http://dx.doi.org/10.3745/KTSDE.2013.2.10.705

Active Learning based on Hierarchical Clustering

Hoyoung Woo† · Cheong Hee Park††

ABSTRACT

Active learning aims to improve the performance of a classification model by repeating the process to select the most helpful unlabeled data and include it to the training set through labelling by expert. In this paper, we propose a method for active learning based on hierarchical agglomerative clustering using Ward's linkage. The proposed method is able to construct a training set actively so as to include at least one sample from each cluster and also to reflect the total data distribution by expanding the existing training set. While most of existing active learning methods assume that an initial training set is given, the proposed method is applicable in both cases when an initial training data is given or not given. Experimental results show the superiority of the proposed method.

Keywords : Active Learning, Clustering, Ward's Method

계층적 군집화를 이용한 능동적 학습

우 호 영† · 박 정 희††

요 약

능동적 학습(active learning)은 소수의 라벨 데이터로 구성된 훈련 집합이 주어진 경우에 분류기 학습에 가장 도움이 될 만한 언라벨드 데이터를 선택하여 전문가에 의한 라벨링을 통해 훈련 집합에 포함시키는 과정을 반복함으로써 분류기의 성능을 향상시키는 것을 목적으로 한다. 본 논문에서는 워드 연결(ward's linkage)을 이용한 계층적 군집화(hierarchical clustering)를 바탕으로 한 능동적 학습 방법을 제안한다. 제안된 방법은 각 군집에서 적어도 하나의 샘플을 포함하도록 초기 훈련 집합을 능동적으로 구성하거나 또는 기존의 훈련 집합을 확장함으로써 전체 데이터 분포를 반영할 수 있게 한다. 기존의 능동적 학습 방법들 중 대부분은 초기 훈련 집합이 주어져 있을 경우를 가정하는 반면에 제안하는 방법은 초기 클래스 정보를 가진 훈련 데이터가 주어지지 않은 경우와 주어진 경우에 모두 적용 가능하다. 실험을 통하여 제안하는 방법이 비교 방법들에 비해 분류기 성능을 크게 향상시킬 수 있는 효과적인 데이터 선택을 수행함을 보인다.

키워드 : 능동적 학습, 군집화, 워즈 방법

1. 서 론

클래스 정보가 없는 데이터(unlabeled data, 언라벨 데이터)에 비해서 클래스 정보가 있는 데이터(labeled data, 라벨 데이터)를 얻기 위해서는 많은 비용과 시간이 들게 된다. 능동적 학습(active learning)은 소수의 라벨 데이터로 구성된 훈련 집합이 주어진 경우에 분류기 학습에 가장 도움이 될 만한 언라벨드 데이터를 선택하여 전문가에 의한 라벨링을 통해 훈련 집합에 포함시키는 과정을 반복함으로써 분류기의 성능을 향상시키는 것을 목적으로 한다[1].

대부분의 능동적 학습 방법들은 초기 훈련 집합이 주어진 경우에 적용 할 수 있다. 예를 들어, 서포트 벡터 머신(support vector machine, SVM)을 기본 분류기로 이용한 능동적 학습인 SVMactive[2]는 초기 훈련 집합으로 SVM을 학습한 후, 분류기에 도움이 되는 데이터로서 현재 결정 경계에 가장 근접한 데이터를 선택한다. 현재 분류 모델을 이용하여 데이터 샘플들의 불확실성(uncertainty)이나 불순도(entropy)를 측정하여 다음 모델 학습에 포함시킬 샘플을 선택하는 것이다.

반면에, 훈련 집합이 주어지지 않았을 경우에 효과적으로 초기 데이터를 선택하는 방법으로 그래프를 기반으로 한 방법들이 최근에 제안 되었다[3,4]. 데이터 샘플들에 대해 주변의 이웃 샘플들의 선형 결합으로 나타내는 계수를 구하고, 계수들을 이용하여 전체 데이터를 가장 잘 재구축 할 수 있는 대표적인 데이터 샘플들을 선택하는 지역적 선형 재구축(Locally Linear Reconstruction, LLR) 방법[3]과 전체 데이터 중 임의의 부분 집합을 선택하여, 가장 불필요한 데이터를 삭제해나가는 방법[4]이 있다. 그러나 이러한 방법들은 작은 수의 데이터로 훈련 집합을 구성할 때 빈약한 정보로 인하여 초기 성능이 떨어지고, 어느 정도 크기를 갖는 훈련 집합을 구성했을 때에야 만족할 만한 성능을 보인다.

※ 이 논문은 2011년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2011-0007779).
† 준 회 원 : 충남대학교 컴퓨터공학과 석사과정
†† 정 회 원 : 충남대학교 컴퓨터공학과 부교수
논문접수 : 2013년 5월 14일
수 정 일 : 1차 2013년 6월 7일
심사완료 : 2013년 7월 2일
* Corresponding Author : Cheong Hee Park(cheonghee@cnu.ac.kr)

훈련 집합이 주어지지 않았을 경우에는 데이터 집합의 분포를 잘 표현(대표적으로 표현)하는 데이터를 선택하여 초기 훈련 집합을 구성하는 것이 중요하다. 이것은 분류기의 만족할만한 성능을 얻기 위해 필요한 훈련 집합의 크기를 줄일 수 있다. 군집화(clustering)는 생물학, 정보 검색, 기후, 심리학 의학 등 여러 방면에서 이용되는 대표적인 무감독 학습(unsupervised learning) 방법이다. 군집화는 클래스 정보가 주어지지 않은 상황에서 유사한 데이터들을 같은 그룹에 속하게 하고 상이한 데이터들은 다른 그룹에 속하도록 전체 데이터를 몇 개의 그룹으로 나누는 방법이다[5]. 우리는 유사한 특징을 갖는 데이터를 그룹으로 묶는 군집화 방법을 이용하여, 초기 훈련 집합이 주어지지 않을 경우와 초기 훈련 집합이 주어졌을 경우 모두 적용이 가능한 능동적 학습 방법을 제안한다. 본 논문은 학술발표논문 [6]에서 제안되었던 초기 훈련 집합이 주어지지 않았을 경우의 능동적 학습 방법을 초기 훈련 집합이 주어졌을 경우까지 확장하고 있다. 또한 군집화에서 결합방법으로 워드 방법을 선택하는 것에 대한 근거를 제시하고 있다.

