능동적 학습을 위한 복수 문의예제 선정[1]

강재호[+°], 류광렬[++]
[+]동아대학교 지능형통합항만관리연구센터, [++]부산대학교 컴퓨터공학과
{jhkang[+°], krryu[++]}@pusan.ac.kr

Selecting Multiple Query Examples for Active Learning

Jaeho Kang[+°] and Kwang Ryel Ryu[++]
[+]Center for Intelligent and Integrated Port Management Systems, Dong-A University
[++]Department of Computer Engineering, Pusan National University

요 약

능동적 학습(active learning)은 제한된 시간과 인력으로 가능한 정확도가 높은 분류기(classifier)를 생성하기 위하여, 훈련집합에 추가할 예제 즉 문의예제(query example)의 선정과 확장된 훈련집합으로 다시 학습하는 과정을 반복하여 수행한다. 능동적 학습의 핵심은 사용자에게 카테고리(category) 부여를 요청할 문의예제를 선정하는 과정에 있다. 효과적인 문의예제를 선정하기 위하여 다양한 방안들이 제안되었으나, 이들은 매 문의단계마다 하나의 문의예제를 선정하는 경우에 가장 적합하도록 고안되었다. 능동적 학습이 복수의 예제를 사용자에게 문의할 수 있다면, 사용자는 문의예제들을 서로 비교해 가면서 작업할 수 있으므로 카테고리 부여작업을 보다 빠르고 정확하게 수행할 수 있을 것이다. 또한 충분한 인력을 보유한 상황에서는, 카테고리 부여작업을 병렬로 처리할 수 있어 전반적인 학습시간의 단축에 큰 도움이 될 것이다. 하지만, 각 예제의 문의예제로써의 적합 정도를 추정하면 유사한 예제들은 서로 비슷한 수준으로 평가되므로, 기존의 방안들을 복수의 문의예제 선정작업에 그대로 적용할 경우, 유사한 예제들이 문의예제로 동시에 선정되어 능동적 학습의 효율이 저하되는 현상이 나타날 수 있다. 본 논문에서는 특정 예제를 문의예제로 선정하면 이와 일정 수준이상 유사한 예제들은 해당 예제와 함께 문의예제로 선정하지 않음으로써, 이러한 문제점을 극복할 수 있는 방안을 제안한다. 제안한 방안을 문서분류 문제에 적용해 본 결과 기존 문의예제 선정방안으로 복수 문의예제를 선정할 때 발생할 수 있는 문제점을 상당히 완화시킬 있을 뿐 아니라, 복수의 문의예제를 선정하더라도 각 문의단계마다 하나의 예제를 선정하는 경우에 비해 큰 성능의 저하가 없음을 실험적으로 확인하였다.

1. 서론

기계학습의 분류기술을 실제 문제에 적용하기 위해서는 카테고리를 부여한 훈련예제를 상당수 준비하여야 한다. 예제에 카테고리를 부여하는 작업에는 무시할 수 없는 시간과 인력이 요구되며, 응용분야에 따라서는 그 비용이 상당할 수 있다. 능동적 학습은 제한된 시간과 인력으로 가능한 정확도가 높은 분류기를 생성하기 위하여 카테고리를 부여할 예제들을 선별하면서 학습하는 전략이다[1][2][3].

능동적 학습은 사용자가 담변할 수 있는 최대예제개수에 도달할 때까지 학습단계와 문의단계를 반복하여 수행한다. 학습단계에서는 학습 알고리즘을 현재 보유한 훈련집합에 적용하여 분류기를 생성한다. 문의단계에서는 학습단계에서 생성한 분류기를 이용하여 카테고리가 부여되지 않은(unlabeled) 예제들을 분류해보고, 이들 예제 중에서 다음 번 학습에 가장 효과가 높을 것으로 추정되는 예제들을 선정하여 사용자에게 카테고리 부여를 요청한다. 문의단계에서 사용자에 의해 카테고리가 부여된 신규 훈련예제들은 기존의 훈련집합에 추가된다.

