군집화 기법을 이용한 능동적 학습의 최초학습예제 선정[1]

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

Selecting Initial Training Set for Active Learning by Clustering

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

요 약

기계학습의 분류(classification) 기술을 실제 문제에 적용하기 위해서는 카테고리(category)를 부여한 학습예제를 상당수 준비하여야 한다. 예제에 카테고리를 부여(labeling)하는 작업에는 무시할 수 없는 시간과 인력을 필요로 한다. 능동적 학습(active learning)은 동일한 수의 학습예제로 최대한의 성능을 달성하기 위하여 카테고리를 부여할 학습예제를 선별하는 전략이다. 능동적 학습은 현재까지 파악된 정보에 기반하여 분류기(classifier)를 생성하고, 생성된 분류기를 활용하여 카테고리를 부여 받았을 때 가장 이득이 큰 예제들을 선정하여 사용자에게 문의하는 과정을 반복하여 수행한다. 만일 능동적 학습의 첫 학습단계에서 학습에 보다 유용한 예제들을 최초학습예제집합으로 선정한다면 같은 수의 학습예제로 더 나은 성능을 달성할 수 있을 것이다. 본 논문에서는 유사한 예제들은 동일한 카테고리에 속할 가능성이 높다는 일반적인 가정에 기반하여 예제들을 군집화(clustering)한 후, 생성된 각 군집을 대표할 수 있는 예제로 최초학습예제집합으로 구성하는 방안을 제안한다. 제안한 방안을 문서분류 문제를 대상으로 실험해 본 결과 최초학습예제들을 임의로 선정하는 방식보다 정확도가 높은 분류기를 생성할 수 있음을 확인하였다.

1. 서론

기계학습의 분류기술을 실제 문제에 적용하기 위해서는 카테고리가 부여된 학습예제를 상당수 준비하여야 한다. 예제에 카테고리 부여작업에는 무시할 수 없는 그리고 응용분야에 따라서는 막대한 시간과 인력을 필요로 한다. 능동적 학습은 동일한 또는 제한된 수의 학습예제로 최대한의 성능을 달성하기 위하여 카테고리를 부여할 학습예제를 선별하는 전략이다[1][2].

능동적 학습기법은 사용자가 답변할 수 있는 최대예제개수에 도달할 때까지 학습단계와 문의단계를 반복적으로 수행한다[2]. 학습단계에서는 현재 보유한 학습예제집합에 학습 알고리즘을 적용하여 분류기를 생성한다. 문의단계에서는 생성된 분류기를 이용하여 카테고리가 부여되지 않은(unlabeled) 예제들을 분류해보고, 이들 예제 중에서 가장 학습에 효과가 높을 것으로 추정되는 예제들을 선정하여 사용자에게 카테고리 부여를 요청한다. 문의단계에서 사용자에 의해 카테고리가 부여된 신규학습예제들은 기존의 학습예제집합에 추가된다.

능동적 학습과 관련한 기존 연구들은 신규학습예제를 선정하는 방안들을 주로 탐구해 왔으며 최초의 학습단계에 필요한 예제들은 임의로 선정하였다. 어떠한 예제들을 최초학습예제집합으로 선정하는 것이 향후 능동적 학습에 보다 유리한지에 대한 연구결과는 제시되지 않았다. 극소수의 예제들로 최초학습예제집합을 구성하고 이후 능동적 학습기법을 적용하는 단순한 전략은 경우에 따라 비효율적일 수 있다. 일반적으로 사용자는 가능한 한번에 많은 수의 예제들을 처리하는 것을 하나씩 처리하는 것보다 선호하며 작업을 보다 효율적으로 처리할 수 있다. 또한, 카테고리 부여작업을 수행할 수 있는 인력이 충분하다면 동시에 많은 수의 예제들에 대한 카테고리 부여작업을 생성하여 병렬로 처리하는 것이 전체 시스템 구축에는 유리할 것이다.

동일한 수의 최초학습예제들을 활용하여 더 높은 정확도를 가진 분류기를 생성할 수 있다면, 계속되는 능동적 학습과정에서도 상대적인 성능의 우위를 유지할 수 있을 것이므로 최초학습예제의 선정은 중요한 문제이다.

능동적 학습을 위한 학습예제들을 선정하는 연구와 유사한 연구로는 데이터 마이닝(data mining) 분야의 데이터 축약(data condensation)이 있다. 마이닝을 위하여 수집한 데이터의 용량은 수십 기가바이트 이상 될 수 있으며, 이들 데이터를 모두 활용하여 기계학습을 적용하는 것은 저장공간이나 수행시간 측면에서 비효율적이다. 데이터 축약은 이러한 상황에서 기계학습을 보다 효율적으로 적용할 수 있도록 원본 데이터의 특성을 가능한 반영하는 부분집합을 생성하는 연구이다[3][4].

