레이블이 없는 문서를 이용한 SVM 기반의 점증적 지도학습

김 수영, 조 성배
연세대학교 컴퓨터 과학과

Incremental Supervised Learning based on SVM with Unlabeled Documents

Soo-Young Kim and Sung-Bae Cho
Computer Science Department, Yonsei University

요 약

컴퓨터가 널리 보급되고 인터넷이 발전함에 따라 수없이 많은 정보가 디지털 형태로 생산되고 있다. 이러한 정보를 사람이 일일이 가공하고 분류하기에는 한계가 있으므로 자동으로 문서를 분류하고자 하는 연구가 대두되었다. 문서를 자동으로 분류하기 위해 기계학습 방법이 많이 이용되고 있다. 기계학습방법을 이용한 문서분류가 좋은 성능을 내기 위해서는 충분한 양의 학습데이터가 필요하다. 학습데이터를 만들기 위해서는 사람이 일일이 분류해야 하므로, 비용이 많이 든다. 본 논문에서는 적은양의 labeled 데이터로부터 시작하여, 점증적으로 unlabeled 데이터를 학습에 참여시킴으로써, 문서분류의 성능을 높이고자 한다. 실험을 통해 Unlabeled 문서데이터를 사용한 것이 좋은 성능을 보였음을 알 수 있다.

1. 서 론

문서의 자동분류는 1960년대의 정보검색의 한 분야로 연구되어지기 시작했고, 1980년대 말 까지는 주로 이론적인 연구에 머물러 있었고, 실제 응용프로그램 또한 전문가의 수작업을 통해 규칙을 생성해 낸 기반으로 문서를 자동분류 하는 방법을 통해 구현되었다. 1990년대에 접어들어 컴퓨터가 널리 보급되고 인터넷이 발전함에 따라 디지털 형태의 정보가 급격히 증가하기 시작하여 정보의 과잉현상이 나타나게 되었다. 이런 이유로 많은 양의 정보를 자동으로 가공하여 분류하는 문서분류분야의 중요성이 널리 인식되기 시작하였다.[1]

최근의 문서자동분류의 연구는 주로 여러 통계적 기계학습 기법을 기반으로 이루어지고 있다. K-최근접이웃, 결정트리, Support Vector Machine(SVM), 신경망, Naïve Baise 등의 기계학습 기법은 도메인 지식에 독립적이며 대량의 정보를 다룰수 있어서 자동분류분야연구에 알맞은 기법이다. 그러나 통계적인 기계학습을 기반으로 하는 문서분류는 이미 분류가 되어 있는 충분한 양의 데이터를 기계학습 알고리즘으로 학습시킨후, 새로운 문서가 들어왔을 때 학습된 데이터를 기반으로 문서를 분류하는 것이다. 기계학습 알고리즘이 좋은 성능을 내기 위해서는 미리 분류되어진 많은 양의 학습데이터가 필요하다. 그러나 이러한 학습데이터를 일일이 분류하는 데에는 많은 비용이 들기 때문에 가능한 모든 데이터를 분류하는 것은 불가능한 경우가 많다. 반면에 분류되지 않은 문서는 쉽게 다량으로 얻을 수 있다.

본 논문에서는 소량의 분류된(labeled) 문서와 많은 양의 분류되지않은(unlabeled) 문서를 이용하여 문서분류의 성능을 향상시킬 수 있는 방법을 제안한다. 문서분류 분야에서 좋은 성능을 보여주는 것으로 알려진 SVM 분류기를 기반으로 하여, Unlabeled 데이터를 점정적으로 학습에 참여시킴으로써 문서를 분류한다. 초기에 Labeled 된 데이터를 SVM으로 학습시킨 후 그 모델을 이용해 unlabeled 문서를 분류하고 그 중에서 일정한 임계치에 해당하는 Unlabeled 문서만을 샘플링하여 label을 붙인후, 다시 재학습하는 과정을 반복적으로 적용하여 문서분류의 성능을 향상시킬 수 있음을 보인다.

2. 관련연구

2.1 문서분류

문서분류란 새로운 문서를 미리 정의된 부류로 대응 시

키는 일련의 작업을 말한다. 미리 정의된 부류의 집합을 $C=\{c_1, c_2, ..., c_n\}$, 새로운 문서의 집합을 $D=\{d_1, d_2, ..., d_m\}$이라 할 때, 알려지지 않은 분류함수 $g: C \times D \rightarrow \{0, 1\}$에 근접한 함수 $f: C \times D \rightarrow \{0, 1\}$을 만들어 내는 작업으로 정의할 수 있다. 여기서 분류함수 g는 임의의 부류 c_i $(1 \le i \le n)$에 대한 임의의 문서 d_j $(1 \le i \le m)$의 멤버쉽함수이다. 즉, $f(c_i, d_j)$가 1일 경우에는 d_j가 c_i에 속하는 것을 의미하며, 0일 경우에는 속하지 않음을 의미한다. f는 분류함수 g에 가능한 한 근접하게 만들어야 하며, 근접한 정도가 f의 성능, 즉 자동분류 성능의 기준이 된다.