능동적 학습의 핵심인 문의예제의 선정작업은 예제의 학습에 대한 효과를 추정할 수 있는 척도를 정의하고, 이를 각 예제에 적용하여 평가한 후 그 평가값이 가장 높은 예제들을 문의예제로 선택하는 방식으로 이루어진다. 초기 연구에서는 이러한 척도로 카테고리 추정의 모호성(불확실성, uncertainty)이 제안되었다[1]. 하지만 모호성에 기반한 방안은 카테고리 정보를 획득하더라도 학습에는 큰 영향을 끼치지 않는 따로 떨어진 예제들(outliers)에 민감하여 능동적 학습의 효율성을 저하시킬 수 있다. 이 후 연구에서는 이러한 문제점을 해결하기 위하여 다른 예제들과의 유사 정도를 반영하여 개별 예제가 위치한 지역의 예제 밀도를 추정하고, 추정된 밀도를 해당 예제의 모호성과 곱하여 적합도를 평가하는 방안이 제안되었다[2]. Roy와 McCallum은 개별 예제마다 카테고리가 부여되었을 때를 가정하여 학습을 수행한 후, 생성되는 분류기의 예상 오류가 최소인 예제를 선정하는 방안을 제안하였다[3].

일반적으로 능동적 학습은 매 문의단계마다 하나의 문의예제를 선정하는데, 이는 기계학습을 적용하는 상황에 따라 비효율적일 수 있다. 예를 들어 여러 예제를 동시에 사용자에게 문의하면, 사용자는 예제들을 서로 비교해 볼 수 있으므로 작업을 보다 빠르고 정확하게 수행할 수 있을 것이다. 또한, 카테고리를 부여할 수 있는 인력을 충분히 보유한 상황이라면, 동시에 많은 수의 문의예제를 선정하여 카테고리 부여작업을 병렬로 처리하는 것이 전체 학습에 소요되는 시간을 줄이는데 좀더 유리할 것이다.

기존의 문의예제 선정방안들은 매 문의단계마다 하나의 예제를 선정하는 경우에 가장 적합하도록 고안되어 있어, 이러한 복수 문의예제 선정작업에는 효과적이지 못할 수 있다. 매우 유사한 예제들을 함께 문의예제로 선정하면 능동적 학습의

[1] 국가지정연구실사업 (과제명: 언어 중심의 지능적 정보처리를 위한 단계적 우리말 분석기술의 개발(M10203000028-02J0000-01510))의 지원을 받아 이루어진 것임.

효율성이 저하될 수 있을 것이라는 것은 쉽게 예측할 수 있다. 모호성과 밀도를 함께 고려하는 방안[2]의 경우, 예제들이 밀집된 특정 영역에서 다수의 문의예제가 동시에 선정될 수 있는 가능성이 높다. 이는 유사한 예제들끼리는 문의예제로써의 적합도 평가결과도 비슷하기 때문이다. 추정오류 최소화 척도[3]를 이용한다면 유사한 예제들이 문의예제로 동시에 선정되는 문제점은 해결 가능할 것이나, 조합 가능한 모든 문의예제 집합에 대하여 분류기를 생성하고 평가하는 과정이 요구되기 때문에 현실적으로 적용하기는 어렵다[2]. 따라서 이러한 복수의 문의예제 선정을 위해서는 간단하면서도 보다 현실적인 접근 방안이 요구된다. 본 논문에서는 각 문의단계에서 특정 문의예제를 선정하면, 이와 일정수준 이상 유사한 예제들은 해당 문의단계에서 동시에 선정될 수 있는 대상에서 제외함으로써 이러한 문제점을 해결하고자 한다.

본 논문의 구성은 먼저 2장에서 능동적 학습의 문의예제 선정을 위한 척도와 복수 문의예제 선정 시 발생할 수 있는 문제점, 그리고 본 제안방안을 자세히 설명한다. 이어지는 3장에서는 제안한 방안을 문서분류 문제에 적용한 실험결과를 분석하고, 4장에서 관련 연구를 소개한다. 마지막 5장에서는 결론 및 향후 연구과제를 제시한다.

