韓國經營科學會誌
第33권 第3號
2008年 9月

미분류 데이터의 초기예측을 통한 군집기반의 부분지도 학습방법*

김응구** · 전치혁***†

A Clustering-based Semi-Supervised Learning through Initial Prediction of Unlabeled Data*

Eung-Ku Kim** · Chi-Hyuck Jun***

■ Abstract ■

Semi-supervised learning uses a small amount of labeled data to predict labels of unlabeled data as well as to improve clustering performance, whereas unsupervised learning analyzes only unlabeled data for clustering purpose. We propose a new clustering-based semi-supervised learning method by reflecting the initial predicted labels of unlabeled data on the objective function. The initial prediction should be done in terms of a discrete probability distribution through a classification method using labeled data. As a result, clusters are formed and labels of unlabeled data are predicted according to the information of labeled data in the same cluster. We evaluate and compare the performance of the proposed method in terms of classification errors through numerical experiments with blinded labeled data.

Keyword : Clustering, Labeled Data, Semi-supervised Learning, Unlabeled Data

1. 서 론

기계학습 분야는 지도학습(Supervised learning), 비지도학습(Unsupervised learning), 부분지도학습(Semi-supervised learning)의 세 영역으로 나누어진다. 지도학습은 입력패턴과 그에 대응하는 출력

논문접수일 : 2007년 12월 27일 논문게재확정일 : 2008년 09월 02일
논문수정일(1차 : 2008년 07월 02일)
* 본 논문은 한국학술진흥재단의 이공계대학교육과정연구지원사업의 연구비로 지원되었음.
** 한국생산성본부 컨설팅본부 CS경영센터 연구원
*** 포항공과대학교 산업경영공학과 교수
† 교신저자

값(범주)을 가진 학습 데이터를 통해 예측모델을 도출하고 이로부터 새로운 입력패턴에 대응한 범주를 예측하는 방법이다. 반면에 비지도학습의 경우 입력패턴만 가지고 있는 데이터를 분석하는 군집분석 등을 일컫는다.

한편, 부분지도 학습은 입력패턴만 있고 출력값은 알려지지 않은 미분류 데이터(Unlabeled data)와 입력 및 출력값이 모두 알려진 분류 데이터(Labeled data)가 혼합된 경우 사용할 수 있는 방법으로 미분류 데이터를 군집화하거나 출력값을 예측하는데 사용한다. 전통적인 지도학습에서 분류 분석은 분류 데이터의 학습을 통하여 분류기(Classifier)를 생성한다. 그러나, 입력패턴에 대한 실제 범주 정보를 얻기 위해서는 해당 분야에 대한 전문적인 지식이 있거나, 고가의 실험 장비를 필요로 하는 등 많은 시간과 비용을 수반하는 것이 통상적이다 [15]. 부분지도학습은 가용한 소수의 분류 데이터에 많은 양의 미분류 데이터를 함께 사용함으로써 미분류 데이터의 출력값을 예측하고자 한다. 따라서 부분지도학습은 문서 분류 [11], 필기체 숫자 인식 [5], 의학 진단 [4] 등과 같이 미분류 데이터는 얻기 쉬우나 분류 데이터를 얻기 위해서 추가적으로 많은 시간과 비용, 인력이 투입되어야 하는 분야에서 효과적으로 사용된다.

부분지도학습은 주목적이 데이터의 분류인지 군집화인지에 따라 다시 부분지도 분류분석(Semi-supervised classification)과 부분지도 군집분석(Semi-supervised clustering)으로 나뉘어진다. 부분지도 분류분석에서는 미분류 데이터를 활용하여 분류성능을 향상시키고자 하며, 부분지도 군집분석은 소수의 분류 데이터를 함께 사용함으로써 군집화의 성능을 향상시키고자 한다. 그러나 부분지도 군집분석에서도 미분류데이터의 범주 예측이 가능하므로 부분지도 분류분석과의 경계가 애매한 점이 있다. 통상적으로 분류/미분류의 혼합 데이터를 다루되 기존의 분류분석 방법론에 근거하고 있으면 부분지도 분류분석, 기존의 군집분석 방법론에 근거하면 부분지도 군집분석이라 일컫는다.

본 연구는 다양한 부분지도 군집분석 중 제약식에 기반한 새로운 부분지도 군집분석 방법을 제안한다. 제 2장에서는 기존의 부분지도 군집분석 방법에 대한 연구 결과들을 소개하며, 제 3장에서 새로운 부분지도 군집분석 방법을 제안하고 있다. 기존의 제약기반 접근법들이 직접적으로 분류 데이터의 범주 정보만을 사용한 것과 달리 제안하는 새로운 방법에서는 미분류 데이터에 대한 범주의 초기 예측결과를 추가적으로 반영한 군집방법을 사용한다. 또한, 제안된 방법은 모든 분류 데이터를 초기 군집 중심으로 사용함으로써 하나의 범주에 대하여 다수의 군집을 형성하는 것이 가능하다는 특징이 있다. 제 4장에서는 다양한 실험 데이터를 통해 본 연구에서 제안된 방법과 기존 연구들의 분류성능을 비교하고 마지막으로 제 5장에선 이에 대한 결론을 내릴 것이다.

2. 기존 부분지도 군집분석

최근의 부분지도 군집분석과 관련된 연구는 학습 데이터 중 분류 데이터로부터 얻은 정보를 군집화 단계에서 활용하는 방안에 대하여 중점적으로 이루어 지고 있다. Bilenko 등 [3]은 군집화 단계에서 분류 데이터가 사용되는 방식에 따라 이를 척도기반 접근법(Metric-based approaches)과 제약기반 접근법(Constraint-based approaches)으로 구분하고 있다.

척도기반 접근법은 분류 데이터의 학습을 통해 얻은 정보를 군집화 과정에서 사용하는 거리척도에 반영하는 방법이다. 척도기반 접근법에서 거리척도의 학습 과정은 미분류 데이터를 배제하고 분류데이터만을 이용한다. Klein 등 [9]은 최단경로 알고리즘(Shortest-path algorithm)에 의해 훈련된 유클리드 거리를 사용하였으며, Xing 등 [14]과 Bar-Hillel 등 [1]은 볼록 최적화(Convex optimization)를 사용하여 훈련된 마할라노비스 거리를 사용하였다.

