One-class 문서 분류를 위한 긍정 자질과 부정 자질의 결합

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

Combining Positive and Negative Features for One-Class Document Classification

Ho-Jin Song[0] In-Su Kang Seung-Hoon Na Jong-Hyeok Lee
Dept. of Computer Science and Engineering, POSTECH
{hojsong[0] , dbaisk , nsh1979, jhlee} @postech.ac.kr

요 약

문서 분류에서의 one class 분류 문제는 오직 하나의 범주를 생성하고 새로운 문서가 주어졌을 때 그 문서가 미리 만들어진 하나의 범주에 속하는가를 판별하는 문제이다. 기존의 여러 범주로 이루어진 분류 문제를 해결할 때와는 달리 one class 분류에서는 학습 시에 관심의 대상이 되는 하나의 범주와 관련이 있는 문서들만을 사용하여 학습을 수행하기 때문에 범주의 경계를 정하는 것은 매우 어려운 작업이다. 이에 본 논문에서는 기존의 연구에서 one class 분류 문제를 해결할 때 관심의 대상이 되는 예제의 일부를 부정 예제로 간주하여 one class 문제를 two class 문제로 변환하고 추가적으로 새로운 가상 부정 예제를 설정하여 학습을 수행하였던 방법에서 더 나아가 범주화를 위한 적절한 부정자질을 선택하고 이를 긍정자질과 함께 사용하여 학습을 수행한 후 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 문서들의 수는 기하급수적으로 늘어나고 있으며 이에 따라 방대한 문서를 체계적으로 분류해야 할 필요성이 점차 증대하고 있다. 그러나 나날이 늘어나는 많은 문서들을 사람이 직접 분류하는 것은 많은 노력과 시간을 필요로 하게 된다. 따라서 방대한 문서들을 자동으로 분류하고자 하는 욕구가 증가하게 되고 그와 더불어 문서의 자동 분류 연구에 대한 관심도 증가하고 있는 실정이다.

많은 문서들을 자동으로 분류하기 위해서는 기계학습 방법을 이용하여 미리 범주를 정하고 새로운 문서가 주어졌을 때 그 문서를 적절한 범주에 배정할 수 있다. 그러나 일반적으로 알려진 많은 기계학습 방법들을 이용하기 위해서는 각각의 범주에 해당하는 문서(긍정 학습 예제)들과 그 범주에 해당하지 않는 문서(부정 학습 예제)들이 모두 학습 데이터로 주어지는 경우에만 적용이 가능하다.

또한 기존의 분류에서는 긍정 학습 예제의 개수와 부정 학습 예제의 개수가 어느 정도 균형을 갖춘 상태에서 학습을 할 때 가장 좋은 성능을 얻을 수 있다.

그러나 현실적으로는 하나의 범주에 속하는 학습 예제의 개수가 다른 범주에 속하는 학습 예제의 개수보다 훨씬 많은 경우가 존재하기도 한다. 이러한 상황에서 기존의 분류 방법을 적용하면 학습 예제의 개수가 적은 범주에서는 낮은 분류 성능을 보일 것이다[1]. 특히 학습 개수가 적은 범주에만 관심이 있는 경우에는 기존의 two class 분류 방법을 통한 낮은 성능은 큰 문제가 될 것이다.

최근에 학습 예제가 특정 범주에 편중된 상황에 대한 분류 기법에 관한 많은 연구가 진행되고 있는데 one class classification은 이러한 연구의 한 방향이라고 볼 수 있다[2][3].

One class classification에서는 관심의 대상이 되는 긍정 학습 예제만을 가지고 학습을 수행하여 관심 대상 문서만을 분류하는 모델을 구축하게 된다. 이는 긍정 학습 예제와 부정 학습 예제를 모두 사용하여 학습을 수행하는 기존의 분

류 방법과 one class classification과의 가장 큰 차이점이라고 할 수 있다.

기존의 분류 문제에서도 분류를 위한 정확한 분류 경계를 정하는 것이 중요한 문제이지만 one class classification에서는 관심의 대상이 되는 예제들만 존재하기 때문에 분류 경계를 정하는 일은 특히 어렵고 중요한 일이다. 본 논문에서는 one class 분류 문제를 해결하는 기존의 방법들을 살펴보고 그 방법이 간과하고 있는 부분이 있음을 지적한 후 적절한 부정 자질을 선별하여 학습에 이용하는 방법을 제시할 것이다.