제안된 방법은 각 군집에서 적어도 하나의 샘플을 포함하도록 초기 훈련 집합을 능동적으로 구성하거나 또는 기존의 훈련 집합을 확장함으로써 전체 데이터 분포를 반영할 수 있게 한다. 2절에서는 능동적 학습에 대한 설명 및 대표적인 방법에 대한 소개를 하고, 3절에서는 계층적 군집화에 대한 설명과 연결 방법 등을 비교하고, 4절에서는 계층적 군집화를 이용한 능동적 학습을 제안한다. 5절에서는 제안한 방법의 성능을 측정하는 실험 결과를 보이고, 6절에서 결론으로 마무리 한다.

2. 관련 연구

능동적 학습은 초기 소수의 라벨 데이터로 구성된 훈련 집합이 주어질 경우에 분류기 학습에 가장 도움이 되는 데이터를 선택하여 전문가의 라벨링에 의해 훈련 집합에 포함시키거나, 초기 훈련 집합이 주어지지 않았을 경우에 전체 데이터 분포를 잘 나타내는 데이터 부분 집합을 선택하여 훈련 집합을 만드는 방법이다. 가장 단순한 방법으로 전체 데이터에 대하여 동일한 확률로서 임의적으로 선택하는 임의 추출(random sampling) 방법이 있다. 굉장히 단순하며 빠르고 성능적으로도 평균은 도달하기 때문에, 여러 연구나 응용에서 사용되어지고 있다. 그러나 이것은 표현 그대로 임의적으로 추출하기 때문에 만족할만한 성능에 도달하기까지 비교적 많은 데이터를 필요로 한다.

능동적 학습 방법은 좀 더 효과적으로 만족스러운 성능에 도달하기까지의 비용과 시간을 감소시켜 준다. 초기 클래스 정보가 주어지지 않을 경우의 능동적 학습 방법으로 Locally Linear Reconstruction(LLR)[3]과 Random Sampling and Backward Deletion[4]이 최근 제안 되었다. LLR[3]은 각 데이터 샘플과 주변 이웃들이 선형 결합 관계로 나타내진다는 가정 하에 전체 데이터에 대한 새로운 공간을 재구성 했을 때, 재구성 에러를 가장 최소화시키는 대표적 샘플들을 선택한다. 그러나 큰 규모의 데이터 집합에 대해 데이터와 이웃 데이터 간의 관계에 대한 반복적 행렬 연산으로 인해 시간적 소모가 매우 크다는 단점이 있다. Random Sampling and Backward Deletion[4]은 가장 단순하면서 대표적으로 데이터를 선택하는 방법인 임의추출법을 이용하여 전체 데이터 중 부분집합을 선택한 후 그 중에서 가장 불필요한 데이터 샘플을 선택하여 제거하는 것을 반복한다. 불필요한 데이터를 제거하기 위해서 먼저 데이터 간의 거리를 계산하고, 주위의 다른 데이터 샘플들과 가장 거리가 작은 데이터를 선택하여 부분집합에서 제거한다. 한 쌍의 데이터가 가장 가까운 거리를 가지면 두 데이터는 같은 클래스 정보를 가질 가능성이 크다는 것에 기반을 두고 있다. 이 논문에서는 거리 측정 방법으로 두가지 방법을 제시하는데, 유클리디안 거리를 이용한 방법과 그래프 기반의 최단 거리를 이용한 측정 방법이다. 그러나, 임의 추출을 사용하여 전체 데이터의 부분 집합을 선택하므로 데이터 전체의 분포를 잘 알 수 없는 경우가 발생한다. 따라서 그에 따른 정보 손실 발생으로 성능이 떨어지는 경우가 생긴다.

기존의 능동적 학습 방법들 중 대부분은 초기 훈련 집합이 주어져 있을 경우를 가정한다. 대표적인 방법으로서, 기존의 학습된 모델에 의해 가장 불확실성이 높다고 판단되는 데이터에 대한 클래스 정보를 얻는 uncertainty sampling 방법[2,7], 다수의 분류기 모델을 구성하여 의견이 가장 불일치하는 데이터를 선택하는 query by committee 방법[8,9], 현재 모델의 일반화 오류(generalization error)를 최대로 감소시킬 수 있는 데이터를 추정하여 선택하는 Expected Error Reduction 방법[10] 등이 있다. uncertainty sampling 방법은 기존의 학습된 모델에서 가장 불확실성이 높은 데이터에 대한 라벨 정보를 얻어 온다. SVM은 클래스를 분할하는 결정 경계인 초평면의 마진을 최대화하는 방법으로 여러 분야에서 활용되고 있다. 능동적 학습에서도 많은 연구방법이 제안되었는데, SVMactive[2]와 SvSB[7]는 초기 클래스 정보를 가진 데이터 집합을 가지고 SVM을 기본 분류기로 학습하고, 클래스 정보가 없는 집합(unlabeled pool)에서 가장 불확실성이 높은 데이터를 선택하는 방법이다. SVMactive[2]는 현재 학습된 분류기로 만들어진 결정 경계에 가장 근접한 언라벨드 데이터를 선택하고, 그 데이터를 분류기에 도움이 되는 데이터로서 학습 집합에 포함시킨다. SvSB[7]는 언라벨드 데이터에 대해 각 클래스에 속하는 정도를 확률적으로 나타내고, 각 클래스에 속할 확률 중에서 가장 높은 확률을 가지는 경우와 두 번째 높은 확률을 가지는 경우의 차가 가장 적은 데이터를 선택한다.