2. 예제간 유사도를 고려한 복수 문의예제 선정 방안

본 장에서는 *k*-NN (*k*-nearest neighbor) 알고리즘을 문서분류 문제에 적용하는 경우를 가정하여 문의예제 선정을 위한 모호성과 예제 밀도를 정의한다. 또한 기존 방안으로 복수의 문의예제를 선정하는 경우 발생할 수 있는 문제점과 본 논문의 제안 방안을 소개하고자 한다. 문서는 예제를 획득하는데 소요되는 비용에 비해 예제에 카테고리를 부여하는 데 필요한 수작업 비용이 높은 도메인(domain)의 하나이며, *k*-NN은 문서를 높은 정확도로 분류할 수 있는 기계학습 기법의 하나이다[4].

k-NN은 문제예제와 가장 유사한 *k*개의 훈련예제를 파악하고, 이들 예제의 카테고리 별로 문제예제와의 유사 정도를 가중치로 삼아 합을 구한 후, 그 값이 가장 높은 카테고리를 문제예제의 카테고리로 추정하는 분류기법이다[3]. *k*-NN을 문서분류 문제에 적용하는 경우, 개별 문서는 *tf*×*idf* 공간상의 벡터로 표현할 수 있으며, 문서간의 유사성은 일반적으로 코사인 유사도(cosine similarity)로 측정한다[5]. 코사인 유사도는 두 벡터 간의 각도의 코사인 값으로 두 문서의 단어 등장 비율이 완전히 일치할 경우에는 1로, 두 문서에 공통으로 등장하는 단어가 하나도 없는 경우에는 0으로 계산된다.

k-NN에 의하여 *j*번째 예제 e_j의 추정되는 카테고리들의 가중치 분포를 W_j라 한다면, W_j는 *c*개의 카테고리가 있는 경우, $<w_{j1}, w_{j2}, \ldots, w_{jc}>$로 표현할 수 있다. e_j는 $p = \arg\max_i w_{ji}$인 카테고리 *p*로 분류된다. 만일 해당 예제에 대한 카테고리 추정이 모호하다면 카테고리 *p*에 대한 가중치 w_{jp}와 나머지 카테고리에 대한 가중치의 합 Σw_{ji}, $(p \neq i)$ 간의 차이는 작을 것이며, 이와 반대의 경우는 그 차이가 클 것이다[4]. 따라서, *k*-NN에서 e_j의 카테고리 추정의 모호한 정도 u_j는 수식 (1)과 같이 표현할 수 있다. 본 논문에서는 u_j를 0과 1사이로 정규화하여 사용하였다.

$$u_j = \left(\sum_{i,i\neq p} w_{ji}\right) - w_{jp}, p = \arg\max_i w_{ji} \quad (1)$$

문의예제를 선정할 때 각 예제가 위치한 지역의 예제 밀도를 반영하기 위해서는 이를 추정할 수 있는 척도를 정의하여야 한다. 예제 밀도는 특정예제에 카테고리가 부여되었을 때, 학습에 미치는 영향 정도를 간접적으로 예상하는 방법이다. 특정 예제와 유사한 예제들이 많은 경우 해당 예제가 위치한 지역의 밀도가 높으며, 반대의 경우에는 밀도가 낮다고 할 수 있다. 따라서, e_j와 모든 예제들간의 평균 유사도를 이용하여 예제 밀도를 추정할 수 있다. 전체 예제의 집합을 *D*라고 할 때, e_j가 위치한 지역의 밀도 d_j는 수식 (2)와 같이 추정할 수 있다. 수식에서 $\mathrm{sim}(e_j, e_i)$는 e_j와 e_i간의 코사인 유사도 값을 의미한다.

