레이블이 없는 데이터로부터의 학습에 의한 자동 문서 분류

박성배[0] 김유환 장병탁

서울대학교 컴퓨터공학부

{sbpark,yhkim,btzhang}@cse.snu.ac.kr

Automatic Text Classification by Learning from Unlabeled Data

Seong-Bae Park[0] Yu-Hwan Kim Byoung-Tak Zhang

School of Computer Science and Engineering, Seoul National University

요 약

본 논문에서는 레이블이 없는 데이터를 이용하는 새로운 자동 문서 분류 방법을 제시한다. 제시된 방법은 적은 수의 레이블이 있는 데이터로부터 학습된 후 많은 수의 레이블이 없는 데이터로 보강되는 일련의 분류기(classifier)에 기반한다. 레이블이 없는 데이터를 활용하기 때문에, 필요한 레이블이 있는 데이터의 수가 줄어들고, 분류 정확도가 향상된다. 두 개의 표준 데이터 집합에 대한 실험 결과, 레이블이 없는 데이터를 사용함으로써 분류 정확도가 증가함을 보였다. 분류 정확도는 전체 데이터의 2/3만 사용하고도 NIPS 2000 워크숍 데이터 집합에 대해서는 약 7.9% 정도, WebKB 데이터 집합에 대해서는 9.2% 증가하였다.

1. 서론

점점 더 많은 수의 텍스트 문서가 전자화되면서, 자동 문서 분류의 중요성이 계속해서 증가하고 있다. 이 문제에 적용된 많은 기계학습 기법은 대량의 레이블이 부여된 학습 데이터를 필요로 한다[8]. 하지만, 각 데이터에 레이블을 부여하는 작업을 사람인 해야 하기 때문에, 그러한 데이터는 일반적으로 매우 비싸고 구하는 데에 많은 시간이 소요된다. 반면에, 레이블이 없는 데이터는 어디에나 있으며 레이블이 있는 데이터에 비해 구하기도 쉽다. 따라서, 자동 문서 분류에서는 레이블이 있는 데이터에 추가로 레이블이 없는 데이터를 활용하는 것은 자연스러운 일이다.

본 논문은 레이블이 있는 데이터와 없는 데이터를 모두 활용해서 문서를 분류하는 방법을 제시한다. 제시된 방법에서는 우선 분류기(classifier)가 레이블이 있는 데이터로 학습되고 분류기의 신뢰도가 결정된다. 그 다음에, 레이블이 없는 데이터로부터 일련의 분류기가 학습된다. 분류기 순서에서 다음 분류기를 결정하기 위해서, 올바르게 분류될 확률이 현재 분류기의 신뢰도보다 큰 모든 학습 예제는 레이블이 있는 데이터 집합과 없는 데이터 집합 양쪽에서 제거된다. 왜냐하면, 이런 학습 예제들은 다음 분류기를 위한 정보를 제공하지 않기 때문이다. 다음 분류기는 남은 레이블이 있는 데이터와 레이블이 없는 데이터 중에서 선택된 중요한 일부에 의해 학습된다. 이 과정은 레이블이 없는 데이터가 모두 다 사용될 때까지 반복된다. 제시된 방법은 두 개의 표준 데이터 집합인 NIPS 2000 워크숍 데이터 집합과 WebKB 데이터 집합에 대해 평가되었다. 레이블이 없는 데이터를 사용함으로써, 자동 문서 분류의 정확도가 증가하였고, 제시된 방법은 co-training[2]보다 더 좋은 성능을 보였다.

2. 분류 문제를 위한 레이블이 없는 데이터

문서가 *bag-of-words* 로 표현된다고 가정하면, 각 문서는 차원이 어휘의 수와 같은 차원 공간 상의 벡터로 표현된다. 문제를 간단하게 하기 위해서, 본 논문에서는 이진 분류 문제만 고려하였다. 주어진 문서 벡터 $\mathbf{x}$ 의 레이블은 $y \in \{-1, +1\}$로 나타내어 지며, 여기서 +1 은 문서가 적절한 문서임을, -1 은 적절하지 않은 문서임을 나타낸다.

