한국지능정보시스템학회논문지 제11권 제1호 2005년 6월(pp.169~189)

능동적 학습을 위한 군집화 기반의 다양한 복수 문의 예제 선정 방법

강재호
부산대학교 컴퓨터공학과 Post-Doc.
(jhkang@pusan.ac.kr)

류광렬
부산대학교 컴퓨터공학과 교수
(krryu@pusan.ac.kr)

권혁철
부산대학교 컴퓨터공학과 교수
(hckwon@pusan.ac.kr)

능동적 학습은 훈련 예제의 수가 제한적일 때 효율적으로 분류기를 생성할 수 있는 학습 방법이다. 이 방법에서는 분류하기 모호한 예제에 카테고리를 부여하기 위한 문의 과정과 이렇게 얻어진 예제들을 추가해 가면서 분류기를 재생성하는 과정을 반복적으로 수행한다. 특히 온라인 환경에서는 반복적으로 예제에 카테고리를 부여해야 하는 사용자의 부담을 줄이기 위해 문의 예제의 총 수뿐만이 아니라 문의 횟수 또한 최소화하여야 한다. 예제 수와 문의 횟수를 줄이면서도 좋은 분류기를 생성하기 위해서는 매 문의 시 사용자에게 다양하면서도 대표성이 높은 복수의 모호한 예제들을 선정하여 제시하는 것이 좋다. 본 논문에서는 다양하면서, 대표적이며, 또한 모호성이 높아 능동적 학습에 효과적인 복수의 문의 예제를 선별하기 위하여 군집화 기법을 활용하는 방안을 제안한다. 문서 분류 문제를 대상으로 본 제안 방안을 실험한 결과 모호성만을 기준으로 복수의 문의 예제를 선정하는 방법보다 우수한 분류기를 생성할 수 있음을 확인하였다.

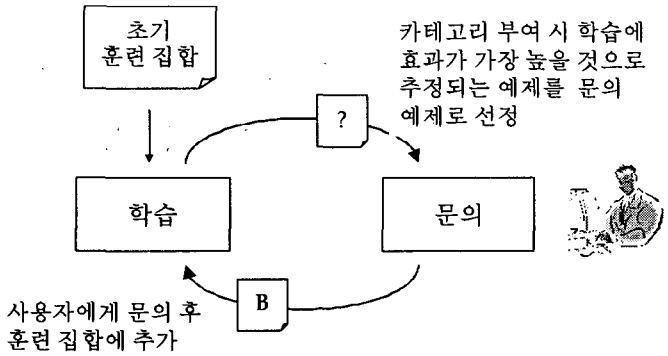
논문접수일 : 2005년 5월 게재확정일 : 2005년 6월 교신저자 : 강재호

1. 서론

기계 학습의 분류(classification) 기술을 현실 문제에 성공적으로 적용하기 위해서는 카테고리(category, class)가 부여된(labeled) 예제, 즉 훈련 예제(training example)를 상당수 미리 준비하여야 한다. 예제에 카테고리를 부여하는 작업에는 무시하기 어려운 시간과 인력이 소요되며, 특히 문서 분류의 경우에는 사용자가 직접 읽어 보아야 개별 문서의 카테고리를 정확하게 부여할 수 있으므로 훈련 예제를 생성하는데 소요되는 비용이 상당하다. 이렇게 과도할 수 있는 훈련 예제 생성 비용을 효과적으로 줄이기 위하여 능동적 학습(active learning) 기법이 제안되었다(Lewis and Gale, 1994). 능동적 학습은 훈련 예제로 사용할 수 있는 예제의 수가 제한되어 있는 상황에서 최대한 정확도가 높은 분류기를 생성하기 위하여, 학습 알고리즘이 스스로 카테고리 부여를 요청할 예제들을 선별하면서 점진적으로 학습한다.

능동적 학습 기법은 [그림 1]과 같이 학습 단계와 문의(query) 단계를 반복하여 수행하는 구조로 이루어져 있다. 학습 단계에서는 현재까지 수집된 훈련 예제들의 집합에 학습 알고리즘을 적용하여 예제들의 카테고리를 추정할 수 있는 분류기

* 이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2004-002-D00350)



[그림 1] 능동적 학습의 흐름

(classifier)를 생성한다. 문의 단계에서는 학습 단계에서 생성한 분류기를 이용하여 아직 카테고리가 부여되지 않은(unlabeled) 예제들을 분류해보고, 이들 중에서 다음 번 학습에 가장 효과가 높을 것으로 추정되는 예제들을 골라, 문의 예제로 선정한 후 이들 예제에 대하여 사용자에게 카테고리 정보를 문의한다. 사용자에 의하여 카테고리가 부여된 문의 예제들(그림에서는 B라고 카테고리가 부여된 문서)은 기존의 훈련 집합에 추가되어 다음 학습 단계에서 새로운 분류기를 생성하는데 활용된다. 능동적 학습은 이러한 학습 단계와 문의 단계를 사용자가 답변할 수 있는 최대 예제 수에 도달할 때까지 반복하여 수행함으로써 학습 단계에서 생성하는 분류기의 정확도를 점진적으로 향상시킨다.

능동적 학습의 기존 연구에서 제안한 여러 가지 문의 예제 선정 방안들은 대부분 매 문의 단계마다 하나의 문의 예제를 선정하는 상황에 최적화되어 있다. 하지만 능동적 학습을 실제 문제에 적용하는 상황에서는 문의 단계마다 하나가 아닌 복수의 예제를 선정하여 사용자에게 문의하는 것이 보다 바람직할 수 있다. 예를 들어 온라인으로 뉴스를 제공하는 업체에서 사용자가 관심 있어 하는 분야를 파악하기 위하여 능동적 학습을 적용한다고 가정하자. 보유하고 있는 많은 기사들 중에서 50건 정도의 기사에 대하여 사용자가 관심 여부를 표현하여야 해당 사용자가 만족할 만한 정확도로 자동 분류가 가능하다면, 한 번에 기사 하나씩 50회에 걸쳐 묻기 보다는 5건 내지 10건의 기사를 한꺼번에 제시하고, 그 중에서 관심 있는 기사들을 선택하게 하는 과정을 10회 내지 5회 반복하는 것이 사용자에게 훨씬 부담이 덜 할 것이다. 또한 매 문의 단계마다 여러 건의 기사를 함께 제시하면, 사용자는 제시된 기사들의 내용을 서로 비교해 볼 수 있어 제시된 기사에 대한 관심 여부, 즉 카테고리를 보다 빠르고 정확하게 부여할 수 있을 것이다.

본 논문에서 제안하는 군집화(clustering) 기반 복수 문의 예제 선정(batch query selection) 방안은 '대표적이며 카테고리를 추정하기 어려운 다양한 예제들'이 능동적 학습의 성능 향상에 보다 효과적일 것이라는 가정에 기반을 둔다. 이러한 조건을 만족하는 복수 문의 예제를 선정하기 위하여 먼저 보유하고 있지만 아직 카테고리가 부여되지 않은 예제들을 대상으로 k-means 군집화 기법을 적용하여 유사한 예제들끼리 모은다. 이 때 각 예

제의 카테고리 추정이 얼마나 어려운지 그 정도를 추정한 모호성(ambiguity, uncertainty)을 가중치로 반영하여 군집화를 수행하면, 생성된 각 군집(cluster)은 모호성이 높은 예제들을 중심으로 유사한 예제들이 모인 집합이 될 것이다.[1] 서로 다른 군집에 속하는 예제들끼리는 상대적으로 상이하므로, 군집 별로 해당 군집에 소속된 예제들을 대표할 수 있으면서 모호성이 높은 예제 하나를 문의 예제로 선정한다면, '대표적이며 카테고리를 추정하기 어려운 다양한 예제들로 이루어진 집합'을 구성할 수 있다.

하나의 군집을 대표하는 예제로는 k-means 군집화 결과로 생성된 각 군집의 중심점이 가장 적합할 것이다. 하지만, k-means에서 군집의 중심점은 실제 존재하는 예제가 아니므로, 이들을 사용자에게 제시하여 직접 카테고리를 부여 받을 수 없다. 따라서 본 논문에서는 각 군집의 예제들 중에서 자신이 소속된 군집의 중심점과 가장 비슷한 동시에 모호성 또한 높은 예제를 해당 군집의 대표 예제(representative example)로 삼고, 이들 대표 예제로 문의 집합을 구성하는 방안을 제안한다. 문서 데이터를 이용한 실험에서 이와 같은 방식으로 구성한 문의 집합이 능동적 학습의 성능 향상에 효과적임을 확인할 수 있었다. 더 나아가 사용자에게 카테고리를 부여 받은 대표 예제는 자신이 소속된 군집의 중심점과 가장 유사하므로, 군집의 중심점을 예제화한 후 이 예제에 대표 예제의 카테고리를 부여함으로써, 사용자에게 추가의 문의 없이 군집을 효과적으로 표현할 수 있는 훈련 예제를 얻는 방안을 함께 제안한다. 군집의 중심점을 훈련 예제로 사용하는 경우 본 논문에서는 이를 모델 예제(model example)로 칭하며, 모델 예제를 대표 예제와 함께 훈련 예제로 활용한 실험에서 추가의 성능 향상을 얻을 수 있었다.