제약기반 접근법은 분류 데이터로부터 생성된 제약식을 통하여 군집화 과정에서 보다 적합한 자료분할이 일어나도록 하는 방법이다. 실제적으로는 목

적식을 수정함으로써 제약식을 만족하도록 하고있다. Demiriz 등 [7]은 군집분석 시에 식 (1)의 목적식을 사용할 것을 제안 하였다.

$$\min \beta \times Cluster\_dispersion + \alpha \times Cluster\_impurity \quad (1)$$

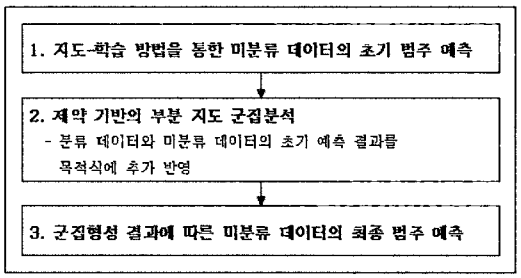
여기서, α와 β는 적절한 양의 계수를 나타내며, Cluster_dispersion는 군집의 퍼짐 정도를 의미하며, Cluster_impurity는 각 군집의 불순도의 척도로 지니 지수(Gini index)를 사용한다. 군집 분석은 유사한 성질을 지닌 데이터들을 같은 군집에 속하도록 하지만 같은 범주의 데이터가 동일한 군집에 속하는 것은 보장하지 않는다. 따라서 식 (1)과 같은 목적식에 지니 지수를 반영함으로써 같은 범주의 분류 데이터들이 가능 한 같은 군집에 속할 수 있도록 하는 것이다.

Wagstaff 등 [13]은 분류 데이터로부터 획득한 정보를 이용하여 제약이 있는 일종의 K-Means 방법(Constraint K-Means라 일컫음)을 제안하였다. 우선 분류 데이터로부터 'Must-link'와 'Cannot-link'의 두 종류 제약 조건을 고려하였다. 그리고 EM (Expectation-Maximization) 알고리즘을 통한 군집화 과정에서 각 관측치를 군집에 할당할 때 위 두 종류의 제약 조건을 적용하도록 한다. Demiriz 등 [7]과의 차이점은 Wagstaff 등 [13]이 제안한 제약식이 각 관측치 단위로 계산되는데 반해 지니 지수는 군집단위로 계산된다는 점이다.

EM 알고리즘을 사용하는 군집분석은 초기 군집중심을 어떻게 설정하느냐에 따라 서로 다른 국소최적해(Local optimal solution)를 갖을 수 있다 [8]. 부분지도 군집분석의 경우 분류 데이터는 사전에 각 관측치들의 범주가 알려져 있는 상태이므로 각 범주별로 분류 데이터의 평균을 취하여 이를 초기 군집중심으로 사용한다. Basu 등 [2]은 분류 데이터를 제약식뿐만 아니라 초기 군집중심을 설정하는 과정에서도 사용할 것을 제안하였다. 그리고 Constraint K-Means에 적용하여 EM 알고리즘 수행시 군집 중심은 갱신 되지만 분류 데이터의 소속 군집은 변경하지 않도록 하였다.

3. 제안된 부분지도 군집방법

기존의 제약기반 부분지도 군집방법에서는 제 2장에서 보듯이 학습데이터 중 분류데이터 정보만을 사용하고 있다. 그러나 미분류 데이터에 대한 범주를 예측하고 이를 동시에 활용하면 군집성능이 좋아질 것으로 예상된다. 본 연구에서는 이러한 아이디어를 바탕으로 [그림 1]과 같은 제약기반의 부분지도 군집분석을 위한 프레임워크를 마련하였다. 우선 전통적인 지도학습 방법을 사용하여 분류 데이터로부터 분류기를 생성하고 이를 미분류 데이터에 적용하여 범주의 초기치를 예측한다. 다음 단계에서 분류 데이터 범주와 미분류 데이터의 범주에 대한 초기 예측 결과를 추가로 목적식에 반영한 부분지도 군집분석을 실시한다. 마지막으로 부분지도 군집분석을 통한 군집형성 결과로부터 미분류 데이터의 최종 범주를 예측 한다.



[그림 1] 제안된 방법의 프레임워크

본 연구에서 대상으로 하는 학습데이터는 총 N개의 관측치로 구성된다. 이 중 일부는 분류 데이터이며 나머지는 미분류 데이터이다. 분류 데이터의 입력값(다차원일 수 있음)을 편의상 X', 이에 대응하는 출력값(범주)을 Y'로 표기한다. 한편, 미분류 데이터의 입력값을 X'', 이에 대응하는 예측된 출력값을 $\hat{Y}''$로 표기한다. 분류/미분류의 구분이 없는 경

우에는 입력값을 X, 출력값을 Y로 표기한다.

3.1 미분류 데이터의 초기 범주 예측

제안된 방법은 우선 분류방법을 사용하여 학습데이터 중 분류 데이터로부터 분류기를 생성하며 이를 미분류 데이터에 적용하여 범주의 초기치를 예측한다. 이를 위하여 기존의 다양한 분류방법들이 사용 가능하나 단순히 미분류 데이터의 범주 예측이 아닌 범주별 사후확률(Posterior probability)을 부여할 수 있는 방법을 사용하도록 한다. 그리고 분류 데이터와 미분류 데이터의 출력값인 범주 정보는 다음과 같이 벡터 형태로 변환 한다. 즉, 총 K개의 범주가 있을 때 분류 데이터의 경우 범주 Y^l를 크기 K의 벡터로 변환하되

$$Y^l = (v_1, v_2, ..., v_K) \quad (2)$$

i번째 범주에 속하면 $v_i = 1$ 나머지는 0을 취한다(i = 1, ⋯, K). 미분류 데이터의 경우 p_i(i = 1, ⋯, K)를 예측된 i번째 범주에 대한 사후확률이라 할 때 예측범주 $\hat{Y}^u$를 다음과 같이 표현한다.

$$\hat{Y}^u = (p_1, p_2, ..., p_K) \quad (3)$$

3.2 새로운 제약기반 군집방법

비지도학습의 군집분석은 군집 형성과정에서 $X = (X^l, X^u)$만을 사용하는데 반해, 기존의 제약기반의 부분지도 군집분석은 Y^l로부터 생성된 제약식의 만족을 위해 목적식에서 (X^l, X^u, Y^l)를 사용하고 있다 [2, 7, 13]. 본 연구에서 제안하는 방법은 목적식에서 기존 연구들에서 사용한 Y^l뿐만 아니라 $\hat{Y}^u$을 추가적으로 사용하고자 한다.

