Co-Trained Support Vector Machines 을 이용한 문서분류

박성배° 장병탁
서울대학교 컴퓨터공학부
{sbpark,btzhang}@scai.snu.ac.kr

Text Categorization Using Co-Trained Support Vector Machines

Seong-Bae Park° Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

대부분의 자동문서분류 시스템은 문서에 사용된 단어의 분포만 고려하고, 또 하나의 중요한 정보인 통사 정보는 무시한다. 본 논문에서는 통사정보와 어휘정보를 모두 사용함으로써 대규모의 비구조 문서를 분류하는 방법을 제시한다. 이를 위해, 학습 데이터에 대해 독립된 두 개의 관점을 요구하는 일종의 부분 감독 학습 알고리즘인 co-training 알고리즘을 사용한다. 어휘정보와 통사정보가 각각 문서의 독립된 관점이 될 수 있으므로, 이 두 정보와 레이블이 없는 문서를 사용하여 문서 분류의 성능을 높일 수 있다. Reuters-21578 문서집합과 TREC-7 filtering 문서집합에 대한 실험 결과는 제시된 방법의 유효성을 보인다.

1. 서론

문서분류(text categorization) 문제는 레이블이 있는 학습 문서 집합으로부터 추출된 정보에 기초하여 향후에 주어지는 레이블이 없는 문서를 미리 정해진 범주로 분류하는 것이다. 자동문서분류는 점점 더 많은 양의 문서가 전자화됨으로써 실용적인 측면에서 그 중요성이 더욱 부각되고 있다. 자동문서분류에 적용된 대부분의 기계학습(machine learning) 알고리즘들은 문서를 학습 자질이 단어인 bag of words 로 표현한다[1,2]. 하지만, 이런 표현은 또하나의 중요한 분류 요소인 언어학적 정보를 무시한다.

각 문서는 그 스타일적인 특성이 있고, 통사정보는 서로 다른 종류의 문서들 사이에 존재하는 스타일 차이를 구분하는 가장 좋은 측정값 중의 하나이다. 통사정보가 문서분류를 위한 많은 정보를 주기는 하지만, 이런 정보가 너무 복잡하며 형식 정의(formal definition)가 없어서 잘 사용되지 않아왔다. 또한, 불행하게도 현재의 자연언어처리 기술도 통사 분석에 있어서 만족할 만큼 정확한 결과를 제공해 주지 못하는 실정이다. 따라서, 통사 정보를 이용하기 위해 완전 구문 분석(full parsing)을 하기 보다는 필요한 통사 분석에 필요한 충분한 정보를 줄 수 있는 문서 단위화(text chunk)[3]를 사용하는 것이 더 현실적이다.

문서분류의 또 다른 문제는 레이블이 있는 문서의 수가 적은데 비해, 레이블이 없는 문서는 수 없이 많다는 점이다. Co-training 알고리즘[4]은 레이블이 없는 데이터를 처리하는 성공적인 알고리즘 중의 하나이다. 이 알고리즘은 일반적으로 각 학습 예제에 대한 두 개의 독립된 관점이 있는 문제에 적용된다. 예를 들어, 웹문서는 내용에 사용된 단어와 링크에 사용된 단어의 두 관점이 있다.

우리는 문서분류를 위한 co-trained Support Vector Machine 을 제안하여 주어진 소수의 레이블이 있는 문서와 대량의 레이블이 없는 문서를 효과적으로 이용한다. Co-training 알고리즘의 두 관점을 위해 통사정보와 어휘정보를 사용한다. 따라서, 링크정보가 없는 비구조 문서를 분류하는데 사용될 수 있다. 우리가 기본 분류기로 SVM 을 사용하기 때문에 레이블이 없는 예제의 레이블을 추정하는 confidence 로 margin 을 사용할 수 있다.

2. Co-Training 알고리즘을 이용한 문서분류

2.1 두 관점

웹페이지가 자연스러운 두 개의 관점을 가지고 있기 때문에 co-training 알고리즘의 응용은 대부분 웹페이지 분류이다. 그러나, 링크정보가 없는 일반 비구조 문서를 위해서는 두개의 독립된 관점을 어떻게 만들 수 있는가 하는 점은 불확실하다.

문서분류를 위한 하나의 관점은 각 문서를 요소가 어휘에 대한 중요도를 나타내는 벡터로 표현하는 것이다. 대부분이 기계학습 방법이 이 방식을 취한다. 이 관점의 약점은 (i) 각 단어들이 서로 독립이어야 하고, (ii) 문서에 포함된 언어학 정보를 무시한다는 점이다.

