능동적 학습을 위한 군집기반 초기훈련집합 선정
(Selection of An Initial Training Set for Active Learning Using Cluster-Based Sampling)

강 재 호 † 류 광 렬 †† 권 혁 철 †††
(Jaeho Kang) (Kwang Ryel Ryu) (Hyuk-Chul Kwon)

요 약 본 논문에서는 능동적 학습이 보다 적은 수의 훈련예제로도 높은 학습성능을 달성할 수 있도록 군집화기법을 이용하여 초기훈련집합을 선정하는 방안을 제안한다. 본 제안 방안은 유사한 예제들보다는 다양한 예제들로 그리고 특수한 예제들보다는 보편적인 예제들로 구성한 집합이 학습에 유리할 것이라는 가정을 바탕으로, 먼저 k-means 군집화 기법으로 예제들을 군집화한 후, 각 군집을 가장 잘 표현하는 대표예제로 개별 군집의 중심점과 가장 가까운 예제를 선정하여 초기훈련집합을 구성한다. 또한 개별 군집의 중심점을 가상의 예제로 가정하여, 이와 연관된 대표예제의 카테고리를 부여함으로써 추가의 훈련예제로 활용하는 방안을 함께 제안한다. 여러 문서 분류 문제를 대상으로 실험한 결과, 본 제안 방안으로 선정한 초기훈련집합에서 출발한 능동적 학습이 임의로 선정한 초기훈련집합에서 출발한 경우에 비해 보다 적은 수의 훈련예제로도 동등한 성능을 달성할 수 있음을 확인하였다.

키워드 : 능동적 학습, 초기훈련집합 선정, 군집화, 문서 분류

Abstract We propose a method of selecting initial training examples for active learning so that it can reach high accuracy faster with fewer further queries. Our method is based on the assumption that an active learner can reach higher performance when given an initial training set consisting of diverse and typical examples rather than similar and special ones. To obtain a good initial training set, we first cluster examples by using k-means clustering algorithm to find groups of similar examples. Then, a representative example, which is the closest example to the cluster's centroid, is selected from each cluster. After these representative examples are labeled by querying to the user for their categories, they can be used as initial training examples. We also suggest a method of using the centroids as initial training examples by labeling them with categories of corresponding representative examples. Experiments with various text data sets have shown that the active learner starting from the initial training set selected by our method reaches higher accuracy faster than that starting from randomly generated initial training set.

Key words : Active learning, Initial training set selection, Clustering, Text classification

1. 서 론

기계학습의 분류(classification)기술을 현실 문제에 성공적으로 적용하기 위해서는 카테고리(category, class)가 부여된(labeled) 훈련예제(training example)를 상당수 준비하여야 한다. 예제에 카테고리를 부여하는 작업에는 무시할 수 없는 시간과 인력이 필요하며, 응용분야에 따라서는 그 비용이 상당할 수 있다. 이러한 문제에 효과적으로 대처하기 위한 방법의 하나인 능동적 학습(active learning)[1,2]은 카테고리를 부여할 수 있는 훈련예제의 수가 제한되어 있는 상황에서 최대한 정확도가 높은 분류기를 생성하기 위하여 학습과정에서 카테고리를 부여할 예제들을 선별하면서 학습하는 전략이다.

능동적 학습기법은 사용자가 답변할 수 있는 최대예제 수에 도달할 때까지 학습단계와 문의(query)단계를 반복적으로 수행한다.[1] 학습단계에서는 학습 알고리즘을 현재 보유한 훈련집합(training set)에 적용하여 분류기

· 국가지정연구실사업(과제명: 언어 중심의 지능적 정보처리를 위한 단계적 우리말 분석기술의 개발(M10203000028-02J0000-01510))의 지원을 받아 이루어진 것임
† 학생회원 : 부산대학교 컴퓨터공학과
jhkang@pusan.ac.kr
†† 종신회원 : 부산대학교 컴퓨터공학과 교수
krryu@pusan.ac.kr
††† 종신회원 : 부산대학교 컴퓨터공학과 교수
hckwon@pusan.ac.kr
논문접수 : 2004년 1월 29일
심사완료 : 2004년 4월 30일

(classifier)를 생성한다. 문의단계에서는 생성된 분류기를 이용하여 카테고리가 부여되지 않은(unlabeled) 예제들을 분류해보고, 이들 중에서 학습에 가장 효과가 높을 것으로 추정되는 예제들을 선정하여 사용자에게 카테고리 부여를 요청한다. 사용자에 의하여 카테고리가 부여된 문의예제들은 기존의 훈련집합에 추가된다. 능동적 학습은 이러한 과정의 반복을 통하여 생성되는 분류기의 정확도를 점진적으로 향상시킨다.

능동적 학습에 관한 기존 연구들은 학습에 효과적인 문의예제들을 판별할 수 있는 다양한 방안을 탐구해 왔으나, 초기학습에 필요한 예제들의 선정문제에는 상대적으로 소홀하였다. 즉, 학습에 보다 효과적인 예제들을 선별하여 초기훈련집합을 구성함으로써 능동적 학습과정 전반에 걸쳐 성능을 향상시키고자 하는 연구는 아직 구체적으로 수행되지 않았다. 초기훈련집합의 중요성을 단거리 경주에 비유하여 설명한다면 능동적 학습이 학습에 보다 유용한 초기훈련집합을 사용할수록, 선수는 좀 더 결승선에 가까운 위치에서 출발하게 되는 셈이라 할 수 있다.