Cramér-Rao 부등식에 의하면, 모수 θ 의 비편향 추정량(unbiased estimator) $T(\mathbf{x})$는 피셔정보(Fisher information)의 역에 한정된다. 즉,

$$\mathrm{var}(T) \geq \frac{1}{I(\theta)},$$

여기서, 피셔정보는 $f(\cdot,\theta)$가 밀도 함수일 때 다음과 같다.

$$I(\theta) = E_\theta\left[\frac{\partial}{\partial\theta}\ln f(\mathbf{x};\theta)\right]^2.$$

피셔정보가 커지면 분산이 작아지고 추정량의 오류는 분산에 비례하므로, 피셔정보가 커지면 추정량의 오류는 작아진다. Shahshahani 와 Landgrebe 는 레이블이 있는 데이터와 없는 데이터를 모두 쓸 때의 피셔정보, $I_{labeled+unlabeled}$ 는

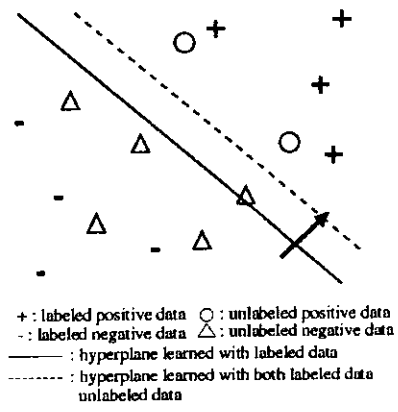
$$I_{labeled+unlabeled} = I_{labeled} + I_{unlabeled}$$

임을 보였다[6]. 따라서, 피셔정보가 분산과 반비례하기 때문에, 레이블이 없는 데이터를 사용함으로써 추정량의 정확도를 높일 수 있다.

그러나, Zhang 과 Oles 는 준모수적 모델(semi-parametric model)에서는 $I_{unlabeled} = 0$ 이라고 주장하였다[9]. 그들은 준모수적 모델에서 레이블이 없는 데이터의 피셔정보를 최대화하기 위해서는 능동학습 기법을 사용해야 하며, 정보가 많은 레이블이 없는 데이터는 다음의 두 가지 기준으로 선택되어져야 한다고 하였다.

1. 추정된 모수로 신뢰도가 낮은 데이터를 선택한다.
2. 중복되지 않은 데이터를 선택한다.

이 기준은 직관과 일치한다. 그림 1 에서, 직선은 주어진 레이블이 있는 데이터로 학습된 초평면(hyperplane)이고,



[그림 1] 레이블이 없는 데이터가 분류기를 더 정확하게 만드는 방법.

점선은 실제 초평면이다. 학습된 초평면을 실제 초평면 쪽으로 옮길 때, 레이블된 음의 예(negative example)로 둘러 쌓인 레이블이 없는 음의 예들은 아무런 정보도 제공하지 않는다. 반면에, 신뢰도가 낮은 레이블이 없는 예제들은 도움이 될 수 있다.

3. 순차적 학습에 의한 레이블이 없는 문서 분류

레이블이 없는 데이터의 신뢰도를 측정하기 위해, 본 논문에서는 SEQUEL (SEQUEnce Learner)[1]의 아이디어를 사용한다. SEQUEL 은 각 분류기의 능력에 대한 간단한 추정에 기반하여 분류기들의 앙상블(ensemble)을 만드는 방법이다.