3. 계층적 군집화(hierarchical clustering)

우리는 4절에서 군집화 알고리즘 중에 계층적 병합 군집화를 이용한 능동적 학습 방법을 제안한다. 우리가 군집화를 능동적 학습을 위해 활용하려는 이유는 잘 형성된 군집

에서 각 군집을 대표하는 데이터를 선택함으로써 최소의 데이터 선택으로 전체 데이터 분포를 잘 반영할 수 있기 때문이다. 우리는 좋은 군집을 형성하는 방법을 찾고, 그 방법을 활용한 능동적 학습 방법을 제안하고자 한다.

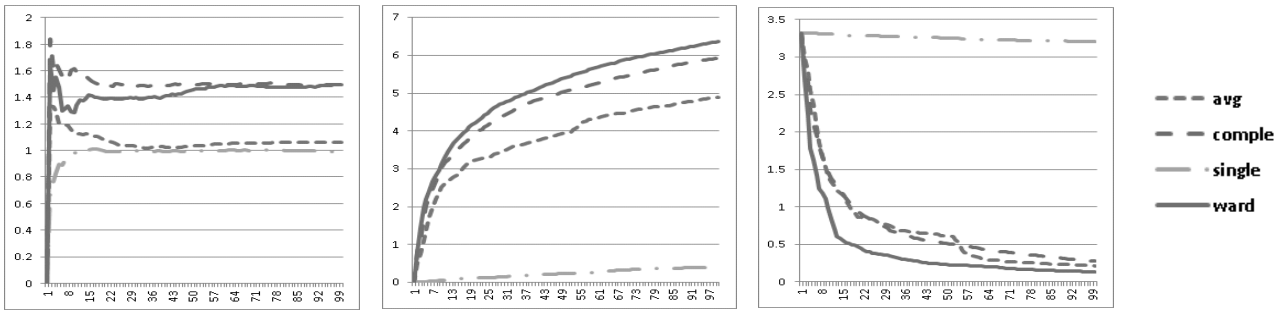
군집화 방법은 계층적 군집화 방법과 분할적 군집화 방법으로 나눌 수 있다[5]. 계층적 군집화 방법은 군집을 계층적으로 형성하여 분열(divisive)하거나 응집(agglomerative)하는 형태의 덴드로그램 혹은 트리 형태의 결과를 출력한다. 특히, 계층적 병합 군집화(hierarchical agglomerative clustering, 이하 HAC)는 데이터들이 단일 군집(singleton cluster)으로 출발하여 연결방법에 따라 두 개의 군집이 결합하여 하나의 군집이 되는 과정을 전체 데이터가 하나의 군집으로 묶여질 때까지 반복적으로 수행한다. 기본적인 계층적 병합 군집화 알고리즘을 Fig. 1에 요약하였다.

Input : data set
Output : dendrogram of clusters
1. Compute distances between data points **2.** Start with the points as individual clusters **3.** Repeat line 4, 5 until only one cluster remains **4.** Merge the closest two clusters **5.** Update the distance between the new cluster and the existing clusters

Fig. 1. Basic agglomerative hierarchical clustering algorithm

Fig. 1의 라인4에서 두 군집간의 거리를 어떤 거리 척도로 계산하는지에 따라 다양한 결과를 도출할 수 있다. 많이 알려져 있는 방법은 크게 4가지로 단일 연결(single linkage 혹은 MIN), 완전 연결(complete linkage 혹은 MAX), 그리고 그룹 평균(average linkage)와 워드 방법(ward's method)이 있다. 단일 연결은 한 군집 안에 포함된 점들과 다른 군집 안에 포함된 점들 사이의 거리를 계산하여, 최소 거리를 가진 점들 사이의 거리를 두 군집의 거리로 사용한다. 완전 연결은 반대로 두 군집 간에 포함된 두 점 사이의 최대거리를 군집 간의 거리로 본다. 그룹 평균은 두 군집 간에 포함된 점들의 거리의 평균을 사용한다. 워드 방법[11,12]은 두 군집이 결합했을 때의 평균과 속한 모든 샘플들과의 오차제곱 합이 결합하기 전의 두 군집 각각에서의 오차제곱합보다 가장 적게 증가하는 두 군집을 결합하는 방법이다. 두 군집을 C_r와 C_s라고 할 때, ward 방법에 의한 거리를 $D(C_r, C_s)$라고 하면,

$$D(C_r, C_s) = \sum_{x \in C_r \cup C_s} \| x - \bar{x} \|^2 - \left(\sum_{x \in C_r} \| x - \overline{x_r} \|^2 + \sum_{x \in C_s} \| x - \overline{x_s} \|^2 \right) \quad (1)$$



Davies-Bauldin　　entropy in cluster [pen digit data]　　weighted entropy with class label

Davies-Bauldin　　entropy in cluster [segment data]　　weighted entropy with class label

Fig. 2. Comparison of three evaluation measures

이다. 이때 $\overline{x}$는 군집 C_r과 C_s가 병합되었을 경우의 중심점이고, $\overline{x_r}$과 $\overline{x_s}$는 각 군집의 중심점, n_r과 n_s는 각 군집의 샘플 수이다. 간단한 계산 과정을 거쳐 식(1)을 두 군집의 중심 사이의 거리를 이용하는 표현으로 나타낼 수 있다.

$$D(C_r, C_s) = \frac{n_r \bullet n_s}{n_r + n_s} \| \overline{x_r} - \overline{x_s} \|^2 \quad (2)$$

3.1 결합 방법에 따른 계층적 군집화 성능 평가

제안하는 능동적 학습 알고리즘을 설명하기 전에 4가지 결합 방법에 의한 군집화 성능을 비교하고자 한다. 계층적 병합 군집화에 사용되는 여러 연결 방법에 대한 군집화 성능을 비교하기 위해 다음의 세 가지 척도를 사용하였다.