본 논문의 구성은 다음과 같다. 제 2장에서는 기존에 제안된 one class 분류 문제 해결 방법들에 대해 소개를 하고, 제 3장에서는 현재까지 연구된 자질 선택 기법들의 특징을 간단히 살펴본 후에 one class classification에서 적용되지 않았던 방법인 적절한 부정 자질을 선택하고 긍정 자질과 함께 사용하여 학습을 수행하는 방법에 대해 기술한다. 제4장에서는 본 논문의 실험에 사용되는 가상 부정 예제의 설정에 대하여 간단히 기술하고 제 5장에서 대표 분류기로 SVM을 사용하여 제 3장에서 제안된 방법의 효과를 확인한 후 제 6장에서 결론을 내리고 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[4][5].

Novelty detection, Outlier detection 문제 해결에 있어서 대표적인 one class classification 방법으로는 Global Gaussian approximation, Parzen density estimation, 1-Nearest neighbour method, Local Gaussian approximation, 신경망 방법이 있다. 신경망을 제외한 네가지 방법들에서의 실제 분류는 자질 공간에서의 거리에 바탕을 두고 있다. 즉, 테스트 예제와 학습 예제들과의 거리를 미리 설정된 임계치와 비교하여 그 거리가 임계치보다 작은 경우에는 관심의 대상이 되는 범주로 분류하고 그 거리가 임계치보다 큰 경우에는 관심의 대상이 아닌 예제로 분류한다[6].

One class classification 신경망 방법에서는 Feedforward Neural Net을 사용하지만 기존의 MLP(Multi Layer Perceptron)와는 다른 Autoassociator 를 사용하고 있다. Autoassociator에서는 입력 층과 출력 층에서의 노드 수가 같으며 이는 입력으로 들어가는 자료와 출력으로 나오는 자료가 같도록 학습하기 위함이다. 즉, 학습 예제들이 신경망을 통과한 후 특정 오차 내에서 자신의 값을 다시 얻을 수 있도록 신경망을 학습하면 실제 분류에 있어서 학습 예제들과 유사한 테스트 예제들은 학습에 의해 설정된 특정 오차 내에서 자신의 값을 복원할 수 있을 것이라는 가정을 사용하여 분류를 수행하게 된다.

문서 분류 문제를 one class classification을 이용하여 해결한 대표적인 연구는 원점을 대표 부정 학습 예제로 간주하여 one class 분류 문제를 일반적인 two class 분류 문제로 변환시키고 이를 SVM에 적용한 Schölkopf의 연구와 신경망을 사용하여 해결한 Manevitz의 연구가 있다[7].

Schölkopf는 SVM 커널을 통하여 자질을 고차원 공간상에 표현한 다음 그 공간 상에서 원래의 학습문서들을 하나의 범주를 설정하고 원점 하나로만 이루어진 또 하나의 범주를 설정하였다. 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

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

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

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

학습을 통해 최적화된 ω와 ρ를 구한 후 다음과 같은 결정 함수를 통하여 test 문서의 분류를 시행할 수 있다.

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

Manevitz는 [8]에서 원점 하나만을 부정 학습 예제로 사용했던 Schölkopf 의 방법을 개선하여 원점 뿐만 아니라 관심의 대상이 되는 범주와 관련이 적은 긍정 학습 예제들도 부정 학습 예제들로 사용하여 학습을 수행하였다. 이는 학습 예제가 어느 하나의 범주에 편중되는 분류 상황을 해결하기 위한 방법 중에 하나인 over sampling 기법을 사용하여 범주간의 학습 예제 수의 균형을 이루도록 만든 것이다. 여기서 관심의 대상이 되는 범주와 관련이 적은 긍정 학습 예제들은 벡터로 표현했을 때 0의 값을 가지는 요소의 개수가 일정 개수 이상인 긍정 학습 예제들 또는 원점과의 거리가 충분히 가까운 긍정 학습 예제들이다.

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

그 밖에 [9]에서는 [8] 의 방법이 관심의 대상

이 되는 예제와 그렇지 않은 예제들간의 분류 경계를 정하는 데 있어서 부족한 부분이 있음을 지적하고 [8]에 방법에 가상 부정 학습 예제를 추가하여 문제를 해결하였다. [9]에서는 이러한 가상 부정 예제들을 추가함으로써 벡터 공간상에서 긍정 학습 예제들로 이루어지는 공간을 최소화하고 관심의 대상이 아닌 잠재적인 부정 예제들과 긍정 예제와의 더 정확한 분류 경계를 얻을 수 있도록 함으로써 분류 성능의 향상을 보고하였다.