$$d_j = \frac{\sum_{e_i \in D} \mathrm{sim}(e_j, e_i)}{|D|} \quad (2)$$

각 예제의 문의예제로써의 적합도는 모호성과 밀도를 곱한 값 $u_j \times d_j$로 평가되며, 이 값이 높은 순서대로 필요한 수만큼 문의예제를 선정한다. 하지만 유사한 예제들끼리는 모호성과 밀도 모두 비슷하여 평가값 또한 큰 차이가 없게 되므로 복수의 문의예제 선정 시 이들이 동시에 선택될 가능성이 높다. 이러한 문제점을 해결하기 위하여 본 논문에서는 하나의 예제 e_r이 가장 높은 적합도를 가지고 있어 이를 문의예제로 선정하고자 한다면, e_r과 가장 유사한 일부의 예제들은 문의예제 선정대상에서 제외하는 방안을 제안한다. e_r과 가장 유사한 예제들 중에서 동시에 문의예제로 선정하지 않아야 하는 예제의 집합 N_r의 크기는 다음과 같은 가정에 기반하여 설정하였다. 만일 전체 *n*개의 예제 중에 *t*개의 훈련예제가 있다면, *t*개의 훈련예제들이 전체 *n*개의 예제들을 대표한다고 할 수 있다. 각 훈련예제는 자신을 포함하여 평균 *n*/*t*개의 예제를 대표하게 되므로, N_r의 크기는 $\lceil n/t \rceil - 1$로 결정할 수 있다. 따라서, *t*개의 훈련예제가 이미 훈련집합에 있는 상황에서 능동적 학습이 *q*개의 문의예제를 선정하고자 하는 경우 N_r의 크기는 $\lceil n/(t+q) \rceil - 1$이다[5].

3. 실험 결과

이상에서 제안한 방안의 효과를 확인하기 newsgroups-20 말뭉치[6]를 활용하여 문서분류 실험을 수행하였다. Newsgroups-20 말뭉치는 문서분류 연구에 자주 활용되는 20개의 유즈넷 뉴스그룹에 올려진 약 20,000건의 기사 모음이다. 이 중에서 분류의 난이도에 따른 본 제안방안의 효과를 확인하기 위하여 [7]에서 사용한 바와 같이, 주제가 다른 세 개의 뉴스그룹 (alt.atheism, rec.sport.baseball, sci.space)에 올려진 기사들로 Different-3 데이터를 생성하였고, 주제가 유사한 세 개의 뉴스그룹 (comp.graphics, comp.os.ms-windows, comp.windows.x)에 올려진 기사들로 Same-3 데이터를 생성하였다. 각 뉴스그룹에 올려진 기사의 수는 약 1,000건으로 매우 균일하다. 선정된 기

---

[2] 카테고리 수가 *c*개, 카테고리가 부여되지 않은 예제가 *n*개인 상황에서 *m*개의 문의예제를 선정하고자 한다면, 조합 가능한 문의예제집합의 수는 총 $_nC_m$이다. 여기에 각 문의예제집합 별로 발생할 수 있는 카테고리의 조합 수 c^m를 고려하여야 하므로, 전체적으로 $_nC_m \times c^m$번 분류기를 생성하고 평가하여야 한다.

[3] 본 논문에서는 *k*-NN 알고리즘을 문서분류 문제에 적용하므로 *k*-NN이 예제간의 유사성에 기반한다고 설명하였다. 예제간의 거리개념으로 *k*-NN 기법을 설명하는 것이 보다 일반적이다.

[4] 이외에도 다양한 방식으로 모호성을 추정할 수 있으며, 본 논문에서는 실험한 방식 중에서 가장 효과가 높았던 척도를 소개하고 있다.

[5] 여기서는 전체 예제와 훈련예제의 분포가 동일(또는 유사)하다는 가정에 기반하여 N_r의 크기를 결정하였다. 하지만 능동적 학습이 어느 정도 진행되면 카테고리를 결정하기 모호한 예제들이 많은 지역(예를 들어 두 카테고리간의 경계)에서 상대적으로 많은 훈련예제들을 선정하는 것이 학습의 효율 측면에서 바람직하므로 $|N_r|$을 $\lceil n/(t+q) \rceil - 1$보다 작은 수로 점차 줄여가는 것이 나을 수 있다.

