하여 SVM과 다른 학습 알고리즘과의 성능을 비교하였다.

1. 개 요

뉴스그룹 문서 집합은 여러 사용자들이 자신의 관심사에 해당하는 그룹에 기고한 문서들로 구성된다는 특징이 있다. 온라인을 통해 자신에게 배달되는 문서를 해당 관심사별로 구분지어 준다면 매우 편리할 것이다. 문서 분류란 어떤 문서를 미리 정의된 범주들로 구분 지어 주는 작업을 말한다. 각 문서는 하나의 범주에 속할 수도 있고, 여러 개의 범주에 속할 수도, 어떤 범주에도 속하지 않을 수도 있다. 뉴스그룹 문서의 경우에는 각 그룹이 하나의 범주에 해당되고, 각 문서는 하나의 범주에 속한다고 생각할 수 있다.

SVM은 구조적 리스크 최소화(Structural Risk Minimization) 원리를 제대로 구현한 알고리즘이다[7]. SVM 학습 알고리즘은 원래 패턴 분리문제를 해결하기 위해 고안되었으며 통계적 학습이론으로 그 성능이 뒷받침된다[6][7].

본 논문에서는 패턴 분리 문제에서 우수한 성능을 보이는 SVM을 이용하여 다중 클래스 문서 분류 시스템을 구현해 보았다. 원래 이진 패턴 분리 알고리즘인 SVM을 다중 클래스의 문제에 적응시키는 방법에 따른 성능을 관찰하고 다른 학습 알고리즘과 비교한다.

논문의 구성은 다음과 같다. 2절에서 뉴스그룹 문서 분류 문제에 대해 설명하고, 3절에서는 SVM 알고리즘에 대해 설명하며, 4절에서는 실험결과를 보이고 5절에서 결론을 맺는다.

2. 데이터의 구성

일반적으로 문서 분류 시스템에서의 학습을 위해서 문서는 단어들에 대한 벡터로 표현된다. 하나의 단어가 벡터의 한 원소를 차지하므로 입력 공간의 차원은 총 단어의 수, 즉 사전의 크기와 같다. 불필요한 정보를 최소로 하기 위해 스테밍(stemming) 알고리즘과 불용어 목록(stop list)를 사용한다. 또한, 뉴스그룹 문서[1] 자체는 본문 전에 해당 프로토콜에 수반되는 여러 헤드 정보를 가지고 있는데 이 정보들도 제외시켰다.

이 전처리 과정을 거친 후 벡터를 tfidf로 표현하는 것이 효율적이라고 알려져 있다[8][9].

---

1) 실험에 사용된 뉴스그룹 문서는 다음 URL에서 다운로드. http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes/20_newsgroups.tar.gz

$$DOC_i(w_j) = TF(DOC_i, w_j) \cdot IDF(w_j)$$

where

$$TF(DOC_i, w_j) = \text{count of } w_j \text{ in } DOC_i$$

$$IDF(w_j) = \log\left(\frac{\text{total number of document}}{DF(w_j)}\right)$$

$$DF(w_j) = \text{count of document having } w_j$$

3. SVM

3.1 SVM 알고리즘

SVM(Support Vector Machines)은 Vladimir Vapnik에 의해 고안되었다[7]. 처음의 SVM은 그림 1처럼 선형 분리 가능한 두 클래스를 구분지으며 마진을 최대로 하는 초평면 $\mathbf{w}\cdot\mathbf{x}+b=0$을 찾는 문제였다.

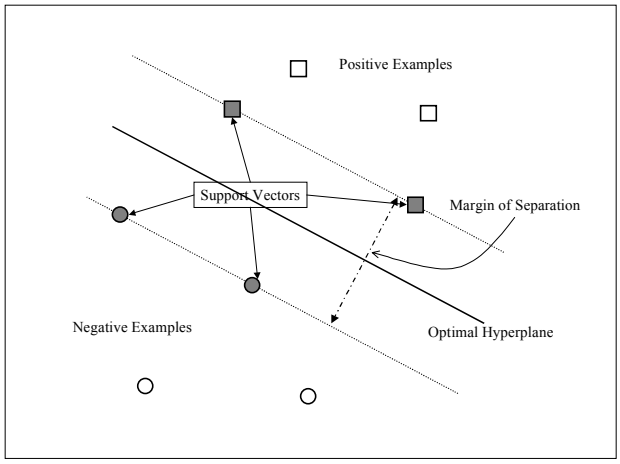

그림 1 선형 SVM

통계적 학습 이론에서, Vapnik과 Chervonenkis는 모델의 복잡도의 측정 수단인 VC 차원(Vapnik Chervonenkis dimension)을 도입했다. 이들은 학습 모델의 수렴 경계 조건을 다음의 식으로 제시했다[6].

$$R(\mathbf{w}) \le R_{emp}(\mathbf{w}) + \sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{4}}{l}}, \quad \forall \mathbf{w} \in \mathrm{W}$$

where

$$R(\mathbf{w}) = \int |d - F(\mathbf{x},\mathbf{w})| dF_{\mathbf{x},D}(\mathbf{x},d)$$

$$R_{emp}(\mathbf{w}) = \frac{1}{l}\sum_{i=1}^{l}|d_i - F(\mathbf{x}_i,\mathbf{w})|$$

여기에서 h는 VC 차원이고, l은 학습 데이터의 수이며, $R(\mathbf{w})$는 기대 리스크(expected risk)이고 $R_{emp}(\mathbf{w})$는 경험적 리스크(empirical risk)이다. 수렴 경계 조건이 뜻하는 바는 최적의 모델을 찾기 위해서 $R_{emp}(\mathbf{w})$ 만을 최소로 하는 모델을 찾는 것만으로는 부족하고 모델 복잡도까지 같이 고려해야 한다는 것이다.

