One-class 문서 분류를 위한 가상 부정 예제의 사용

송호진[0] 강인수 나승훈 이종혁
포항공과대학교 컴퓨터공학과
{hojsong[0], dbaisk, nsh1979, jhlee}@postech.ac.kr

One-Class Document Classification using Pseudo Negative Examples

Ho-Jin Song[0] In-Su Kang Seung-Hoon Na Jong-Hyeok Lee
Dept. of Computer Science and Engineering, POSTECH

요 약

문서 분류에서의 one class classification 문제는 오직 하나의 범주를 생성하고 새로운 문서가 주어졌을 때 미리 만들어진 하나의 범주에 속하는가를 판별하는 문제이다. 기존의 여러 범주로 이루어진 분류 문제를 해결할 때와는 달리 one class classification에서는 학습 시에 이미 정해진 하나의 범주와 관련이 있는 문서들만을 사용하여 학습을 수행하기 때문에 범주의 경계를 정하는 것이 매우 어려운 작업이며 또한 분류기의 성능에 있어서도 매우 중요한 요소로 작용하게 된다. 본 논문에서는 기존의 연구에서 one class classification 문제를 해결할 때 관심의 대상이 되는 예제의 일부를 부정 예제로 간주하여 one class 문제를 two class 문제로 변경시켜 학습을 수행했던 것에서 더 나아가 추가적으로 새로운 가상 부정 예제를 설정하여 학습을 수행하고, SVM을 통하여 범주화 성능을 확인해 보기로 한다.

1. 서 론

인터넷이 대중화되고 네트워크 기술이 발전함에 따라 과거의 많은 문서들이 전자 문서로 대체되고 있다. 따라서 전자 문서의 수는 기하급수적으로 늘어나고 있으며 이에 따라 방대한 양의 문서에 대한 체계적인 분류의 필요성은 점점 늘어나고 있다. 그러나 많은 문서를 사람이 직접 분류하기에는 너무 많은 노력과 시간이 필요하기 때문에 문서의 자동 분류에 대한 필요성이 증가하고 있다.

문서를 자동분류하기 위해서 다양한 기계학습 방법이 사용될 수 있다. 그러나 일반적으로 알려진 기계 학습 방법들은 각각의 범주에 해당하는 문서(positive 학습 예제)와 그 범주에 해당하지 않는 문서(negative 학습 예제)들이 모두 학습 데이터로 제공될 경우에 적용될 수 있다..

그러나 현실적으로 positive 학습 문서들의 수가 많고 negative 학습 문서의 수가 극히 드물거나 혹은 그 반대의 경우가 발생할 수도 있다. 그리고, 의학적인 문서들의 경우처럼 negative 학습 문서를 획득하기가 어려운 상황도 존재한다. 이러한 상황에서 negative 학습 문서에 대한 별도의 처리가 없이 단순히 일반적인 two class 문제와 동일하게 취급하여 분류를 시도하는 것만으로는 근본적인 해결책이 되지 못한다[1]. one-class classification은 negative 학습 예제가 편중된 상황에 대한 분류 기법에 관한 연구로 최근 많은 관심을 보이는 분야이다.

One class classification에서는 관심의 대상이 되는 문서(positive 학습 문서)들만 가지고 학습을 수행하여 관심 문서만을 분류하는 모델을 구축한다. 이는 positive와 negative 학습 문서 모두를 사용하는 기존의 분류방법과 가장 큰 차이점이라고 할 수 있다. One class classification에서 얻어진 학습 모델은 새롭게 발생한 문서가 관심 대상인지 아닌지를 판단하는데 사용된다.

정확한 분류 경계를 정하는 것은 classification에 있어서 가장 중요한 문제인데 one-class classification에서는 관심이 대상이 되는 예제들만 존재하기 때문에 그 분류경계를 설정하는 것이 쉽지 않다. 본 논문에서는 기존 one- class 분류 문제 해결에 있어서 분류 경계를 정하기 위한 방법으로 관심의 대상이 되는 예제의 일부를 부정 예제로 간주하여 학습을 수행하는 기존의 방법이 분류 경계 설정에 있어서 간과하고 있는 부분이 있음을 지적하고 기존의 방법에서 더 나아가 추가적으로 새로운 가상 부정 예제를 설정하여 학습을 수행하는 방법을 제시할 것이다.