3.2.1 목적함수

전통적인 K-Means 군집분석은 각 군집중심에서 관측치까지의 거리 합이 최소가 되도록 하는 방법으로 거리 척도로 제곱 유클리드거리(Squared Euclidean distance)를 사용하는 경우 식 (4)의 목적식 (J)를 최소화하는 방법이다.

$$\text{Min } J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|X_n - \mu_k\|^2 \quad (4)$$

where

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk} X_n}{\sum_{n=1}^{N} r_{nk}}$$

$$r_{nk} = \begin{cases} 1 & \text{if n - th observation is assigned to } k \text{ - th cluster} \\ 0 & \text{otherwise} \end{cases}$$

본 연구에서는 분류 데이터와 미분류 데이터의 초기 예측결과를 목적식에 추가로 반영하는 방법을 새로이 제안하며 목적식은 식 (5)와 같다. 이 목적식에는 군집별 입력값들의 거리 뿐만아니라 출력값의 거리가 포함되어 있다. 여기서 Y는 분류데이터의 경우 기분류된 범주를, 미분류데이터의 경우 예측된 범주 초기치를 사용한다.

$$\text{Min } J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}(\lambda\|X_n - \mu_{xk}\|^2 + (1-\lambda)\|Y_n - \mu_{yk}\|^2) \quad (5)$$

where

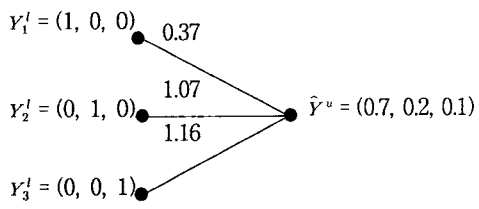
$$\mu_{xk} = \frac{\sum_{n=1}^{N} r_{nk} X_n}{\sum_{n=1}^{N} r_{nk}}, \quad \mu_{yk} = \frac{\sum_{n=1}^{N} r_{nk} Y_n}{\sum_{n=1}^{N} r_{nk}}$$

$$r_{nk} = \begin{cases} 1 & \text{if n - th observation is assigned to } k \text{ - th cluster} \\ 0 & \text{otherwise} \end{cases}$$

목적식에서 λ는 입력값의 거리 및 출력값의 거리 간의 가중치를 부여하기 위한 파라미터로 0에서 1 사이의 값을 갖는다. 특히, λ가 1일 경우 식 (5)의 목적식은 식 (4)의 목적식과 같아지게 되며 전통적인 K-Means와 동일한 결과를 제공한다. 반대로 λ

가 0인 경우 식 (5)의 목적식에서 입력패턴으로부터 계산되는 거리의 영향이 없어지게 되므로 미분류 데이터는 이전 단계에서 계산된 사후 확률이 가장 큰 범주를 갖는 군집에 할당된다.

출력값 간의 거리를 산출할 수 있는 것은 제 3.1 절에서 언급한 바와 같이 분류 데이터와 미분류 데이터의 범주 정보를 벡터 형태로 변환하였기 때문이다. 동일한 범주에 속한 분류 데이터간의 거리는 0이 되며 서로 다른 범주에 속한 분류 데이터들간의 거리는 모두 동일한 값을 갖는다. 또한, 미분류 데이터는 어떤 범주에 대한 사후확률이 커질수록 해당 범주와 거리가 가까워 지며 반대로 사후확률이 작아질수록 먼 거리를 갖는다. [그림 2]의 예에서 미분류 데이터 예측범주 $\hat{Y}^u = (0.7, 0.2, 0.1)$는 범주 1의 사후확률이 가장 크므로 $Y_1^l = (1, 0, 0)$과의 거리가 $0.37(=\sqrt{0.3^2+0.2^2+0.1^2})$로 가장 가까우며 범주 3의 사후확률이 가장 작으므로 $Y_3^l = (0, 0, 1)$와의 거리가 가장 멀다.



[그림 2] 분류 데이터와 미분류 데이터의 범주간 유클리드 거리

3.2.2 초기 군집중심 설정

Basu 등 [2]은 분류 데이터의 각 범주별 입력치들의 평균 값을 초기 군집중심으로 사용함으로써 보다 안정적인 군집 형성 결과를 얻을 수 있음을 보여 주었다. 그러나 미분류 데이터가 혼합되는 경우 분류데이터의 중심을 한 군집의 중심으로 시작하는 것이 반드시 좋을 수 없다. 텍스트 데이터의 경우와 같이 하나의 범주에 대하여 다수의 혼합 성분을 허용하는 것이 좋을 수 있다[11].