- Davies-Bauldin[13]

$$\frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j}\left(\frac{\mu_i + \mu_j}{d(c_i, c_j)}\right) \quad (3)$$

- Entropy in cluster[14]

$$-\sum_{i=1}^{n} \frac{|x_i|}{|x|} log_2 \frac{|x_i|}{|x|} \quad (4)$$

- Weighted entropy of labeling samples in cluster[15]

$$-\sum_{i=1}^{n} \frac{x_i}{x} \sum_{k=1}^{K} \frac{x_{ik}}{x_i} log_2 \frac{x_{ik}}{x_i} = -\frac{1}{x}\sum_{i=1}^{n}\sum_{k=1}^{K} x_{ik} \log_2 \frac{x_{ik}}{x_i} \quad (5)$$

Davies-Bauldin은 내부 군집에 대한 평가 방법 중에 하나이며, n은 군집의 수이며, c_i는 i번째 군집의 중심점(centroid)이고, μ_i는 i번째 군집 안에 속하는 샘플과 c_i간의 평균거리를 나타낸다. 군집 안의 거리가 작고, 군집 간의 거리가 크면, Davies-Bauldin의 값은 작아진다. 이 척도 값이 작을수록 군집 성능이 더 좋다고 판단한다. Entropy는 랜덤 변수가 가질 수 있는 불확실성에 대한 척도로서, 측정값이 높을수록 혼돈 상태가 높다는 것을 나타낸다. 식(4)에서 x는 전체 샘플의 수이고, x_i는 i 군집에 속한 샘플의 수이다. Entropy 값이 클수록 군집들의 크기가 균일함을 의미한다. 세 번째 평가 척도는 클래스 정보와 같은 외적인 정보를 이용한 외부 평가 방법이다. n은 군집의 수이고 K는 클래스의 개수이다. x_{ik}는 i 군집에 속한 데이터 중에 클래스 k에 속하는 데이터의 수이다. 이것은 구성된 군집 내에서 클래스 분포가 얼마나 혼돈 상태인지(클래스 정보가 한쪽에 기울어진 상태인지, 균일한 상태인지)를 측정한다. 결과 값이 작을수록 좋은 군집이라 평가한다.

Fig. 2는 세 개의 평가 척도에 대한 실험결과를 나타낸다. 실험은 Table 1에 설명된 UCI 데이터를 이용하였고, 본 논문에는 그 중 pandigit과 segment데이터에 대한 실험결과를 나타낸다. Table 1에 있는 다른 데이터 집합에 대해서도 유사한 결과를 얻었다. 우리는 단일 연결, 완전 연결, 평균 연결, 워드 연결을 이용한 덴드로그램을 구성하고, 각 군집화 방법에서 1부터 100까지 군집의 수를 변화시켜가면서 수치를 측정하였다. Davies-Bauldin 에 대한 결과에서 워드 방법은 가장 높은 값을 가지고, 단일 연결 방법은 가장 낮은 값을 가졌다. 이 평가 법은 군집간 거리는 멀고, 군집내 데이터간의 거리는 좁을 수록 작은 척도 값을 가진다. 군집간의 엔트로피를 측정한 결과는 워드 방법이 가장 높은 값을 나타냈다. 값이 높다는 것은 불순도가 가장 높은 결과를 나타내며, 이것을 군집에 적용하였을 경우, 균일한 크기의 군집들을 형성하는 것이 워드 방법이라는 것을 알 수 있다. 클래스 정보를 이용한 군집 평가 방법에서는 워드 방법이 가장 낮은 값을 가지는 것을 알 수 있다. 이것은 각 군집에 속한 데이터들이 가장 동일한 클래스 정보를 가지고 있는 연결 방법이 워드 방법이라는 것을 나타낸다. 세 개의 평가 방법을 통하여 우리가 하고자 하는 방향을 제일 수치적으로 잘 표현한 워드 방법을 이용하여 능동적 학습에 적용하고자 한다.

4. 계층적 군집화(hierarchical clustering)

4.1 초기 훈련 집합이 없는 경우의 능동적 학습

라벨 데이터로 구성된 초기 훈련 데이터가 주어지지 않을 경우에 초기 훈련 데이터로서 효과적인 데이터를 선택하기 위해서는 전체 데이터의 분포를 잘 표현하는 데이터 선택이 중요하다. 우리는 워드 연결을 이용한 군집화를 수행하고, 선택하고자 하는 데이터 수와 같은 개수의 군집이 형성된 지점을 찾는다. 그리고 각 군집에서 대표적인 하나의 데이터를 선택한다. 각 군집에 속한 데이터들은 서로 유사한 특성을 가지기 때문에 하나의 군집으로 묶여 있으며, 각 군집에서 대표적인 하나의 데이터를 선택함으로써 전체 데이터 분포를 잘 표현하는 데이터 집합을 구성하고자 한다. 얻고자 하는 데이터의 수를 k라 하고, 선택되는 샘플의 집합을 $\{s_1,...,s_k\}$라 할 때, 각 군집에서 선택하는 데이터를 다음의 식으로 나타낼 수 있다.

$$s_i = \arg min_{x \in C_i} \| x - \overline{x_i} \| \quad (6)$$

식 (6)에서 x는 군집 C_i에 속하는 데이터, $\overline{x_i}$는 그 군집의 중심점을 의미한다. 우리는 각 군집에서 식(6)처럼 중심점에 가장 근접한 데이터를 선택한다. 이것은 우리가 선택하려고 하는 가장 대표적인 샘플, 가장 정보력 있는 샘플과 일치한다. 제안하는 방법을 Fig. 3에 요약 한다.

4.2 초기 훈련 집합이 주어진 경우의 능동적 학습

초기 훈련 집합이 주어진 상황에서 덴드로그램을 활용하여 능동적 학습에 적용하고자 한다. 주어진 초기 훈련 집합을 L이라고 하자. Fig. 4와 같은 덴드로그램에서 세개의 군집이 형성되는 지점을 찾았을 때 L의 원소들이 세개의 군집에 골고루 퍼져 있을 수도 있고 몇 개의 군집에 편중되어 있을 수도 있다. 만약 L의 원소를 전혀 포함하고 있지 않은 군집이 있다면 그 군집에서 원소를 선택하여 L에 포함시키

Input k : the number of samples to select

Output $\{s_1,...,s_k\}$: the set of the selected samples

1. Construct a dendrogram Z by performing hierarchical clustering using Eq. (2)
2. From Z, find the location where k clusters $C_1,...,C_k$ are composed
3. for $i=1$ to k
4. $s_i = \arg min_{x \in C_i} \|x - \overline{x_i}\|$ ($\overline{x_i}$ is a centroid of C_i)
5. end for