학습을 수행할 수 있는 최소한의 예제들로 초기훈련집합을 구성하고 이후 과정에서 하나씩 문의하고 학습하는 과정을 반복하는 단순한 전략은 현실에서 비효율적일 수 있다. 사용자는 한번에 여러 개의 예제에 대한 문의요청을 받을 경우 유사한 작업을 반복하여 처리함으로써[2] 숙련될 수 있으며, 작업순서에 대한 유연성을 발휘할 수 있어 보다 효율적으로 문의를 처리할 수 있다. 또한, 사용자는 서로 다른 여러 예제들을 상호 비교하면서 작업이 가능하므로 카테고리를 보다 정확히 부여할 수 있을 것이다. 특히 문의에 답변할 수 있는 인력이 충분하다면 동시에 많은 수의 예제들에 대한 카테고리 부여작업을 생성하여 병렬로 처리하는 것이 시간 절약 측면에서 바람직할 것이다. SVM(support vector machine)[3]과 같이 학습에 소요되는 시간이 상당한 학습 알고리즘을 사용하고자 하는 경우에는 학습 시간의 단축 효과도 얻을 수 있다. 이렇게 본다면, 능동적 학습은 시간과 인력이라는 제한된 자원을 최대한 활용하여 가장 높은 성능을 달성하는 것이 목표가 되며, 이를 위해서는 초기훈련집합 선정을 포함한 전반적인 학습 과정에서 학습에 사용할 복수의 예제들을 효과적으로 선별할 수 있어야 한다.

본 논문에서는 이러한 인식하에 능동적 학습의 초기 성능과 직결되는 초기훈련집합을 효과적으로 선정하기 위하여 군집화(clustering)기법을 이용하는 방안을 제안한다. 본 제안 방안은 유사한 예제들보다는 다양한 예제들로 그리고 특수한 예제들보다는 보편적(typical)인 예제들로 구성한 훈련집합이 학습에 유리할 것이라는 가정에 기반한다. 카테고리가 부여되지 않은 상태의 예제들을 대상으로 군집화를 수행하면, 유사한 예제들을 모아 군집의 집합을 생성할 수 있다. 서로 다른 군집에 속하는 예제들은 상대적으로 상이하므로 각 군집별로 해당 군집을 대표할 수 있는 예제를 하나씩 추출한다면 다양하면서 보편적인 예제들로 이루어진 초기훈련집합을 구성할 수 있다.

하나의 군집을 대표하는 예제로는 군집화 결과로 생성되는 군집별 모델(model)[3]이 가장 적합할 것이다. 하지만, 일반적으로 군집의 모델은 실제 존재하는 예제가 아니므로 사용자에게 제시하여 직접 카테고리를 부여받을 수 없다. 본 논문에서는 실제 존재하는 예제들 중에서 군집의 모델과 가장 유사한 예제를 해당 군집의 대표예제로 삼아 사용자에게 문의하고, 이들 대표예제들을 이용하여 초기훈련집합을 구성하는 방안을 제안한다. 여러 문서 데이타를 이용한 실험에서 이와 같은 방식으로 구성한 초기훈련집합이 능동적 학습의 초기 성능 향상에 매우 효과적임을 확인할 수 있었다. 더 나아가 카테고리를 부여받은 대표예제는 해당 군집의 모델과 가장 흡사하므로, 군집의 모델을 예제화한 후 대표예제의 카테고리를 부여함으로써 사용자에게 추가의 문의 없이 군집을 효과적으로 표현할 수 있는 훈련예제를 얻을 수 있는 방안을 함께 제안한다. 군집의 모델을 훈련예제로 사용하는 경우 이들을 모델예제로 칭하며, 이러한 모델예제를 대표예제와 함께 초기훈련예제로 활용함으로써 추가의 성능 향상을 얻을 수 있었다.

본 논문의 구성은 먼저 2장에서 군집화 기법을 이용하여 초기훈련집합을 선정하는 방안에 대하여 보다 구체적으로 설명하고, 3장에서 본 제안 방안을 여러 문서 분류 문제에 적용하여 실험한 결과를 정리하여 분석한다. 4장에서는 관련된 연구를 소개하고, 마지막 5장에서 결론과 향후 연구방향을 기술한다.

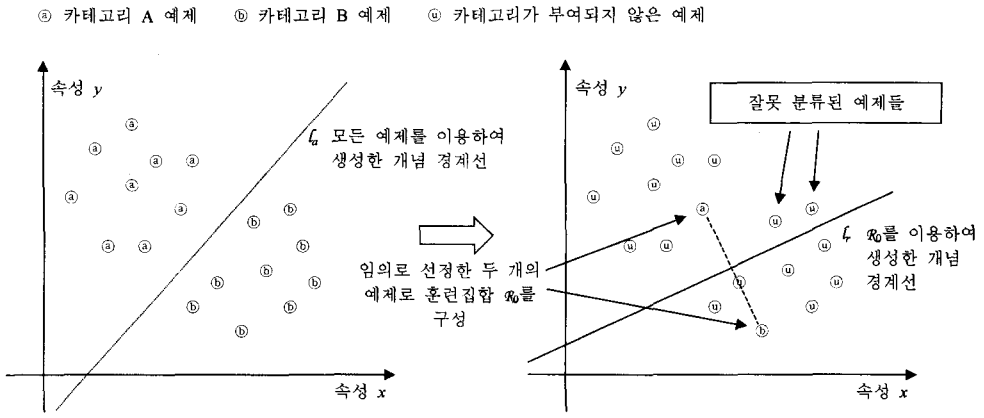
2. 군집화 기법을 이용한 초기훈련집합 선정 방안

본 장에서는 군집화 기법을 이용하여 초기훈련집합을 선정하는 방안에 대하여 기술한다. 본 논문에서 제시하는 방안을 직관적으로 설명하기 위하여 먼저 간단한 예

1) 엄밀하게는 능동적 학습 중에서도 선별적 추출(selective sampling)을 의미한다. 본 논문에서는 용어의 이해도 측면에서 능동적 학습으로 통일하여 기술하였다.

2) 능동적 학습은 현재 문의와 다음 문의 사이에 재학습과정이 필요하므로 학습에 소요되는 시간이 상당한 알고리즘이 사용되는 경우, 사용자는 예제에 대한 문의의 연속성이 결여된 것처럼 느낄 수 있다.

3) 본 논문의 실험에 사용한 k-means 군집화 알고리즘의 경우 중심점이 모델에 해당된다.