본 논문의 구성은 먼저 2장에서 능동적 학습과 관련한 기존 연구들을 소개하고 이어지는 3장에서는 군집화 기반 복수 문의 예제 선정 방안에 대하여 보다 구체적으로 설명한다. 4장에서는 제안한 방안을 적용하여 실험한 결과를 정리하여 분석하고, 마지막 5장에서 결론과 향후 연구 방향으로 매듭을 짓는다.

2. 관련 연구

능동적 학습과 관련한 기존 연구들은 문의할 예제들을 효과적으로 선정할 수 있는 방안에 대해 주로 탐구해 왔다. 능동적 학습 과정의 핵심인 '문의 예제 선정 작업'은 카테고리가 부여되지 않은 각 예제에 대하여 문의 예제로 적합한 정도를 평가하고, 그 평가 값이 가장 높은 예제들을 문의 예제로 선정하는 방식으로 이루어진다. 각 예제의 문의 예제로의 적합 정도를 평가하기 위해서는 이를 추정할 수 있는 척도가 필요하다.

능동적 학습의 초기 연구에서는 이러한 척도로 카테고리 추정이 모호한 정도를 사용하는 방안이 제안되었다(Lewis and Gale, 1994). 카테고리 추정의 모호성이란 분류기가 해당 예제를 어떠한 카테고리로 분류할 지 판단하기 어려운 정도이다. Lewis의 연구에서는 문서를 정례(positive example)와 반례(negative example) 두 가지 카테고리로 구분하는 분류 문제에 나이브 베이즈(Naive Bayes) 학습 알고리즘으로 능동적 학습을 적용하였는데, 학습 단계에서 생성한 분류기로 정례에 속할 확률이 0.5에 가장 가까운, 즉 정례인지 반례인지 추정하기 가장 애매한 예제들을 문의 예제로

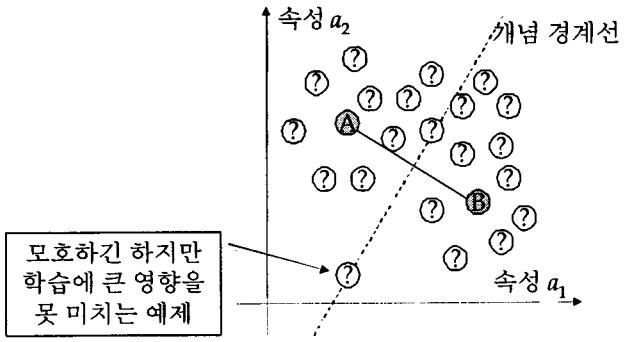
1) 본논문에서는 군집화 시 사용자에게 동시에 문의하고자 하는 예제의 만큼군을 생성한다.

선정하였다.

모호함에만 기반을 둔 문의 예제 선정 방식은 카테고리 정보를 획득하더라도 학습에는 크게 긍정적인 영향을 끼치지 못하는 따로 떨어진 예제들(outliers)을 문의 예제로 선정할 가능성이 높아, 능동적 학습의 효율을 저하시킬 수 있다는 단점이 있다(McCallum and Nigam, 1998). [그림 2]는 이렇게 따로 떨어진 예제의 예를 보이고 있다. 그림에서 현재 학습한 분류기의 개념 경계선 상에 위치한 따로 떨어진 예제는 카테고리 추정이 매우 모호하긴 하지만, 자신과 유사한 예제가 거의 없어 사용자에게 문의하여 카테고리를 부여받아 훈련 예제로 사용되더라도 새로운 예제를 분류하는데 크게 도움이 되지 못한다. 이 후 연구에서는 이러한 문제점을 해결하기 위하여 예제의 대표성을 모호성과 함께 고려하는 방안이 제안되었다(McCallum and Nigam, 1998). 예제의 대표성은 다양한 방식으로 추정이 가능하며, 간단하게는 예제가 대표적일수록 해당 예제와 유사한 예제들이 많을 것이라 예상할 수 있다. 이러한 가정 하에 다른 예제들과의 평균적인 유사 정도나, 자신과 k번째 가까운 예제와의 유사 정도 또는 거리를 이용하여 해당 예제의 대표성을 추정할 수 있다.

또 다른 문의 예제 선정 척도의 하나인 위원회 기반 문의 예제 선정 방법(committee-based sampling)은 학습 단계에서 하나가 아닌 복수의 분류기를 생성하고 이들 간의 의견이 가장 불일치한 예제를 문의 예제로 선정하는 기법이다 (Seung et al. 1992)의 연구에서는 신문 기사와 같이 예제들이 지속적으로 생성되어 제공되는 상황에서 위원회 기반 문의 예제 선정 방법을 적용한 능동적 학습을 제안하였다. 이 방법은 하나의 새로운 예제가 제시되면, 현재 보유한 훈련 집합과 모순이 없는(consistent) 두 개의 분류기를 임의로 생성하고, 새로운 예제에 대한 이 두 분류기간의 의견이 불일치한 경우 해당 예제를 사용자에게 문의한다.

(Abe and Mamitsuka, 1998)의 연구에서는 능동적 학습 시 동일한 훈련 집합을 기반으로 복수의 서로 다른 분류기를 생성하기 위하여 bagging (Breiman, 1994)과 boosting(Freud and Schapire, 1995) 기법을 적용하였다. 이 방법은 bagging이나 boosting 기법으로 생성된 복수의 분류기 간의 의견이 가장 불일치한 예제를 문의 예제로 선정한다. 위원회 기반 문의 예제 선정 방법은 결정 나무(decision tree)나 규칙 집합(rule set)과 같이 개별 예제에 대한 카테고리 추정의 모호성을 직접적으로 표현하기 어려운 학습 알고리즘으로도 능동적



[그림 2] 따로 떨어진 예제의 예

학습을 적용할 수 있게 한다.

(Muslea et al. 2000)의 연구에서는 (Blum and Mitchell, 1998)의 연구에서 제안된 협동 훈련(co-training) 기법을 문의 예제 선정에 응용한 협동 검사(co-testing) 기법을 제안하였다. 협동 검사는 예제들을 묘사하는 속성의 집합을 서로 겹치지 않는 두 개의 부분 집합(view)으로 나누고, 학습 단계에서는 각 부분 집합 별로 독자적으로 두 개의 분류기를 생성한다. 문의 예제로는 두 분류기 간의 카테고리 예측이 가장 불일치한 예제가 선정된다. 이후 연구에서 이들은 능동적 학습의 학습 단계에 Co-EM(Nigam and Ghani, 2000)을 활용하고 문의 단계에서는 협동 검사 기법을 적용한 Co-EMT를 제안하였다(Muslea et al. 2002). 이들은 제안한 Co-EMT가 다양한 실험 환경에서 협동 검사를 포함한 여러 준감독(semi-supervised) 학습 기법에 비해 안정적(robust)인 성능을 발휘함을 보였다.

휴리스틱에 기반을 둔 접근 방법 이외에 최적의 문의 예제를 선정하고자 한 보다 이론적인 접근으로는 (Cohn et al. 1996; Roy and McCallum, 2001)이 있다. (Roy and McCallum, 2001)의 연구에서는 개별 예제에 카테고리가 부여되었을 때를 가정하여 학습을 수행하고, 이 때 예상되는 전체 오류가 가장 낮은 예제를 문의 예제로 선정하는 방안을 제시하였다. 이 방법은 문의 예제 후보가 될 수 있는 예제 별로 사용자가 부여 가능한 카테고리 수만큼 학습을 수행하여야 하므로, 점진적(incremental)인 학습이 가능한 알고리즘을 사용하더라도 문의 예제 선정에 소요되는 비용이 매우 높다는 단점이 있다. 카테고리가 부여되지 않은 예제의 수를 n, 카테고리의 수를 c라 하면, 이 방법으로 하나의 문의 예제를 선정하기 위해서는 n^c 횟수만큼 학습을 수행하여 각 예제들의 적합 정도를 평가하여야 한다.

능동적 학습과 관련한 기존 연구들은 일반적으로 임의로 선정한 초기 훈련 예제들로 초기 학습을 수행한 후 매 문의 단계마다 하나의 문의 예제를 효과적으로 선정하는 방안에 주된 초점을 맞추어 왔다. (Kang et al. 2004)의 연구에서는 k-means 군집화 기법을 이용하여 복수의 초기 훈련 예제들을 주의 깊게 선정함으로써 능동적 학습을 적용한 문서 분류의 성능을 향상 시킬 수 있음을 보인 바 있다. 본 연구는 초기 훈련 집합 선정에만 군집화 기법을 적용한 (Kang et al. 2004)의 연구를 능동적 학습의 초기 이후 문의 단계에도 적용할 수 있도록 일반화하였다.

기존 능동적 학습 연구에서 제안된 문의 예제 선정 방안들은 대부분 매 문의 단계마다 하나의 예제를 선정하는 경우에 가장 적합하도록 고안되어 있어, 복수 문의 예제 선정 작업에는 효과적이지 못할 수 있다. 기존의 방안들을 단순히 문의 예제로의 평가 결과가 높은 순으로 여럿 선정하도록 확장하게 되면 예제들이 밀집된 특정 영역에서 서로 비슷한 다수의 문의 예제가 함께 문의 예제로 선정될 가능성이 높다. 이는 유사한 예제들의 경우 문의 예제로서의 적합도 평가 결과도 엇비슷할 것이기 때문이다.