제안된 방법은 초기 군집중심 설정 과정에서 모든 분류 데이터의 개별 관측치를 별도의 잠재적 군집중심으로 사용한다. 따라서 알고리즘 초기 단계에서 군집 수는 분류 데이터의 수와 동일하다. 그리고 군집화 과정에서 각각의 군집들은 특별한 조건을 만족 시킬 경우 삭제되며 삭제되는 군집에 속한 데이터들은 각각 다른 가까운 군집에 재할당된다. 최종 군집 수는 전체 범주의 수보다 크거나 같고, 분류 데이터의 수보다 작거나 같은 값을 갖게 된다. 따라서 제안된 방법은 하나의 범주에 대하여 다수의 군집 형성을 가능하게 하는 특징이 있다. 전통적인 군집분석의 경우 하나의 범주를 다수 군집으로 나눌 경우(K > 범주 수) 각 군집 중 동일한 범주에 속하는 군집들을 구분하는데 어려움이 있다. 하지만 부분지도 군집분석의 경우 각 군집에 속한 분류 데이터 정보를 이용하여 해당 군집의 범주를 예측하는 것이 가능하므로 이러한 문제가 발생하지 않는다.

3.2.3 군집 삭제

제안된 방법에 따른 군집화 과정은 시작 단계에서 각 군집별로 하나의 분류 데이터를 가지고 있다. 따라서 군집화 과정에서 서로 다른 군집에 속했던 분류 데이터가 같은 군집에 할당될 경우 분류 데이터가 존재하지 않는 군집이 발생하게 된다. 이 경우 해당(분류 데이터가 존재하지 않는) 군집을 삭제하고 이에 속한 미분류 데이터들은 각각 가까운 군집으로 재할당하는 과정을 거치도록 한다.

3.2.4 파라미터(λ) 최적화

본 연구에서 군집화 과정에 사용한 목적식에 추가된 출력값 간의 거리는 같은 범주에 속한 분류 데이터간에는 항상 0이다. 반면에 분류 데이터와 미분류 데이터간, 미분류 데이터들간에는 항상 이보다 멀다. 따라서 입력패턴으로부터 계산되는 거리가 동일할 지라도 같은 범주에 속한 분류 데이터 간의 거리는 다른 데이터들에 비해 상대적으로 작다. 그러므로 군집화 과정에서 같은 범주에 속한 분류 데이터 간에는 미분류 데이터와 비교하여 상대적으로

같은 군집으로 뭉치려는 경향이 강하게 나타난다. 이러한 경향은 λ값이 0에 가까울수록 더욱 커지며 1에 가까울수록 약하게 나타난다.

전통적인 군집분석이 초기 군집중심에 따라 다른 군집형성 결과를 제공하는데 반해 제안된 방법은 λ가 고정되어 있을 경우 항상 동일한 결과를 제공한다. 본 연구에서는 5-fold 교차타당성(Cross validation) 검증을 통해 적절한 λ값을 설정하고자 한다. 5-fold 교차타당성 검증은 우선 분류 데이터를 5개의 집단으로 나눈 다음 4개 집단만을 분류 데이터로 간주하고 나머지 한 집단의 범주는 감춰 미분류 데이터와 동일하게 취급하여 제안된 방법을 통해 예측을 실시한다. 감춘 범주와 예측범주를 비교하여 예측오차를 계산하며, 이때 예측오차를 최소화 하는 λ값을 최적값으로 선정하도록 한다.

3.3 미분류 데이터 범주의 최종 예측

제 3.2절의 과정을 통해 최종적으로 분류 데이터와 미분류 데이터의 혼합물에 대한 군집 형성 결과를 얻을 수 있다. 이때 최종 군집의 수는 전체 데이터의 범주 수 보다 크거나 같으며 각각의 군집은 최소한 한 개 이상의 분류 데이터를 포함하고 있다. 따라서 미분류 데이터 범주에 대한 예측은 동일한 군집에 속한 분류 데이터를 이용할 수 있다. 본 연구에서는 단순한 방법으로 각 군집에 속한 분류 데이터중 가장 다수의 범주를 해당 군집내의 미분류 데이터의 범주로 최종 예측하는 방법을 사용한다.

3.4 제안방법의 알고리즘 및 수치 예

군집과정은 기본적으로 EM 알고리즘을 사용한다. 따라서 제안된 부분지도 군집방법의 알고리즘을 다음과 같이 단계별로 정리할 수 있다.

단계 1 : 미분류 데이터 초기범주 예측
미분류 데이터의 범주에 대한 초기 예측을 실시하며, 분류/미분류 데이터의 범주들을 벡터화한다.

단계 2 : 제약기반 군집
(파라미터 λ는 5-fold 교차타당성 검증을 통해 최적화 한다. 이는 λ값을 변화시키면서 단계 2와 단계 3을 반복하여 분류오차를 최소로 하는 것이다).

단계 2-1 : 분류데이터의 각 관측치를 초기 군집 중심으로 설정한다.

단계 2-2 : 각 관측치를 입력치 및 출력치의 가중거리가 최소인 군집에 할당한다.

단계 2-3 : 분류 데이터가 존재하지 않는 군집 발생시 해당 군집을 삭제한다.

단계 2-4 : 군집할당 결과로부터 새로운 군집 중심을 계산한다.

단계 2-5 : 식 (5)의 목적식이 최소가 될 때 까지 단계 2-2~2-4를 반복한다.

단계 3 : 미분류 데이터의 최종 범주 예측
각 군집의 분류 데이터중 다수의 범주를 동일한 군집에 속한 미분류 데이터의 최종 범주로 예측한다.