Fig. 3. AL-HAC Algorithm for composing an initial training set

는 것이 필요하다. Fig. 4에서는 L이 붉은 원으로 표시된 세 개의 데이터 샘플을 포함하고 있다고 할 때, 첫 번째 군집에 모두 속해 있는 경우를 나타낸다. 따라서 추가 하고자 하는 원소의 수가 k개라고 할 때, k+1개의 군집이 형성되는 지점부터 시작하여 L의 원소를 포함하지 않는 군집이 k개가 될 때까지 덴드로그램의 분할 방향을 따라 군집의 수를 늘리면서 탐색을 한다. L의 원소를 포함하고 있지 않은 군집의 수가 k개가 되었을 때 각 군집에서 하나씩 대표 샘플을 선택하여 L의 집합에 추가한다. Fig. 4의 예에서 보여주는 것처럼 |L|=3이고 k=3개의 데이터를 추가하고자 할 때 (a)지점에서 L의 원소를 포함하고 있지 않은 세 개의 군집을 찾을 수 있고, 붉은 사각형으로 표시된 세 개의 데이터 샘플이 L에 추가된 것을 알 수 있다.

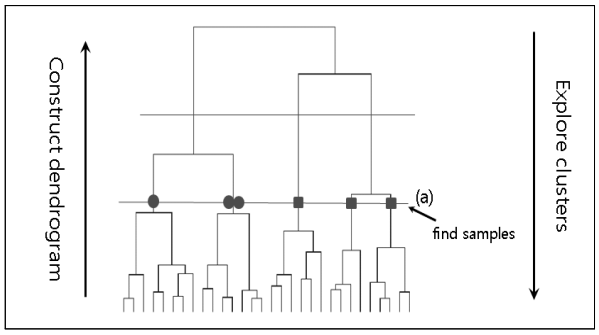


Fig. 4. dendrogram of clustering using ward's linkage

제안하는 방법은 계층적 병합 군집화가 구성된 역방향으로 군집의 개수를 증가시키면서, 초기 훈련집합의 원소를 포함하고 있지 않은 군집에서 데이터를 선택하는 방법이다. 제안하는 계층적 군집화를 이용한 능동적 학습 방법(Active Learning using Hierarchical Agglomerative Clustering, AL-HAC)을 Fig. 5에 요약하였다.

5. 실험결과

제안한 방법의 성능 평가를 위해 실제 데이터를 이용한 분류 정확도를 비교하였다. 5.1절에서는 초기 훈련 집합이 주어지지 않을 경우의 실험 결과를 나타내고, 5.2절에서는 초기 훈련 집합이 주어진 경우의 실험 결과를 나타낸다.

UCI[16]의 4개의 데이터와 LLR[3]에서 사용된 USPS 이미지 데이터를 실험 데이터로 사용하였다. Table 1은 실험에 사용된 데이터 정보이다. 각 데이터 집합에 대해 전체 데이터를 6:4의 비율로 클래스 정보가 없는 집합(unlabeled pool)과 테스트 집합으로 나누었다. 성능측정을 위한 기본 분류기로 선형 커널을 사용한 libSVM[17]과 WEKA[18]의 다층 퍼셉트론(Multi-Layer Perceptron, MLP)을 이용하였고, 10번 실험을 반복하여 측정한 분류 정확도의 평균을 결과로 나타냈다.

Table 1. data set

data	samples	features	classes
usps	11,000	256	10
pendigit	10,992	16	10
segment	2,308	19	7
optdigits	5,620	64	10
isolet	7,798	256	26

5.1 초기 훈련 집합이 없는 경우의 실험 비교

초기 훈련 집합이 없는 경우의 제안한 방법을 평가하기 위해서, 선택하는 데이터 샘플의 수를 10부터 100까지 10개씩 증가시키며 선택하고, 선택한 데이터를 가지고 SVM과 MLP를 이용하여 테스트 데이터에 대한 분류 정확도를 측정하였다. 다음과 같은 4개의 알고리즘의 성능을 비교 하였다.

Input k : the number of samples to select
L : an initial training set

Output $\{s_1,...,s_k\}$: a set of the selected samples

1. Construct a dendrogram Z by performing hierarchical clustering using Eq. (2)
2. From Z, find the location where k+1 clusters $C_1,...,C_{k+1}$ are composed
3. add = 0
4. while add < k
5. Remove the cluster to be splitted and add two new clusters to S
6. if any one among two added clusters does not include the elements of L,
   then add = add + 1
7. end while
8. select a representative element in each cluster of S that does not include any element of L by using Eq. (6)

Fig. 5. AL-HAC algorithm for extending an initial training set

- random sampling : 임의의 데이터를 선택하는 방법
- LLR[3] : 지역 정보를 이용한 재구성 에러를 최소화하는 데이터 선택 방법
- graphed 3nn[4] : 그래프를 구축하고, 임의의 부분 집합을 선택하여, 가장 근접한 데이터 쌍 중 하나를 삭제해나가는 방법
- AL-HAC : 제안한 방법

Fig. 6은 SVM을 이용한 분류 정확도를 나타내고 Fig. 7은 MLP에 의한 결과를 보여 준다. x축은 훈련 집합에 추가되는 샘플의 수, y축은 분류 정확도를 나타낸다. segment 데이터를 제외한 나머지 모든 데이터에서, 제안한 방법이 분류 정확도가 높은 것을 알 수 있다. 이것은 초기 클래스 정보를 가진 데이터가 주어지지 않은 상황에서 제안된 방법이 효과적이라는 것을 나타낸다.