본 논문의 구성은 다음과 같다. 제 2장에서는 기존에 제안된 One class classification에서의 분류 경계를 결정하는 방법에 대해 소개를 하고, 제 3장에서 이를 개선한 방법인 가상 부정 예제를 사용한 방법에 대해 기술한다. 제 4장에서는 대표 분류기로 SVM을 사용하여 제안된 방법의 효과를 확인하고 제 5장에서 결론을 내린 후에 향후의 연구 계획에 대해서 간단히 기술한다.

2. 관련 연구

현재 One-class classification는 패턴인식, novelty detection, Outlier detection 문제의 해결에 적용되고 있고 이에 따른 많은 연구가 진행되고 있다[2][3].

문서 분류 문제를 one-class classification을 이용하여 해결한 기존연구는 원점을 대표 negative 학습 예제로 간주하고 이를 SVM에 적용한 Schölkopf의 연구[3]와 신경망을 사용하여 해결한 Manevitz 의 연구[4]가 있다. 그리고, 원점뿐만 아니라 원점과 가까운 몇 개의 positive 학습 문서도 같이 negative 학습 문서로 사용하고 이를 SVM에 적용한 Manevitz의 연구[5]가 있다

Schölkopf는 SVM 커널을 통하여 feature를 high dimensional space 상에 표현한 다음 그 공간 상에서 원래의 학습 문서들을 하나의 범주를 설정하고 원점 하나로만 이루어진 또 하나의 범주를 설정하였다[3]. 그러면 이 문제는 미리 설정한 범주에 속하는 문서들의 집합을 원점과 분리시키는 two class 문제이며, 이것은 l 개의 학습 문서 $\{(x_1,y_1), \cdots, (x_l,y_l)\}$ 가 주어졌을 때

$$< \omega \bullet \omega > + b \geq \rho - \zeta_i \quad \zeta_i \geq 0, b = 0, \ i=1,\ldots,l \quad (1)$$

이라는 제약하에서 다음과 같은 QP(quadratic problem) 문제를 해결하는 문제로 볼 수 있다.

$$\min \frac{1}{2}\|\omega\|^2 + \frac{1}{vl}\sum_{i=1}^{l}\zeta_i - \rho \quad (2)$$

학습을 통해 최적화된 ω 와 ρ 를 구한 후 다음과 같은 결정함수를 통하여 test 문서를 가지고 분류를 시행할 수 있다.

$$f(x) = sign(<\omega \bullet \omega> - \rho) \quad (3)$$

Manevitz는 [5]에서 원점 하나만을 negative 예제로 사용했던 Schölkopf 의 방법을 개선하여 원점 뿐만 아니라 실제로 다른 positive 예제들과 관련이 적은 positive 예제들을 negative 예제들로 사용하여 학습을 수행하였다. 다른 positive 예제들과 관련이 적은 positive 예제는 학습 예제들을 벡터로 표현했을 때 0의 값을 가지는 요소의 개수가 일정 개수 이상인 학습 예제들 또는 원점에서의 거리가 충분히 가까운 학습 예제들이다.

Manevitz는 이러한 기준에 따라서 positive 학습 예제와 negative 학습 예제를 구성하여 학습을 수행하고 SVM을 통한 분류 성능을 보고하였다.

3. 범주의 분류 경계 설정을 위한 가상 부정 예제의 추가

기존의 연구 중에 Manevitz는 [5]에서 기존의 schölkopf 의 실험과는 다른 방법을 제안하여 실험을 하고 성능의 변화를 보고하였다. 그러나 그가 설정한 negative 학습 예제들은 실제의 negative 예제들을 대표하기에는 부족하다.

그림 1에서 보는 것처럼 실제적으로 A 부분에 대한 분류 경계가 제대로 설정되지 않았다. 만일 이 부분에 negative 예제가 존재한다면 이미 알려진 positive 예제들로 이루어진 공간 B를 B가 아닌 공간과 잘 구분해 줄 것이다.