데이터 축약과 관련한 기존 연구들은 카테고리가 미리 부여된 상황을 가정한다. 일부 연구에서는 이러한 가정이 필요 없는 밀도기반(density-based) 데이터 축약 방안을 제안하였다[4][5]. 데이터 축약 연구들에서 축약 결과를 기계학습에 적용하여 그 성능을 평가한 바 있지만 능동적 학습에 미치는 효과를 분석한 연구는 아직 제시되지 않았다.

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

[2] 엄밀하게는 능동적 학습 중에서도 선별적 표본추출 (selective sampling)을 의미한다. 본 논문에서는 용어의 이해도 측면에서 능동적 학습으로 통일하여 기술하였다.

본 논문에서는 유사한 예제들은 동일한 카테고리에 속할 가능성이 높다는 일반적인 가정에 기반하여 예제들을 군집화한 후, 생성된 각 군집을 대표할 수 있는 예제들로 최초학습예제 집합으로 구성하는 방안을 제안한다.

본 논문의 구성은 먼저 2장에서 군집화 기법을 이용한 최초 학습예제집합 선정 방안에 대하여 자세히 설명한다. 이어지는 3장에서 로이터(Reuter) 말뭉치를 대상으로 본 접근방안을 문서분류 문제에 적용한 실험결과를 분석하고 4장에서 결론 및 향후 연구과제를 제시한다.

2. 군집화 기법을 이용한 최초학습예제 선정 방안

본 장에서는 군집화 기법을 이용한 최초학습예제 선정방안에 대하여 기술한다. 본 논문에서 제시하는 방안의 효과를 설명하기 위하여 간단한 예를 들고자 한다.

그림 1에는 이차원 상에 분포된 예제들을 보이고 있다. 각각의 예제들은 두 가지 카테고리 A, B 가운데 하나에 속한다. 더하기(+) 기호로 표시된 예제들은 카테고리 A에 속하며, 빼기(-) 기호로 그려진 예제들은 카테고리 B에 해당된다. 물음표(?) 모양의 예제들은 사용자가 카테고리를 부여하지 않은 예제들이다. 점선으로 표기된 타원형은 각각 카테고리 A와 B의 공간을 나타낸다. 기계학습 알고리즘에 카테고리가 부여된 예제들을 충분히 제공한다면 두 카테고리를 높은 정확도로 구분할 수 있는 분류기를 얻게 될 것이다. 하지만 수집된 예제들을 학습에 사용하려면 카테고리를 부여하는 작업이 선행되어야 하고 이러한 작업에는 상당한 시간과 인력이 소요되기 때문에 가능한 여건이 허용하는 한도 내에서 많은 학습예제를 제공하고자 할 것이다. 그림 1은 임의로 선정한 2개의 학습예제로 기계학습을 수행한 경우이다. 본 예에서는 최근접 이웃 찾기 알고리즘 (nearest neighbor algorithm)을 가정하였다. 최초학습예제를 임의로 선정하였을 때 학습예제들의 위치가 효과적이지 못한 경우 그림에서와 같이 개념(concept)을 온전히 학습하지 못하여 일부 예제들이 잘못 분류되는, 즉 정확도가 떨어지는 결과가 나타날 수 있다.

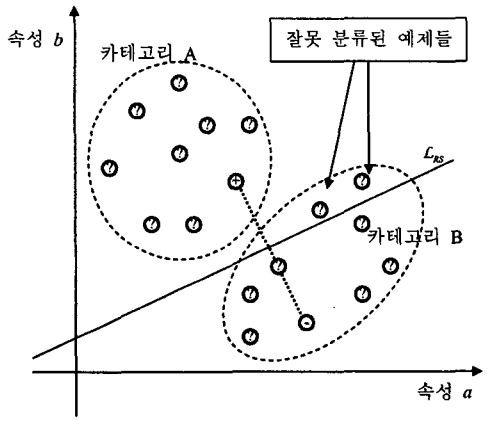

그림 1. 임의 선정된 학습예제들을 이용하여 학습한 경우

그림 2는 본 논문에서 제시하는 군집화 기법을 이용한 최초학습예제 선정방안을 설명하고 있다. 그림 1과 동일한 예제들(아직 사용자가 카테고리를 확인해 주지 않은 상태임)을 각 군집의 대표예제(*k*-means 군집화 기법의 경우 최초중심점)로 삼아 군집화를 수행한다. 본 예에서는 *k*-means 알고리즘을 가정하였다. 최종적으로 안정화된 각 군집은 유사한 예제들의 모음이므로 하나의 군집 내 예제들은 동일한 카테고리에 속할 가능성이 높다. 학습을 위한 최초학습예제들로 이들 각 군집을 가장 잘 표현할 수 있는 예제를 대표로 선정하여 사용자에게 제시한다.