(a) 모든 예제를 이용하여 학습한 결과 (b) 훈련집합 $\mathcal{R}_0$를 이용하여 학습한 결과 (임의로 선정한 두 개의 예제)

그림 1 임의로 선정한 두 개의 훈련예제를 이용하여 학습한 결과의 예

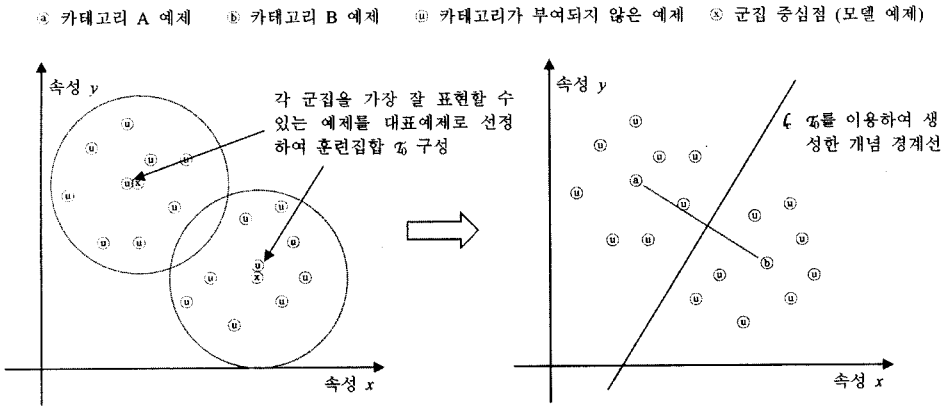
를 들고자 한다.

그림 1에는 이차원 상에 분포된 예제들을 보이고 있다. 각각의 예제는 두 가지 카테고리 A 또는 B 가운데 하나에 속한다. 기호 ⓐ로 표시된 예제들은 카테고리 A에 속하며, 기호 ⓑ가 그려진 예제들은 카테고리 B에 해당된다. ⓤ 모양의 예제들은 학습 시 카테고리가 부여되지 않은 예제들이다. 분류 알고리즘에 카테고리가 부여된 예제들을 충분히 제공한다면 그림 1(a)의 직선 ℓ_a와 같이 두 카테고리를 높은 정확도로 구분할 수 있는 분류기를 생성할 수 있다. 하지만 예제를 학습에 사용하려면 카테고리를 부여하는 작업이 선행되어야 하고 이러한 작업에는 상당한 시간과 인력이 소요될 수 있기 때문에 가능한 여건이 허용하는 한도 내에서 많은 예제에 카테고리를 부여하고자 할 것이다. 그림 1(b)는 임의로 선정한 2개의 예제에 카테고리를 부여하여 학습을 위한 집합 $\mathcal{R}_0$를 구성한 후 학습을 수행한 결과를 보이고 있다. 본 예에서는 예제들의 공간을 직선으로 나누는 분류기를 생성하는 학습 알고리즘을 가정하였다. 선정된 훈련예제들의 위치가 효과적이지 못한 경우 그림 1(b)에서와 같이 개념(concept)을 온전히 학습하지 못하여 일부 예제들을 잘못 분류하는, 즉 정확도가 다소 떨어지는 분류기가 생성될 수 있다. 그러므로 초기훈련집합을 보다 신중하게 선정함으로써 능동적 학습의 효과를 최대화할 수 있는 방안이 필요하다.

그림 2는 본 논문에서 제안하는 군집화 기법을 이용하여 초기훈련집합을 구성하는 예를 보이고 있다. 카테고리 개수만큼 군집을 생성하는 일반적인 군집화와는 달리, 본 문제에서의 군집화는 유사한 예제들끼리 모으고 개별 군집별로 학습에 사용할 예제를 하나씩 선정하는 것을 목적으로 하므로, 초기학습에 사용할 예제의 수만큼 군집을 생성한다. 본 예에서는 군집화 기법으로 k-means 군집화 알고리즘을 가정하였다. 먼저 임의로 선정한 예제들(여기서는 그림 1(b)에서 선정한 예제들과 동일하나 아직 사용자가 카테고리를 확인해 주지 않은 상태)을 각 군집의 초기중심점(initial centroid, seed)으로 삼아 k-means 알고리즘으로 군집화를 수행한다. 그림 2(a)는 군집화가 완료된 상황을 보이고 있다.

각 군집은 유사한 예제들의 모음이므로 동일한 군집 내 예제들은 동일한 카테고리에 속할 가능성이 높다. 초기훈련예제로 이들 군집을 가장 잘 표현할 수 있는 군집별 모델에 대하여 사용자가 카테고리를 부여하는 것이 가장 바람직하나, 모델은 실제 존재하는 예제가 아니므로 직접 카테고리를 부여하기 어렵다. 따라서, 군집별로 해당 군집을 가장 잘 표현할 수 있는 실제 존재하는 예제를 대표예제(representative example)로 선정하여 사용자에게 문의한다. k-means 알고리즘의 경우 군집의 모델은 중심점이며, 해당 군집의 중심점과 가장 가까운 예제를 대표예제로 선정할 수 있다.

선정된 대표예제들을 사용자에게 문의하면 카테고리를 부여 받을 수 있으며 이를 이용하여 초기훈련집합을 다음의 두 가지 방안으로 구성할 수 있다. 첫 번째 방안은 카테고리가 부여된 대표예제들만으로 초기훈련집합을 구성하는 방안이다. 두 번째 방안은 개별 군집 모델의 카테고리를 연관된 대표예제의 카테고리로부터 유추하고, 이를 가상의 예제로 삼아 대표예제들과 함께 초기훈련집합을 구성하는 방안이다. 대표예제에 의하여 카테고리가 부여된 모델을 훈련예제로 취급하는 경우 본 논문에서는 이를 모델예제(model example)라 정의한다.