추정 오류 최소화 척도(Roy and McCallum, 2001)를 복수 문의 예제 선정을 위해 확장하여 적용한다면 유사한 예제들이 문의 예제로 함께 선정되는 문제점은 해결 가능할 것이나, 조합 가능한 모든 문의 예제 집합에 대하여 분류기를 생성하고 평가하는 과정이 요구되기 때문에 현실적으로 적용하기 어렵다. 이 방법을 이용할 경우 카테고리 수가 c개, 카테고리가 부여되지 않은 예제가 n개인 상황에서 b개의 문의 예제를 선정하고자 한다면, 조합 가능한 문의 예제 집합의 수는 ${}_nC_b$이다.

여기에 각 문의 예제 집합 별로 부여 가능한 카테고리의 조합 수 c^b을 고려하여야 하므로, 모두 ${}_nC_b \times c^b$번 분류기를 생성하고 평가하여야 한다. 따라서 복수의 문의 예제를 효율적으로 선정하기 위해서는 간단하면서도 보다 현실적인 접근 방안이 요구된다.

이러한 관점에서 능동적 학습을 위하여 복수의 문의 예제를 선정할 수 있는 효율적인 방안에 대한 연구가 수행된 것은 최근의 일이다. (Brinker, 2003)의 연구에서는 SVM을 이용한 능동적 학습에서 다양성과 모호성을 동시에 고려하여 복수 문의 예제를 선정하는 휴리스틱을 제안한 바 있다. SVM은 그 정확도는 높으나 학습에 소요되는 계산 비용이 매우 높은 학습 알고리즘이다. 이 연구는 문의 예제를 복수로 선정함으로써 SVM을 이용한 능동적 학습의 소요 시간을 줄일 수 있고, 동일한 수의 문의 예제를 한꺼번에 선정하는 경우에는 다양성을 모호성과 함께 고려함으로써 모호성만을 이용하는 경우보다 능동적 학습의 성능을 개선할 수 있음을 보였다. 하지만 이 연구는 기본적으로 SVM을 이용한 학습 자체에 소요되는 시간을 줄이는데 초점을 맞추었다. 또한 제안한 다양성 추정 방식은 SVM을 학습 알고리즘으로 사용할 경우에만 적합한 방식이라 할 수 있으며, 본 논문에서 능동적 학습을 활용하고자 하는 문서 분류 문제에 적용한 실험 결과는 제시되지 않았다.

훈련 예제들을 선별한다는 측면에서 본 논문과 연관성을 가진 다른 연구 분야로는 데이터 축약(data condensation, reduction)에 관한 연구가 있다. 데이터 마이닝을 위하여 수집한 데이터의 용량은 그 응용 분야에 따라 수십 기가바이트(giga-bytes) 이상 될 수 있다. 방대한 양의 데이터를 모두 활용하여 마이닝 기술을 적용하는 것은 저장 공간이나 기계 학습의 수행 시간 측면에서 비효율적이며, 대부분의 기계 학습 기법들은 학습에 필요한 데이터를 메모리에 상주시켜 사용하므로, 경우에 따라서는 적용 자체가 불가능할 수 있다. 데이터 축약은 이러한 상황에서 마이닝 기술을 효율적으로 적용할 수 있도록 원본 데이터의 특성을 유지할 수 있는 일부의 데이터를 선택하는 것이다(Mitra et al. 2002, Provost and Kolluri, 1999).

데이터 축약과 관련한 많은 연구는 예제에 카테고리가 미리 부여된 상황에서 출발하므로 본 연구와 근본적으로 차이가 있다. 최근 (Mitra et al. 2002)의 연구에서는 카테고리 정보를 사용하지 않는 밀도 기반(density-based) 데이터 축약 방안을 제안하였다. 이 방법은 다음과 같이 동작한다. ① 모든 예제에 대하여 k번째 가장 가까운 이웃과의 거리를 측정하여 그 거리가 가까울수록 해당 예제의 대표성이 높은 것으로 추정한다. ② 대표성이 높은 예제 순으로 대표 예제를 선택하되, 이 때 선택된 대표 예제와 k번째 가장 가까운 이웃과의 거리를 r이라 한다면, 해당 예제와 $2r$ 범위 이내의 예제들은 선택의 대상에서 제거한다. ③ 모든 예제들이 선택 또는 제거될 때까지 ②의 과정을 반복한다. [그림 3]에 이 알고리즘을 제시하고 있다.

이들은 이렇게 선정한 대표 예제들이 다른 데이터 축약 기법들을 이용하여 훈련 집합을 축약하는 경우에 비해 분류나 군집화 문제와 같은 기계 학습 응용에 보다 효과적으로 사용될 수 있음을 보였다. 하지만 이들에 의해 제안된 방법 역시 대부분의 데이터 축약 연구와 마찬가지로 학습 기법을 적용할 수 있는 수준까지 데이터를 축약하는 것이 주된 목적이므로, 최소한의 훈련 예제만을 사용하려는 능동적 학습과는 접근 관점에서 차이가 있다. 알고리즘의 특성상 원하는 예제의 축약 비율(또는 선별할 예제의 수)을 얻기 위해서는 여러 번의 실

밀도 기반 데이터 축약 알고리즘

용어
D 보유한 모든 예제들로 이루어진 집합
k 예제의 대표성을 추정하기 위하여 사용할 가까운 예제의 순번 (데이터 축약 비율과 연관이 있음)

방법
1. D에 포함된 각 예제 x에 대하여
   x의 대표성 r_x를 x와 k번째 가까운 예제 y와의 거리를 이용하여 추정한다. r_x (distance(x,y)
2. D에서 가장 높은 대표성을 가진 예제 x를 선정한다.
   x와 $2r_x$ 거리 내에 있는 예제들을 D에서 제거한다.
3. D가 공집합이 될 때까지 2의 과정을 반복한다.

[그림 3] 밀도 기반 데이터 축약 알고리즘

험을 통해 적당한 k값을 결정하여야 하는 부담이 있으며, 축약 이후 응용으로 문서 분류 문제 및 능동적 학습을 적용한 사례는 없다. 또한 추가의 비용 없이 학습의 성능을 상당히 향상시킬 수 있는 본 논문의 서론에서 언급한 모델 예제와 같은 개념은 제시되지 않았다.

(Shih et al. 2003)의 연구에서 문서 데이터 축약 방안으로 제안한 문서 뭉치화(text bundling) 기법은 본 연구에서 제안하는 모델 예제와 유사한 특성을 지니고 있다. 이들은 동일한 카테고리에 속하는 비슷한 문서들을 묶어 하나의 가상 문서로 변환하는 방법을 통해 데이터를 축약하였다. 하지만 문서 뭉치화 기법은 SVM과 같이 수행 시간이 상당한 학습 기법을 위하여 데이터를 축약하는데 그 목적이 있으며, 많은 데이터 축약 연구에서와 같이 모든 예제의 카테고리 정보가 주어진 상황에서 각 카테고리 별로 문서 뭉치화를 수행한다는 점에서 본 논문에서 제시하고자 하는 모델 예제와 차이가 있다.

데이터 축약 연구들은 근본적으로 능동적 학습과 같이 소수의 예제에만 카테고리가 부여된 상황에서 추가의 복수 문의 예제를 선정하기 위한 목적으로 제안된 방법은 아니나 일부 방안은 문서 분류를 위한 능동적 학습에 적용할 수 있도록 변형이 가능하다. 본 논문에서는 (Mitra et al. 2002)에서 제안된 밀도 기반 데이터 축약 기법을 수정 보완하여 제안 방안과 실험에서 그 성능을 비교하였다.

3. 군집화 기법을 이용한 복수 문의 예제 선정 방안

본 장에서는 본 논문에서 제안하는 군집화 기법을 이용하여 복수 문의 예제를 선정하는 방안에 대하여 보다 자세히 소개하고자 한다. 먼저 군집화 기법을 이용하여 복수 문의 예제를 선정하는 예를 보이고, k-NN 알고리즘을 문서 분류 문제에 적용하기 위하여 문서를 표현하는 방법과 문서 간 유사 정도를 추정하는 척도를 소개한다. 마지막으로 능동적 학습의 문의 예제 선정과 본 제안 방안의 가중치 반영 군집화에 필요한 카테고리 추정의 모호성 척도를 정의하고, 모호성을 가중치로 반영하여 군집화를 수행하는 가중치 반영 k-means 군집

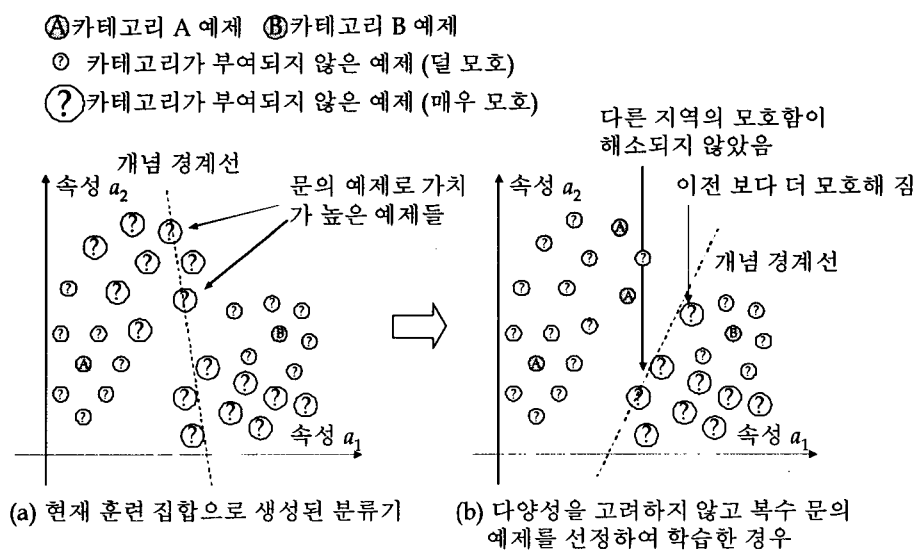
화 알고리즘을 제시한다.