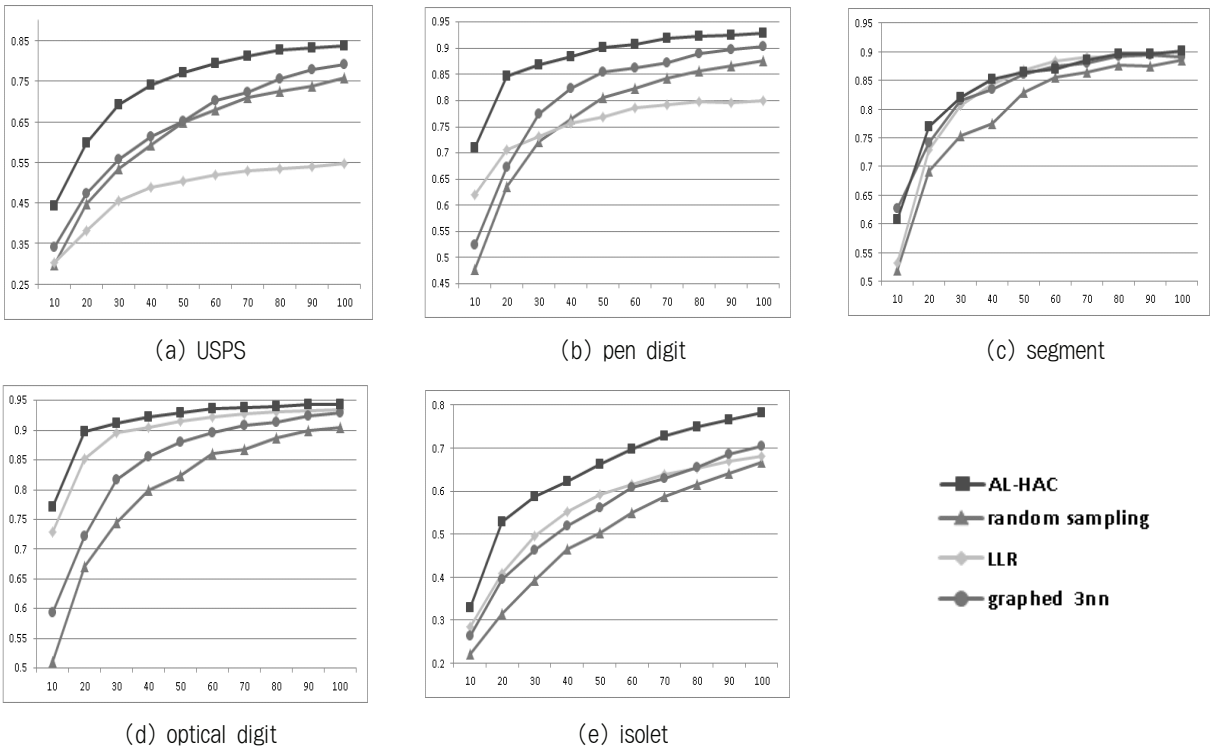
5.2 초기 훈련 집합이 주어진 경우의 실험 비교

각 클래스에서 한 개씩의 데이터 샘플을 임의적으로 선택하여 초기 훈련 데이터를 구성하였다. Fig. 5의 알고리즘에서 k=1로 하여 한 개씩의 샘플을 추가하는 것을 100번 반복하여 총 100개의 샘플을 훈련집합에 더하면서 SVM과 MLP를 이용하여 분류 정확도를 측정하였다. 우리는 비교를 위해서, 2절에서 설명한 SVMactive[2], BvSB[7]와 임의 추출(random sampling)을 사용하여 실험하였다.

Fig. 8과 Fig. 9는 SVM과 MLP를 이용한 실험에 대한 결과를 보여준다. Fig. 8에서는 segment 데이터의 경우는 제안한 방법이 SVMactive와 비슷한 성능을 보이거나 미세하게 높은 결과를 나타낸다. 나머지 데이터에 대해서는 다른 방법들에 비해서 뛰어난 결과를 나타낸다. MLP의 결과인 Fig. 9에서도 비슷한 결과를 나타낸다.

6. 결 론

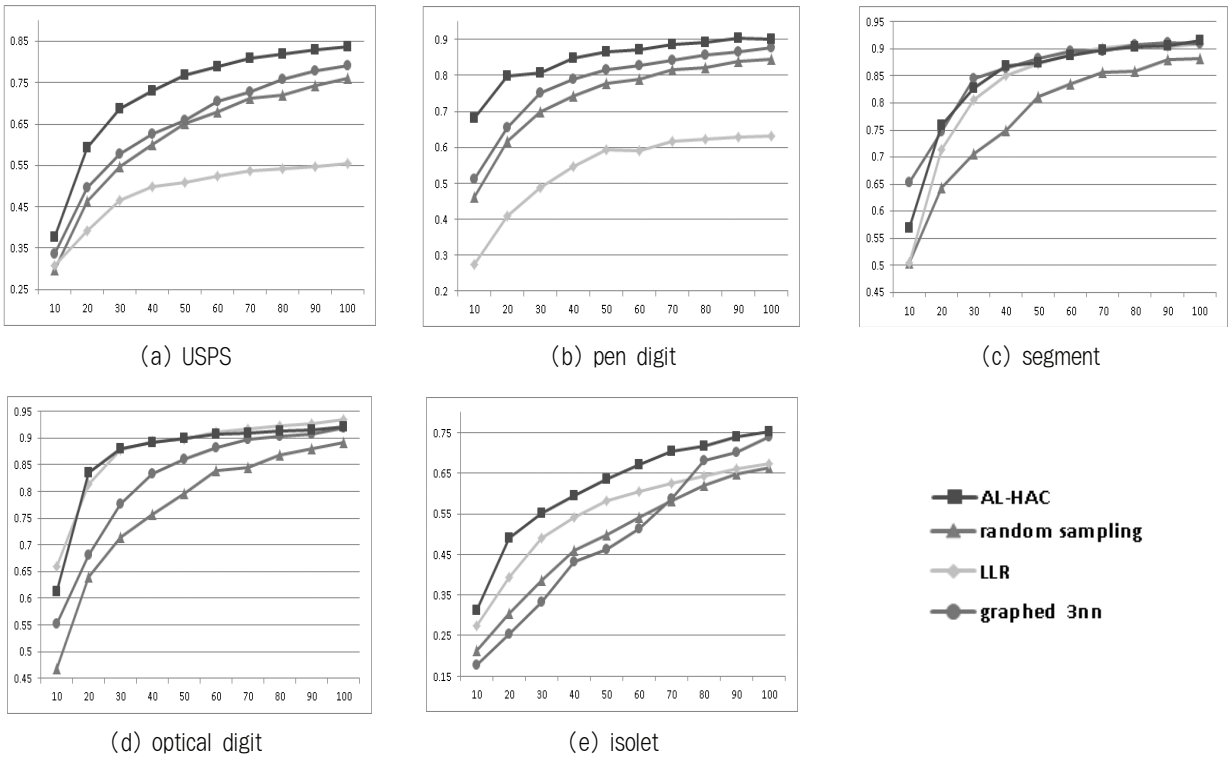
본 논문에서는 계층적 군집화를 이용하여, 초기 클래스 정보를 가진 훈련 데이터가 주어지지 않은 경우와 주어진 경우에 모두 적용 가능한 능동적 학습 방법을 제안하였다. 제안하는 방법은 군집화를 활용하여 잘 형성된 군집에서 각 군집을 대표하는 데이터를 선택함으로써 최소의 데이터 선택으로 전체 데이터 분포를 잘 반영할 수 있다. 초기 훈련 데이터가 없을 경우는 워드 방법을 이용한 군집화를 통해 데이터의 군집을 나누고 각 군집을 대표하는 데이터를 선택하여 전체 데이터 분포를 고려한 효과적인 데이터를 선택한다. 초기 훈련 데이터가 있을 경우, 형성된 덴드로그램을 이용하여 초기 훈련 데이터를 포함하지 않은 군집들에서 대표적인 데이터를 선택한다. 이것은 전체 데이터 분포와 초기 훈련 집합으로 주어진 클래스 정보를 가진 데이터 분포를 모두 고려한 방법으로 실험을 통하여 만족스러운 분류 정확도를 보였다. 또한 대부분의 능동적 학습 방법들이 사용되는 분류기에 의존하여 데이터 선택을 하지만 제안하는 방법은 사용하는 분 분류기와 독립적으로 사용될 수 있다는 장점이 있다.