2.2 Labeled, Unlabeled 데이터를 이용한 학습

기계학습방법을 이용한 문서분류기가 최대의 성능을 발휘하려면 충분한 양의 Labeled 데이터가 존재해야 하는데, 그것을 얻기위해서는 많은 비용이 든다. 따라서 최근에는 Unlabeled 데이터를 이용해서 문서분류의 향상을 높이려는 연구가 이루어지고 있다. Nigam은 unlabeled 데이터가 문서분류의 성능을 향상시킬 수 있는 이유는 unlabeled 데이터들이 문서에 있는 단어들 사이의 결합확률분포에 관한 정보를 줄 수 있기 때문이라는 것을 증명 한 바 있다[2]. 학습분류기에 labeled 데이터와 unlabeled데이터를 결합하는 연구의 한가지 부류는 우도 최대화 접근으로, 문서분류를 mixture model 의 한 형태로 보고 파라미터를 EM알고리즘으로 조절하여, 문서분류에 이용하는 것이다. 또한가지 부류는 선택적 샘플링으로, Unlabeled 데이터를 지도학습(supervised learning)에 통합시키는 방법이다.

3. Unlabeled 문서의 점증적 지도학습

본 논문에서는 SVM분류기의 특성을 이용하여, unlabeled 문서를 점증적으로 학습에 추가시킴으로써, 문서분류의 성능을 향상시킬 수 있는 방법을 제안한다. 그림1은 unlabeled 데이터를 이용한 문서 분류기의 그림이다. 문서자동분류분야에 사용되는 통계적인 기계학습방법중 비교적 최근에 개발된 학습방법으로 V.Vanpic에 의해 제안된 Support Vector Machine이 있다. SVM의 기본 아이디어는 구조적 리스크 최소화를 통해 벡터

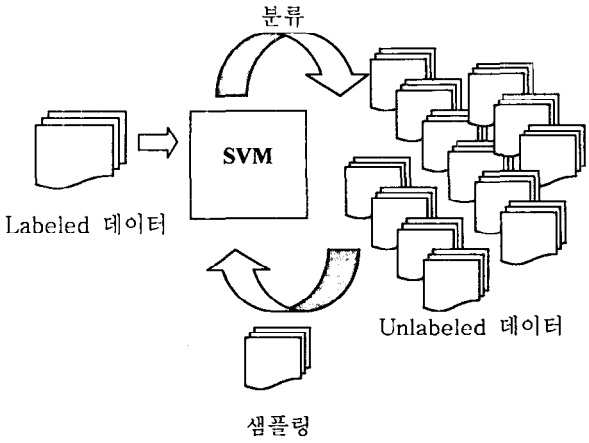

그림 1. unlabeled 데이터를 이용한 문서분류기

공간에서의 최적의 결정경계영역을 찾아내는 것으로 이진분류문제를 푸는 방법으로 이용되고 있다. [3]

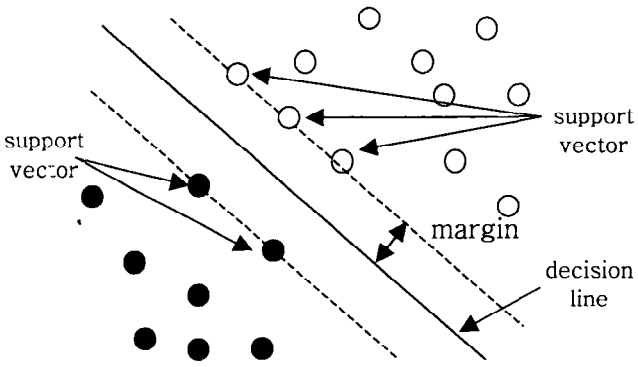

그림 2. 2차공간에서의 SVM의 결정경계영역

SVM은 결정경계영역을 구함으로써, 분류를 수행하게 되는데, Unlabeled 데이터가 들어오면 unlabeled데이터에 맞게 결정경계면이 변화해야만 분류의 성능이 높아지게 된다. Labeled 데이터로 학습된 결정경계면을 이상적인 결정경계면으로 옮기는데 필요한 Unlabeled 데이터를 뽑을수 있다면, 분류의 성능이 높아질 수 있을것이다. 점증적 학습과정은, 먼저 Labeled 문서를 가지고 SVM을 학습시킨 후에, 그 SVM을 가지고 Unlabeled 문서들에 레이블링을 한다. Unlabeled문서중에서 일정 임계치를 가지는 문서들만 샘플링 한다. 샘플링 된 문서들에 SVM으로부터 얻은 레이블을 붙여서 labeled 문서집합에 편입시킨다. 재구축된 Labeled 문서집합을 가지