3. 자질 선별

분류 문제를 해결하는 데 있어서 주어진 학습 예제에 나타나는 모든 자질들이 분류에 도움이 되는 것은 아니다. 자질 중에는 오히려 분류를 방해하는 자질들도 존재하고 이러한 자질들을 학습에 사용하였을 경우에는 오히려 분류 성능이 저하될 수도 있다. 따라서 분류에 도움이 되는 자질들을 찾아내서 그 자질들을 가지고 분류를 수행하는 것은 분류 문제 해결에 있어서 중요한 요소 중에 하나이다.

3.1 two class 분류 문제 해결을 위한 자질 선별 기법

일반적인 two class 이상의 분류 문제들을 해결하는데 있어서 현재까지 다양한 자질 선택 기법들이 연구되었고 그러한 방법들을 통한 분류 성능의 향상도 보고되었다[10].

분류 문제에 있어서 자질들은 각 범주 별로 긍정 자질과 부정 자질로 나눌 수 있는데 긍정 자질들은 그 범주를 잘 표현해 줄 수 있는 자질들이고 부정 자질들은 그 범주를 잘 표현하지 못하는 자질들이다. 자질 선택 기법에 있어서 긍정 자질과 부정 자질들을 모두 선택하는 기법을 사용하여 자질을 선별하고 학습을 수행하는 경우에 높은 분류 성능을 얻을 수 있는데 이는 긍정 자질만을 선택하는 자질 선택 기법들(CC, OR, GSS)과 긍정 자질과 부정 자질을 모두 선택하는 자질 선택 기법(CHI-square)을 비교해 보아도 알 수 있다. 긍정 자질만을 선택하는 자질 선택 기법들(CC, OR, GSS)은 하나의 범주에 대하여 자질들의 관련성을 나타낼 때에 자질 선별 기법에 의해 부여된 값이 클수록 그 자질이 범주와 관련이 많은 긍정 자질임을 나타내고 반대로 값이 작을수록 그 자질이 범주와의 관련이 적은 부정 자질임을 나타낸다. 이러한 자질 선택 기법에서는 상위 일부의 자질들을 선택하여 긍정 자질들만을 선택하게 된다. 그러나 긍정 자질과 부정 자질을 모두 선별하는 자질 선택 기법(CHI-square)에서는 자질 선별 기법에 의해 각각의 자질에 부여된 값만 가지고 그 자질이 긍정 자질인지 부정 자질인지 판단할 수 없다. 이러한 자질 선택 기법에서 자질 선택 기법에 의해 부여된 값이 큰 자질들은 긍정 자질일 수도 있고 부정 자질일 수도 있다. 즉, 하나의 범주에 대하여 가장 관련이 있는 자질들뿐만 아니라 가장 관련이 없는 자질들도 긍정 자질과 부정 자질을 모두 선별하는 자질 선택 기법(CHI-square)에서는 높은 값을 부여 받아 선택되는 것이다.

긍정 자질만을 사용하여 학습을 수행하는 경우에는 새로운 문서의 분류 시에 범주와 관련이 없는 문서가 관련 있는 문서로 잘못 판단되는 경우가 많아지는 문제가 발생하게 된다. 긍정 자질만을 선택하는 자질 선택 기법들이 긍정 자질과 부정 자질을 모두 선택하는 자질 선택 기법보다 분류 성능이 낮게 나오는 것도 바로 이러한 면에서 발생하는 문제이다.

또한 긍정 자질과 부정 자질들의 수에 있어서 어느 정도 균형을 이루어 존재할 때 가장 좋은 분류 성능을 얻을 수 있다. 그러나 학습 예제들이 특정 범주로 편중된 경우에는 긍정 자질과 부정 자질을 모두 선별하는 자질 선택 기법을 사용하여도 범주의 긍정 자질과 부정 자질의 수가 균형을 이루지 못하고 한쪽으로 치우치게 된다. 최근에는 이러한 문제를 해결하기 위하여 각각의 범주에서 긍정 자질들을 선택하고 이러한 자질들을 의도적으로 거의 같은 개수로 결합하여 긍정 자질과 부정 자질의 수에 있어서 균형을 맞추어 주고 이를 학습에 사용하기도 한다[10].