3.1 복수 문의 예제 선정의 예

[그림 4]는 a_1과 a_2 두 가지 속성으로 기술되는 예제들을 이차원 상에 나타내고 있다. 각각의 예제는 두 가지 카테고리 A 또는 B 중 하나에 속한다. 본 예에서는 45도 각도로 그어진 직선을 기준으로 개념이 나누어진다고 가정하였다. 기호 로 표시된 예제들은 카테고리 A에 속하며, 기호 가 그려진 예제들은 카테고리 B에 해당된다. 모양의 예제들은 학습 시 카테고리가 부여되지 않은 예제들이다. 그림 4의 (a)는 각각 카테고리 A와 B에 속하는 회색으로 표시된 와 두 개의 초기 훈련 예제를 받아 능동적 학습이 초기 학습을 수행한 상태를 나타내고 있다. 생성된 분류기로 아직 카테고리가 부여되지 않은 나머지 예제들의 카테고리를 추정해 보면, 그 추정의 모호한 정도를 평가할 수 있다. 그림에서 카테고리가 부여되지 않은 예제들은 모호한 정도에 따라 그 크기를 다르게 표현하였다.

작은 물음표 동그라미로 표현된 예제들은 분류기의 개념 경계선에서 멀리 떨어져 있으면서 카테고리가 부여된 훈련 예제가 가까이 있어, 분류기가 상당한 확신을 가지고 카테고리를 추정할 수 있는 예제들이다. 큰 물음표 동그라미로 표현된 예제들은 개념 경계선 가까이에 위치하거나 카테고리가 부여된 훈련 예제가 주변에 없어 분류기가 어떤 카테고리라고 확신하여 추정하기 어려운 예제들이다.

카테고리 추정이 모호한 예제를 문의 가치가 높은 예제라 한다면, 능동적 학습은 그림 4의 (a)와 같이 경계선 상에 위치한 두 개의 예제를 복수 문의 예제로 선정하게 될 것이다. 이 경우에 비슷한 두 개의 예제가 문의 예제로 함께 선정되었는데, 이러한 상황은 단순히 문의 가치가 높은 순으로만 복수의 문의 예제를 선정하는 경우 빈번하게 발생할 수 있다. 그 이유는 유사한 예제들의 경우 문의 가치의 평가 결과도 서로 비슷할 것이기 때문이다.

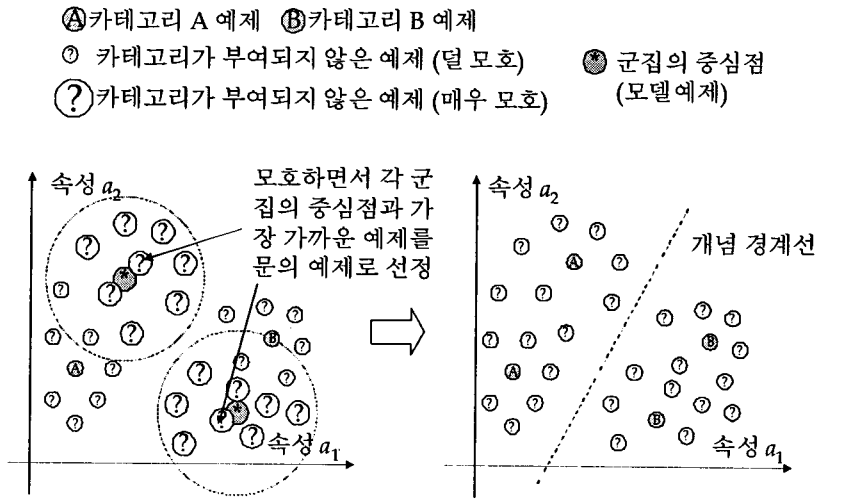


[그림 4] 다양성을 고려하지 않고 복수 문의 예제를 선정하여 학습한 예

[그림 4]의 (b)는 선정한 문의 예제들을 사용자에게 제시하여 카테고리를 부여 받아 기존 훈련 집합에 추가한 후 다시 학습한 결과를 보이고 있다. 그림에서 알 수 있듯이, 이전 상태보다는 분류기의 정확도가 다소 향상되었으나, 특정 지역에서 집중적으로 문의 예제를 선정하였기 때문에, 다른 지역에 위치한 예제들의 모호성은 크게 해소되지 않았다. 또한, 일부 예제는 이전보다 카테고리 추정이 더 모호해진 것을 볼 수 있다. 이와 같은 예는 복수 문의 예제를 선정할 때, 개별 예제의 문의 가치뿐 아니라 예제들 간의 다양성도 함께 고려되어야 함을 시사해 준다.

[그림 5]는 본 논문에서 제안하는 가중치 반영 군집화 기법을 이용하여 복수 문의 예제를 선정하는 예를 보이고 있다. 각 예제의 모호성을 가중치로 반영하여 군집화를 수행하면 모호한 예제들이 밀집한 지역을 중심으로 군집이 생성될 것이다. 카테고리 개수만큼 군집을 생성하는 일반적인 군집화와는 달리, 본 문제에서는 유사한 예제들끼리 모으고 군집 별로 학습에 사용할 예제를 하나씩 선정하기 위하여 군집화를 수행하므로, 능동적 학습이 사용자에게 문의할 예제의 수만큼 군집을 생성한다. 본 예에서는 군집화 기법으로 k-means 군집화 알고리즘을 가정하였다. 임의로 선정한 예제들을 각 군집의 초기 중심점(initial centroid, seed)으로 설정하고 각 예제의 모호성을 가중치로 삼아 k-means 알고리즘으로 군집화를 수행한다. [그림 5]의 (a)는 군집화가 완료된 상황을 보이고 있다.

군집화가 완료되면 이제 군집 별로 해당 군집을 대표하는 예제를 하나씩 선별한다면, 다양한 예제로 이루어진 복수 문의 집합을 구성할 수 있다. 각 군집을 대표하는 예제는 해당 군집을 잘 표현하면서 모호성 높은 예제가 적절할 것이다. 따라서 군집의 중심점과 가까운 동시에 모호성이 높은 예제를 군집의 대표 예제로 선정한다. 즉, 소속된 군집의 중심점과의 유사 정도와 예제의 모호성 두 값



(a) 가중치를 고려하여 군집화 수행 후 군집 별 문의 예제 선정

(b) 가중치 반영 군집화 기법으로 선정한 복수 문의 예제를 추가하여 학습

[그림 5] 가중치 반영 군집화 기법으로 선정한 복수 문의 예제로 학습한 예

을 곱하여, 그 값이 최대인 예제를 군집의 대표 예제로 선정한다. 각 군집의 중심은 획득한 대표 예제의 카테고리를 이용하여 학습에 사용할 수 있는 모델 예제로 변환할 수 있다.

[그림 5]의 (b)는 선정한 대표 예제들만을 기존의 훈련 집합에 추가하여 다시 학습한 결과를 보이고 있다. 앞서 다양성을 고려하지 않고 모호성만으로 복수의 문의 예제를 선정한 [그림 4]의 (b)보다 높은 정확도를 가진 분류기가 생성된 것을 알 수 있다.

3.2 문서의 표현과 문서 간 유사도 척도

보유한 전체 문서들의 집합을 D = {d_1, d_2, …, $d_{|D|}$}라고 하자. D에 포함된 하나 이상의 문서에 등장하는 모든 단어들의 집합을 T_D = {t_1, t_2, t_3, …, $t_{|TD|}$}라고 둔다. 문서 d_i에 단어 t_j가 등장하는 횟수(term frequency; tf)를 f_{ij}로 표현하면, 문서 d_i는 <f_{i1}, f_{i2}, f_{i3}, …, $f_{i|TD|}$>와 같이 벡터의 형태로 나타낼 수 있다.

