군집화 기법을 이용한 준감독 군집화의 훈련예제 선정[1]

김종성+ , 강재호++, 류광렬+
+부산대학교 컴퓨터공학과, ++동아대학교 지능형통합항만관리연구센터
{kimjs1 , jhkang, krryu}@pusan.ac.kr

Selecting Examples to Be Labeled for Semi-Supervised Clustering Using Cluster-Based Sampling

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

요 약

기계학습의 군집화(clustering) 기법은 예제들 간의 유사성에 근거하여 주어진 예제들을 무리 짓는 방법이다. 준감독(semi-supervised) 군집화는 카테고리가 부여된(labeled) 소수의 예제들을 적극적으로 활용하여 군집형태가 보다 자연스럽게 형성되도록 유도하는 군집화 방법이다. 준감독 군집화 문제에서 예제에 카테고리를 부여하는 작업은 현실적으로 극히 제한적이거나 카테고리를 부여하는데 소요되는 비용이 상당하므로, 제한된 자원 내에서 군집화에 효용성이 높을 예제들을 선정하여 카테고리를 부여하는 것이 필요하다. 본 논문에서는 기존 연구에서 능동적 학습의 초기 훈련예제 선정을 위해 제안된 군집기반 훈련예제 선정 방법을 준감독 군집화에 적용하여 군집 결과의 질을 향상시키고자 한다. 군집화를 이용한 예제 선정 방법은 유사한 예제들은 동일한 카테고리에 속할 가능성이 높다는 가정하에 전체 예제를 활용하여 선정하고자 하는 예제 수 만큼 군집을 생성 한 후, 각 군집의 중심점에 가장 가까운 예제들을 대표 예제로 선정하여 훈련 집합을 구성하는 방법이다. 본 논문에서는 문서를 대상으로 하는 준감독 군집화 실험을 통해, 카테고리를 부여할 예제를 임의로 선정한 경우에 비해 군집화를 이용한 훈련 예제들로 준감독 군집화를 수행한 경우가 보다 좋은 군집을 형성함을 확인하였다.

1. 서론

기계학습의 군집화 기법은 예제들 간의 유사성에 근거하여 주어진 예제들을 유사한 예제들로 무리 짓는 방법이다. 일반적으로 군집화 기법은 예제들의 카테고리 정보가 전혀 없는 상황에서 군집화를 수행하는 비감독 군집화(unsupervised clustering)를 말한다. 최근 적은 수의 카테고리가 부여된 예제나 예제들 간의 관계와 제약조건을 고려하여 군집화를 수행하는 준감독 군집화(semi-supervised clustering)[1][2]에 관한 연구가 수행 되었다.

준감독 군집화는 카테고리가 부여된 소수의 예제들을 적극적으로 활용하여 보다 자연스러운 군집이 형성되도록 유도하는 방법으로, 상대적으로 카테고리 정보를 전혀 이용하지 않는 비감독 군집화에 비해 나은 군집 결과를 도출할 수 있다.

카테고리가 부여된 예제들을 충분히 제공하여 준감독 군집화를 수행한다면, 보다 좋은 군집 결과를 얻을 수 있지만, 군집화 문제에서 예제들에 카테고리를 부여하는 작업은 일반적으로 상당한 비용과 노력이 소요되며 이러한 작업은 현실적으로 극히 제한 된다.[2] 따라서 동일한 수의 카테고리가 부여된 예제를 사용하더라도, 어떠한 예제들에 카테고리를 부여 하는가에 따라 준감독 군집화 결과에 영향을 미치게 되므로, 카테고리를 부여할 예제는 신중하게 선정되어야 한다.

준감독 군집화에 관련된 기존 연구로는 제약된 k-means (Constrained k-means)[1]를 비롯하여 여러 연구[3][4]가 있었으나, 카테고리를 부여할 예제를 능동적으로 선정하여 준감독 군집화를 수행한 연구는 아직 제시되지 않았다.

학습에서 카테고리를 부여할 예제 선정과 관련된 연구로는 강재호, 류광렬이 제안한 능동적 학습에서 동일한 수의 학습 예제로 최대한의 분류 성능을 달성하기 위하여 훈련 예제를 선별하는 전략[5]이 있다. 이 방법은 유사한 예제들은 동일한 카테고리에 속할 가능성이 높다는 일반적인 가정에 기반하여 카테고리를 부여할 예제의 수 만큼 군집을 생성한 후, 생성된 각 군집을 대표할 수 있는 예제에 카테고리를 부여함으로써 임의로 예제를 선정하는 방식보다 정확도가 높은 분류기를 생성할 수 있음을 실험적으로 확인하였다. 본 논문에서는 이러한 방법을 준감독 군집화에 적용하여 군집화 성능을 개선하고자 한다.