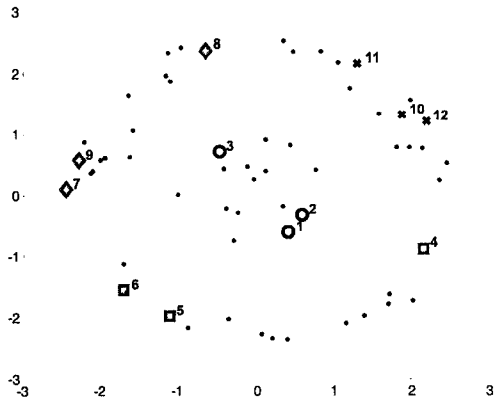
[그림 3]은 제안된 방법에 따라 군집화를 실시한 예이다(λ = 0.5적용). 본 예에 사용된 데이터는 [그림 3(a)]에서 보듯이 두 개 변수에 60개의 관측치이며, 4개의 범주(중앙, 하변, 우상귀, 좌상귀에 각각 동일한 범주가 몰려있음)로 구성되어 있다. 분류 데이터는 각 범주별로 3개씩(총 12개)이며 각각의 번호가 부여되어 있다. 또한 동일한 범주의 분류 데이터들은 같은 기호로 표시하였다. 미분류 데이터는 각 범주별로 12개씩(총 48개)이며 점으로 표시되었다. [그림 3 (b)]는 [그림 3 (a)]의 모든 분류 데이터를 초기 군집 중심으로 사용하여 제안된 방법에 따른 군집화 과정을 1회 수행한 결과이다. 총 12개의 군집이 형성되었으며 미분류 데이터들은 동일한 군집에 속한 분류 데이터와 같은 번호와 기호로 표시하였다. [그림 3 (c)]는 군집화 과정을 2회 수행한 결과이다. 4개의 군집(1, 5, 7, 12)은 [그림 3 (b)]에서 해당 군집에 속한 분류 데이터가 이웃한 다른 군집에 속함에 따라 미분류 데이터만으로 구성되어

있다. 따라서 [그림 3(d)]에서 보듯이 4개 군집들은 삭제되었으며 이에 속했던 데이터들은 각각 다른 가까운 군집으로 재할당 되어 최종적으로 8개의 군집이 형성됨을 확인할 수 있다.

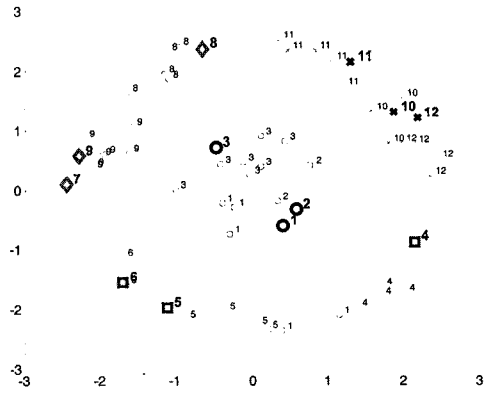
4. 실험 결과

4.1 실험 데이터 설명

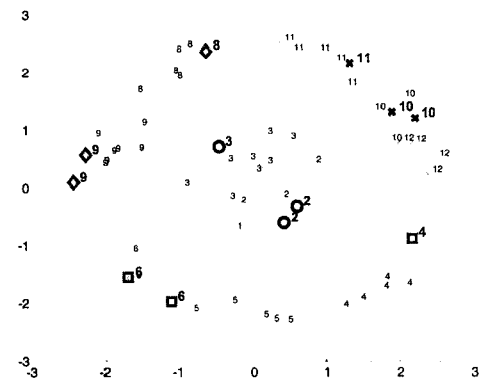
제안된 방법의 성능, 특히 미분류 데이터의 범주 예측 성능을 평가하기 위하여 6개 종류의 데이터가 실험에 사용 되었으며 이를 <표 1>에 정리하였다. coil20과 uspst는 Chapelle and Zien [5]의 연구에서 사용된 실제 사례 데이터들이다. coil20은 20개의 서로 다른 대상을 여러 각도에서 촬영한 흑백 이미지 데이터이며 uspst는 필기체 숫자 인식에서 널리 사용되는 USPS(United States Postal Service) 데이터의 테스트 데이터 부분이다. sonar, segment는 분류 문제에서 자주 사용되는 데이터들로 UCI repository [16]로부터 구할 수 있다. sonar의 경우 변수간 척도의 차이가 심하기 때문에 입력패턴에 대한 정규화를 하였으며 나머지 데이터는 그대로 사용하였다. tae와 g50c는 인공적으로 생성된 데이터들로 tae는 Lee and Lee [10]의 연구에서 사용 되었으며 g50c는 Chapelle and Zien [5]에서 사용되었다.



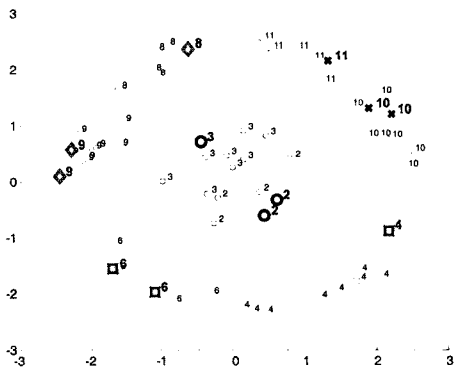
(a) 예제 데이터



(b) 1회 수행



(c) 2회 수행



(d) 3회 수행

[그림 3] 제안된 방법에 따른 군집화 예

<표 1>에 각 데이터별로 차원수(변수수 ; dimension), 범주수(#classes), 학습데이터의 관측수(#train), 학습에 사용한 분류데이터 관측수(#labeled), 그리고 학습에 참여하지 않은 테스트 데이터 관측수(#test)를 차례로 기술하였다. 부분지도 학습은 동일한 데이터를 사용하여 실험을 한다 하더라도 분류 데이터의 양에 따라 예측 성능이 다르게 나타난다. 따라서 본 연구에서 제안한 방법과 다른 연구 문헌의 결과를 비교하기 위하여 실험에 사용한 분류데이터 관측수는 Lee and Lee [10]의 실험과 동일하게 설정하였다. 그리고 테스트 데이터는 미분류 데이터에 대한 예측 완료 후 새로운 입력패턴에 대한 분류 성능을 평가하기 위한 데이터로 사용 되었다. 모든 데이터는 실제 범주가 알려져 있으나 본 실험을 목적으로 학습데이타의 경우 <표 1>에 선정된 분류데이터 수만 범주를 알고 나머지는 모르는 것으로 간주하였으며, 테스트 데이터의 경우 모든 범주를 모르는 것으로 설정하였다.