그런데 대부분의 학습 기법에서 VC 차원의 측정은 어렵다. 그러나 SVM에서는 다음의 정리에 의해 VC 차원의 직접적인 제어가 가능하다[6][7].

> 입력 차원이 m_0인 학습 데이터 전체를 포함할 수 있는 구의 반지름을 R이라고 하고, 분류 함수인 초평면의 식을 다음과 같이 표현하면
>
> $$F_k(\mathbf{w},\mathbf{x}) = \{\mathbf{w}^{\mathrm{T}}\cdot\mathbf{x}+b : \|\mathbf{w}\|^2 \le A_k\}$$
>
> F_k에 대한 VC 차원 h_k는 다음 상한을 갖는다.
>
> $$h_k \le \min\{R^2A_k^2, m_0\}+1$$

즉, 선형 분리 가능 문제의 경우 $R_{emp}(\mathbf{w})$는 항상 0이 되게 할 수 있으므로 VC 차원이 최소인 모델을 찾으면 최적의 $R(\mathbf{w})$를 찾게 된다[7][6]. SVM에서 이것은 마진을 최대로 하는 즉, $||\mathbf{w}||^2$을 최소로 하는 모델을 찾는 것으로 귀결된다.

$$\min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to } d_i(\mathbf{w}\cdot\mathbf{x}_i - b) \ge 1, \forall i$$

선형 분리 가능이라는 제약 조건은 Cortes와 Vapnik에 해결되었다[1]. 그들은 슬랙 변수(slack variable)를 두어 에러를 허용하고 C 파라미터를 두어 마진과 에러의 트레이드오프를 조절했다.

$$\min_{\mathbf{w},b,\boldsymbol{\xi}}\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi_i \quad \text{subject to } d_i(\mathbf{w}\cdot\mathbf{x}_i+b) \ge 1-\xi_i, \forall i$$

이를 라그랑쥐 듀얼 문제로 변환하면,

$$\min_{\lambda}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}d_id_jK(\mathbf{x}_i,\mathbf{x}_j)\lambda_i\lambda_j - \sum_{i=1}^{l}\lambda_i$$

$$\text{subject to}$$

$$0 \le \lambda_i \le C, \forall i$$

$$\sum_{i=1}^{l}d_i\lambda_i = 0$$

의 최적화 문제가 된다. 여기에서 $K(\mathbf{x}_i,\mathbf{x}_j)$는 커널 함수로서 비선형 함수에까지 일반화한 것이다. 이 최적화 문제의 해로 얻어지는 $0<\lambda_i<C$에 해당하는 데이터$(\mathbf{x}_i,d_i)$가 마진 상에 존재하는 데이터이고 이 데이터를 SV(Support Vector)라고 한다. 초평면의 식은 SV만을 이용하여

$$\sum_{i=1}^{l_s}d_i\lambda_iK(\mathbf{x},\mathbf{x}_i)+b=0$$

로 나타내어 진다. 이 때 l_s는 SV의 개수이다.

결국 SVM은 QP(Quadratic Programming)으로 정형화가 된다[3]. 일반적인 QP 라이브러리는 SVM에 적용이 용이하지가 않다[2]. SVM 학습을 위한 효과적인 QP 휴리스틱으로, Chunking, Decomposition, SMO(Sequential Minimal Optimization) 기법 등이 연구되었다[2][4][7].

3.2 다중 클래스로의 확장

SVM은 원래 이진 패턴 분리 문제를 위한 알고리즘이었으므로 만일 K개의 클래스가 있는 문제를 학습시키려면 여러개의 SVM을 조합해야하는 문제가 있다.

첫 번째 방식으로 승자독식(winner-take-all) 방식이 있다. 이 경우 다음과 같이 각 클래스당 하나씩 총 K개의 SVM 분류기를 학습 시킨다. 그리고 SVM의 출력으로 ±1 값을 얻는 것이 아니라 실수값을 얻어서 각 분류기가 선거를 치루게 된다. 만일 어떤

테스트 문서에 대하여 k번째 분류기의 출력이 가장 큰 값이라면 이 문서는 k번째 클래스에 속한다고 결론을 내린다.

두 번째 방식은 각 쌍마다(pairwise) 분류기를 만드는 것이다 [5]. 즉 $K(K-1)/2$개의 모든 조합에 대하여 다음과 같이 SVM 분류기를 학습시킨다.