(a) USPS (b) pen digit (c) segment

(d) optical digit (e) isolet

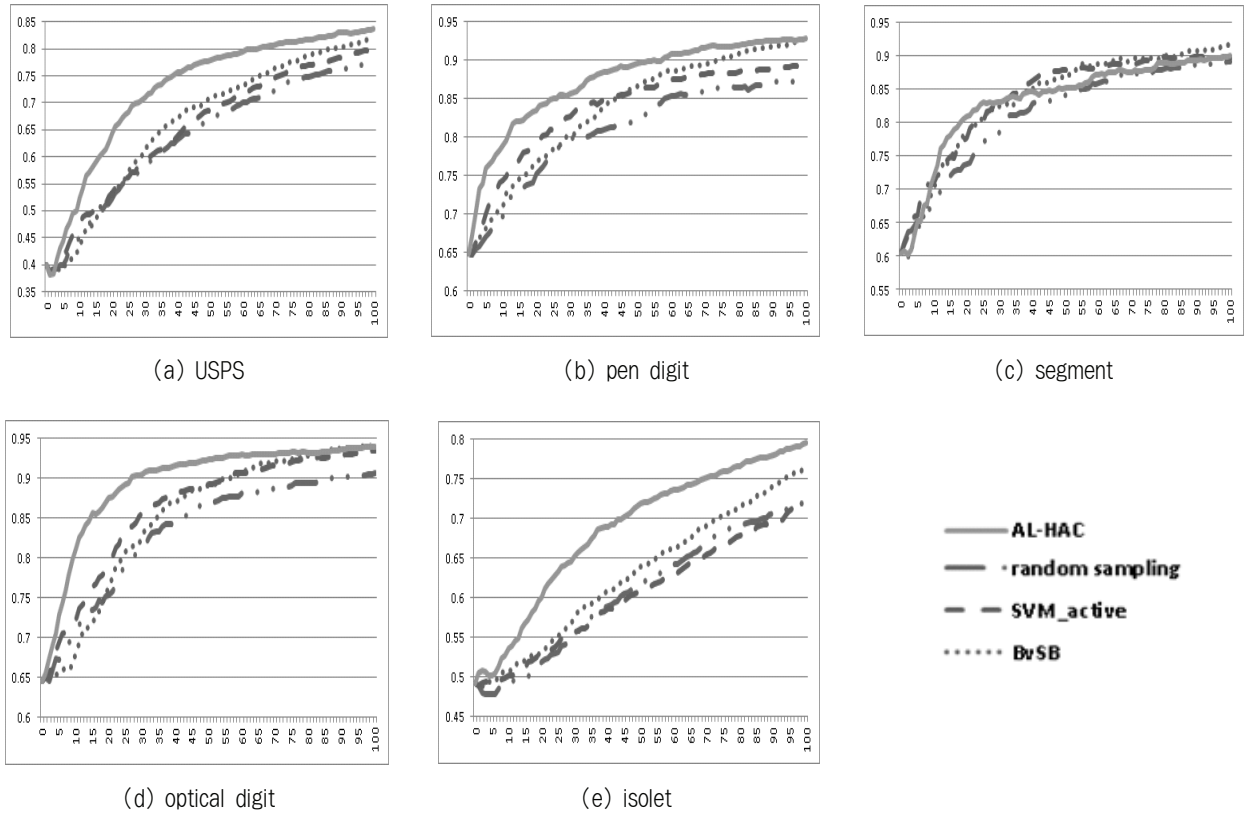
Fig. 6. Prediction Accuracy using SVM when an initial labeled set is not given