분류기 f_t가 확률 추정량이라고 하면, $f_t(\mathbf{x})$는 문서 $\mathbf{x}$ 가 적합할 확률이다. 각 분류기는 임계치 τ_t를 가진다. 이 임계치는 가장 확률이 높은 음의 예제의 확률로 설정된다. 주어진 데이터에 대해서, 만약 분류기의 출력이 이 임계치보다 높으면, 분류기는 예측을 할 능력이 있다고 생각되어진다. 각 분류기의 신뢰도는 이 임계치를 넘는 전체 예제의 수를 $\mathbf{x} : f_t(\mathbf{x}) \geq \tau_t$인 양의 예제의 수로 나눈 값이다.

SEQUEL 에서는 앙상블을 위한 분류기의 집합이 자질을 다르게 함으로써 만들어 진다. SEQUEL 을 학습할 때, 첫번째 분류기가 데이터 중의 일부를 확실한 예제로서 레이블을 정한다. 이 데이터는 최소한 τ_t 이상의 확률을 갖는다. 첫번째 분류기가 레이블을 정한 모든 확실한 예제가 제거되고 남은 예제들을 가장 잘 분류할 가능성이 있는 분류기가 다음 분류기로 정해진다. 이 과정은 분류기의 신뢰도가 정해진 임계치보다 낮아질 때까지 반복된다.

그림 2 는 레이블이 없는 데이터를 사용하기 위해 변경된 SEQUEL 을 보이고 있다. 먼저, 우리는 레이블이 있는 데이터 집합 L 을 이용해서 첫번째 분류기 f_0 를 학습한다. Zhang 과 Oles 는 앞 절에서 제시된 두 조건을 따를 때 레이블이 없는 데이터의 피셔정보가 최대화될 수 있다고 하였다. 레이블이 없는 데이터를 사용하기 위해서, 분류기의 집합은 레이블이 없는 데이터에서 두 조건을 만족하도록 중요한 예제들을 추출해서 구성된다. 각 분류기 $f_t(\mathbf{x})$의 신뢰도는 τ_t로 결정된다.

Given unlabeled example set $U = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$
and labeled example set $L = \{(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_l, y_l)\}$
Train a classifier f_0 with $L_0 = L$.
Set $t = 0$ and $\tau_{-1}=1$.
Train each base classifier $C_j (1 \leq j \leq M)$ from $S_t^{(0)}$.
Do
1. **Calculate** $\tau_t = \tau_{t-1} \times \tau$.
   where τ is the probability given to the negtive example in L_t with the highest probability.
2. **Sort** data in L_t according to $f_t(\mathbf{x} \in L_t)$.
3. **Sort** data in U_t according to $f_t(\mathbf{x} \in U_t)$.
4. **Delete** data in L_t and U_t such that
   $f_t(\mathbf{x}) > \tau_t$.
5. **Set** $s = |L_t|$.
6. **Set** U_{add} such that
$$U_{add} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_{l-s}, y_{l-s}) \mid \mathbf{x} \in U_t, y = f_t(\mathbf{x})\}.$$
7. **Set** $L_{t+1} = L_t + U_{add}$
8. **Train** f_{t+1} with L_{t+1}.
9. **Set** $t = t + 1$.

While($|U_{add}| > 0$ and $\tau_t > 0.5$)
Output the final classifier:
$$f^*(\mathbf{x}) = \left(\prod_{i=1}^{k_x-1} \tau_i\right) f_{k_x}(\mathbf{x}).$$

[그림 2] 레이블이 없는 데이터를 이용하는 자동 문서 분류 알고리듬. k_x는 $\mathbf{x}$ 의 가장 좋은 분류기에 대한 인덱스이다.