〈표 1〉 실험 데이터

data set	data set description				
	dimension	# classes	# train	# labeled	# test
coil20	1024	20	1090	40	350
uspst	256	10	1518	50	489
sonar	60	2	104	32	104
segment	19	7	1540	307	770
tae	2	2	600	30	200
g50c	50	2	425	50	125

4.2 초기범주예측 및 파라미터 설정

미분류 데이터의 범주를 초기 예측을 위해 사용하는 분류방법은 범주별 사후확률 부여만 가능하다면 어떤 방법을 사용하여도 무방하나, 본 연구에선 wKNN(weighted K-Nearest Neighbor) [12] 방법을 사용하였다. KNN 방법은 미분류 데이터의 입력패턴과 거리가 가장 가까운 K개의 이웃을 분류 데이터 중에서 선정하고, 이에 속한 데이터들의 분류정보를 이용하여 미분류 데이터의 범주를 예측하는 방법이다. 이때, 예측하고자 하는 데이터와의 거리를 이용하여 가중치를 부여함으로써 예측 결과가 K에 덜 민감하게 할 수 있다. 본 연구에선 유클리드거리를 가중치로 사용하여 식 (6)과 같이 입력패턴 x를 같는 미분류데이터의 범주를 예측한다.

$$\hat{Y}^u(\boldsymbol{x}) = \frac{\sum_{j\in S} w(\boldsymbol{x}, \boldsymbol{x}_j) Y^l(\boldsymbol{x}_j)}{\sum_{j\in S} w(\boldsymbol{x}, \boldsymbol{x}_j)} \tag{6}$$

여기서 S는 K-nearest neighbor 집합을 나타내며, $w(\boldsymbol{x}, \boldsymbol{x}_j)$는 입력패턴 x와 분류 데이터 $\boldsymbol{x}_j$와의 가중치로서 아래와 같다.

$$w(\boldsymbol{x}, \boldsymbol{x}_j) = 1 \Big/ \left\| \boldsymbol{x} - \boldsymbol{x}_j \right\|^2 \tag{7}$$

그리고, K값은 식 (8)를 만족하는 최대의 정수를 사용하였다. KNN 방법의 경우 최적 K값을 구하기 위해선 교차타당성 검증을 사용하여야 하나, 위에서 언급한 바와 같이 wKNN 방법은 예측 결과가 K값에 덜 민감하게 된다. 본 연구에서는 K값을 정함에 있어 수행시간의 단축과 과적합을 방지하기 위하여 범주당 분류데이터의 수를 사용하였으며, 그 최대값이 10을 넘지 않도록 하였다.

$$K \le \min\ (10,\ N_l/c) \tag{8}$$

단, N_l은 분류 데이터의 관측수이며 c는 범주수

〈표 2〉 파라미터(λ) 선정 결과

data set	Selected λ
coil20	0.20
uspst	0.38
sonar	0.67
segment	0.06
tae	0.35
g50c	0.33

이다.

한편, 제안 방법과 관련된 파라미터 λ의 최적값을 구하기 위하여 0부터 0.01단위로 증가 시켜가면서 5-fold 교차타당성 검증을 사용실시 하였다. <표 2>는 교차타당성 검증 결과 구해진 최적 λ값을 데이터 별로 정리한 결과이다.

4.3 결과분석

제안된 방법과 기존의 부분지도 학습 방법들의 성능 평가를 위하여 LDS(Low density separation) [5], semi-SVC [10], Constrained K-Means [2] 등 3개의 방법과 비교 하였다. LDS와 semi-SVC는 Lee and Lee [10]의 실험에서 가장 분류 성능이 좋은 것으로 보고된 방법이다. Constrained K-Means는 분류 데이터로부터 생성된 제약식을 사용하는 부분지도 군집분석 방법으로 별도의 파라미터를 필요로 하지 않으며, 제안된 방법과 밀접하게 관련이 있다.

각 방법의 성능평가를 위해 학습데이터 중 분류 데이터를 <표 1>에서 선정된 갯수만큼 무작위로 추출하여 실험을 수행하되 이런 과정을 100번 반복하여 학습데이터(train) 및 테스트 데이터(test)의 미분류 데이터에 대한 오분류율의 평균과 표준편차를 계산하였으며 이를 <표 3>에 정리 하였다. 단, LDS와 semi-SVC의 결과는 Lee and Lee [10]의 실험 결과이다. 참고로 wKNN 및 LDS의 경우 테스트 데이터에 대한 결과는 산출할 수 없다.

<표 3>에서 볼 때, 제안된 방법의 경우 모든 데이터에서 wKNN보다 뛰어 나거나 비슷한 예측 성능을 보여주고 있다. 이는 제안 방법이 범주의 초기 예측시 wKNN을 사용함에 따라 당연한 결과로 여겨진다. 그러나 다른 방법의 경우 wKNN보다 항상 성능이 좋지는 않음을 볼 수 있다. 예를 들어, segment와 tae의 경우 제안된 방법을 제외한 어떤 부분지도 학습방법 보다도 오히려 wKNN이 좋은 예측 성능을 보여주고 있다. 이러한 현상은 미분류 데이터의 활용을 통한 분류성능 향상 이라는 부분지도 학습의 근본 취지에서 벗어난다. 즉, 학습 과정에서 미분류 데이터를 추가적으로 사용하는 것이 분류 데이터만을 사용하는 것 보다 항상 좋은 결과를 보장하는 것은 아니다는 사실([6])을 입증하고 있다. 이런 관점에서 제안된 방법의 가치를 찾을 수 있겠다.

또한, 제안된 방법은 인공 데이터인 g50c의 경우를 제외하고 Constrained K-Means 보다 항상 좋은 분류성능을 보여준다. g50c는 변수들이 다변량 정규분포를 따르도록 인공적으로 생성된 데이터로서 Constrained K-Means와 같이 실제 범주수와 동일한 수의 혼합 모델을 사용하는 방법이 가장 적합하다. 하지만 다른 실제 사례 데이터들이나 정규분포를 따르지 않는 인공 데이터의 경우 Constrained K-Means는 다른 방법들과 비교하여 성능이 떨어짐을 확인할 수 있다.