Stamatatos 등은 여러가지 언어학 정보 중에서 통사정보가 문서분류를 위해 신뢰할 수 있는 단서가 됨을 실험적으로 보였다[5]. 통사정보를 co-training 알고리즘에 사용할 때 얻을 수 있는 장점 중 하나는 이 정보가 어휘정보와는 어느 정도 독립적이라는 점이다. 불행히도, 현재의 자연언어처리 기술은 정확한 통사정보를 제공하지 못한다. 그러나, *text chunk* 가 통사정보를 제공하기 위한 좋은 자질이 될 수 있다.

따라서, 우리는 비구조 문서를 위한 두 개의 관점을 다음과 같이 정의한다.

- **어휘 정보**

  대부분의 기계학습 알고리즘은 *tf·idf* 에 기초한다. 본 논문에서도 어휘정보를 이를 사용한다.

- **통사 정보**

  각 문서를 요소가 통사 자질인 벡터로 표현한다.

2.2 문서분류를 위한 통사정보

문서를 요소가 통사정보인 벡터로 표현하기 위해서 각 문서는 단위(chunk) 정보로 표현된다. 문서 내의 모든 문장을 단위로 나누기 위해서, 문맥 단어들의 품사, 어휘

등을 사용한다. 각 단어의 단위는 CoNLL-2000 데이터로 학습된 SVM 에 의해 결정된다.

CoNLL-2000 데이터에는 12 종류의 구가 있지만, 우리는 그 중 다섯 가지만 사용한다: NP, VP, ADVP, PP 와 O. 여기서 O 는 NP, VP, ADVP, PP 가 아닌 경우이다. O 를 제외한 각 구는 다시 두 종류의 단위 표지(B-XP 와 I-XP)를 가질 수 있다. 따라서, 총 11 종류의 단위 표지가 있다.

단위화를 위한 SVM 의 사용은 아주 좋은 성능을 보였다[3]. 단어 w_i 의 단위 표지를 결정하기 위해서 사용된 자질은 다음과 같다.

$$w_j, POS_j \quad (j = i-2, i-1, i, i+1, i+2)$$
$$c_j \quad (j = i-2, i-1)$$

여기서, *POS* 와 c 는 각각 품사 표지와 단위 표지이다. SVM 이 기본적으로 이진 분류기이고 11 종류의 단위 표지가 있기 때문에, 쌍 분류(*pairwise classification*)[6]로 학습하였다.

표 1 은 문서분류를 위해 본 논문에서 사용된 통사정보 자질이다. 이 자질들은 단위화된 문서로부터 기계적으로 계산할 수 있는 수치값을 갖는다. 상위 5 개의 자질은 구들이 얼마나 자주 나타나는지를 나타내고, 하위 5 개는 얼마나 긴지를 나타낸다. 그리고 마지막 자질은 평균적인 문장의 길이이다.

자질	설명
SF1	Detected *NP*s / total detected chunks
SF2	Detected *VP*s / total detected chunks
SF3	Detected *PP*s / total detected chunks
SF4	Detected *ADVP*s / total detected chunks
SF5	Detected *O*s / total detected chunks
SF6	Words included in *NP*s / detected *NP*s
SF7	Words included in *VP*s / detected *VP*s
SF8	Words included in *PP*s / detected *PP*s
SF9	Words included in *ADVP*s / detected *ADVP*s
SF10	Words included in *O*s / detected *O*s
SF11	Sentences / words

표 1. 문서분류를 위해 사용된 통사정보 자질 (feature).

3. 실험

3.1 실험 데이터

Reuters-21578

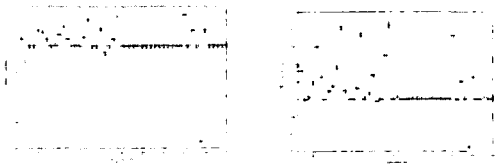
Reuters-21578 말뭉치는 135 개의 토픽을 가지는데, 우리는 이중 주요한 10 개의 토픽만 실험대상으로 삼았다. 이 말뭉치를 학습집합과 테스트집합으로 나누는 데에는, 세 종류의 방법이 있는데, "ModLewis", "ModApte", "ModHayes"가 그들이다. 이 중에서, 우리는 가장 널리 쓰이는 "ModApte"를 사용하였다. 따라서, 9,603 개의 학습 문서와 3,299 개의 테스트 문서를 사용하였다.