f_t 가 레이블이 있는 데이터로부터 학습된 후, L_t 를 위한 임계치 τ 를 계산하고, f_t의 신뢰도를 $\tau_t = \tau_{t-1} + 1$ 로 갱신한다. L_t와 U_t에 있는 데이터는 마진(margin)에 따라 정렬된다. 레이블된 예제 $(\mathbf{x}, y)$의 마진은 $y \bullet f_t(\mathbf{x})$로 계산된다. 여기서, 현재 분류기 f_t 에 의해 레이블된 레이블이 없는 예제의 레이블이 옳다고 가정한다. 즉, $\mathbf{x} \in U$ 에 대해서,
$$y \cdot f_t(\mathbf{x}) > 0.$$
여기서, y 는 $\mathbf{x}$ 의 실제 레이블이다. 따라서, 레이블이 없는 예제에 대해서는, f_t 의 출력을 마진으로 생각한다. 확률이 τ_t 보다 큰 L_t 와 U_t 에 있는 예제는 초평면을 옮기는데 아무런 정보도 제공하지 않으므로 데이터 집합에서 제거된다. 이는 앞 절에서 제시한 두 기준 중 두번째에 해당된다. 남은 레이블된 데이터는 $(t+1)$번째 과정에서의 레이블된 데이터의 수가 t 번째와 같아지도록 중요한 레이블이 없는 데이터에 의해 보충된다. 이 새 레이블된 데이터로 새 분류기 f_{t+1}이 학습된다.

이 과정은 레이블이 없는 데이터가 다 사용되거나 τ_t 가 0.5 보다 작아질 때까지 반복된다. 모르는 문서 $\mathbf{x}$ 에 대해, $\mathbf{x}$ 가 적합할 확률은 다음과 같이 결정된다.
$$f^*(\mathbf{x}) = \left(\prod_{i=1}^{k_x-1} \tau_i\right) f_{k_x}(\mathbf{x}).$$
여기서, k_x 는 x 를 분류할 가장 좋은 분류기의 인덱스이다.

4. 데이터 집합

첫번째 데이터 집합은 NIPS 2000 의 "Using Unlabeled Data for Supervised Learning" 워크숍에서 사용되었던 데이터 집합이다. 이 데이터 집합에는 두 종류의 웹 문서 분류 데이터가 포함되어 있다. 첫번째 문제(P2)는 http://www.microsoft.com과 http://www.linux.org의 홈페이지를 구분하는 문제이고, 두번째 문제(P6)는 http://www.mit.edu와 http://www.uoguelph.ca의 홈페이지를 구분하는 문제이다. 두번째 문제는 'MIT', 'Institute',

'Guelph'와 같이 중요한 단어들이 제거되었기 때문에 첫 번째 문제보다 어렵다.

표 1 은 이 데이터 집합을 보인다. 각 문서는 bag-of-words 로 표현되었으며, 각 자질은 단어 빈도수로 표현되었다. 이 표현을 좀더 정교하게 하기 위해서, 본 논문에서는 각 자질을 *tf•idf* 로 표현하였다.

|표 1| NIPS 2000 워크숍 데이터 집합.

Data Set	P2	P6
No. of Labeled Data	500	50
No. of Unlabeled Data	5,481	3,952
No. of Test Data	1,000	100
No. of Terms	200	1,000

두번째 집합은 co-training 에서 사용된 CMU text learning group 의 "The 4 Universities Data Set"중 일부이다. 이 집합은 Cornell, University of Washington, University of Wisconsin, University of Texas 에서 수집된 1,051 개의 웹문서로 구성된다. 이 1,051 개의 문서는 수작업으로 *course* 와 *non-course* 로 분류되었다. 표 2 는 각 대학의 웹문서 수를 보인다. 기준 정확도는 각 예제에 대해 non-couse 라고 대답했을 때 얻어지는 분류 정확도이다.

|표 2| WebKB 데이터 집합.