3.2 one class 분류 문제 해결을 위한 자질 선별 기법

One class classification에서 우리가 알 수 있는 범주는 관심의 대상이 되는 학습 예제들로 구성된 하나의 범주이다. One class 문제에 기존의 two class 분류 방법이나 자질 선택 방법들을 적용하기 위해서는 관심의 대상이 되는 하나의 범주와 이 범주에 속하지 않는 모든 것들을 다른 하나의 범주로 가정하여 two class 분류 문제로 변환하는 과정이 필요하다. 그러나 관심의 대상이 아닌 모든 것을 하나의 범주로 가정하는 경우 이러한 범주에 속하는 모든 부정 예제를 대변해 줄 수 있는 최상의 부정 학습 예제들을 찾는다는 것은 어려운 일이고 그러한 부정 학습 예제를 잘 표현해 줄 수 있는 자질을 찾는다는 것 또한 어려운 일이다

따라서 기존의 One class 분류 문제 해결에 있어서의 자질 선택은 범주와 관련이 높은 긍정 자질만을 선택하여 학습 시 사용하였고 긍정 자질

을 찾는 기준으로는 자질이 범주의 속하는 문서들 중에서 몇 개의 문서에 나타나는가를 표현하는 문서 빈도수(document frequency)를 주로 사용하였다. 즉, 많은 긍정 학습 예제들에서 발생하는 자질들은 관심의 대상이 되는 범주를 잘 표현할 수 있기 때문에 문서 빈도수가 높은 자질들만을 사용하여 긍정 학습 예제들을 표현하였다.

3.3 one class 분류 성능 향상을 위한 적절한 부정 자질의 선택

One class 분류 문제에서는 관심의 대상이 되는 범주에 속하는 긍정 학습 예제들에서 가장 발생 빈도가 낮은 자질들은 관심 대상 범주와 관련이 거의 없다고 볼 수 있다. One class 분류에서는 관심 대상 예제로 이루어진 관심 대상 범주와 관심 밖의 모든 예제로 이루어진 또 하나의 범주로 나누어 질 수 있기 때문에 관심의 대상이 되는 범주와 관련성이 낮은 자질들은 관심이 대상이 아닌 범주와 관련성이 높다고 할 수 있다.

본 실험에서는 그러한 개념을 이용하여 문서 빈도수를 기준으로 상위 10개의 긍정 자질만을 선택하고 선택된 자질들 상에서 학습을 수행하는 경우와 문서 빈도수 하위 10개의 부정 자질들과 문서 빈도수 상위 10개의 긍정 자질들을 함께 선택하여 선택된 20개의 자질들 상에서 학습을 수행하는 경우로 나누어 실험을 수행하였다.

4. 가상 부정 예제의 사용

관심 밖의 범주의 정의가 어려운 상황에서 가상 부정 예제를 설정하여 학습 시에 사용하는 방법은 one class 분류의 성능을 높일 수 있으며 가상 부정 예제의 설정 방법은 다음과 같다[9].

우선 긍정 학습 예제들을 자질 값을 사용하여 벡터로 표현했을 때 긍정 예제 벡터들은 선택된 자질의 개수와 같은 수의 원소를 가지는 벡터가 되는데 모든 긍정 학습 예제 벡터들에서 각 원소별로 가장 큰 값에 일정한 값을 더한 값으로 구성된 하나의 벡터 A와 A를 구성하는 벡터 원소들의 개수만큼의 또 다른 벡터들이 가상 부정 예제가 된다.

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

: (m = 선택된 자질의 개수)

: (max(i)=범주의 i 번째 자질의 최대 가중치)

여기서 벡터 원소의 개수만큼의 가상 부정 예제들은 하나의 원소만 A의 대응되는 원소의 값을 가지고 나머지 원소는 0의 값을 가지는 벡터이고 이는 다음과 같은 벡터로 표시될 수 있다.

B_i = i 번째 긍정 자질의 가중치는 max(i) + 0.02
나머지 긍정 자질의 가중치는 0) (5)

즉 n개의 자질이 사용된다면 n + 1 개의 가상 부정 예제가 추가되어 학습에 이용된다.

[9]에서는 Manevitz가 [8]에서 제안한 outlier methodology를 그대로 적용하고 가상 부정 예제를 추가하였다. Manevitz의 outlier methodology는 2장의 관련 연구 소개에서 이미 소개한 바 있다.