임의의 두 문서 간의 유사 정도를 추정하는 가장 간단한 방법은 얼마나 많은 단어들이 두 문서에 공통으로 나타나는지 세는 것이다. 하지만 이 방법은 모든 단어를 아무런 구분 없이 동등하게 취급하므로 문서간의 유사 정도를 추정하는데 그다지 효과적이지 못하다. 예를 들어 질문과 답변 게시판에 올려진 게시물에는 '질문'이나 '답변'이라는 단어가 해당 게시물의 카테고리(예를 들어 쇼핑몰의 경우라면 배송, 결재, 회원가입 등)와 관계없이 대부분의 게시물에 포함되므로, 문서간의 유사 정도를 추정하는데 큰 도움이 되지 못한다. 따라서 k-NN 분류 알고리즘과 k-means 군집화 알고리즘을 이용한 텍스트 마이닝 연구에서는 정보검색 분야에서 문서 검색을 위하여 제안된 tf×idf 공간상에 문서를 벡터로 표현하는 방법이 일반적으로 사용된다(Yetes and Neto, 1999). 이 방법은 개별 단어의 등장 횟수 f_{ij} 대신 f_{ij} × idf_j를 원소로 하는 벡터로 문서를 표현한다. 단어의 idf 값은 전체 문서군에서 해당 단어가 등장하는 문서수의 역빈도수(inverse document frequency)로서 소수의 문서에 등장할수록 중요한 단어로 취급하여 높은 가중치를 준다. 단어 t_j의 idf 값 idf_j는 수식 (1)과 같이 계산되며, 수식에서 df_j는 집합 D에서 단어 t_j가 나타나는 문서의 수(document frequency)다. |D|는 보유한 전체 문서의 수다. 이러한 방식으로 문서 d_i는 tf×idf 공간상에서 d_i = <e_{i1}, e_{i2}, e_{i3}, …, $e_{i|TD|}$>, e_{ij} = f_{ij} × idf_j로 표현된다.

$$idf_j = \log_2(|D| / df_j) \qquad \text{수식 (1)}$$

두 문서 d_i와 d_j간의 유사 정도 sim(d_i,d_j)는 tf×idf 공간상에 표현된 두 문서 벡터 간의 코사인 유사도(cosine similarity)를 이용하여 추정한다(Yetes and Neto, 1999). 코사인 유사도는 두 벡터 간 각도의 코사인 값으로 수식 (2)와 같이 정의된다. 수식에서 |d_i|는 문서 d_i의 벡터 크기이다. 코사인 유사도는 두 문서에 등장하는 단어들의 비율이 완전히 일치한 경우에는 1의 값을, 두 문서에 공통으로 등장하는 단어가 하나도 없는 경우에는 0의 값을 갖는다.

$$\text{sim}(d_i,d_j) = d_i \cdot d_j / (|d_i||d_j|) \qquad \text{수식 (2)}$$

3.3 k-NN 알고리즘을 이용한 문서 분류

최근접 이웃 찾기(nearest neighbor; NN) 알고리즘은 분류 대상인 시험 예제와 각각의 훈련 예제와의 유사 정도를 측정하고, 그 중에서 가장 비

슷하다고 판단되는 훈련 예제의 카테고리를 시험 예제(test example)의 카테고리로 추정하는 학습 알고리즘이다(Cover and Hart, 1967). NN 알고리즘은 새로운 훈련 예제를 효율적으로 추가하여 학습할 수 있는 점진적인(incremental) 알고리즘이며, 간단명료하여 다른 학습 알고리즘에 비해 구현이 쉽다. 또한 훈련 예제를 수정 없이 저장하므로 빠르게 학습할 수 있다. 반대로 훈련 예제 수에 비례하여 저장 공간이 요구되며, 각각의 훈련 예제와 비교하여 가장 유사한 훈련 예제를 선정하여야 하므로 많은 훈련 예제를 사용하는 경우 상대적으로 다른 학습 알고리즘에 비해 분류 결과를 도출하는데 소요되는 시간이 길다. NN 알고리즘은 카테고리가 잘못 부여된 훈련 예제(noisy example)에 대해 민감한 것으로 알려져 있다(Witten and Frank, 1999). k-NN 알고리즘은 이러한 잘못된 예제에 의하여 분류의 정확도가 저하되는 현상을 최소화하기 위하여 시험 예제와 가장 유사한 하나의 훈련 예제가 아닌 k개의 훈련 예제를 찾아 각각의 훈련 예제와 유사한 정도에 따라 가중 투표 방식으로 시험 예제의 카테고리를 구한다. k-NN 알고리즘은 문서 분류 문제에서 정확도가 높은 분류기를 생성하는 알고리즘 중의 하나로 평가 받고 있다(Yang, 1999). [그림 6]에 k-NN 문서 분류 알고리즘을 구체적으로 소개하고 있다.

k-NN 알고리즘을 이용하여 훈련을 수행하고 문서를 분류하는 과정은 다음과 같다. 먼저 훈련 단계에서는 훈련 집합 L의 문서들을 이용하여 단어 집합 T_L과 T_L에 포함된 단어들의 df와 idf값을 구한다. 시험 단계에서 시험 예제 x가 주어지면, 앞 절에서 소개한 코사인 유사도 척도를 이용하여 x와 가장 유사한 k개의 훈련 예제들의 집합 $R = \{r_1, r_2, r_3, \cdots, r_k\}$, R (L을 찾는다. 카테고리의 집합을 $C = \{1, 2, 3, \cdots, |C|\}$, R_h를 R에 속하는 훈련

k-NN 문서 분류 알고리즘

용어
  L 카테고리가 부여된 예제들로 이루어진 훈련 집합
    k 시험 예제의 카테고리 추정에 사용할 가장 유사한 훈련 예제의 수
  $C = \{1, 2, 3, \cdots, |C|\}$ 카테고리의 집합

훈련 단계
  1. L 내의 문서에 존재하는 모든 단어들의 집합 T_L을 생성한다.
  2. 각각의 단어 t_j (T_L 에 대하여
    t_j의 등장 문서 수 df_j를 L내에서 구하고 수식 (1)을 이용하여 idf_j를 구한다.

시험 단계
  1. 시험 예제 x와 가장 유사한 k개의 훈련 예제를 찾아 집합 R을 구성한다.
    $R = \{r_1, r_2, r_3, \cdots, r_k\}$ (L
    x와 훈련 예제와의 유사 정도는 수식 (2)의 코사인 유사도를 이용한다.
  2. 가능한 카테고리 h (C 에 대하여
    2a. R 내의 훈련 예제 중 카테고리가 h인 모든 예제들을 모아 R_h를 구성한다.
    2b. R_h를 이용하여 x가 카테고리 h일 가중치 w_{xh}를 계산한다.
      $w_{xh} = (\mathrm{sim}(x, r_i), r_i$ (R_h
  3. 가장 높은 가중치를 받은 카테고리 p를 x의 카테고리로 추정한다.
    $p = \arg\max_h w_{xh}$

[그림 6] k-NN 문서 분류 알고리즘

예제들 중에서 카테고리가 h인 모든 예제들의 집합이라 정의하자. k-NN 알고리즘은 훈련 예제 x가 카테고리 h일 가중치 w_{xh}를 구한다. w_{xh}는 (sim(x,r_i), r_i (R_h와 같이 계산된다. 시험 예제 x는 가장 높은 가중치를 받은 카테고리 p = arg max$_h$ w_{xh}로 분류된다.

3.4 카테고리 추정의 모호성 척도

능동적 학습이 문의 예제를 선정하는데 사용되는 예제의 카테고리 추정의 모호 정도는 학습 알고리즘에 따라 그 적합한 척도가 달라진다[2]. k-NN 알고리즘의 경우에는 분류 결과 해당 예제가 각 카테고리에 속할 가중치가 생성된다. 예제 d_i에 대한 카테고리 추정의 가중치 분포는 W_i = {w_{i1}, w_{i2}, ..., $w_{i|C|}$}, $\forall k\ w_{ik} \geq 0$과 같이 표현될 수 있다. w_{ik}는 d_i가 카테고리 h에 소속될 가중치이다. 앞에서 서술한 바와 같이 d_i는 p = arg max$_h$ w_{ih}인 카테고리 p로 분류된다. 만일 d_i에 대한 카테고리 추정이 모호하다면 카테고리 p에 대한 가중치 w_{ip}에서 나머지 카테고리에 대한 가중치의 합 (w_{ih}, ($h \neq p$)를 빼면 그 값이 작을 것이며, 반대의 경우는 그 값이 클 것이다. 따라서 예제 d_i의 카테고리 추정의 모호성 uncertainty(d_i)는 수식 (3)과 같이 정의될 수 있다. 실험에서는 uncertainty(d_i) 값을 0과 1사이로 정규화 하여 사용하였다.

$$\text{uncertainty}(d_i) = \left(\sum_{h \in C} w_{ih}\right) - 2w_{ip}, p = \text{argmax}_h\, w_{ih} \quad \text{수식 (3)}$$

위에서 제시한 척도 이외에도 k-NN을 능동적 학습에 사용할 경우 다양한 방법으로 모호성을 정의할 수 있다. 본 연구에서는 실험에 앞서 몇 가지 모호성을 추정하는 척도를 비교 실험하여 가장 좋은 성능을 보인 수식 (3)을 모호성 척도로 사용하였다.

초기 훈련 예제를 선정할 때는 아직 분류기가 생성되지 않은 상황이므로 카테고리가 부여되지 않은 모든 예제들의 모호한 정도가 동일하게 추정된다. 또한 적은 수의 훈련 예제로 도출한 분류기를 사용하여 문서들을 분류하면 모호한 정도가 동일한 경우가 많이 발생할 수 있다. 하나의 문서는 보유한 전체 문서의 집합에서 1회 이상 나타난 단어들 중 극히 일부분에 해당되는 단어들로 구성되기 때문이다[3]. 이러한 단어 등장의 희박성 (sparseness) 때문에 소수의 훈련 예제만을 이용하여 분류기를 생성하면 훈련 예제와 공통되는 단어가 거의 없는 예제들이 많이 발생하게 된다. 따라서 모호성 평가 값이 동일한 예제들도 많아지게 된다. 본 논문에서는 동일한 모호성 평가 값을 가진 예제들 중에서는 임의로 문의 예제를 선정할 수 있도록 제안한 척도로 모호성을 추정한 후 그 값에 임의로 생성한 매우 작은 양수 ε를 더하였다.