TREC-7

TREC-7 filtering track 에 사용된 데이터는 1988 년부터 1990 년까지의 AP 뉴스들로 구성되었다. 1988 년 뉴스를 학습집합으로 사용하고, 1989 년과 1990 년 뉴스는 테스트 집합으로 사용한다. 학습집합에는 79,898 개의 기사가 있는데 이중 약 12%인 9572 개만이 레이블이 되어있다. 따라서, 우리는 이 9572 개의 기사를 레이블이 있는 데이터로, 나머지 88%를 레이블이 없는 데이터로 사용한다.

Class	어휘정보	통사정보	모두
Earn	2876	2504	**2895**
Acq	1587	1028	**1642**
Money-fx	188	147	**193**
Grain	**26**	14	**26**
Crude	292	150	**312**
Trade	153	114	**155**
Interest	130	112	**141**
Ship	113	88	**116**
Wheat	98	70	**108**
Corn	86	68	**87**

표 2. 단순히 어휘정보와 통사정보를 더하여 학습했을 때의 문서분류 성능 향상. LF1 이용.



(a) LF1 : Both – Lexical (b) LF1 : Both - Syntactic

그림 0. 두 정보를 모두 사용함으로써 얻은 성능의 향상.

3.2 성능 평가 수단

제시된 방법의 분류 성능을 평가하기 위해서, 우리는 *utility measure* 를 사용한다. R_+를 relevant 하고 추출된 문서의 수, N_+를 irrelevant 하고 추출된 문서의 수, R_-를 relevant 하고 추출되지 않은 문서의 수, N_-를 irrelevant 하고 추출되지 않은 문서의 수라고 하자. 그러면, linear utility 를 다음과 같이 정의된다.

$$\text{Linear Utility} = aR_+ + bN_+ + cR_- + dN_-.$$

여기서, a, b, c, d 는 상수 계수이다. 특별히 LF1 과 LF2 는 다음과 같다.

$$\text{LF1} = 3R_+ - 2N_+$$
$$\text{LF2} = 3R_+ - N_+$$

Linear utility 의 단점은 몇 개의 토픽에 의해 성능이 좌우되므로 평균값이 의미가 없다는 점이다. 따라서, 우리는 scaled linear utility 를 사용한다.

$$\text{Scaled Linear Utility} = \frac{\max\{u(S,T), U(s)\} - U(s)}{MaxU(T) - U(s)}$$

여기서, $u(S, T)$는 시스템 S의 토픽 T에 대한 linear utility 이고, $MaxU(T)$는 토픽 T 의 최대 utility 점수이다. 그리고, $U(s)$는 s irrelevant 한 문서를 추출하는 utility 이다.

3.3 실험 결과

Reuters-21578

레이블이 없는 데이터를 고려하지 않았을 때, 통사적인 정보를 사용해서 얻은 이득이 표 2 에 주어져 있다. 두 정보를 다 사용했을 때의 분류 성능이 둘 중 하나만 사용했을 때보다 훨씬 좋다.

TREC-7

그림 1 은 어휘정보와 통사정보를 모두 사용한 효과를 보이고 있다. 이 결과는 레이블이 있는 데이터와 없는 데이터를 모두 사용했을 때 얻은 것이다. X 축은 토픽번호이고, Y 축은 두 정보를 사용함으로써 얻는 성능 차이

이다. 따라서, 이 값이 0 보다 크면 성능이 개선되었다는 뜻이다. Both-Syntactic 이 Both-Lexical 보다 큰 이유는, 통사정보만 사용하는 것이 어휘정보만 사용하는 것보다 성능이 좋기 때문이다. 표 3 의 평균 성능은 표 3 에 나와 있다.

Measure	Lexical	Syntactic	Both
LF1	0.2005	0.1680	**0.2192**
LF2	0.2010	0.2010	**0.2155**

표 3. 제시된 방법의 TREC-7 데이터에 대한 결과.