(a) USPS (b) pen digit (c) segment

(d) optical digit (e) isolet

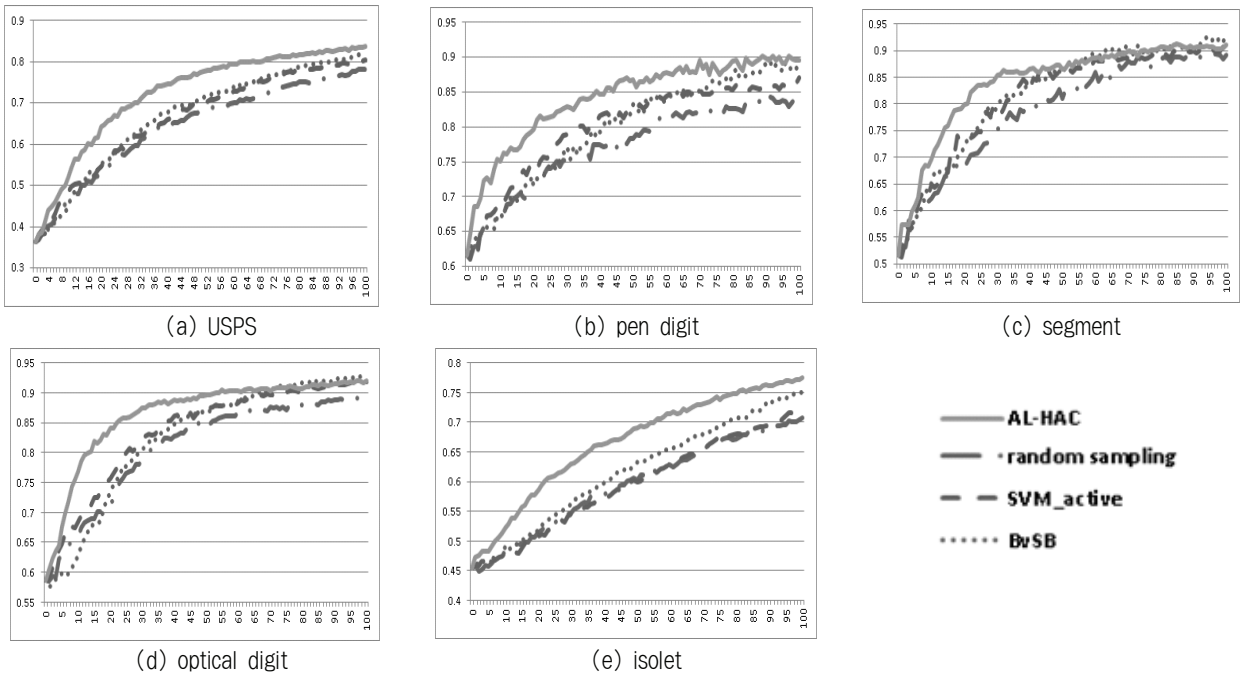
Fig. 7. Prediction Accuracy using MLP when an initial labeled set is not given



(a) USPS (b) pen digit (c) segment

(d) optical digit (e) isolet

Fig. 8. Prediction Accuracy using SVM when an initial labeled set is given(one sample per class)



(a) USPS (b) pen digit (c) segment

(d) optical digit (e) isolet

Fig. 9. Prediction Accuracy using MLP when an initial labeled set is given(one sample per class)

참 고 문 헌

[1] B. Settles, "Active learning literature survey: Computer sciences technical report 1648", *University of Wisconsin-Madison*, 2009

[2] S. Tong and D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification", *J. Machine Learning Research*, Vol.2, pp.45-66, 2002.

[3] L. Zhang, C. Chen, J. Bu, D. Cai, X. He, T. S. Huang, "Active Learning Based on Locally Linear Reconstruction", *IEEE Trans. Pattern Anal. Machine Int.*, Vol.33, No.10, pp. 2026-2038, 2011.

[4] Hoyoung Woo, Cheong Hee Park, "Efficient Active Learning Method Based on Random Sampling and Backward Deletion", *LNCS* Vol.7751, 2013.

[5] P. Tan, M. "Steinbach, and V. Kumar, Introduction to Data Mining", *Addison Wesley*, Boston 2006.

[6] Woo H, C. H. Park, "Active Learning using Hierarchical Clustering and stratified sampling", *KISSE proceeding*, Vol.39, No.2(B), pp.216-218, 2012.

[7] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification", *in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognition*, pp.2372-2379, 2009.

[8] Y. Freund, H.S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm", *Machine learning*, Vol.28(2{3), 1997.

[9] P. Melville and R. Mooney. "Diverse ensembles for active learning", *In Proceedings of the International Conference on Machine Learning (ICML)*, pp.584 - 591. Morgan Kaufmann, 2004.

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

[11] Ward, J. H., Jr., "Hierarchical Grouping to Optimize an Objective Function", *Journal of the American Statistical Association*, 48, 236 - 244, 1963.

[12] N. Semmar, B. Bruguerolle, N. Simon, "Cluster Analysis: An Alternative Method for Covariate Selection in Population Pharmacokinetic Modeling", *Journal of Pharmacokinetics and Pharmacodynamics*, Vol.32, 2005.

[13] Davies, David L. Bouldin, Donald W. A "Cluster Separation Measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1 (2): 224-227. 1979.

[14] Ihara, Shunsuke., "Information theory for continuous systems", *World Scientific*. p. 2. ISBN 978-981-02-0985-8. 1993.

[15] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schutze. "Introduction to Information Retrieval". *Cambridge University Press*. ISBN 978-0-521-86571-5, 2008.

[16] UCI Machine Learning Repository [Internet], http://archive.ics.uci.edu/ml

[17] A Library for Support Vector Machines [Internet], http://csie.ntu.edu.tw/~cjlin/libsvm/

[18] Machine Learning Group at University of Waikato [Internet], http://www.cs.waikato.ac.nz/ml/weka/

우 호 영

e-mail : hywoo@computing-bridge.com
2011년 충남대학교 컴퓨터공학과(학사)
2011년~현 재 충남대학교 컴퓨터공학과 석사과정
관심분야 : Data Mining, High Performance Computing

박 정 희

e-mail : cheonghee@cnu.ac.kr
1998년 연세대학교 수학과(박사)
2004년 University of Minnesota, Computer Science & Engineering (박사)
2005년~현 재 충남대학교 컴퓨터공학과 부교수
관심분야 : Data Mining, 패턴인식