[9]와 마찬가지로 본 실험에서도 Manevitz의 outlier methodology를 적용하고 난 후에 가상 부정 예제를 추가하는 방법을 사용하였다. 그리고 더 나아가 3.3 절에서 소개한 부정 자질 추가 실험을 수행하였다. 즉, 긍정 자질들만을 선택하여 outlier methodology를 적용하고 가상 부정 예제를 추가하여 학습한 경우와 긍정 자질들과 부정 자질들을 함께 선택하여 outlier methodology를 적용하고 가상 부정 예제를 추가하여 학습한 경우의 성능에 있어서 어떠한 변화를 보이는가를 살펴보고자 하였다. 또한 본 실험에서는 긍정 자질과 부정 자질을 함께 선택한 경우에 가상 부정 예제를 설정함에 있어서 부정 자질들은 없는 것으로 간주하고 긍정 자질 상에서만 설정하는 경우와 가상 부정예제를 긍정 자질과 부정 자질을 모두 존재하는 상태에서 설정하는 두 가지 경우로 나누어 실험을 수행하고 SVM을 통하여 분류 성능을 살펴 보았다.

5. 실험 및 평가

5.1 데이터 구성

본 실험에서는 문서 분류의 데이터로 가장 많이 쓰이고 있는 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 TOPIC 범주 집합 내의 135개의 범주들 중에서 각 범주에 속하는 문서수의 개수를 기준으로 상위 10개의 범주만을 사용하여 이루어지고 있다.

본 실험에서도 다른 분류 실험에서 사용하는 것과 마찬가지로 TOPIC 범주 집합내의 문서 수 상위 10개의 범주를 사용하였으며 실험에 사용된 총 데이터의 수는 표 1에 나타내었다.

이러한 데이터에는 실제로 각 범주를 잘 표현하지 못하는 단어들도 존재하게 된다. 이러한 단어들은 문서 분류의 성능을 감소시키는 결과를 가져올 수도 있으므로 학습 시에 사용하지 않는 것이 좋은 방법인데 이를 위하여 One class 문서 분류에서 자질을 선택하는 기준으로 가장 많이

쓰이는 문서 빈도수(document frequency) 자질 선택 기법을 각 범주 별로 적용하여 이 점수가 높거나 낮은 일정량의 단어를 자질로 선정하였다. 본 실험에서는 문서 빈도수 점수가 높은 10개의 단어만 선택하여 학습을 위한 자질로 사용하거나 문서 빈도수 점수가 높은 10개의 단어와 문서 빈도수 점수가 낮은 10개의 단어를 선택하여 총 20개의 단어를 학습을 위한 자질로 사용하였다

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
Int.	129	372
Ship	95	195
Wheat	48	252
Corn	39	204

각 범주 별로 자질로 선별된 10개 또는 20개의 단어에 대하여 Okapi' s representation 을 사용하여 각 범주에 속하는 모든 예제들을 문서 벡터로 표현하였다.

5.2 기계학습 방법의 선택

다양한 기계 학습 방법이 존재하고 많은 연구에서 여러 기계 학습 방법이 사용되고 있지만 본 실험에서는 최근에 높은 성능을 보이고 있는 기계 학습 방법인 SVM을 사용하였다. SVM 학습을 위해 LIBSVM (ver. 2.71)을 사용하여 기계학습을 수행하였다. SVM에 사용되는 kernel 은 LIBSVM(ver. 2.71)에서 default kernel 인 RBF kernel 을 사용하였으며 다른 매개변수 값도 별다른 설정 없이 LIBSVM(ver. 2.71)에서 기본적으로 설정한 값을 사용하였다.

5.3 실험 결과 및 분석

표 2는 One class classification에서 학습 예제들 상에서 문서 빈도수에 의하여 긍정 자질만을 선별한 후 범주 별로 원점에서 먼 n 개의 긍정 학습 예제만을 긍정 학습 예제로 간주하고 나머지 긍정 학습 예제들은 원점과 함께 부정 학습 예제로 간주하고 가상 부정 예제를 추가하여 학습을 수행한 결과이다. 표에서 비율이 의미하는 것은 학습 시 긍정예제로 사용할 예제 문서의 개수와 전체 긍정 예제 개수의 비율을 나타낸다. 즉, 25%라 함은 전체 긍정 예제 중에 25%에 해당하는 개수의 긍정 예제만을 학습시의 긍정 예제로 간주하고 나머지 75%는 실제로는 긍정 학습 예제이지만 범주와 관련이 적다고 판단하여 학습시 부정 학습 예제로 사용함을 의미한다. 또한 원점에서 멀다고 하는 것의 의미는 학습 예제들의 가중치 표현에 의한 벡터 공간상에서의 원점과의 거리를 의미하는데 본 실험에서는 각 긍정 학습 문서에 대하여 가중치의 합을 기준으로 거리의 개념을 표현하였다.

부정 예제와 긍정 예제의 개수에 있어서 균형을 이루는 경우에 가장 높은 분류 성능을 보인다는 기존의 연구 결과를 표2에서도 볼 수 있다.