LDS는 다른 방법들과 비교할 때 6개 데이터 중 2개 데이터에서 가장 우수한 성능을 보이고 있으며

〈표 3〉 방법에 따른 오분류율의 평균 및 표준편차

Data sets	wKNN	LDS	Constrained K-Means		semi-SVC		Proposed	
	train	train	train	test	train	test	train	test
coil20	27.5±2.0	13.9±1.8	29.9±2.7	30.1±3.0	23.9±1.2	22.0±2.1	25.9±3.3	26.5±3.7
uspst	28.0±2.6	15.4±1.9	29.5±5.0	28.0±4.9	29.6±4.4	28.2±3.1	21.8±3.0	19.9±3.7
sonar	40.3±5.7	38.8±5.6	50.1±4.4	44.0±6.4	34.9±4.9	27.6±5.1	40.2±5.2	31.2±6.0
segment	8.2±0.9	28.9±2.2	27.7±3.9	28.8±3.5	13.0±1.4	17.0±1.9	7.8±0.8	10.2±1.1
g50c	12.6±2.5	8.2±1.6	5.1±0.7	5.4±0.8	8.7±3.4	6.9±3.0	10.7±1.9	11.8±2.7
tae	2.8±1.2	3.1±1.6	10.3±0.4	9.8±0.4	3.7±2.1	3.5±1.9	2.4±1.2	2.9±1.1

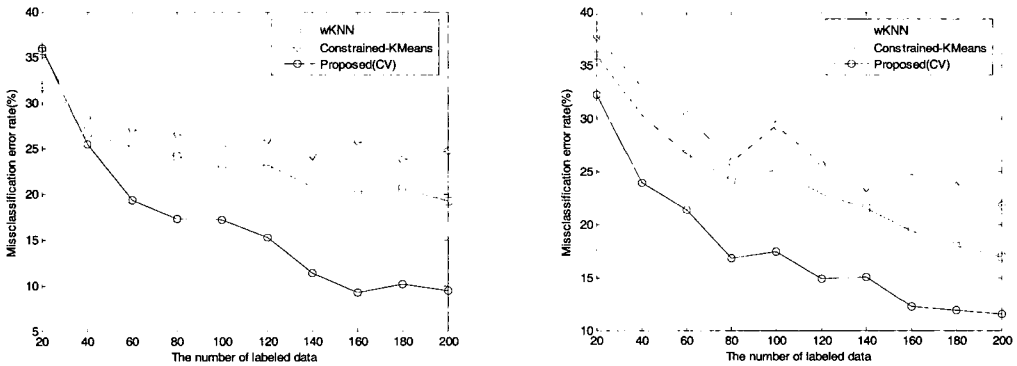
〈표 4〉 데이터별 던컨 사후검정 결과

Data sets	방법별 오분류율 순서
coil20	LDS < semi-SVC < Proposed < wKNN < Constrained K-Means
uspst	LDS < Proposed < wKNN < (Constrained K-Means, semi-SVC)
sonar	semi-SVC < (LDS, Proposed, wKNN) < Constrained K-Means
segment	(Proposed, wKNN) < semi-SVC < Constrained K-Means < LDS
g50c	Constrained K-Means < (LDS, semi-SVC) < Proposed < wKNN
tae	Proposed < (wKNN, LDS) < semi-SVC < Constrained K-Means

나머지 데이터에 대해서도 대체로 좋은 성능을 보여준다. 그러나 LDS는 새로운 입력패턴(테스트 데이터)에 대한 예측을 위해서 전체 알고리즘을 새로 수행해야 한다는 단점이 있다. 반면에, semi-SVC는 평형점(Equilibrium)을 이용한 전체 입력공간의 분할을 통하여, Constrained K-Means와 제안된 방법의 경우와 같이 최종 결과로부터 얻은 군집중심을 이용하여 새로운 입력패턴에 대한 예측을 실시할 수 있다는 장점이 있다.

semi-SVC와 제안된 방법은 서로 6개 데이터 중 3개 데이터에서 좋은 예측 성능을 보이고 있으며 전반적으로 비슷한 예측 성능을 보여준다. 실제 사례 데이터(coil20, uspst)와 UCI Repository 로부터 얻은 데이터(sonar, segment)의 경우 범주당 분류데이터의 수가 많고 차원수가 작은 경우 제안된 방법의 성능이 좋지만 반대의 경우 semi-SVC의 성능이 제안된 방법보다 좋다. semi-SVC는 분류데이터의 수보다 많은 평형점의 적용이 가능하며 최종 도출되는 평형점의 수는 분류데이터의 수보다는 전체 데이터의 분포형태에 주로 영향을 받는다. 반면에 제안된 방법의 경우 최종 도출되는 군집수가 분류데이터의 수에 의해 영향을 받기 때문에, 차원수가 크고 비선형적으로 분포되어있는 데이터의 경우 상대적으로 많은 분류데이터를 필요로 한다.

<표 4>는 <표 3>의 실험결과에 대하여 분산분석후 던컨(Duncan) 사후검정(Post-hoc Analysis)을 실시하여 분류성능의 통계적 유의성을 확인한 결과이다. 방법별로 유의차가 있는 경우 작은 순으로 나열하였으며, 유의차가 없는 경우 괄호로 묶어서 표시하였다. LDS와 semi-SVC의 경우 Lee and Lee [10]의 실험에서 보고된 오분류율의 평균과 표준편차값을 이용하여 정규분포를 따르는 수를 100



(a) coil20 (b) uspst

[그림 4] 분류 데이터의 수에 따른 제안된 방법의 오분류율

개 생성하여 던컨 통계량을 산출하였다.