사들은 제목을 제외한 유즈넷 헤더를 모두 제거하고, 불용어 처리(stop word removal)와 표준형 변환(stemming)을 거쳐 실험에 사용할 데이터로 구축하였다. *k*-NN의 *k*는 실험에 사용할 훈련예제의 수(10~50)와 카테고리 수(3)를 감안하여 5로 고정하였다. 각 실험은 10분할 상호검증 (10-fold cross-validation) 방식으로 10회 실시 후 정확도의 평균값을 취하였다.

능동적 학습은 임의[6]로 선정된 10개의 훈련예제로부터 시작하여 최대 50개까지 훈련예제를 사용할 수 있다고 가정하였다. 매 문의단계마다 하나씩 예제를 선정하는 경우(Q1)와 5개 또는 10개씩 동시에 선정하는 경우를 실험하였다. 복수의 문의예제를 선정하는 경우 문의예제로써의 적합도만을 이용하는 방안(Q5, Q10)과, 본 논문에서 제안하는 유사한 예제는 함께 문의하지 않는 방안(LQ5, LQ10)을 구현하여 비교 실험하였다.

그림 1과 그림 2는 각각 Different-3와 Same-3 데이터를 이용하여 실험한 능동적 학습의 성능 그래프이다. 가로축은 학습에 사용한 훈련예제의 수를 의미하며 세로축은 생성된 분류기의 정확도를 나타낸다. 그림에서 Random은 능동적 학습의 효과를 확인하기 위하여 훈련예제를 임의로 선정하여 학습한 경우이다. 각 그림에서 왼쪽의 그래프에는 문의예제로의 적합도만을 이용하여 복수의 문의예제를 선정한 경우(Q5, Q10)의 성능이 표시되어 있으며, 오른쪽은 본 논문에서 제안한 방안을 이용한 경우(LQ5, LQ10)의 성능이 포함되어 있다.

Q5와 Q10은 유사한 예제들이 함께 문의예제를 선정될 가능성이 높아, Q1에 비해 특히 학습 초기에 정확도가 상당히 낮은 것을 볼 수 있다. 이는 기존 문의예제 선정방안을 그대로 복수문의 선정에 이용하는 경우 능동적 학습의 효율이 저하될 수 있음을 보여주는 것이다. 이에 비해 LQ5와 LQ10은 이러한 문제점이 해소되어 Q1에 비해 큰 성능의 격차를 보이고 있지 않다. 특히 난이도가 높은 Same-3 데이터를 이용한 실험에서도 LQ5의 경우 Q1과 거의 동등한 수준의 성능을 보이고 있는데, 이러한 결과는 하나씩 예제를 문의할 경우 달성할 수 있는 능동적 학습의 성능 희생을 최소화하면서도 복수의 문의예제를 선정할 수 있음을 보여준다 할 수 있다[7].

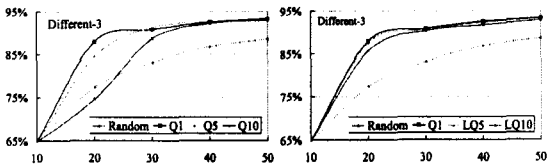

그림 1. 복수 문의를 허용하는 능동적 학습 (Different-3)

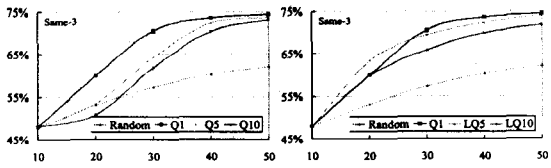

그림 2. 복수 문의를 허용하는 능동적 학습 (Same-3)

[6] 각 카테고리 별로 최소 1개씩의 훈련예제는 보장하였다.