표 2. 범주 별 전체 긍정 학습 문서의 일정 비율만을 긍정 학습 문서로 사용하고 가상 부정 예제를 추가한 경우의 분류 성능

(긍정 자질만을 사용)

(okapi' s tf representation, Dimension = 10)

	25%	40%	50%	75%	85%	95%
범주	F_1	F_1	F_1	F_1	F_1	F_1
ship	0.05	0.16	0.2	***0.29***	0.09	0.06
Corn	0	0.17	***0.21***	0.16	0.07	0.06
wheat	0	0.21	***0.51***	0.39	0.16	0.07
Int.	0.42	0.45	***0.45***	0.19	0.17	0.1
Crude	0.32	***0.45***	0.41	0.25	0.19	0.11
Trade	0.1	0.14	0.19	***0.23***	0.17	0.12
Grain	0.26	***0.32***	0.3	0.25	0.17	0.13
Money	0.28	0.47	***0.57***	0.25	0.22	0.15
acq	0.29	0.43	0.49	***0.51***	0.49	0.4
Earn	0.13	0.28	0.37	0.6	***0.69***	0.65
Avg	0.16	0.31	***0.37***	0.26	0.24	0.19

표 3. 범주 별 전체 긍정 학습 문서의 일정 비율만을 긍정 예제 문서로 사용하고 가상 부정 예제를 추가한 경우의 분류 성능

(긍정 자질과 부정 자질을 모두 사용)

(okapi' s tf representation, Dimension = 10)

	25%	40%	50%	75%	85%	95%
범주	F_1	F_1	F_1	F_1	F_1	F_1
ship	0	0.16	0.2	***0.29***	0.1	0.06
Corn	0	0	0	**0.39**	0.39	0.09
wheat	0	0	*0.23*	**0.46**	0.39	0.13
Int.	0.17	0.39	***0.45***	0.22	0.2	0.1
Crude	0.29	***0.46***	***0.46***	0.25	0.2	0.2
Trade	0.06	0.15	0.19	***0.21***	0.17	0.13
Grain	0.1	***0.29***	***0.29***	0.24	0.18	0.14
Money	0.25	0.46	***0.55***	0.25	0.25	0.22
Acq	0.29	0.43	0.47	***0.51***	0	0.
Earn	0	0	0	0.	0.69	***0.72***
Avg	0.12	0.24	***0.29***	0.28	0.26	0.18

표 3에서는 표 2에 사용된 모두 방법이 적용되었지만 표 2에서는 긍정 자질만 사용된 것과 달리 표 3에서는 긍정 자질과 부정 자질을 선택하여 학습 시에 이용한 결과를 보여주고 있다. 표 2와 표 3에서 개수 비율의 각 지점에서의 분

류 성능 비교에서는 부정 자질을 사용하는 경우 성능이 저하된 것처럼 보인다. 하지만 표 5에서 보는 것처럼 각 범주 별 최상의 성능으로 이루어진 구성에서는 긍정 자질과 부정 자질을 함께 사용하는 것이 약간의 분류 성능의 향상을 가져왔다고 할 수 있다.

표 4. 범주 별 전체 긍정 학습 문서의 일정 비율만을 긍정 예제 문서로 사용하고 가상 부정 예제를 추가한 경우의 분류 성능
(긍정 자질과 부정 자질을 모두 사용하지만 긍정 자질에 대해서만 가상 부정 예제 설정)
(okapi' s tf representation, Dimension = 10)

	25%	40%	50%	75%	85%	95%
범주	F_1	F_1	F_1	F_1	F_1	F_1
ship	0	0.16	0.2	***0.29***	0.1	0.06
Corn	0	0	0	**0.39**	0.4	0.1
wheat	0	0	*0.23*	**0.45**	0.4	0.13
Int.	0.17	0.4	***0.45***	0.22	0.2	0.1
Crude	0.29	***0.46***	***0.46***	0.25	0.2	0.2
Trade	0.06	0.15	0.19	***0.2***	0.17	0.13
Grain	0.1	***0.29***	***0.29***	0.24	0.18	0.14
Money	0.25	0.45	***0.55***	0.25	0.22	0.22
acq	0.13	0.42	0.47	***0.5***	0.49	0.4.
Earn	0.11	0.29	0.37	0.63	0.69	***0.72***
Avg	0.12	0.26	0.35	***0.36***	0.31	0.22

표 4에서도 표 3에서의 방법과 마찬가지로 긍정학습 예제들의 벡터 표현을 위해 긍정 자질과 부정 자질을 모두 사용하지만 가상 부정 예제를 설정함에 있어서는 긍정 자질만 존재할 경우와 마찬가지 방법으로 설정을 하고 학습을 수행하여 얻은 학습 모델에서의 분류 성능을 나타낸다.