우선 각 범주 별로 벡터 공간 상에서 원점에서 멀리 있는 n % 개의 positive 학습 예제들을 새로운 positive 학습 예제로 선정하고 남은 positive 예제들은 원점과 함께 negative 예제로 선정하여 one-class 분류 문제를 two-class 분류 문제로 변형시킨 후에 학습을 수행하였다. Positive 예제로 사용할 예제들의 수를 늘려가면서 학습을 수행하고 따로 마련된 평가 문서들을 가지고 테스트를 하면서 성능을 비교하면 범주에서 원점 방향으로의 최상의 성능을 보이는 구성을 찾을 수 있고 이 구성은 two-class classification을 위한 positive 예제들과 원점을 중심으로 하는 negative 예제들로 구성된다.

그림 1의 A 부분에는 특별히 negative 학습 예제로 간주할 만한 지점이 존재하지 않기 때문에 Reuter 문서 집합상의 10개의 범주 각각에 대하여 다음과 같은 가상 부정 예제를 만들어 주었다.

$$N = (\max(1)+0.02, \max(2)+0.02, \cdots, \max(m)+0.02) \quad (4)$$

여기서 m 은 키워드로 선택된 단어의 개수를 의미하고 max(i)는 각 범주의 i 번째 키워드의 최대 가중치를 의미한다. 또한 다음과 같은 m개의 가상 부정 예제를 구성하여 범주의 분류 경계를 더욱 확실하게 설정해 주었다.

B_i = i 번째 keyword 의 가중치는 max(i) + 0.02
나머지 keyword의 가중치는 0 (5)

그림 2에서 X 로 표시된 벡터들이 이러한 가상 부정 예제들이며 이전에 얻어진 최적의 구성의 negative 예제 집합에 식 (4)의 N과 식 (5)의 m 개의 벡터 B_i 를 추가시켜 학습에 사용하였다.

4. 실험 및 평가

4.1 데이터 구성

본 실험에서는 문서 분류의 데이터로서 Reuters-21578 문서 집합을 사용하였다. Reuters-21578 문서 집합은 Reuter-newswire에 실린 기사들의 모음이다. Reuters-21578은 총 다섯 개의 범주 집합으로 구성되어 있고, 이 중에서 문서 분류에 대한 연구는 주로 135개의 범주로 구성된 TOPIC 범주 집합내의 10개의 범주를 사용하여 이루어지고 있다.

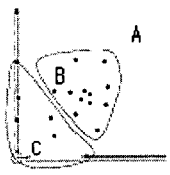

그림 1. Manevitz의 분류

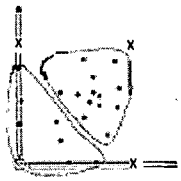
그림 2. 가상 부정 예제 추가

본 실험에서도 TOPIC 범주 집합내의 문서 수 상위 10개의 범주를 사용하였으며 실험에 사용된 데이터의 수는 표 1에 나타내었다.

이러한 데이터에는 실제로 각 범주를 잘 표현하지 못하는 단어들도 존재하게 된다. 이러한 단어들은 문서 분류의 성능을 감소시키는 결과를 가져올 수 있으므로 각 범주마다 각 단어에 다음과 같은 점수를 부여하고 이 점수가 높은 일정량의 단어들을 가지고 실험을 수행하였다. 본 실험에서는 각 범주마다 가장 많이 발생한 10개 또는 20개의 단어를 선별하고 이 단어들을 키워드로 사용하였다.