본 논문은 먼저 2장에서 준감독 군집화 기법 중 제약된 k-means에 관해 살펴보고, 3장에서 군집화를 이용하여 카테고리를 부여 할 예제를 선정하는 방안에 대하여 설명한다. 4장에서는 본 접근방안을 문서 군집화 문제에 적용한 실험결과를 분석하고 5장에서 결론 및 향후 연구를 제시한다.

2. 준감독 군집화

준감독 군집화는 카테고리가 부여된 예제나 예제들 간의 관계와 제약조건을 고려하여 보다 자연스러운 군집이 형성되도록 유도하는 군집화 방법으로, 본 논문에서는 제약된 k-means

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

[2] 일반적으로 카테고리가 부여된 예제를 충분하게 확보 할 수 있다면 분류 문제로 접근하며, 군집화 문제에 사용되는 예제는 카테고리 구분 기준이 없거나(불명확하거나), 카테고리 부여하는 비용이 상당하다.

기법을 사용한다. 제약된 k-means 방법은 카테고리가 부여된 예제들을 이용하여 각 군집의 초기 씨드를 설정하고, 군집 수행 시 카테고리가 부여된 예제는 초기에 할당된 군집에 고정시킴으로써 예제의 카테고리가 변경되는 것을 막는다. 알고리즘 1에 제약된 k-means 방법을 정리하였다.

1. 카테고리가 부여된 예제 집합 T_0로부터 각 군집의 최초중심점을 계산
2. 모든 예제들은 각 군집 중심점과 가장 가까운 군집에 할당하여 군집화를 수행, 단 예제 집합 T_0에 속한 예제는 최초 군집에 고정하여 할당
3. 형성된 각각의 군집으로부터 새로운 중심점 계산
4. 모든 군집의 안정화 될 때 까지 단계 2, 3을 반복 수행

알고리즘 1. 제약된 K-means

3. 군집화를 이용한 카테고리 부여 예제의 선정 방안

본 장에서는 군집화 기법을 이용하여 카테고리를 부여할 예제를 선정하는 방안에 대하여 자세히 기술한다.

그림 1과 그림 2는 두 가지 속성을 가진 예제들을 2차원상에 나타낸 것이다. 각각의 예제들은 두 가지 카테고리 A, B 중 하나에 속하며, Ⓐ 기호로 표시된 예제들은 카테고리 A에 속하는 예제들이며, Ⓑ 기호로 표시된 예제들은 카테고리 B에 속하는 예제들이다. ⑦ 기호로 표시된 예제들은 준감독 군집화에 사용자가 카테고리를 부여하지 않은 예제들이다. 점선으로 표시된 타원형태의 영역은 각각 카테고리 A와 B의 영역을 나타내며, 실선으로 표시된 타원형태의 영역은 카테고리가 부여된 예제를 사용하여 준감독 군집화를 수행한 결과를 나타낸다.

그림 1은 임의로 선정한 4개의 카테고리가 부여된 예제를 감안하여 준감독 군집화를 수행한 경우로, 2장에서 설명한 제약된 k-means 알고리즘을 가정하였다. 카테고리가 부여될 예제를 임의로 선정할 경우 예제의 위치가 효과적이지 않다면 그림에서와 같이 일부 예제들이 잘못 군집화되는 결과를 나타낼 수 있다.

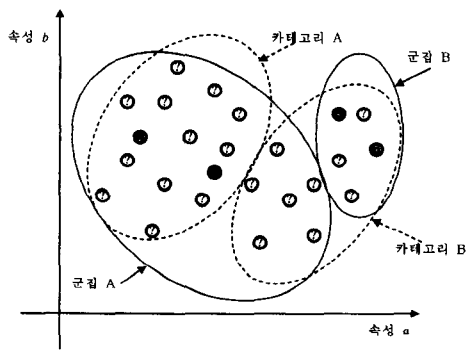

그림 1. 임의 선정된 예제들로 군집화한 경우

그림 2는 본 논문에서 적용하고자 하는 군집화 기법을 이용하여 카테고리를 부여할 예제를 선정하는 방안을 설명하고 있다. 음영색으로 표시된 타원형태의 영역은 카테고리를 부여할 예제를 선정하기 위하여 군집화를 수행한 결과로서, 본 예에서는 4개의 예제를 선정하기 위하여 4개의 군집으로 군집화를 수행한 후 각 군집의 중심점(+)과 가장 가까운 예제들을 카테고리 부여 대상으로 선정하였다. 사용자는 선정된 4개의 예제에 카테고리를 부여한 후, 앞에서와 동일한 준감독 군집화 알고리즘을 적용하여 군집화 수행한다. 그림 2의 예에서는 보다 자연스러운 군집화 결과를 보여 주고 있다. 알고리즘 2에서는 본 적용 방법을 정리하였다.