(a) 군집으로부터 대표예제의 선별

(b) 훈련집합 𝒯를 이용하여 학습한 결과 (군집화 기법으로 선정한 두 개의 예제)

그림 2 k-means 군집화 기법을 이용하여 대표예제를 선정하는 과정과 선정된 대표예제를 이용하여 학습한 결과의 예

그림 2(b)는 대표예제들만으로 구성된 초기훈련집합을 이용하여 학습한 결과를 보이고 있다. 앞서 임의로 선정한 예제들을 초기훈련집합으로 사용한 그림 1(b)의 경우보다 안정적으로 학습이 된 것을 알 수 있다.

카테고리가 부여되었을 때 카테고리가 부여되지 않은 예제들에 대하여 보다 신뢰성 있게 카테고리를 추정할 수 있게 하는 예제를 학습에 효과가 높은 예제로 정의한다면[2], 학습에 효과가 가장 높은 예제집합이란 집합 내의 모든 예제들에 카테고리가 부여되었을 때 이러한 신뢰 정도가 최대화되는 집합이라 할 수 있다. 따라서 [2]에서 제시한 개별 예제에 카테고리가 부여되었을 때 예상되는 오류를 계량화할 수 있는 척도를 집합에 대하여 추정할 수 있도록 확장한다면, 최적으로 추정되는 초기훈련집합을 생성할 수 있을 것이다. 하지만 구성 가능한 모든 초기훈련집합을 생성하고 이들 각각을 평가하여 최적의 집합을 선택하는 것은 그 가짓수가 방대하여 제한된 시간 내에 처리하기가 현실적으로 어렵다. 예를 들어 전체 예제의 수가 n, 카테고리의 수가 c인 문제에서 k개의 최적 훈련예제를 선정하고자 한다면 조합 가능한 집합의 수는 ${}_nC_k$이다. 따라서, 빠른 시간 내에 상대적으로 학습에 유용한 예제집합을 생성하기 위해서는 휴리스틱적인 접근이 필요하다.

본 논문에서 이러한 휴리스틱적 접근을 위하여 설정한 가정들은 다음과 같다. 첫 번째는 특정 예제에 카테고리를 부여하면 학습 알고리즘은 상대적으로 이와 유사한 예제들의 카테고리를 해당 예제와 동일한 카테고리로 보다 신뢰성 있게 추정할 수 있다. 따라서, 자신과 유사한 예제들이 많을수록 학습에 유용한 예제로 예상할 수 있다. 두 번째는 유용한 훈련집합의 예제들은 충분히 서로 달라서 선정된 집합외의 많은 예제들에 대하여 학습 알고리즘이 신뢰성 있게 카테고리를 추정할 수 있게 한다.

위와 같은 가정하에서 유사한 예제들끼리 무리지어 군집간의 차별성을 최대화한 후 각 군집별로 하나씩 예제를 선정하는 휴리스틱적 접근방법으로 초기훈련집합을 구성할 수 있다. 따라서 먼저 군집화 기법을 적용하여 유사한 예제들을 묶고, 각 군집을 대표할 수 있는 예제를 선별한 후, 사용자에게 문의하여 집합을 구성하면 능동적 학습의 초기학습에 유용한 초기훈련집합을 얻을 수 있을 것이다. 각 군집을 대표하는 예제로는 자신이 포함된 군집내 다른 예제들과 평균적으로 가장 유사한 예제가 적절할 것이며, k-means의 경우 군집의 중심점이 이러한 예제에 해당된다.

이와 같은 방법으로 초기훈련예제를 선정한다면 동일한 수의 훈련예제를 임의로 선정하는 방안에 비해 학습에 보다 효과적일 것이다. 또한 각각의 군집을 대표하는 훈련예제는 다른 군집(경우에 따라서는 다른 카테고리)과의 경계선 상에 위치한 예제들에 비해 사용자가 카테고리를 부여하기 다소 수월한 전형적인 예제일 가능성이 높을 것이다.

군집화 기법을 이용하여 초기훈련예제를 선별하는 알고리즘의 복잡도(complexity)는 다음과 같이 추정할 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 반복 횟수를 t라고 하면 복잡도는 $O(nkwt)$이다.[4] 본 연구에서 실험한 결과로는 n이 3,000인 경우, 50개의 초기훈련예제를 선별하는데

알고리즘: 군집 기반 선정 (cluster-based sampling)

입력

k 초기훈련예제로 사용할 예제의 수
$\mathcal{D}$ 카테고리가 부여되지 않은 예제집합

출력

$\mathcal{T}_0$ 초기훈련집합

방법

1. $\mathcal{T}_0$를 Ø로 초기화한다.
2. 군집화 기법을 $\mathcal{D}$에 적용하여 서로 겹치지 않는 군집의 집합을 생성한다.
   X_i (i = 1, ···, k), $\mathcal{D}$ = $\cup X_i$
3. 각각의 군집 X_i에 대하여
   3a. X_i의 대표예제 r_i를 선정한다.
   (k-means의 경우 X_i의 중심점에 가장 가까운 예제를 대표예제로 선정한다)
   3b. r_i를 $\mathcal{T}_0$에 추가한다.
4. $\mathcal{T}_0$를 구성하는 모든 예제를 사용자에게 문의하여 각 예제의 카테고리를 부여 받는다.
5. 각각의 군집 X_i에 대하여
   5a. X_i의 모델예제 m_i를 생성한다.
   (k-means의 경우 X_i의 중심점을 모델예제로 사용한다)
   5b. m_i의 카테고리를 r_i의 카테고리로 설정한다.
   5c. m_i를 $\mathcal{T}_0$에 추가한다.

그림 3 군집화 기반 초기훈련집합 선정 알고리즘

반복 횟수 t는 수십회 이내였으며, 시간상으로는 50개를 선별하는 경우에도 수분[5]을 넘지 않아 실용성이 있음을 확인할 수 있었다. 그림 3에는 본 논문에서 제안하는 군집화 기법을 이용하여 초기훈련집합을 선정하는 방안을 정리하였다.