3.5 가중치 반영 k-means 군집화 알고리즘

이상에서와 같이 모호성과 유사성 척도가 정의되면 가중치 반영 군집화를 위한 기본 척도들은 모두 마련된 셈이다. 가중치 반영 군집화 방안은 각 예제를 동등하게 취급하는 일반적인 군집화와는 달리 개별 예제의 모호성을 가중치로 반영하여 군집화를 수행하는 방법이다. 즉 능동적 학습의 현 분류기로 모호하지 않다고 판단되는 예제들은 적

2) 심지어는 능동적 학습 수행 과정에서 매 상황마다 가장 적절한 문의 예제 선정 척도도 다를 수 있다고 알려져 있다(Baam et a 2003).

3) Similar-3 데이터의 경우 전체 2,999 건의 문서들 중에서 1회 이상 나타난 단어의 수는 66,760개인데 비해, 개별 문서는 평균적으로 92개의 단어를 포함하고 있다.

은 가중치를 가지고 군집화에 참여하게 된다. 이러한 가중치 반영 군집화를 위해서는 기존 군집화 방안을 개선할 필요가 있다. 본 논문에서는 널리 알려진 군집화 기법의 하나인 k-means 알고리즘을 보완하여 적용하였다. 각 예제의 가중치를 무게와 같은 개념으로 설정한다면, k-means 군집화 과정에서 군집의 중심점으로 군집에 소속된 예제들의 무게 중심, 즉 가중 평균을 사용하면 예제들의 가중치를 반영할 수 있다. k-means 군집화 기법 이외에 유사한 예제들끼리 모을 수 있는 다른 군집화 알고리즘을 사용하더라도 가중치를 반영할 수 있도록 확장하여 본 논문에서 제안하는 방안을 적용할 수 있을 것이다.

k-means 알고리즘은 다음과 같이 동작한다. ① 먼저 k 개의 군집을 형성하기 위하여 (일반적으로 임의로 선정한 예제를 이용하여) 각 군집 별로 하나씩 모두 k개의 초기 중심점(initial centroids, seeds)을 생성한다. ② 각 예제는 k개의 중심점들 중 자신과 가장 가까운 중심점이 있는 군집에 할당된다. ③ 모든 예제에 대한 군집 할당이 끝나면 군집 별로 소속된 예제들의 평균을 계산하여 새로운 중심점을 생성한다. ④ 모든 중심점이 안정화될 때까지 ②에서 ③의 과정을 반복한다. 여기서 군집의 중심점과 예제 간의 유사성은 문서의 경우 앞에서 설명한 코사인 유사도를 이용한다. 군집 별로 새로운 중심점을 계산할 때 각 예제의 모호성을 가중치로 반영하여 가중 평균을 구함으로써 가중치 반영 군집화를 수행할 수 있다.

전체 예제의 수를 n, 선별하고자 하는 예제의 수를 k, 각 예제의 평균 등장 단어 수를 w, k-means 군집화 알고리즘 적용 시 군집이 안정화될 때까지 군집의 할당과 새로운 중심점 계산의 평균 반복 횟수를 t라고 하면 k-means 군집화 알고리즘의 복잡도(complexity)는 이들의 곱에 비례한다. 즉,

가중치 반영 k-means 문서 군집화 알고리즘

용어
- D 모든 예제들의 집합
- k 생성할 군집의 수
- weight(d_i) 문서 d_i의 가중치

방법
1. D 내의 문서에 존재하는 모든 단어들의 집합 T_D을 생성한다.
2. 각각의 단어 t_j (T_D 에 대하여
   t_j의 등장 문서 수 df_j를 D내에서 구하고 수식 (3)을 이용하여 idf_j를 구한다.
3. D 내에서 k개의 문서를 임의로 선택하여 초기 중심점의 집합 S를 생성한다.
   $S = \{s_1, s_2, \cdots, s_k\}$
4. 각 문서 d_i (D에 대하여
   4a. d_i가 소속될 군집 p를 구한다. $p = \arg\max_h (\mathrm{sim}(d_i, s_h))$
   4b. d_i의 군집 cluster(d_i)를 p로 설정한다.
5. 각 군집 h에 대하여
   cluster(d_i)가 h인 모든 문서를 가중치를 반영하여 새로운 중심점 s_h를 구한다.
   $s_h = ((z_i \times \mathrm{weight}(d_i) \times d_i) / (z_i,\ z_i = 1, \text{ if cluster}(d_i) = p, \text{ else } z_i = 0$
6. 모든 군집 중심점의 변동이 없을 때까지 4부터 5의 과정을 반복한다.

[그림 7] 가중치 반영 k-means 문서 군집화 알고리즘

복잡도는 $O(nkwt)$라 할 수 있다[4]. 실험에서 k-means를 이용한 군집화는 수십 회 이내에 안정화되었으며, 이는 모든 예제 간의 유사 정도를 계산하여야 하는 계층적 군집화(Hofmann, 1999)나 복잡한 모델과 계산이 필요한 밀도 기반 군집화에 비해 학습 시간 단축에 보다 유리하다. [그림 7]에 이상에서 소개한 가중치 반영 k-means 문서 군집화 알고리즘을 정리하였다.

3.6 군집 별 단일 문의 예제 선정

예제들을 대상으로 가중치 반영 군집화를 수행하면 각 군집은 서로 유사한 예제들의 모음이 되므로, 군집 별로 가장 적절한 하나의 예제를 문의 예제로 선정함으로써 본 논문에서 제안하는 복수 예제 선정이 마무리 된다. 각 군집 별로 학습에 가장 효과적일 것으로 추정되는 예제를 선정하는 방법으로는 기존에 연구된 능동적 학습의 다양한 문의 예제 선정 방안들을 활용할 수 있다. 하지만 본 논문에서는 가중치를 반영한 군집 결과를 최대한 활용할 수 있는 방안을 제안한다.

군집을 생성하는 과정에 모호성을 함께 반영하기 때문에 각 군집의 중심점이 대표성 및 모호성 측면을 동시에 고려하여 가장 적합한 지점이라 추정할 수 있다. 이렇게 군집의 중심점이 명시적인 결과로 도출된다는 점이 본 논문에서 계산 비용에 대한 고려와 함께 k-means 군집화 기법을 사용한 이유의 하나이다. 따라서 해당 군집의 중심점에 대하여 카테고리를 부여할 수 있다면, 바로 그 중심점이 가장 적절한 문의 예제가 될 것이다. 하지만 중심점은 해당 군집에 소속된 문서들의 가중 평균이므로 실제 존재하지 않는 문서이며, 사용자는 이에 대한 카테고리를 직접적으로 부여할 수 없다. 대신 중심점과 가장 가까우면서 문의 가치가 높은 예제를 군집의 대표 예제로 선정하여 문의 예제로 사용할 수 있다. 더 나아가 이렇게 문의된 예제와 해당 군집의 중심점은 서로 충분히 유사하므로, 문의 결과 예제에 부여된 카테고리를 군집의 중심점에도 그대로 부여하여 능동적 학습의 훈련 예제로 사용할 수 있다. [그림 8]에 군집 기반 복수 문의 예제 선정 알고리즘을 제시하고 있다.

4. 실험 결과

이상에서 제안한 방안의 효과를 확인하기 위하여 다음과 같은 몇 가지 문서 분류 문제를 대상으로 제안한 방안의 효과를 실험하였다. 문서 분류를 위한 대상 데이터로 기존 문서 분류 연구에서 자주 활용되어 온 Newsgroups-20 유즈넷(USENET) 뉴스 기사 모음과 Reuters-21578 신문 기사 말뭉치를 사용하였다(UCI).

Newsgroups-20은 20가지 서로 다른 주제의 유즈넷(USENET) 뉴스 그룹에 올려졌던 약 20,000건의 기사로 이루어져 있으며 분류의 난이도에 따른 본 제안 방안의 효과를 확인하기 위하여 (Basu et al. 2002)에서 실험한 바와 같이 주제가 상당히 다른 세 개의 뉴스 그룹 {alt.atheism, rec.sport.baseball, sci.space}으로 Different-3 문서 데이터를 생성하고, 주제가 매우 유사한 3개의 뉴스 그룹 {comp. graphics, comp.os.ms-windows, comp.windows.x}으로 Similar-3 문서

4) k-means 군집화에 가장 많은 시간이 소요되는 중심점과 개별 문서간의 유사 정도를 계산하기 위하여 문서군에 등장하는 모든 단어를 고려할 필요는 없다. $tf \times idf$ 공간상에서 코사인 유사도를 이용하여 문서간 유사 정도를 추정하는 경우, 비교하고자 하는 문서에 등장하는 단어에 대해서만 각 중심점과의 유사 정도를 계산하면 되므로 복잡도는 문서의 평균 단어 수 w에 비례한다.

군집 기반 복수 문의 예제 선정 알고리즘

용어
D 모든 예제들의 집합
k 문의 예제로 사용할 예제의 수
Q 문의 집합

방법
1. Q를 Ø로 초기화한다.
2. 가중치 반영 군집화 기법을 D에 적용하여 군집의 집합을 생성한다.
   X_h (i = 1, ···, k), D = $\cup X_h$
3. 각각의 군집 X_h에 대하여
   3a. X_h의 대표 예제 r_h를 선정한다.
   (k-means의 경우 X_h의 중심점과 가까우면서 카테고리 추정의 모호성이 높은 예제를 선정한다.
   즉, max {uncertainty (d_i) × sim (d_i, s_h)}인 예제를 선정한다)
   3b. r_h를 Q에 추가한다.
4. Q를 구성하는 모든 예제를 사용자에게 문의하여 카테고리를 부여 받는다.
5. 각각의 군집 X_h에 대하여
   5a. X_h의 모델 예제 m_h를 생성한다.
   (k-means의 경우 X_h의 중심점 s_h을 모델 예제로 사용한다)
   5b. m_h의 카테고리를 r_h의 카테고리로 설정한다.
   5c. m_h를 Q에 추가한다.