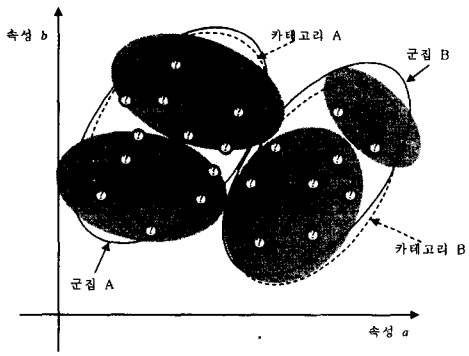

그림 2. 군집화 후 생성된 예제들로 군집화한 경우

1. 전체 예제집합 D 중에서 t개의 예제를 임의 선정하여 S_0 생성
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} 생성
4. S_{final}의 예제들에 대한 카테고리를 사용자에게 문의한 후 카테고리가 부여된 예제 집합 T_0 생성
5. T_0를 기반으로 준감독 군집화 수행

알고리즘 2. 군집화를 이용한 카테고리를 부여할 예제 선정 방안

4. 실험 결과

3장에서 설명한 방안을 준감독 군집화 문제에 적용하였을 때 그 성능을 확인하기 위해 아래와 같은 실험을 수행하였다.

실험에 사용한 newsgroups-20 말뭉치[6]는 20개의 유즈넷 뉴스그룹에 올려진 약 20,000건의 기사 모음으로 구성되어 있다. 각 기사들은 제목과 본문만 사용하였고, 불용어 처리(stop word removal) 및 표준형 변환(stemming)으로 전처리하였다. 실험 데이터의 난이도에 따른 효과를 확인하기 위하여 [2]에서 사용한 방법과 같이 기사의 주제가 상이한 3개의 뉴스그룹 (alt.atheism, rec.sport.baseball, sci.space)의 기사들로 구성된 Different-3의 데이터와, 주제가 유사한 3개의 뉴스그룹 (comp.graphics, comp.os.ms-windows, comp.windows.x)의 기사들로 구성된 Same-3 데이터를 생성하여 각각 실험하였다.

카테고리를 부여할 예제를 선정하기 위하여 군집화 기법으로는 k-means 알고리즘을 사용하였다. 카테고리를 부여할 예제의 수 만큼 군집을 생성한 후, 중심점과 가장 가까운 예제에 카테고리를 부여하였다. 준감독 군집화 기법으로 제약된 k-means 알고리즘을 적용하였으며 이때 군집의 수는 카테고리 수(3개)만큼 생성하였다. 예제의 표현은 정보검색분야에서 널리 사용되는 tf×idf 공간상의 벡터로 표현하였으며, 예제와 군집 중심점간의 유사도 계산방법으로는 코사인 유사도를 사용하였다[7].

실험은 임의로 카테고리를 부여할 예제를 선정하는 방안(RS)과 군집화 후 대표성이 있는 예제에 카테고리를 부여하는 방안(CS)을 비교하였다. RS 방안은 각각의 카테고리 마다 최소 하나의 예제를 보장해 주었다. Different-3와 Same-3 각각 두 가

지 선정방법에 대해서 100회 실험을 수행 후 정확도와 Rand Index[3] 척도로 평가 하였다.

그림 3의 실험 결과에서 Different-3는 CS 방안이 동일한 수의 카테고리가 부여된 예제를 이용한다면 RS 방안보다 높은 정확도와 Rand Index(RI)값을 가진 군집을 생성할 수 있음을 알 수 있다.