$$Score_i = tf_i * df_i \quad (6)$$

< i = 각 범주에서 i 번째 term >

각 범주별로 선별된 10개 또는 20개 term에 대하여 변형된 Hadamard representaion을 사용하여 각 범주에 속하는 예제들을 vector로 표현하였다. 변형된 Hadamard representation은 다음과 같은 식으로 표현될 수 있다

$$H_E(e_i) = e_i \times v_{Ei} \quad (7)$$

e_i : 한 문서에서 i 번째 term의 okapi TF

v_{Ei} : 범주에 속하는 전체 문서에서 i 번째 term의 okapi TF

표 1. 실험에 사용된 데이터의 수

범주	학습 문서 개수	테스트 문서 개수
Earn	822	3010
Acq	663	1692
Money-fx	217	560
Grain	119	488
Crude	220	398
Trade	150	378
Interest	129	372
Ship	95	195
Wheat	48	252
Corn	39	204

4.2 기계학습 방법의 선택

본 실험에서는 LIBSVM (ver. 2.71)을 사용하여 기계학습을 수행하였다[7]. Kernel 은 default kernel 인 RBF를 사용하였으며 다른 매개변수 값도 기본적으로 주어진 값을 사용하였다.

4.3 실험 결과 및 분석

표 2는 각 문서의 단어 가중치 값의 합을 기준으로 범주별 전체 positive 문서 중에 상위 n %을 positive 문서로 사용하였고 나머지 positive 문서들을 negative 문서라고 간주하고 학습했을 때의 문서분류 성능을 나타낸다. 10개의 범주에 대하여 n 이 커짐에 따라 성능이 향상되다가 어느 시점부터는

표 2 . 범주별 전체 positive 문서의 일정 비율만을 positive 문서로 사용하여 학습했을 때의 문서분류 성능
(변형된 Hadamard representation, Dimension = 10)

	25%	50%	70%	80%	90%	95%
	F_1	F_1	F_1	F_1	F_1	F_1
Earn	0.17	0.35	0.48	0.44	0.6	*0.73*
Acq	0.26	0.48	*0.54*	0.49	0.48	0.4
Money	0.24	*0.49*	0.44	0.4	0.15	0.15
Grain	0.11	0.36	*0.42*	0.23	0.03	0.2
Crude	0.48	*0.51*	0.31	0.27	0.11	0.11
Trade	0.05	0.18	*0.19*	0.17	0.11	0.1
Interest	0.11	0.44	*0.51*	0.3	0.1	0.1
Ship	0.06	*0.18*	0.06	0.06	0.06	0.06
Wheat	*0.36*	0.08	0.07	0.07	0.07	0.07
Corn	*0.07*	0.06	0.05	0.06	0.06	0.06
Avg	0.19	*0.3*	*0.3*	0.25	0.18	0.2

성능이 저하되는 것을 볼 수 있다. 10개의 범주 모두에 대하여 일관성 있게 적용되지는 않았지만 많은 범주에서 n이 70 또는 50에서 최대의 성능을 나타내었다. 이 같은 결과는 10개 범주의 평균 성능을 보아도 알 수 있다.

표 2에서 각 범주별로 가장 최상의 성능을 보이는 최적의 positive 예제, negative 예제 구성에 가상 부정 예제들을 추가하였을 때는 표 3과 같은 성능을 보였고 이는 표 2와 비교하였을 때 분류 성능의 향상을 가져왔다고 할 수 있다. 또한 이 성능은 같은 데이터 집합을 사용하고 비슷한 학습 문서의 개수를 가지고 실험을 한 [5]의 성능과 비교해 볼 수 있는데 Manevitz의 방법이나 Schölkopf 방법보다 낮은 Macro 평균이 보였지만 Micro 평균에서는 두 방법보다 약간 더 좋은 성능을 보였다.