3. 실험 결과

이상에서 제안한 방안의 효과를 확인하기 위하여 몇 가지 문서 분류 문제를 대상으로 그 성능을 실험하였다. 실험에는 문서 분류 연구에 자주 사용되는 Reuters-21578 신문기사 말뭉치와 Newsgroups-20 USENET 뉴스 기사 모음을 활용하였다[3]. Reuters-21578 말뭉치는 1987년부터 1991년 사이에 생성된 Reuters사의 경제 기사 21,578건으로 이루어져 있다. 이들 문서 중에서 주제를 기준으로 단일 카테고리만 부여된 문서들 중 빈도수로 상위 2개의 주제에 해당되는 문서들을 추출하여 실험에 활용하였다. Newsgroups-20는 20가지 USENET 뉴스 그룹에 올려졌던 약 20,000건의 기사로 이루어져 있으며 군집화(또는 분류)의 난이도에 따른 본 제안 방안의 효과를 확인하기 위하여 [5]에서 실험한 바와 같이 주제가 상당히 다른 세 개의 뉴스 그룹 Different-3(alt.atheism, rec.sport.baseball, sci.space)와 주제가 매우 유사한 3개의 뉴스 그룹 Same-3(comp.graphics, comp.os.ms-windows, comp.windows.x)의 기사를 사용하였다. 모든 문서는 USENET 헤더 제거(Differnet-3와 Same-3의 경우), SGML 태그 제거(Reuters의 경우), 표준형 변환(stemming)과 불용어 제거(stop word removal) 과정을 거쳐 실험에 사용할 데이타로 구축하였다. 표 1은 실험에 사용한 문서 데이타의 특성을 나열하고 있다.

학습기법으로는 여러 문서 분류 연구[6]에서 그 효과가 입증된 기법의 하나인 k-NN (k-nearest neighbor) 알고리즘을 적용하였으며, 예제간 유사정도 비교척도로는 정보검색분야에서 일반적으로 활용하는 $tf \times idf$ 벡터

4) k-means 군집화에 가장 많은 시간이 소요되는 군집의 중심점과 개별 문서간의 유사정도를 계산하기 위하여 모든 단어를 고려할 필요는 없다. $tf \times idf$ 공간에 코사인 유사도(cosine similarity)를 이용하는 경우 해당 문서에 등장하는 단어에 대해서만 계산하면 되므로 복잡도는 문서의 평균 단어수 w에 비례한다.

5) 2GHz로 동작하는 UNIX기반의 PC 상에서 실험하였다.

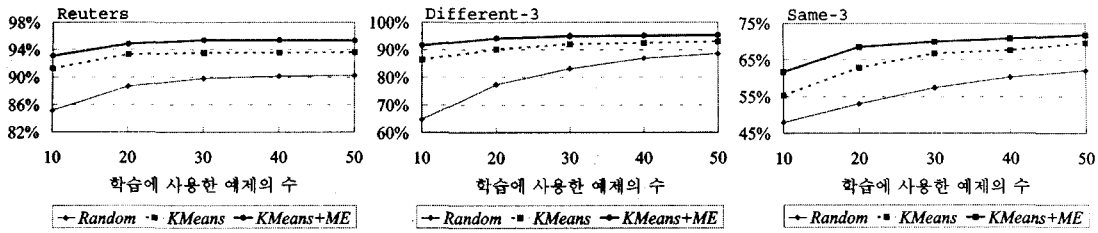


그림 4 다양한 크기의 초기훈련집합으로 학습한 분류기의 초기 성능

표 1 실험 문서 데이타의 특성

문서 데이타 이름	카테고리 수	문서 수	난이도	카테고리 분포
Reuters	2	5,627	쉬움	불균형 (66% : 34%)
Different-3	3	3,000	쉬움	균형
Same-3	3	2,999	어려움	균형

공간상의 코사인 유사도[7]를 사용하였다. 문제예제의 카테고리 예측은 문제예제와 가장 유사한 k개의 훈련예제의 카테고리를 그 유사정도에 따라 가중 평균을 취하여 투표방식으로 결정하였다. 초기훈련예제 선정을 위한 군집화 기법은 k-means 알고리즘을 사용하였으며 k-NN과 동일한 유사도 척도를 적용하였다.

훈련예제를 임의로 선정하는 방안(*Random*), 본 논문에서 제시하는 군집화 후 군집의 대표예제로 초기훈련집합을 구성하는 방안(*KMeans*) 그리고 군집의 대표예제와 군집의 모델예제를 함께 사용하는 방안(*KMeans+ME*) 이 세 가지 방안을 비교 실험하였다. 모든 실험결과치는 초기훈련집합의 크기 및 방법 별로 10 분할 상호검증(10 fold cross-validation)을 10회씩 수행하여 결과의 평균을 취하였다.

그림 4는 각 방안으로 훈련예제를 최소 10개에서 최대 50개까지 선정하였을 때 선정된 예제들만으로 학습한 분류기의 정확도를 보이고 있다. 이 실험은 제안 방안에 의하여 선정된 예제가 학습에 얼마나 효과적인지를 확인하기 위하여 수행하였다. k-NN 알고리즘은 k값에 따라 성능의 편차가 있으므로 적절한 값으로 설정할 필요가 있다. 본 실험에서 k값은 5로 고정하였다.[6] 그림에서 각 그래프의 세로축은 분류의 정확도를 백분율로 나타낸다. 보다 많은 수의 예제를 초기훈련예제로 선정할수록 분류의 정확도는 자연히 증가한다. 세 가지 문서 분류 문제 모두 본 논문에서 제안한 *KMeans*와 *KMenas+ME* 방안으로 선정한 초기훈련집합을 이용하여 학습한 경우가 *Random* 방안으로 선정한 집합을 이용하는 경우에 비해 분류의 정확도가 현격히 높은데, 이는 본 논문에서 제안한 방안이 학습에 유용한 예제를 효과적으로 선별할 수 있기 때문이다. *KMeans+ME* 방안으로 생성한 10개[7]의 훈련예제는 *Random* 방안으로 생성한 50개의 훈련예제와 학습에서의 효과가 비슷함을 알 수 있으며, *KMeans+ME* 방안에서 훈련예제로 생성한 모델예제가 문제의 난이도에 관계없이 학습에 긍정적인 효과를 미치고 있음을 확인할 수 있다.