[7] 결과적으로 50개의 훈련예제를 사용한 후에는 어떠한 방법이나 비슷한 성능을 달성한 것이 아닌가 하는 의문이 있을 수 있다. 하지만, 능동적 학습의 완료 시점은 문제를 해결하는 자원의 상황에 따라 다르다. 예를 들어 30개의 훈련예제만을 사용한 후 종료될 수 있다.

4. 관련 연구

능동적 학습에서 복수의 문의예제를 선정하는 방안에 대한 연구가 수행된 것은 최근의 일이다. Brinker[8]는 support vector machine(SVM)을 이용한 능동적 학습에서 다양성과 모호성을 동시에 고려하는 방안을 제안한 바 있다. SVM은 그 정확도는 높으나 학습에 소요되는 시간이 상당한 분류기법이다. 이 연구에서 문의예제를 복수로 선정함으로써 능동적 학습에 소요되는 시간을 줄일 수 있고, 동일한 수의 복수 문의예제를 선정할 경우에는 다양성과 모호성을 적절히 고려함으로써 보다 학습을 효율적으로 수행할 수 있음을 보였다. 하지만 각 문의단계마다 하나씩 문의예제를 선정하는 경우와 그 성능을 비교하지 않았으며, 제안된 방안은 SVM을 분류기로 사용할 경우에 적합한 방안이다.

5. 결론 및 향후 연구

능동적 학습을 적용할 때 그 전제조건인 제한된 시간과 인력을 보다 효율적으로 활용하기 위해서는 하나 이상의 문의예제를 효과적으로 선정할 수 있어야 한다. 기존에 제안된 문의예제 선정방안들은 복수의 문의예제를 선정하는 상황을 직접적으로 고려하고 있지 않기 때문에, 유사한 예제들이 동시에 선정되어 학습의 효율이 저하될 수 있다. 본 논문에서는 각 예제와 일정 수준 이상 유사한 예제들은 동시에 문의예제로 선정될 수 없도록 함으로써, 보다 다양한 예제로 문의예제집합을 구성하는 방안을 제안하였다. 본 접근방안을 문서분류 문제에 적용하여 실험한 결과 기존 방안의 직접적인 적용 시 발생할 수 있는 문제점을 완화시킬 수 있을 뿐 아니라, 복수의 문의예제를 선정하더라도 문의예제를 하나씩 선정하는 경우에 비해 능동적 학습의 효율이 크게 저하시키지 않아 복수 문의예제 선정의 현실성이 있음을 실험적으로 확인하였다.

향후 본 제안방안을 문서 분류문제에 효과적인 또 다른 학습기법의 하나인 나이브 베이즈(naive Bayes)를 이용하는 능동적 학습에도 적용할 수 있는 방안을 연구할 필요가 있다.

6. 참고문헌

[1] Lewis D., and Gale, W., A sequential algorithm for training text classifiers. *In Proc. of 17th ACM-SIGIR Conference*, 3-12, 1994.

[2] McCallum, A., and Nigam, K., Employing EM in pool-based active learning for text classification. *In Proc. of 15th International Conference on Machine Learning*, 359-367, 1998

[3] Roy N. and McCallum, A., Toward optimal active learning through sampling estimation of error reduction. *In Proc. of 18th International Conference on Machine Learning*, 441-448, 2001

[4] Yang, Y., An evaluation of statistical approaches to text categorization, *Journal of Information Retrieval*, vol. 1, nos. 1/2, 67-88, 1999.

[5] Yates, B. and Neto, R., *Modern Information Retrieval*, Addison-Wesley, 1999

[6] *UCI Knowledge Discovery in Databases Archive*: http://kdd.ics.uci.edu/

[7] Basu, S., Banerjee, A., and Mooney, R., Semi-supervised clustering by seeding, *In Proc. of 15th International Conference on Machine Learning*, 19-26, 2002

[8] Brinker, K., Incorporating Diversity in Active Learning with Support Vector Machines, *In Proc. of 20th International Conference on Machine Learning*, 2003