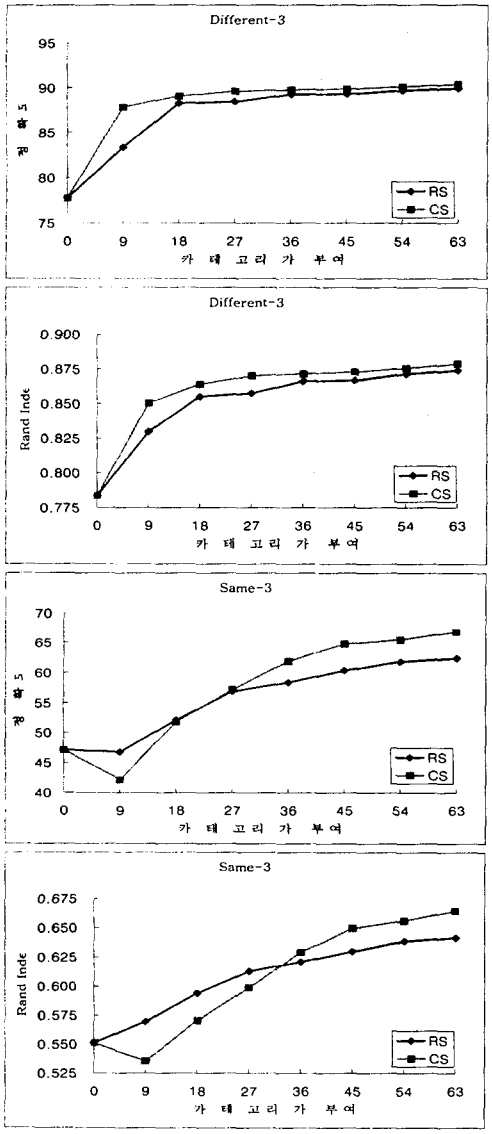

그림 3. 임의 및 군집기반 예제 선정 방안의 성능

Same-3는 카테고리가 부여된 예제 수가 적은 경우 RS에 비해 CS가 좋지 않은 정확도와 RI를 보이고 있다. 그 원인은 선정된 예제들을 분석한 결과 특정 카테고리에 편향되어 예제가 선정되기 때문이었다.

표 1은 Same-3와 Different-3 데이터에서 카테고리가 부여된 예제 수가 각각 9개와 27개인 경우 카테고리 별로 선정된 예제 개수의 평균 비율을 구한 것으로, Same-3에서는 예제 수가 9개와 27개인 경우 CS 방안이 RS 방안보다 불균형하게 예제를 선정함을 볼 수 있다. 반면 Different-3는 CS와 RS 모두 3개의 카테고리에 예제가 균형 있게 선정되었음을 알 수 있다. 이와 같은 결과는 Same-3 예제들의 분포 형태가 특정 카테고리에 편향되어 있어서, 예제 선정을 위한 군집화 수행 시 특정 카테고리를 중심으로 군집이 형성되는 것으로 보인다.

표 1. 카테고리가 부여된 예제의 카테고리 별 평균 비율(%)

데이터 / 카테고리 \ 카테고리가 부여된 예제 수		9 개		27 개	
		RS	CS	RS	CS
Different-3	alt.atheism	31.56	34.89	33.63	35.90
	rec.sport.baseball	35.22	30.56	32.00	30.19
	sci.space	33.22	34.56	34.37	33.98
Same-3	comp.graphics	33.33	**56.34**	33.89	**45.34**
	comp.os.ms-windows	32.67	**23.12**	31.81	**30.48**
	comp.windows.x	34.00	**20.53**	34.30	**24.18**

5. 결론 및 향후 연구

본 논문에서는 군집화를 이용한 카테고리를 부여할 예제 선정시 임의 예제 선정보다 준감독 군집화 결과를 어느 정도 개선 할 수 있음을 보였다. 그러나 특정 데이터에서 비교적 적은 수의 예제를 선정할 때 군집화를 이용한 예제 선정방안이 일부 카테고리로 편향되는 현상이 있었다. 향후 연구로 예제들이 특정 카테고리로 편향되어 선정되는 현상을 억제하기 위한 방안으로, 군집 유효화 지수 같은 군집 평가 지수를 예제 선정에 활용하는 연구가 필요하다.

참고문헌

[1] Kiri Wagstaff, and Claire Cardie. "Clustering with Instance-level Constraints". In Proceedings of 17th International Conf. on Machine Learning, pp. 1103-1110, 2000

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

[3] Sugato Basu, Arindam Banerjee, and Raymond J.Mooney. "Semi-supervised clustering by seeding". In Proceedings of 19th International Conf. on Machine Learning, pp. 19-26, 2002

[4] Kiri Wagsta, Claire Cardie, Seth Rogers, and Stefan Schroedl. "Constrained K-means clustering with background knowledge". In Preceedings of 18th International Conf. on Machine Learning, pp. 577-584, 2001

[5] 강재호, 류광렬. "군집화 기법을 이용한 능동적 학습의 최초 학습예제 선정". 정보과학회 2003 추계학술발표회, pp. 16-18,

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

[7] Ricardo Baeza-Yates, Berthier Rivero-Neto. Modern Information Retrieval, 1999

[3] 예제 쌍의 모든 경우에 대해 어떤 한 쌍의 예제가 군집화 후 같은 혹은 다른 군집에 속하는 경우의 수로 군집결과간의 유사도를 측정하는 척도로, 군집결과가 완전히 일치하는 경우 1의 값을 가진다.