그림 2와 경우에는 각 군집의 중심점과 가장 가까운 예제들을 선정하여 이를 사용자에게 최초학습예제로 추천하였다. 사용자가 최초학습예제로 추천된 이 두 예제에 카테고리를 부여하면, 앞에서와 동일한 학습 알고리즘을 적용하여 학습을 수행할 수 있으며, 그림의 예에서는 학습의 결과가 보다 안정적임을 알 수 있다. 알고리즘 1에 본 제안방법을 정리하였다.

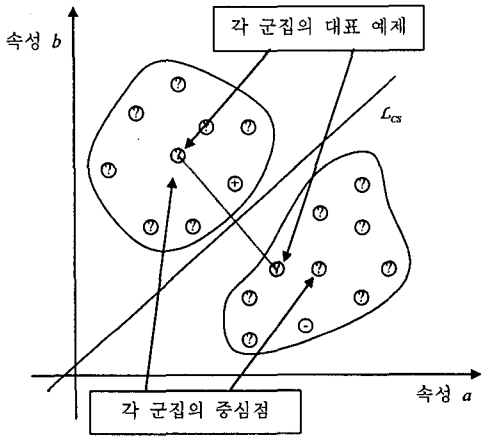

그림 2. 군집화 후 생성된 대표예제들로 학습한 경우

1. 전체 예제집합 E 중에서 t 개의 예제를 임의 선정하여 S_0생성
2. S_0 집합의 각 예제를 최초중심점으로 *k*-means 군집화를 수행
3. 생성된 군집들을 대표할 수 있는 예제집합으로 각 군집의 중심점과 가장 유사한(가까운) 예제들로 S_{final}로 생성
4. S_{final}의 예제들에 대한 카테고리를 사용자에게 문의한 후 최초학습예제 집합 T_0 생성
5. T_0를 기반으로 능동적 학습 수행

알고리즘 1. 군집화를 이용한 학습예제 선정 방안

3. 실험 결과

이상에서 제안한 방안의 효과를 확인하기 위하여 문서분류 문제를 대상으로 그 성능을 실험하였다. 실험에는 문서 분류 연구에 자주 사용되는 Reuters-21578 신문기사 말뭉치[6]를 활용하였다. Reuters-21578 말뭉치는 1987년부터 1991년 사이에 생성된 로이터사의 경제기사 21,578건으로 이루어져 있다. 이들 문서 중에서 주제를 기준으로 단일 카테고리만 부여된 문서들만 우선 추렸다. 로이터 문서들은 각 카테고리별로 문서수의 편차가 심하므로 등장 빈도수로 상위 10개 카테고리에 속하는 문서들 6,744건만 실험대상으로 활용하였다. 선정된 신문기사는 불용어처리(stop word removal)와 표준형 변환(stemming)을 거쳐 실험에 사용할 데이터로 구축하였다.

학습기법으로는 *k*-NN (*k* nearest neighbor) 알고리즘을 적용하였으며 예제간 비교척도로는 정보검색분야에서 일반적으로 활

용하는 코사인 유사도를 사용하였다[8]. 카테고리 예측은 문제 예제와 가장 유사한 k개의 학습예제를 그 유사정도에 따라 가중 평균을 취하였다. 군집화 기법은 k-means 알고리즘을 사용하였으며 역시 동일한 유사도 기준을 적용하여 실험하였다. k-NN 학습 알고리즘의 경우 학습예제의 수에 따라 가장 효과적인 k 값이 변화하기 때문에 이에 대한 최적화가 필요하며, 본 실험에서는 10번의 독립적인 반복실험 후 각 학습예제집합의 크기 별로 가장 우수한 성능을 나타낸 k 값을 사용하였다.

실험에서 초기학습예제집합을 구성하기 위하여 임의로 예제들을 선정하는 방안(RS)과 본 논문에서 제시한 군집화 후 대표성 있는 예제들을 선정하는 방안(CS)을 비교하였다. 먼저 각 방안으로 선정한 학습예제들의 질을 평가하기 위하여 선정된 학습예제들만으로 분류기를 생성하고 그 성능을 비교하였다. 실험은 학습예제집합의 크기별로 10회 반복 실험하여 그 평균을 취하였다.