표 4에서도 표 3과 마찬가지로 개수 비율의 각 지점에서의 분류 성능 비교에서는 성능의 저하를 가져온 것처럼 보인다. 그러나 표 4의 결과 역시 범주 별 최상의 성능으로 이루어진 구성에서는 표 5에서 보면 표 2의 결과보다 높은 성능을 보이고 있음을 알 수 있다.

이러한 결과는 긍정 자질과 부정 자질이 함께 선택하여 그 자질들 상에서 학습 문서를 표현하고 학습하는 경우가 긍정 자질만을 선택하여 그 자질들 상에서 학습 문서를 표현하고 학습하는 경우보다 높은 분류 성능을 얻을 수 있음을 보여준다. 또한 긍정 자질과 부정 자질 그리고 가상 부정 예제를 같이 사용하는 경우에는 가상 부정 예제의 설정 시에 긍정 자질들과 부정 자질들 상에서 가상 부정 예제들을 설정하는 방법과 긍정 자질들만 존재하는 것으로 가정하고 가상 부정 예제를 설정하는 방법을 비교해 볼 수 있는데 F_1 macro 평균에서는 전자와 후자가 거의 비슷한 성능을 보이는 반면에 F_1 micro 평균에서는 후자의 방법이 약간 더 높은 성능을 보이는 것으로 나타났다.

표 5는 각 범주 별로 최고의 성능을 이용하여 시스템의 최대 성능을 얻은 결과이다. 각 범주에서의 최대 성능으로만 구성되었을 경우에 방법의 최대 성능을 얻을 수 있을 것이고 얻어진 최대 성능을 가지고 방법들의 성능을 비교해 볼 수 있다.

각 범주의 최대 성능으로 구성된 성능 비교는 긍정자질과 부정자질 그리고 가상 부정 예제를 사용하되 가상 부정 예제 설정 시 긍정 자질과 부정 자질을 모두 이용하여 가상 부정 예제를 설정하는 방법이 가장 높은 macro average 성능을 보여주었고 긍정 자질과 부정 자질 그리고 가상 부정 예제를 사용하되 긍정 자질만 존재한다고 가정하고 가상 부정 예제를 설정하는 방법이 가장 높은 micro average 성능을 보여 주었다.

표 2, 3, 4, 5 를 종합적으로 볼 때 two class 이상의 분류 방법에서와 마찬가지로 one class classification에서도 긍정 자질만을 사용하여 학습을 수행하고 분류를 시행하는 방법보다는 긍정 자질과 부정 자질을 함께 사용하여 학습을 수행하고 분류를 수행하는 방법이 분류 성능의 향상을 가져올 수 있음을 보여 주었다.

표 5. 각 범주 별 최대 성능에 대한 성능 분석 (F_1-measure)

	macro Avg. F_1	Micro Avg. F_1
긍정자질 가상부정예제	0.4206	0.5514
긍정자질 부정자질 가상부정예제	**0.4326**	0.549
긍정자질 부정자질 긍정자질기반 가상부정	0.4313	**0.555**

6. 결론 및 향후 과제

One-class classification 에서는 주어진 하나의 범주에 대하여 최적의 분류 경계를 정하는 것만이 분류의 성능을 높이는 중요한 요소로 인식되어 왔고 현재까지 그에 대한 많은 연구들이 진행되고 있다. 최적의 분류 경계를 정한다는 것은 관심의 대상이 되는 하나의 범주와 관심의 대상이 아닌 관심 밖의 범주에 대한 분류 경계를 정한다는 것인데 이는 관심의 대상이 되는 긍정 학습 예제들을 모두 포함하면서 범주의 부피는 최소화함으로써 행해질 수 있다.

범주 간의 분류 경계를 정하는 일은 모든 분류 문제에 있어서 중요한 요소이다. 그러나 분류 문제 해결에 있어서 분류 성능을 높이기 위한 또 다른 중요한 요인은 적절한 자질의 선정이다. 그

러나 one class 문제 해결에 있어서는 관심 밖의 범주에 대해서는 학습 예제가 주어지지 않기 때문에 학습에 사용 가능한 정보는 오직 관심의 대상이 되는 범주에 대한 정보뿐이다. 기존의 자질 선택 기법들이 모두 two class 이상의 분류 문제에서 각각의 범주와 관련이 있는 자질들을 선택하는 방법이기 때문에 one class 분류 문제에 직접 적용할 수가 없다. 따라서 one class 분류 문제에 있어서는 비교적 간단하면서도 높은 분류 성능으로 보여왔던 문서 빈도수 자질 선택 방법이 관심의 대상이 되는 범주를 위한 자질들을 선별하는 방법이 되어 왔다.