그림 5와 그림 6은 초기훈련예제로 10개 또는 30개[8] 예제를 선정한 후 능동적 학습을 적용한 실험의 결과로 초기훈련예제를 포함하여 최대 50개까지 예제를 문의할 수 있다고 가정하였다. 능동적 학습 시에는 카테고리가 부여되지 않은 예제들 중에서 예측한 카테고리들의 분포가 가장 모호한(uncertain)[9] 예제를 하나씩 문의하고 재학습하는 과정을 반복하였다. k-NN의 k값은 앞의 실험에서와 같이 5로 고정하였다.

훈련예제가 추가됨에 따라 성능의 격차는 점차 줄어드는데 이는 능동적 학습이 진행되어가면서 나타나는 자연스러운 현상이라 할 수 있다. 하지만 상대적으로 본 논문에서 제안한 *KMeans*와 *KMeans+ME* 방안으로 생성한 초기훈련집합에서 시작한 능동적 학습의 성능이 *Random* 방안으로 생성한 초기훈련집합에서 시작한 경우에 비해 지속적으로 우위에 있음을 확인할 수 있다. 이는 초기훈련집합이 능동적 학습의 성능에 큰 영향을 미치며, 본 논문에서 제안한 방안이 유용한 초기훈련집합을 선정할 수 있음을 입증하는 결과라 하겠다.

그림 5와 그림 6의 Same-3 데이타를 이용한 실험결과를 자세히 살펴보면 독특한 현상을 발견할 수 있다. 그림 7에 이를 보다 자세히 나타내었다. *KMeans+ME*

6) 값을 5로 둔 이유는 분류기의 정확도가 대체적으로 k값이 5일 때 안정적이며 우수하였고, 이 수치가 학습예제의 수 (10~50)와 카테고리의 수 (2~3)를 고려할 때 적절하다고 판단하였기 때문이다. 다른 k값을 사용한 경우에도 상대적인 경향은 그림 4의 경우와 동일하였다.

7) *KMeans+ME* 방안은 이 경우 20개의 예제로 초기훈련집합을 구성한다. 10개의 대표예제는 사용자에 의하여 카테고리가 부여되며, 10개의 모델예제는 대표예제에 의하여 자동으로 카테고리가 부여된다.

8) 초기훈련예제로 20개 또는 40개의 예제를 선정한 경우도 유사한 양상이 나타났다.

9) 본 연구에서는 k-NN 분류기가 가장 확신 있게 예측한 카테고리에 대한 투표 가중치와 나머지 카테고리들에 대한 투표 가중치의 합간의 차이를 사용하였다.

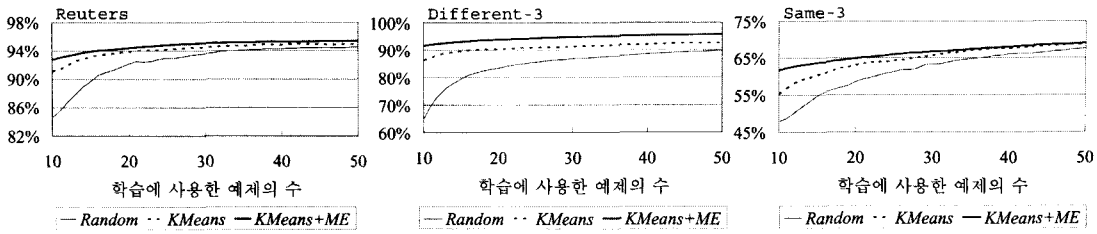


그림 5 10개의 초기훈련예제를 이용한 능동적 학습의 성능

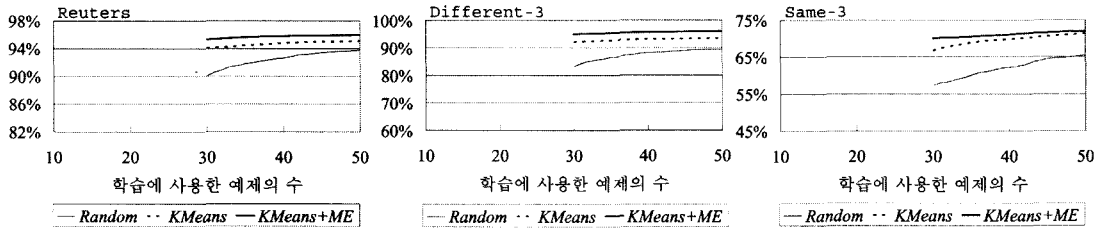


그림 6 30개의 초기훈련예제를 이용한 능동적 학습의 성능

에 의하여 생성된 초기훈련집합을 사용한 실험에서, 30개의 초기훈련예제로 시작한 능동적 학습이 10개의 초기훈련예제로 시작한 경우보다 총 50개의 훈련예제를 사용한 후 보다 높은 정확도를 달성하였다. 이 결과는 *KMenas+ME(30)*에 의하여 선정된 30개의 초기훈련예제가 *KMeans+ME(10)*과 능동적 학습에 의하여 선정된 30개의 훈련예제보다 학습에 효과가 더 높았음을 의미하는 것이다. 이 현상은 다음과 같이 설명될 수 있다. 본 논문에서 제안하는 방안은 군집화에 기반하므로 유사한 예제들이 많은 예제가 초기훈련예제로 선정될 가능성이 높다. 이에 비해 실험에서 문의 예제를 선정하는 기준으로 활용하였던 추정 카테고리의 불명확성은 모호한 정도는 높으나 유사한 예제들이 적어 학습에 끼치는 영향은 낮은 예제들도 선정될 가능성이 높다. 또한, Same-3 데이타는 분류하기가 까다로운 문제에 속하는

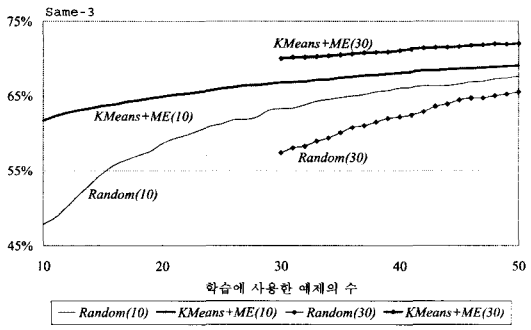


그림 7 Same-3 문제에서 10개 또는 30개의 초기훈련 예제로부터 시작한 능동적 학습의 성능

데 이는 유사한 예제들이 많아 능동적 학습의 관점에서는 비슷하게 모호한 예제들이 많다는 점도 그 이유의 하나가 될 수 있을 것이다.