그림 3은 각 학습예제 수에 따른 두 가지 학습예제 선정방법의 평균적인 성능을 보여주고 있다. 본 논문에서 제안하는 군집화 기반 최초학습예제 선정 방안이 동일한 수의 학습예제를 이용한다면 임의학습예제 선정 방안보다 더 우수한 정확도를 가진 분류기를 생성할 수 있음을 쉽게 알 수 있다.

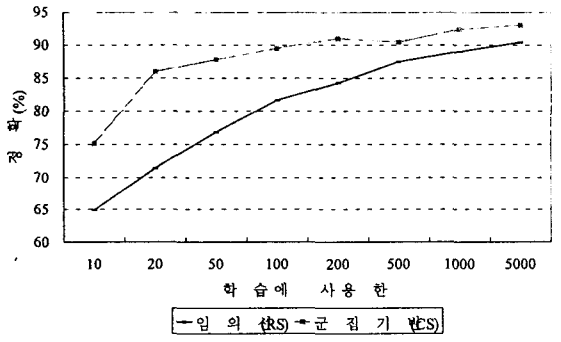

그림 3. 임의 및 군집기반 학습예제 선정 방안의 성능

그림 4에는 최초학습예제집합을 선정한 후 능동적 학습을 수행하였을 때 정확도의 변화를 보여주고 있다. 최초학습예제를 각각 10개, 20개, 50개씩 먼저 선정한 후 최초학습예제들을 포함하여 최대 200개까지 사용자에게 문의할 수 있다고 가정하였다. 능동적 학습의 문의단계에서는 상위 두 카테고리가 가장 모호하게 예측되는 예제 하나를 문의 대상으로 선정하였다. 능동적 학습의 효과로 두 가지 방안 모두 능동적 학습을 적용하지 않은 경우(그림에서 직선으로 표기)에 비해 성능의 향상을 이루었다. 추가적으로 학습예제들을 요청해 나감에 따라 두 방안의 정확도 차이는 점차 줄어들지만 상대적으로 본 논문에서 제안한 군집기반 선정방안과 능동적 학습을 결합한 방안이 더 나은 성능을 보였다. 특히 보다 많은 최초학습예제들을 사용하는 경우 최초의 정확도의 차이가 쉽사리 줄어들지 않는데, 이는 최초학습예제집합을 선정하는 문제가 중요함을 말해준다.

4. 결론 및 향후 연구

본 논문에서는 능동적 학습 시 보다 효과적인 학습이 가능하도록 효과적인 예제들로 최초학습예제집합을 선정할 수 있는 방안을 제시하였다. 유사한 예제들은 카테고리도 동일할 가능성이 높다는 일반적인 가정에 기반하여 먼저 군집화기법을 적용하여 유사한 예제들을 모으고, 생성한 각 군집의 대표예제로 최초학습예제집합을 구성하는 방안을 제안하였다. 제안한 방안이 기존의 임의 선정방안에 비해 능동적 학습 수행 시 보다 적은 수의 학습예제로도 우수한 성능을 발휘할 수 있음을 실험적으로 확인하였다. 향후 본 문서분류 이외에 일반적인 분류문제에도 적용하여 그 효과를 검증하고 기존의 데이터 축약 방안들과 비교하는 연구가 수행되어야 할 것이다.

참고문헌

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

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

[3] Mitra, P. Murthy, C.A. and Pal, S. K. Density Based Multiscale Data Condensation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24: 6, 2002

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

[5] Astrahan, M. M., *Speech Analysis by Clustering, or the Hyperphoneme Method*, Stanford A. I. Project Memo, Stanford University, Califonia, 1970

[6] Lewis, D. D., Reuters-21578 Text Categorization Test Collection, http://www.daviddlewis.com/resources/testcollections/reuters21578/

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

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

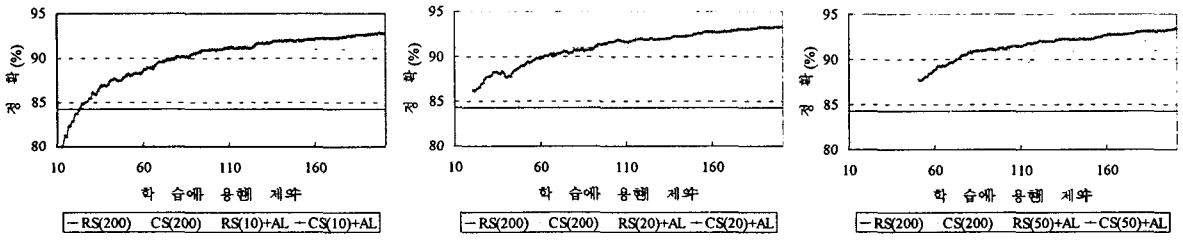

그림 4. 최초학습예제집합 선정방안과 그 크기에 따른 능동적 학습의 성능