고 SVM을 다시 학습시킨다. 이 과정을 종료조건이 만족할 때 까지 만족시킨다. Unlabeled문서 중에서 신뢰할 만한 데이터를 뽑기 위해서 Zhang과 Oles의 준 모수적 모델을 사용하였다. 그들은 준 모수적 모델에서 레이블이 없는 데이터의 피셔정보를 최대화하기 위해서는 능동학습 기법을 사용해야 하며, 정보가 많은 레이블이 없는 데이터를 뽑는 기준은 추정된 모수로 신뢰도가 낮은 데이터를 선택하되 중복되지 않아야 한다고 하였다. [2]. 분류에 도움이 되는 Unlabeled 데이터를 샘플링하기 위해 SVM분류기의 특성을 이용한다. SVM은 학습한 후에 두개의 클래스를 나누는 결정경계면 $\vec{w}\cdot\vec{x}-b=0$ 을 구할 수 있다. 분류해야 할 데이터를 $\vec{x}_i$ 라고 하면, $\vec{x}_i$ 가 결정경계면으로부터 거리가 크면 클수록 제대로 분류될 확률이 높아지고, 결정경계면과의 거리가 가까울수록 제대로 분류할 확률이 적어진다. 먼저 labeled문서로 학습한 SVM분류기를 이용하여 결정경계면을 구하고 하이퍼공간에서 unlabeled문서와 결정경계면 사이의 거리를 구하여, 그 거리가 0보다 크고 margin의 2배보다는 작은 Unlabeled문서만을 샘플링한다.

그림 3은 unlabeled 문서를 이용한 점증적 학습과정을 알고리즘으로 나타낸 것이다.

Given labeled example set $L=\{(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_l, y_l)\}$

and unlabeled example set $U=\{\mathbf{x}_1, \ldots, \mathbf{x}_u\}$

and test example set $T=\{(\mathbf{s}_1, y_1), \ldots, (\mathbf{s}_s, y_s)\}$;

Initialize SVM classifier f_0 by training with $L_0=L$, $U_0=U$

Do

1. Set $margin_t$ as margin of f_t
2. Set $\tau = margin_t * 2$
3. Set dist(x) = distance between
4. Set $U_{add} = \{(x_1,y_1),\ldots,(x_n,y_l) \mid x \in U_t, y=f_t(x), dist(x) \leq \tau\}$
5. Set $L_{t+1} = L_t + U_{add}$
6. Set $U_{t+1} = U_t - U_{add}$
7. Train f_{t+1} with L_{t+1}
8. Classify test set T with f_{t+1}
9. Set PR_t as classification precision of T
10. Set $t = t+1$

While($|U_{add}|>0$ & $PR_t > PR_{t-1}$)

Store the final SVM classifier

그림 3. 점증적 학습 알고리즘

4. 실험

Reuter-21578 코퍼스를 데이터로 하여 실험을 하였다. ModApte 토픽을 사용했으며, 총 135개 카테고리중 빈번히 나타나는 7개의 토픽을 실험대상으로 하였다. 각 카테고리당 각각 20개의 positive샘플과 20개의 negative 샘플을 추출하여 labeled문서로 사용하였다. 분류의 성능분석을 위해 정확률/재현률, F1 measure를 사용하였다. 표 1은 실험결과임.

	Labeled데이터만으로 학습후 분류결과			점증적학습후 분류결과		
	정확률	재현률	F1	정확률	재현률	F1
Grain	33.60	61.94	43.56	64.35	55.22	59.44
Earn	36.21	99.41	53.08	93.91	88.09	90.91
Acq	49.96	89.32	64.08	51.08	91.49	65.56
Money-fx	19.63	89.54	32.20	22.33	88.89	35.69
Corn	11.27	48.00	18.25	51.92	54	52.94
Crude	28.87	50.60	36.76	69.32	36.75	48.03
Wheat	13.94	62.50	22.80	61.54	62.50	62.02
평균	27.64	71.62	38.67	59.21	68.13	59.22

표 1. 분류 성능

5. 결론

실험결과 Labeled문서로만 학습하여 분류한 결과보다 Unlabeled 문서를 이용하여 점증적 학습한 후의 분류결과가 좋은 성능을 보였음을 알 수 있었다.

6. 참고문헌

[1] 정영미, 정보검색론, 구미무역출판부, 1993

[2] K. Nigam, A. McCallum, S. Thrun and T. Mitchell, "Learning to Classify Text from Labeled and Unlabeled Documents," In Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence, pp. 792-799, 1998.

[3] V. Vanpik, "The Nature of Statistical Learning Theory," Springer-Verlag, 1995