5. 결론 및 향후 과제

one-class classification 문제 해결을 위해서는 주어진 하나의 범주에 대한 분류 경계를 정하는 것이 매우 중요한 일이다. 이러한 분류 경계를 정하기 위해서는 주어진 하나의 범주와 가장 관련 있는 단어를 추출하는 것도 매우 중요하지만 다른 범주와의 차별을 줄 수 있는 단어를 찾는다는 것도 중요하다. 하지만 one-class 분류 문제에서는 관심의 대상이 되는 하나의 범주에 대한 학습 예제들만 주어지기 때문에 주어진 범주와 관심 밖의 범주를 차별화 시키는 단어를 추출하는 것은 어려운 일이다. 따라서 one-class 문제 해결에 있어서는 관심의 대상이 되는 범주를 가장 잘 표현하는 용어들을 추출하고 그러한 벡터 공간 상에서 범주에 속하는 positive 예제들은 모두 포함시키면서 범주의 부피를 최소화 하는 것이 분류 성능을 높일 수 있는 최상의 방법이다. 그리고 만일 여러 기계학습 방법을 사용하여 one-class classification 문제를 해결할 때에는 그 기계학습의 특징을 잘 이해하고 one-class classification을 다루어야 더욱 높은 분류 성능을 얻을 수 있다. SVM을 예를 든다면 여러 kernel과 여러 매개 변수 값이 존재한다. 실제로 이런 설정이 분류 성능의 변수로 작용하기 때문에 높은 분류 성능을 위해서는 기계 학습의 특징에 대한 정확한 이해와 더불어 많은 실험을 통하여 적합한 매개변수 값을 찾는 것이 요구된다

이에 본 논문에서는 범주 내에서 단어의 빈도수에 의하여 최적의 단어들을 결정하였고 최상의 분류 경계를 설정하기 위하여 가상 부정 예제들을 구성하여 학습에 이용하였다

앞으로 범주의 부피는 최소화하고 positive 문서들은 모두 포함하는 최적의 분류 경계를 찾기 위한 연구가 계속 되어야 할 것이며 이러한 최적의 분류 경계상에서 기존의 문서 분류

표 3. 최적의 구성에 인공 벡터 추가했을 때 분류성능과 다른 system과의 성능 비교 (measure : F_1)

	Our approach	Schölkopf	Manevitz	Neural Net
Trade	0.19	0.6	0.42	0.57
Money	0.49	0.51	0.57	0.64
Grain	0.44	0.59	0.52	0.47
Ship	0.18	0.54	0.40	0.40
wheat	0.31	0.47	0.39	0.39
corn	0.35	0.30	0.36	0.36
interest	0.53	0.49	0.47	0.47
crude	0.52	0.54	0.47	0.48
Acq	0.54	0.48	0.50	0.62
earn	0.84	0.68	0.75	0.71
Macro avg.	0.44	0.52	0.48	0.51
Micro avg.	0.61	0.57	0.59	0.62

방법들이 적용되어야 할 것이다. 또한 분류 문제의 해결에 있어서 현재 활발하게 연구되고 있는 dimension reduction technique이 one-class 분류 문제 해결에 적용된 사례를 찾아보기가 어려운데 일반적인 분류 문제 해결에 있어서 dimension reduction technique이 좋은 분류 성능을 보이고 있으므로 one-class classification 문제에도 dimension reduction technique을 적용하는 연구가 이루어져야 할 것이다.

참고문헌

[1] Tax, D.M.J, " One-class classification : Concept-learning in the absence of counter-examples," Ph.D. Thesis, TU, Delft, 2001

[2] Tax, D.M.J. and Duin, R.P.W., "Outliers and data descriptions", *7th Annual Conf. of the Advanced School for Computing and Imaging*, pp. 234-241, ASCI, Delft, 2001.

[3] B. Scholkopf, R. Williamson, A. Smola,J.Shawe-Taylor, and J.Platt. " Support vector method for novelty detection," In Advances in Neural Information Processing Systems, pages 582-588. MIT Press, 2000.

[4] Larry M. Manevitz, Malik Yousef. " Document classification on neural networks using only positive examples." SIGIR Conference on Research and Developm
ent in Information Retrieval, July 24-28, 2000, pp. 304-306

[5] Larry M. Manevitz, Malik Yousef. " One-Class SVMs for Document Classification," Journal of Machine Learning Research, vol. 2, no. 2, pp. 139-154(16), 1 May 2002

[6] Yiming Yang and Xin Liu. A re-examination of text categorization methods," Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, page 42-49

[7[LIBSVM: a Library for Support Vector Machines, (http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html)

[8] C.J.C. Burges, " A Tutorial on Support Vector Machines for Pattern Recognition" , Kluwer Academic Publishers, 1998