[그림 8] 군집 기반 복수 문의 예제 선정 알고리즘

데이터를 구성하였다. Reuters-21578 말뭉치는 1987년부터 1991년 사이에 생성된 로이터(Reuters)사의 경제 기사 21,578건으로 이루어져 있다. Reuters-21578 말뭉치에서 주제를 기준으로 단일 카테고리만 부여된 기사들 중 빈도수로 상위 2개의 주제인 earn과 acq에 해당되는 문서들만 추출하여 Reuters-2 문서 데이터를 구성하였다.

모든 문서는 제목을 제외한 유즈넷 헤더 제거(Differnet-3와 Similar-3의 경우), SGML 태그 제거(Reuters-2의 경우), 표준형 변환(stemming)과 불용어 제거(stop word removal) 과정을 거쳐 실험에 사용할 데이터로 최종 구축하였다. <표 1>에 이 세 가지 문서 데이터의 특성을 나열하고 있다.

학습 기법으로 사용한 k-NN 알고리즘은 k 값에 따라 약간의 성능 편차가 있으므로 적절한 값으로 설정할 필요가 있다. 본 실험에서 k 값은 5로 고정하였다. k 값을 5로 둔 이유는 실험에서 k-NN 분류기의 정확도가 대체적으로 k 값이 5일

<표 1> 실험 대상 문서 데이터의 특성

데이터 이름	카테고리수	문서수	난이도	카테고리 분포
Different-3	3	3,000	쉬움	균형
Similar-3	3	2,999	어려움	균형
Reuters-2	2	5,627	쉬움	불균형

때 안정적이며 우수하였고, 이 수치가 훈련 예제의 수(최대 36)와 카테고리의 수(2~3)를 고려할 때 적절하다고 판단하였기 때문이다. 모든 실험 결과는 10 분할 상호검증(10 fold cross-validation)을 10회씩 수행하여 정확도를 평균한 것이다.

문의 예제들을 임의로 선정하는 방안(*Rand*), 본 논문에서 제시하는 군집화 후 군집의 대표 예제로 문의 예제 집합을 구성하는 방안(*KM*), 군집의 대표 예제와 군집의 모델 예제를 함께 문의 예제로 사용하는 방안(*KM+ME*), 모호성만을 기준으로 복수 문의 예제를 선정하는 방안(*UC*), 그림 3의 밀도 기반 데이터 축약 기법(Mitra et al. 2002)을 변형한 방안(*DC*) 이상의 다섯 가지 방안을 구현하여 비교 실험하였다.

비교 대상 방안들 중 *DC* 방안은 본래 능동적 학습의 문의 예제 선정을 위하여 제시된 방안이 아니나 본 논문에서 제안하는 방안과의 성능 비교를 위하여 다음과 같이 변형하였다. 먼저 데이터 축약 비율을 직접적으로 결정할 수 있게 하기 위하여, k개의 문의 예제를 선정한다고 할 때, 선정된 대표 예제와 (|D|/k)-1번째까지 가까운 예제들을 문의 예제 선정 대상에서 제외하였다. 예제의 대표성을 추정하는 방안으로 각 예제와 (|D|/k)-1번째 가까운 예제와의 거리 대신 모든 다른 예제들과의 평균 유사도를 사용하였다. 이는 적은 수의 문의 예제를 선정하는 경우 상대적으로 (|D|/k)-1값이 커지게 되는데, 경우에 따라서는 (|D|/k)-1번째 가까운 예제와는 공유하는 단어가 거의 없어 예제의 대표성을 제대로 추정하지 못할 수 있기 때문이다. 수정된 알고리즘은 [그림 9]에 제시하였다.

[그림 10]은 Different-3 데이터에 대하여 문의 예제 선정 방법 별로 한 번에 문의하는 예제의 수를 3개에서 18개까지 변화시켜 가며 36개의 예제를 문의한 후 평가된 실험 결과이다. 그림에서 가로축은 능동적 학습이 한 번에 문의한 예제 수를 나타낸다. 세로축은 생성한 분류기의 정확도를 나타낸다. 본 논문에서 제안하는 *KM* 방안과 *KM+ME* 방안이 가장 우수한 성능을 보였다. 전체적으로 *KM+ME* 방안의 성능이 가장 우수하였고 *KM*과 *DC* 방안이 각각 그 뒤를 이었다. *DC* 방안의 경우에도 좋은 성능을 보여 주었으나, *KM* 방안보다 정확도가 높은 분류기를 생성하지는 못하였다. 다양성을 고려하지 않고 복수 문의

수정된 밀도 기반 데이터 축약 알고리즘

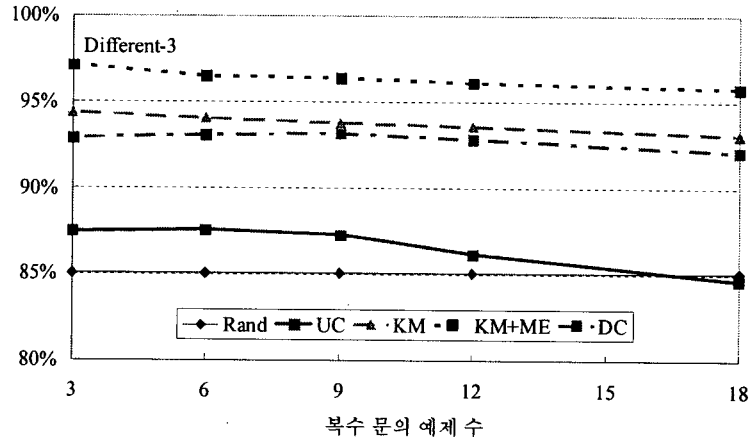
용어
- D 모든 예제들의 집합
- k 문의 예제로 사용할 예제의 수
- weight(d_i) 문서 d_i의 가중치

방법
1. D에 포함된 각 예제 d_i에 대하여
   d_i의 대표성 r_i를 D에 포함된 모든 예제들과의 평균 유사도를 계산하여 추정한다.
2. 대표성 r_i와 가중치 weight(d_i)를 곱하여 그 값이 가장 높은 예제 x를 선정한다.
   x와 (|D|/k)-1 이내로 가까운 예제들을 D에서 제거한다.
3. D가 공집합이 될 때까지 2의 과정을 반복한다.

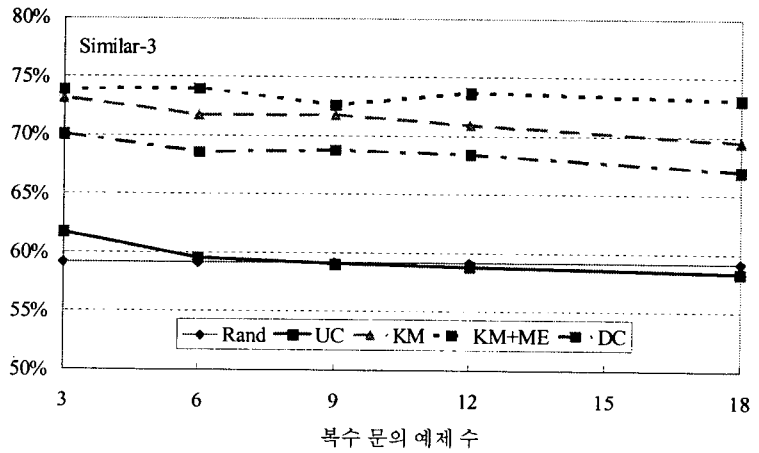
[그림 9] 수정된 밀도 기반 데이터 축약 알고리즘

예제를 선정하는 *UC* 방안의 경우 한 번에 문의하는 예제의 수가 많아질수록 성능이 급격히 저하됨을 알 수 있다. 이에 비해 대표성과 다양성을 함께 고려하여 복수 문의 예제를 선정하는 *KM, KM+ME, DC* 방안의 경우 한 번에 문의하는 예제 수를 증가시키더라도 성능 저하가 적은 것을 확인할 수 있다.

[그림 11]은 Similar-3 데이터에 대하여 동일한 방법으로 수행한 실험 결과이다. Different-3와 마찬가지로 *KM+ME* 방안의 성능이 가장 우수한 성능을 보였고, *KM, DC* 방안 순으로 각각 그 뒤를 이었다. Different-3 데이터를 이용한 실험 결과와 같이 *UC* 방안은 한 번에 문의하는 예제의 수가 늘어나면 임의로 문의 예제를 선정하는 경우보다도 더 성능이 떨어질 수 있음을 확인할 수 있다.