Data Set	Course	Non-Course	Baseline
Cornell	40	203	83.5%
Texas	38	216	85.0%
Washington	74	220	71.1%
Wisconsin	78	220	73.8%
Total	230	821	78.1%

5. 실험 결과

본 논문에서는 분류기로 MLP 를 사용하였고, Pentium III 550MHz 와 256MB 메모리를 가진 PC 에서 Linux 하에서 실험하였다. 표 3 은 NIPS 2000 워크숍 데이터 집합에 대한 실험 결과를 보인다. 이 테이블에 있는 TSVM 은 [4]에서 제시된 Transductive SVM 으로 현재 레이블이 없는 데이터를 이용한 문서 분류에서 가장 좋은 성능을 보이는 알고리듬이다. 표 3 에 따르면, 본 논문에서 제시된 방법이 TSVM 과 거의 비슷한 성능을 보이며, 레이블이 없는 데이터를 사용하지 않았을 때보다 P2 에 대해서 0.7%, P6 에 대해서 15.0% 정도의 성능 향상이 있었다.

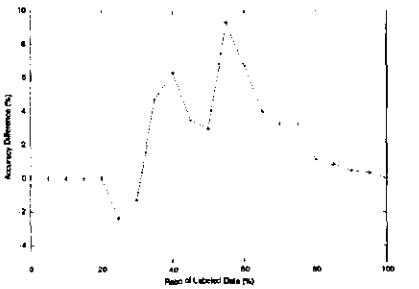
|표 3| NIPS 2000 워크숍 데이터에 대한 실험 결과.

Method	Using Only Labeled Data		Using Both Labeled and Unlabeled Data	
	P2	P6	P2	P6
Our Method	98.8%	60.0%	99.5%	75.0%
TSVM	N/A	N/A	**99.7%**	**80.0%**

WebKB 데이터에 대한 실험에서는 제시된 방법으로 레이블이 있는 데이터만 사용했을 때보다 평균적으로 약 1.5% 정도 높은 분류 정확도를 얻었다(표 4). 이는 기준 정확도보다 16.2% 정도 높은 결과이며, co-training 보다 0.4% 높다. 그림 3 은 레이블이 없는 데이터를 추가적으로 사용함으로써 얻을 수 있는 정확도의 이득을 그래프로 나타낸 것이다. 이 그림에 따르면, 55%의 레이블이 있는 데이터만 사용하고도 9.2%의 정확도 향상을 얻을 수 있다.

|표 4| WebKB 데이터에 대한 실험 결과.

Data Set	Using Partially Labeled Data	Using All Labeled Data	Co-Training
Cornell	**94.2%**	93.4%	N/A
Texas	**97.1%**	96.5%	N/A
Washington	**91.4%**	89.9%	N/A
Wisconsin	**94.2%**	91.3%	N/A
Average	**94.2%**	92.8%	93.8%



|그림 3| 레이블이 없는 데이터를 사용해서 얻어지는 정확도의 차이.

6. 결론

본 논문에서는 제한된 레이블이 있는 데이터의 수를 보충하기 위해서 레이블이 없는 데이터를 사용하는 자동 문서 분류 방법을 제시하였다. 제시된 학습 방법에서는 소수의 레이블이 있는 데이터로 분류기를 학습한 후, 대량의 레이블이 없는 데이터로 다음 분류기를 만들어, 앙상블을 구현하였다. 이를 통해, 지식 습득의 병목을 해소하였다. 실험 결과, 레이블이 없는 데이터를 추가적으로 사용했을 때에 사용하지 않았을 때보다 문서 분류의 정확도가 증가하였다.

감사의 글

이 논문은 정보통신부 대학기초연구(과제번호 : 00-023)와 교육부 BK 21 사업에 의하여 지원되었음.

참고 문헌

[1] L. Asker and R. Maclin, "Ensembles as a Sequence of Classifiers," In *Proceedings of IJCAI-99*, pp. 860-865, 1999.

[2] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," In *Proceedings of COLT-98*, pp. 209-214, 1998.

[3] T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," In *Proceedings of ICML-99*, pp. 200-209, 1999.

[4] B. Shahshahani and D. Landgrebe, "The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 32, pp. 1087-1095, 2000.

[5] Y. Yang and J. Pederson, "Feature Selection in Statistical Learning of Text Categorization," In *Proceedings of ICML-97*, pp. 412-420, 1997.

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