4. 관련연구

능동적 학습과 관련한 기존 연구들은 문의할 예제를 효과적으로 선정하는 방법을 주로 탐구해 왔다. 대표적인 문의예제 선정방법으로 현재 학습된 분류기로 카테고리 추정이 가장 모호한 예제를 선정하는 불명확성 기반 선정방법(uncertainty sampling)[1]과 카테고리를 밝혔을 때 훈련예제들에 부합하는 가설(hypotheses)들을 가능한 절반에 가깝게 배제할 수 있는 예제를 선정하는 위원회 기반 선정방법(committee-based sampling)[8-10]을 들 수 있다. 위원회 기반 방법의 하나로 Abe와 Mamitsuka[10]는 동일한 훈련집합을 기반으로 복수개의 서로 다른 분류기를 생성하기 위하여 bagging과 boosting기법을 능동적 학습에 적용하였다. 이 방법은 생성된 분류기들간의 의견이 가장 불일치한 예제를 문의예제로 선정한다.

Muslea, Minton 그리고 Knoblock[11]은 Blum과 Mitchell[12]이 제안한 협동학습(co-training)기법을 문의예제 선정에 응용한 협동검사(co-testing)기법을 제안하였다. 협동검사는 예제들을 기술하는 속성들의 집합을 서로 겹치지 않는 두 개의 부분집합(view)으로 나누고, 각 부분집합만을 이용하여 분류기를 생성한다. 문의예제로는 두 분류기간의 예측이 가장 불일치한 예제를 선정한다. 이후 연구에서 이들은 능동적 학습의 문의단계 사이에 Co-EM[13]을 활용하고 문의예제는 협동검사 기법

을 적용한 Co-EMT[14]를 제안하였다. 이들은 제안한 Co-EMT가 다양한 실험 환경에서 협동검사를 포함한 여러 준감독(semi-supervised) 학습 기법에 비해 안정적(robust)인 성능을 발휘할 수 있음을 보였다.

휴리스틱적인 접근 이외에 최적의 문의 예제를 선정하는 방안에 관한 보다 이론적인 접근으로는 [2][15]이 있다. Roy와 McCallum[2]은 예제에 카테고리가 부여되었을 때 예상되는 전체 오류가 가장 낮을 것으로 추정되는 예제를 문의예제로 선정하는 방안을 제시하였다.

능동적 학습과 관련한 기존 연구들은 임의로 선정한 초기훈련예제들로 초기학습을 수행한 후 문의예제를 하나씩 선정하는 방안에 주된 초점을 맞추어 왔다. 군집화 기법의 하나인 준감독 EM[13]을 능동적 학습과 병행함으로써 보다 성능을 개선한 연구들[14,16]이 수행되었지만 이들 역시 초기학습 이후 단계를 염두에 두었다.

능동적 학습을 위하여 복수의 문의예제를 선정(batch query selection)할 수 있는 효율적인 방안에 대한 연구가 수행된 것은 최근의 일이다. Brinker는 SVM을 이용한 능동적 학습에서 다양성과 모호성을 동시에 고려하는 방안을 제안한 바 있다[3]. SVM은 그 정확도는 높으나 학습에 소요되는 계산 비용이 매우 높은 학습 알고리즘이다. 이 연구는 문의예제를 복수로 선정함으로써 능동적 학습에 소요되는 시간을 줄일 수 있고, 동일한 수의 복수 문의예제를 선정할 경우에는 다양성을 추가로 고려함으로써 모호성만을 이용하는 경우보다 학습을 효율적으로 수행할 수 있음을 보였다. 하지만 이 연구는 기본적으로 SVM을 이용한 학습 자체에 소요되는 시간을 줄이는데 초점을 맞추었으며, 제안된 방안은 SVM을 분류기로 사용할 경우에 적합한 방법이다. 초기훈련예제는 일반적인 능동적 학습에 관한 연구에서와 같이 임의로 선정하는 방법을 사용하였다.

예제들을 선별한다는 측면에서 본 논문과 연관성을 가진 다른 연구 분야로는 간결한 부분훈련집합 선정(concise subset selection)에 관한 연구와 데이타 축약(data condensation, reduction)에 관한 연구가 있다. 간결한 부분훈련집합 선정은 주어진 훈련집합에서 동일한 학습 효과를 낼 수 있는 가능한 작은 크기의 부분집합을 선택하는 것이다[17,18]. 유사한 예제들을 많이 수집할 수 있으며 인공신경망과 같이 학습에 소요되는 시간이 상당한 기법을 사용하는 음성인식이나 영상인식과 같은 분야에서 주로 연구되었다. 일부 연구에서는 능동적 학습 방식을 응용하여 부분집합을 선정하였다. 이 분야의 연구들은 훈련예제에 카테고리가 미리 부여된 상황에서 출발하기 때문에 본 연구와 근본적인 차이가 있다.

데이타마이닝과 같은 분야에서는 마이닝을 위하여 수집한 데이타의 용량이 수십 기가바이트(giga-bytes) 이상 될 수 있다. 방대한 양의 데이타를 모두 활용하여 마이닝 기술을 적용하는 것은 저장공간이나 수행시간 측면에서 비효율적이며 경우에 따라서는 적용이 불가능할 수 있다. 데이타 축약은 이러한 상황에서 마이닝 기술을 효율적으로 적용할 수 있도록 원본 데이타의 특성을 가능한 유지할 수 있는 일부의 데이타를 선택하는 것이다[19,20].