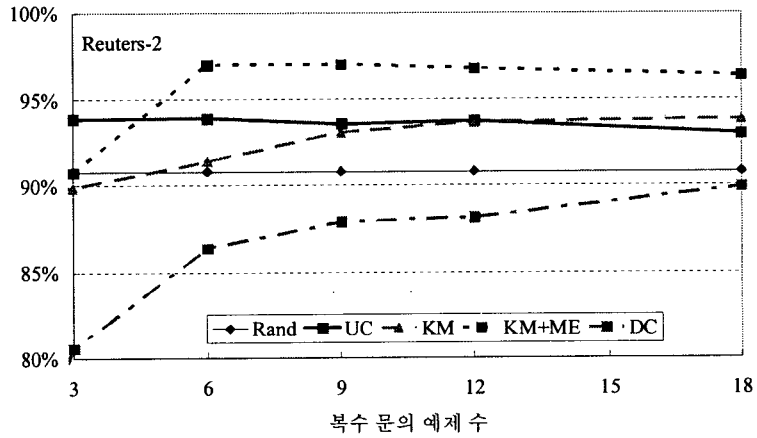
[그림 10] 실험 결과 (Different-3, 36개 훈련 예제 사용 시)



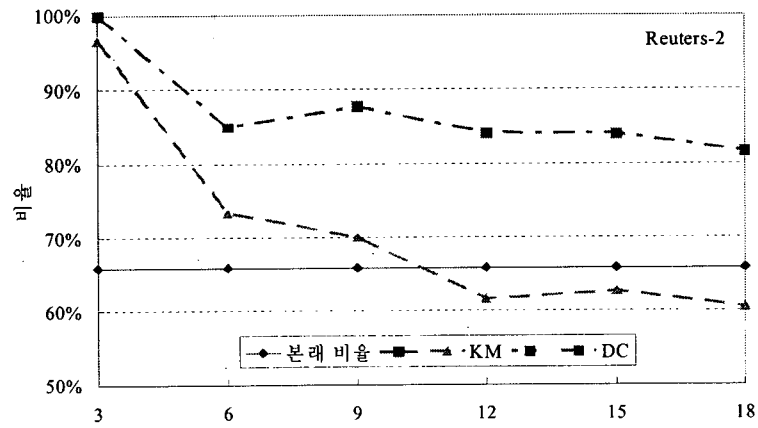
[그림 11] 실험 결과 (Similar-3, 36개 훈련 예제 사용 시)

[그림 12]는 Reuters-2 데이터에 대하여 수행한 실험 결과이다. *KM* 방안과 *KM+ME* 방안은 문의 예제를 3개씩 선정한 경우에는 임의 선정과 비슷한 성능을 보였지만, *KM* 방안의 경우 12개 이상, *KM+ME* 방안의 경우 6개 이상의 문의 예제를 한꺼번에 선정하면 *UC* 방안으로 문의 예제를 선정하는 경우보다 더 나은 결과를 얻을 수 있었다. *DC* 방안의 경우 Reuters-2 데이터에 대하여 불안정한 수행 결과를 보였으며, 임의로 문의 예제를 선정하는 경우보다 우수하다고 판단할 수 있는 경우는 없었다.

Different-3 및 Similar-3 데이터와는 달리 Reuters-2 말뭉치를 대상으로 한 실험에서 한 번에 적은 수의 문의 예제를 선정할 때 *KM*, *KM+ME*, *DC* 방안의 성능이 우수하지 못한 이유를 분석해 본 결과 한 번에 문의 예제를 적게



[그림 12] 실험 결과 (Reuters-2, 36개 훈련 예제 사용 시)



[그림 13] 선정된 문의 예제들 중 earn 카테고리의 비율 (Reuters-2)

선정할수록 [그림 13]과 같이 보다 많은 문서를 보유한 earn 카테고리에 편향되어 문의 예제들이 선정된 것을 확인할 수 있었다. 이러한 결과는 Reuters-2의 경우 본 논문에서 사용한 k-means 군집화 방안으로는 이 두 카테고리를 잘 구분하는 소수의 군집을 생성하기 어려웠기 때문으로 풀이된다. 본 논문에서 제안한 방안은 k-means 이외의 다른 군집화 방안에도 적용이 가능하므로 보다 효과적으로 문서를 군집화 할 수 있는 방안을 적용한다면 이러한 문제점이 해소될 것으로 예상된다.

5. 결론 및 향후 연구

본 논문에서는 능동적 학습을 문서 분류 문제에 보다 효과적으로 적용할 수 있도록 가중치 반영 군집화 기법을 이용하여 복수 문의 예제를 선정하는 방안을 제안하였다. 본 제안 방안을 적용한 능동적 학습 기법을 여러 문서 분류 문제에 적용하여 실험한 결과 복수의 문의 예제 선정 시 능동적 학습이 생성하는 분류기의 성능을 향상시킬 수 있었다. 특히 사용자에 대한 추가의 문의 없이 생성이 가능한 모델 예제를 활용함으로써 얻은 성능의 개선은 문의하는 예제의 수를 최소화하면서 분류기의 성능을 향상시키고자 하는 능동적 학습에 있어 그 가치가 크다 할 수 있다.

향후 본 제안 방안을 확장하여 문서 분류 문제에 많이 활용되는 또 다른 분류 알고리즘의 하나인 나이브 베이즈와 SVM에 적용하는 연구를 수행할 필요가 있으며 능동적 학습 과정에서 매 문의 단계마다 적절한 문의 예제 수를 적응적으로 결정할 수 있는 방안에 대한 연구가 필요하다.

참고문헌

[1] Abe, N. and H. Mamitsuka, "Querying learning using boosting and bagging," *In Proceedings of the 15-th International Conference on Machine Learning*, (1998), 1-10.

[2] Basu, S., A. Banerjee, and R. Mooney, "Semi-supervised clustering by seeding," *In Proceedings of the 19-th International Conference on Machine Learning*, (2002), 19-26.

[3] Baram, Y., R. El-Yaniv, and K. Luz, "Online Choice of Active Learning Algorithm," *In Proceedings of the 20-th International Conference on Machine Learning*, (2003), 595-609.

[4] Blum, A. and T. Mitchell, "Combining labeled and unlabeled data with co-training," *In Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann Publishers, (1998), 92-100.

[5] Breiman, L., *Bagging predictors,* Technical Report 421, University of California at Berkeley, 1994.

[6] Brinker, K., "Incorporating Diversity in Active Learning with Support Vector Machines," *In Proceedings of the 20-th International Conference on Machine Learning*, (2003), 59-66.

[7] Cohn, D., Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, Vol. 4, (1996), 129-145.

[8] Cover, T. and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, 13, (1967), 21-27.

[9] Freud, Y. and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *In Proceedings of the Second European Conference on Computational Learning Theory*, (1995), 23-37.

[10] Hofmann, T., "The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data," *In Proceedings of the 16-th International Joint Conference on Artificial Intelligence IJCAI-99*, (1999), 682-687.

[11] Kang, J., K. R. Ryu, and H.-C. Kwon, "Using Cluster-Based Sampling to Select Initial Training Set for Active Learning in Text Classification," *In Proceedings of PAKDD-2004 Conference*, (2004), 384-388.

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

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

[14] Muslea, I., S. Minton, and C. Knoblock, "Selective sampling with redundant views," *In Proceedings National Conference on Artificial Intelligence*, (2000), 621-626.

[15] Muslea, I., S. Minton, and C. Knoblock, "Active + Semi-Supervised Learning = Robust Multi-View Learning," *In Proceedings of the 19-th International Conference on Machine Learning*, (2002), 435-442.

[16] Mitra, P., C. A. Murthy, and S. K. Pal. "Density Based Multiscale Data Condensation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 6, (2002), 734-747.

[17] Nigam, K. and R. Ghani, "Analyzing the effectiveness and applicability of co-training," *In Proceedings of Information and Knowledge Management*, (2000), 86-93.

[18] Provost, F. and V. Kolluri, "A survey of methods for scaling up inductive algorithms," *Data Mining Knowledge Discovery*, Vol. 2, (1999), 131-169.

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

[20] Seung, H. S., M. Opper, and H. Sompolinsky, "Query by committee," *In Computational Learing Theory*, (1992), 287-294.

[21] Shih, L., J. D. M. Rennie, Y.-H. Chang, and D. R. Karger, "Text Bundling: Statistics-Based Data Reduction," *In Proceedings of the 20-th International Conference on Machine Learning*, (2003), 696-703.

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

[23] Witten, I. H. and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 1999.

[24] Yang, Y., "An evaluation of statistical approaches to text categorization," *Journal of Information Retrieval*, Vol. 1, Nos. 1/2, pp. (1999), 67-88.

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

Abstract

Cluster-Based Selection of Diverse Query Examples for Active Learning

Jaeho Kang* · Kwang Ryel Ryu* · Hyuk-Chul Kwon*

In order to derive a better classifier with a limited number of training examples, active learning alternately repeats the querying stage for category labeling and the subsequent learning stage for rebuilding the classifier with the newly expanded training set. To relieve the user from the burden of labeling, especially in an on-line environment, it is important to minimize the number of querying steps as well as the total number of query examples. We can derive a good classifier in a small number of querying steps by using only a small number of examples if we can select multiple of diverse, representative, and ambiguous examples to present to the user at each querying step. In this paper, we propose a cluster-based batch query selection method which can select diverse, representative, and highly ambiguous examples for efficient active learning. Experiments with various text data sets have shown that our method can derive a better classifier than other methods which only take into account the ambiguity as the criterion to select multiple query examples.

Key words : Active learning, Text classification, Batch query selection, Clustering

* Department of Computer Engineering, Pusan National University