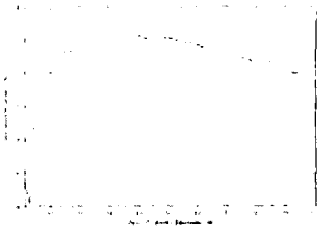
3.4 분석

많은 선행 연구에서 레이블이 없는 문서를 사용하면 문서 분류의 성능이 높아짐을 보였다[7]. 우리의 실험 결과도 레이블이 없는 문서가 문서 분류에 도움이 됨을 보였다. 그림 2 는 co-training 알고리즘에 사용된 레이블이 없는 문서의 reuters-21578 말뭉치에 대한 효용성을 보이고 있다. X 축은 전체 문서에 대한 레이블이 있는 문서의 비율이고, Y 축은 이 레이블이 없는 문서에 의해 향상된 정확도이다. Earn 에 대해, 레이블이 없는 문서는 10% 이상의 문서가 레이블이 있으면 긍정적인 역할을 한다. 이는 Acq 에 대해서도 7% 이상이 레이블되면 마찬가지이다.

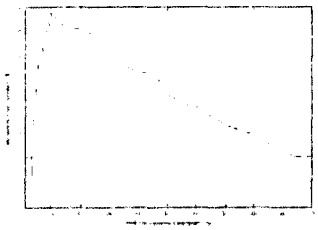
그러나, 레이블이 없는 문서로 최고의 성능 향상 효과를 얻었을 때에도 전체 문서의 레이블을 미리 알고 있을 때의 성능에 미치지 못한다. 예를 들어, Acq 에서 10%의 데이터가 레이블이 있을 때 정확도 향상은 5.81%이다. 이 경우의 정확도는 89.93%이다. 반면에 전체 데이터가 레이블이 있을 때는 정확도가 95.21%이다. 이는 co-training 알고리즘을 학습하는 중 레이블이 없는 데이터의 일부에 대해 레이블이 잘못 추정되어서 부정적인 효과를 미쳤기 때문이다. 레이블이 없는 문서의 효과는 그 수가 적을 때 최대화된다. 따라서, 이 정확도 사이의 차이를 매울 필요가 있다. 이 차이를 매우기 위해서는, 사람의 간섭이 필요하다. 하지만, 학습 중 언제 사람이 개입할 것인지는 결정하는 것은 매우 어려운 문제이다.

4. 결론

우리는 문서 분류를 위한 co-trained Support Vector Machines 를 제시하였다. 이 방법은 co-training 알고리즘을 위해 전통적인 어휘정보 외에도 통사정보도 사용하여 이 알고리즘이 웹페이지 분류 뿐만 아니라 비구조 문서 분류에도 사용될 수 있도록 하였다. 또한, 제시된 방법으로 우리는 수없이 많은 레이블이 없는 문서들을 학습에 활용하여 소수의 레이블이 있는 데이터를 보충하였다. Reuters-21578 말뭉치와 TREC-7 filtering 문서에 대한 실험은 통사정보를 전통적인 어휘정보와 함께 사용하는 것이 분류 성능을 높이고 레이블이 없는 문서가 제한된 수의 레이블이 있는 문서를 한계를 극복할 수 있는 좋은 자원임을 보였다. 이는 소수의 문서만으로도 높은 성능을 보이는 문서분류 시스템을 만들 수 있는 가능성을 보인 것이다. 레이블이 없는 문서의 효용성이 실험적으로 보여졌으나, 학습 도중 레이블이 잘못 추정되는 것을 극복하여야 하는 또 다른 문제가 생겼다. 향후, 이를 해결하기 위한 연구가 필요하다.



(a) Earn



(b) Acq

그림 2. 추가의 레이블이 없는 문서를 사용함으로써 얻은 정확도의 향상.

감사의 글

이 논문은 과기부 BrainTech 프로그램과 교육부 BK 21 사업에 의하여 지원되었음.

참고문헌

[1] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," In *Proceedings of ECML 98*, pp. 137-142, 1998..

[2] Y.-H. Kim, S.-Y. Hahn, and B.-T. Zhang, "Text Filtering by Boosting Naïve Bayes Classifiers," In *Proceedings of ACM SIGIR 2000*, pp. 168-175, 2000.

[3] T. Kodoh and Y. Matsumoto, "Use of Support Vector Learning for Chunk Identification," In *Proceedings of CoNLL-2000 and LLL-2000*, pp. 142-144, 2000.

[4] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," In *Proceedings of COLT 98*, pp. 92-100, 1998.

[5] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic Text Categorization in Terms of Genre and Author," *Computational Linguistics*, 26(4):471-496, 2000.

[6] B. Scholkopf, C. Burges, and A. Smola, *Advances in Kernel Method: Support Vector Machines*, MIT Press, 1999.

[7] T. Zhang and F. Oles, "A Probability Analysis on the Value of Unlabeled Data for Classification Problems," In *Proceedings of ICML 2000*, pp. 1191-1198, 2000.