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

이렇게 얻어진 분류기는 테스트 문서에 대하여 $f_k = \sum_l f_{kl}$ 의 값이 $K-1$인 k로 할당을 하면 된다. 그런데 $K-1$인 f_k가 존재하지 않는다면 SVM의 실수값 출력에 대해 음의 투표(negative vote)가 제일 적은 k로 할당하면 된다.

4. 실험

20개의 유스넷 뉴스그룹 데이터에 대해 실험을 하였다. 데이터는 20,000개의 문서로 되어 있으며 이중에서 70%를 학습에, 나머지 30%를 테스트에 이용하였다. SVM 학습에 사용된 사전의 크기는 6,000개로 구성하여 사용하였다. 나이브 베이지안 모델은 Rainbow 소프트웨어[2]를 이용하여 실험을 하였다. SVM은 C=1000으로, 에러 임계치=0.001, 5차 다항식 커널, σ^2=1.0인 RBF 커널을 사용하여 훈련하였다. 표 1에 결과를 보였다.

범주 집합	NB	SVMPP	SVMPR	SVMWR
0	81.00	69.67	70.67	74.00
1	86.67	83.00	82.00	90.00
2	23.33	81.08	82.43	94.26
3	82.67	76.25	75.92	89.97
4	78.67	85.23	85.67	97.32
5	85.33	85.62	85.95	91.30
6	57.67	79.60	79.93	85.95
7	85.33	88.29	88.96	96.99
8	93.67	95.99	94.65	99.33
9	91.33	93.31	93.31	97.66
10	96.33	95.32	95.99	99.33
11	94.00	92.31	92.98	99.67
12	78.33	85.67	84.67	97.33
13	92.67	92.00	91.33	98.00
14	92.33	94.00	94.00	98.33
15	90.67	89.00	88.33	97.67
16	89.00	85.33	84.33	89.00
17	92.33	91.33	92.00	95.67
18	71.91	73.58	73.24	71.57
19	42.81	50.5	50.17	53.18
평균	80.30	84.35	84.32	90.83

표 1 성능비교. NB: Naive Bayesian, PP: Pairwise Polynomial, PR: Pairwise RBF, WR: Winnertakeall RBF

표에서 각 수치는 정확도(accuracy)를 의미한다. 결과를 보면 나이브 베이지안 모델에 비해 SVM이 매우 뛰어난 성능을 보임을 알 수 있다. 본 논문의 실험에서는 SVMWR이 최고의 성능을 보였다. 나이브 베이지안 모델의 경우 2번째 범주의 결과가 매우 좋지 않은데 이는 comp.os.ms-windows.misc 범주로서 1~5번까지의 범주가 컴퓨터 관련 범주인데 여기에 많이 분류되게 되어 단순한 단어 통계만 가지고는 그 특징을 잘 잡아내지 못함을 알 수 있다.

2) http://www.cs.cmu.edu/~mccallum/bow/rainbow/ 참조

5. 결론

SVM 알고리즘을 뉴스그룹 문서 분류 문제에 적용해 보았다. 승자 독식 모델과, 모든 쌍마다 분류기를 만드는 모델에 대해서 SVM 분류기를 학습한 결과 나이브 베이지안 모델보다 월등한 성능을 보였다.

SVM은 학습 결과의 적용시 모든 SV들과의 커널 내적값이 계산되어야 한다는 단점이 있다. 특히 모는 쌍마다 분류기를 만들었을 경우 190개의 모델이 만들어 지는데 이를 이용해 테스트 데이터 6,000여개를 분류하는데 걸리는 시간이 학습 시간을 초과할 정도였다. 적용시간에 걸리는 시간 및 계산 복잡도의 개선이 향후 연구되어야 할 것이다.

감사의 글: 본 연구는 정보통신부가 시행하고 있는 대학기초 연구기술 지원사업에 의해 지원되었음.

참고문헌

[1] C. Cortes and V. Vapnik. 1995, Support Vector Networks, *Machine Learning*, 20:273-297.

[2] John C. Platt. 1998, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Techinical Report MSR-TR-98-14.

[3] Stephen G. Nash and Areila Sofer. 1997, *Linear and Nonlinear Programming.* McGraw-Hill.

[4] Thorsten Joachims. 1999, Making Large-Scale Support Vector Machine Learning Practical, *Advances in Kernel Methods.* MIT Press. pp. 169-184.

[5] Ulrich H.-G. Kressel. 1999, Pairwise Classification and Support Vector Machines, *Advances in Kernel Methods.* MIT Press. pp. 255-268.

[6] V. Cherkassky and Filip Mulier. 1998, *Learning From Data Concepts, Theory, and Methods.* John Wiley&Sons, Inc.

[7] V. Vapnik. 1995, The Nature of Statistical Learning Theory. Springer-Verlag.

[8] William B. Frakes and Richard Baeze-Yates. 1997, *Information Retrieval Data Structures & Algorithms.* Prentice-Hall, Inc.

[9] Yiming Yang. 1997, An Evaluation of Statistical Approaches to Text Categorization, Technical Report CMU-CS-97-127.