본 논문에서는 긍정 자질과 부정 자질을 함께 사용하는 것이 분류 성능을 높일 수 있다는 기존 연구를 바탕으로 one class 분류에서의 부정 자질의 사용을 시도하였다. One class 분류에서는 관심의 대상이 되는 학습 예제들에서 나타나는 자질들을 볼 때 관심의 대상이 되는 학습 예제들에서 거의 나타나지 않는 그러한 자질들은 관심 밖의 범주와 더 많은 관련성을 가지고 있을 것이라 생각하여 그러한 자질들을 부정 자질로 정하고 긍정 자질과 함께 사용하여 학습을 수행하고 분류를 시행하여 one class 분류에서도 부정 자질의 사용이 분류 성능의 향상을 가져올 수 있음을 보였다.

기계학습 방법을 사용하여 one class 문제를 해결하는 경우에는 사용하는 기계학습 방법의 특징에 대하여 이해하고 학습에 필요한 최적의 매개 변수 값을 찾아서 기계학습 시 적용하는 것이. 분류의 성능을 높이기 위한 또 다른 중요한 요소가 될 수 있다.

앞으로도 one class 분류 성능의 향상을 위해서는 관심의 대상이 되는 범주를 가장 잘 표현해 줄 수 있는 최적의 긍정 자질들과 분류 시에 관심의 대상이 아닌 예제들을 잘 배제시켜 줄 수 있는 최적의 부정 자질을 찾는 연구가 계속 되어야 할 것이다. 또한 관심의 대상이 되는 범주의 부피는 최소화하고 긍정 학습 예제들은 모두 포함하는 최적의 분류 경계를 찾기 위한 연구 또한 계속 되어야 할 것이며 이러한 최적의 분류 경계 상에서 기존의 문서 분류 방법들을 응용한 새로운 방법들이 고안되어야 할 것이다.

현재 일반적인 분류 문제의 해결에 있어서는 dimension reduction 기술들을 적용하여 다른 차원의 작은 자질 공간 상에서 높은 성능을 보이는 연구도 활발히 진행되고 있다. LSI, LDA 등 현재 많이 쓰이고 있는 dimension reduction 기술들을 직접적인 방법이 아니더라도 one class 분류 문제 해결에 적용하여 one class 분류 성능을 높이려는 시도도 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] N. Japkowicz. "Learning from imbalanced data sets: A comparison of various strategies, Learning from imbalanced data sets," The AAAI Workshop 10-15. Menlo Park, CA: AAAI Press. Technical Report WS-00-05, 2000

[3] Nitesh V. Chawla , Nathalie Japkowicz , Aleksander Kotcz, "Editorial: special issue on learning from imbalanced data sets" , ACM SIGKDD Explorations Newsletter, v.6 n.1, June 2004

[4] 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.

[5] 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.

[6] De Ridder Dick, Tax D.M.J. and Duin Robert P.W. An experimental comparison of one-class classification methods Proceedings ASCI' 98, 4th Annual Conference of the Advanced School for Computing and Imagine

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

[8] 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

[9] 송호진, 강인수, 나승훈, 이종혁, "One-class 문서 분류를 위한 가상 부정 예제의 사용" , KCC 2005 학술 발표회, 2005. 7

[10] Forman, G., "An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3:1289-1305, 2003.

[11] Z. Zheng and R. Srihari, "Optimally combining positive and negative features for text categorization" In Proceedings of the ICML' 03 Wrkshop on Learning from Imbalanced Data Sets, 2003

[12] 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
[13] LIBSVM: a Library for Support Vector Machines,
(http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html)
[14] Z. Zheng, X.Wu and R. Srihari, "Feature selection for text categorization on imbalanced data" SIGKDD Exploration, 6(1):80-89, 2004.
[15] P.Juszczak and R. P. W. Duin. Uncertainty sampling methods for one-class classifiers. In proceedings of the ICML' 03 Workshop on Learning from Imbalanced Data Sets, 2003.
[16] N. Japkowicz. Supervised versus unsupervised binary learning by feedforward neural networks. Machine Learning, 42(1/2):97-122, 2001
[17] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", Kluwer Academic Publishers, 1998