데이타 축약과 관련한 많은 연구는 간결한 부분훈련집합 선정 연구와 마찬가지로 예제에 카테고리가 미리 부여된 상황에서 출발하므로 본 연구와 근본적으로 차이가 있다. 최근 Mitra, Murthy 그리고 Pal[20]은 카테고리 정보를 사용하지 않는 밀도기반의(density-based) 데이타 축약 방안을 제안하였다. 이 방법은 다음과 같이 동작한다. (1) 모든 예제에 대하여 k번째 가장 가까운 이웃과의 거리를 측정하여 그 거리가 가까울수록 해당 위치에서의 밀도가 높은 것으로 추정한다. (2) 밀도가 가장 높은 예제부터 순차적으로 선택하되, 이 때 선택된 예제와 k번째 가장 가까운 이웃과의 거리를 r이라 한다면, 해당 예제와 $2r$ 범위 이내의 예제들을 함께 제거한다. (3) 모든 예제들이 선택 또는 제거될 때까지 (2)의 과정을 반복한다. 이들은 이렇게 추출한 대표예제들이 다른 데이타 축약 기법들을 이용하여 훈련예제를 추출하는 경우에 비해 여러 기계학습 응용에 보다 효율적으로 사용될 수 있음을 보였다. 하지만 이들에 의해 제안된 방법 역시 대부분의 데이타 축약연구와 마찬가지로 학습기법을 적용할 수 있는 수준까지 데이타를 축약하는 것이 주된 목적이므로, 최소한의 예제만을 사용하려는 능동적 학습과는 접근관점에서 차이가 있다. 알고리즘의 특성상 원하는 예제의 축약 비율(또는 선별할 예제의 수)을 얻기 위해서는 여러 번의 실험을 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 논문에서 제안한 모델예제와 같은 개념은 제시되지 않았다.

Shih, Rennie, Chang 그리고 Karger[21]가 최근 연구에서 문서 데이타 축약 방안으로 제안한 문서 뭉치화(text bundling) 기법은 본 연구에서 제안한 모델예제와 유사한 특성을 지니고 있다. 이들은 동일한 카테고리에 속하는 유사한 문서들을 묶어 하나의 가상 문서로 변환하는 방법으로 데이타를 축약하였다. 하지만 문서 뭉치화 기법은 SVM과 같이 수행시간이 상당한 학습기법을 위하여 데이타를 축약하는데 그 목적이 있으며, 많은 데이타 축약 연구에서와 같이 예제의 카테고리 정보가 주어진 상황에서 각 카테고리 별로 문서 뭉치화를 수행한다는 점에서 본 논문에서 제시한 모델예제와 근본적으

로 다르다.

5. 결론 및 향후 과제

본 논문에서는 유사한 예제들을 군집화한 후 각 군집을 대표할 수 있는 대표예제와 모델예제를 추출하여 초기훈련집합을 구성함으로써 능동적 학습의 성능을 보다 향상시킬 수 있는 방안을 제안하였다. 여러 문서 분류 문제에 적용하여 실험한 결과 제안 방안으로 선정한 초기훈련집합에서 출발한 능동적 학습이 임의로 선정한 초기훈련집합에서 출발한 경우에 비해 보다 적은 수의 예제로 동등한 성능을 달성할 수 있음을 확인하였다.

향후 본 제안 방안을 확장하여 능동적 학습의 복수 문의예제선정에 적용할 수 있는 방안에 관한 연구와 문서 분류 이외의 다양한 분류 문제에 적용하여 그 효과를 검증하는 연구가 수행되어야 할 것이다.

참 고 문 헌

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

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

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

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

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

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

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

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

[9] Freund, Y., Seung, H. S., Shamir, E. and Tishby. N., "Selective sampling using the query by committee algorithm," *Machine Learning*, Vol. 28, Nos. 2-3, pp. 133-168, 1997.

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

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

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

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

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

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

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

[17] Plutowski, M. and White, H. "Selecting Concise Training Sets from Clean Data," *IEEE Trans. Neural Networks*, Vol. 4, No. 2, pp. 305-318, 1993.

[18] Jung, G. and Opper, M. "Selection of examples for a linear classifier," *Journal of Physics* A, 29, pp. 1367-1380, 1996.

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

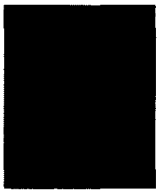
[20] Provost, F., and Kolluri, V., "A survey of methods for scaling up inductive algorithms," *Data Mining Knowledge Discovery*, Vol. 2, pp. 131-169, 1999.

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

강 재 호

1995년 부산대학교 컴퓨터공학과 학사 1997년 부산대학교 컴퓨터공학과 석사 1997년 3월~현재 부산대학교 컴퓨터공학과 박사과정. 1999년 3월~1999년 12월 해동EMC 연구원. 2000년 2월~현재 동아대학교 지능형통합항만관리연구센터 연구원. 관심분야는 인공지능, 기계학습, 정보검색, 데이타마이닝, 최적화 등

류 광 렬
정보과학회논문지 : 소프트웨어 및 응용
제 31 권 제 2 호 참조



권 혁 철
1982년 서울대학교 공과대학 전산학 학사. 1984년 서울대학교 공과대학 전산학 석사. 1987년 서울대학교 공과대학 전산학 박사. 1988년~현재 부산대학교 정보컴퓨터 공학부 교수. 1988년~현재 한국정보과학회 프로그래밍언어 연구회 운영위원. 1990년~현재 한국정보과학회 한국어정보처리 연구회 운영위원. 1992년~1993년 미국 Stanford 대학 CSLI연구소 연구원. 1992년~1993년 Xerox Palo Alto Research Center 자문위원. 2003년~현재 BK21 산업자동화 및 정보통신 분야 인력양성사업단 단장. 2003년~현재 한국정보과학회 한국어정보처리연구회 위원장. 관심분야는 한국어 정보처리, 정보검색, 프로그래밍언어, 인공지능