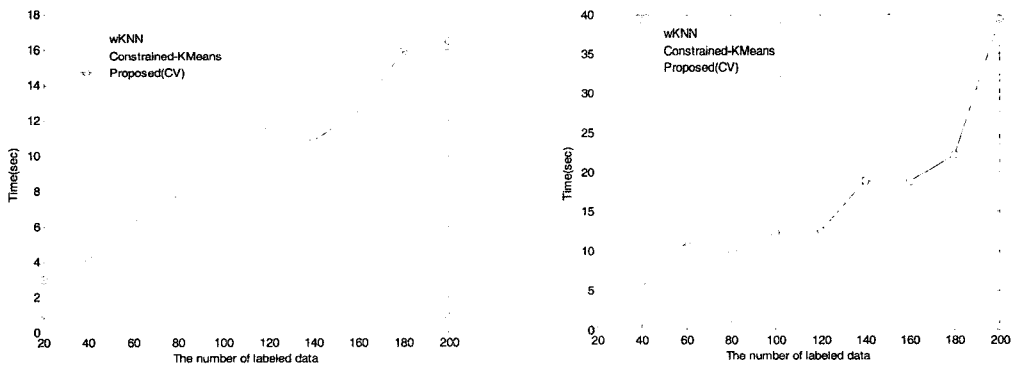
가용한 분류 데이터의 수에 따른 제안된 방법의 성능을 분석하기 위하여 coil20과 uspst 두 개의 데이터를 사용하여 추가적으로 실험을 수행하였다. [그림 4]는 분류 데이터의 수에 따른 오분류율의 변화를 보여준다. 분류 데이터의 수가 늘어남에 따라 제안된 방법뿐만 아니라 wKNN, Constrained K-Means 모두 분류 성능이 향상됨을 확인할 수 있다. 또한 분류 데이터의 수가 적은 경우 각 방법들간의 성능 차이가 적은 편이지만 분류 데이터의 수가 늘어 날수록 제안된 방법의 성능 향상 정도가 다른 방법들보다 큼을 확인할 수 있다.

[그림 5]는 분류 데이터의 수에 따른 알고리즘 수행 시간을 보여준다. Constrained K-Means는 분류 데이터의 수와 관계없이 거의 일정한 시간이 소요되는 반면에 wKNN과 제안된 방법의 경우 분류 데이터의 수가 증가할수록 알고리즘 수행시간이 증가하고 있으며 특히 제안된 방법이 wKNN 보다 수행시간의 증가 정도가 크다. 이러한 현상은 제안된 방법의 경우 전체 데이터의 수가 일정하더라도 분류 데이터의 수가 증가하면 알고리즘 수행시간이 증가할 수 있음을 의미한다. 이는 제안된 방법이 모든 분류 데이터를 초기 군집 중심으로 사용하기 때문에 알고리즘 수행시간이 분류 데이터의 수에 민감하게 영향을 받는 것이다. 그러나, 증가율이 분류 데이터 수에 따라 선형적이므로 현실적인 수행시간 면에서 크게 심각한 것은 아니라 판단된다. Constrained K-Means의 경우 분류 데이터의 수가 증가하더라도 고정된 K값을 적용하므로 수행시간의 차이가 거의 없다.

5. 결 론

본 연구는 미분류 데이터의 범주에 대한 초기 예측치를 추가로 목적식에 반영한 부분지도 군집분석 방법을 제안하고 이를 다양한 실제 데이터에 적용하였다. 부분지도학습의 근본 목적은 대량의 미분류 데이터를 사용하여 분류데이터 만을 이용하는 것보다 좋은 분류기를 생성하는데 있다. 그러나 많은 기존 방법들이 데이터 특성에 따라서는 분류 데이터만을 사용한 분류 방법보다 낮은 성능을 보이는 경우가 있다. 반면에 제안된 방법은 최소한 초기 추정에 사용된 분류 방법보다는 우월한 성능을 제공한다는 점에서 가치가 있다. 또한, 제안된 방법은 대체적으로 Constrained K-Means 보다 좋은 분류 성능을 보여준다.

제안된 방법은 모든 분류 데이터를 초기 군집 중심으로 사용함으로써 하나의 범주에 대하여 다수의 군집을 형성하는 것이 가능하다. 이러한 특성은 하나의 범주가 여러 개의 하위 범주로 구성되어 있는



(a) coil20 (b) uspst

[그림 5] 분류 데이터의 수에 따른 제안된 방법의 수행시간

경우 각 하위범주 별로 별도의 군집을 형성하게 함으로써 분류성능을 향상 시킬 수 있다.

반면에 알고리즘의 수행시간이 분류 데이터의 수에 민감하게 영향을 받으므로 분류 데이터의 수가 많아질 경우 알고리즘의 수행시간이 지나치게 오래 걸릴 우려가 있다. 이와 관련하여, 분류 데이터의 수가 많은 경우 모든 분류 데이터를 초기 군집 중심으로 사용하지 않고 일부만을 사용하거나 분류 데이터에 대한 초기 군집화를 통해 적은 수의 초기 군집 중심을 선택하는 등의 연구가 추가적으로 이루어질 필요가 있다. 또한, 본 연구에서는 미분류 데이터의 초기 예측을 위하여 wKNN 방법을 사용하였으나, 사용되는 분류방법에 따라서 전체 성능이 영향을 받을 수 있으므로 다양한 다른 분류방법에 따른 범주 초기예측을 적용한 실험이 필요하다.

참 고 문 헌

[1] Bar-Hillel, A., T. hertz, N. Shental, and D. Weinshall, Learning distance functions using equivalence relations. *Proceedings of 20th International Conference on Machine Learning*, Washington, USA, 2003, pp.11-18.

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

[3] Bilenko, M., S. Basu, and R. Mooney, Integrating constraints and metric learning in semi-supervised clustering. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004, pp.81-88.

[4] Bouchachia, A. and W. pedrycz, Data clustering with partial supervision. *Data Mining and Knowledge Discovery*, Vol.12, No.1(2006), pp. 47-78.

[5] Chapelle, O. and A. Zien, Semi-supervised classification by low density separation, *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 2005, pp. 57-64.

[6] Cozman, F., I. Cohen, and M. Cirelo, Semi-Supervised learning of mixture models. *Proceedings of the 20th International Conference on Machine Learning*, 2003, pp.99-106.

[7] Demiriz, A., K. Bennett, and M. Embrechts, Semi-Supervised clustering using genetic algorithms. *Intelligent Engineering Systems*, Vol.9(1999), pp.809-814.

[8] Dempster, A.P., N.M. Laird, and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, Vol.39(1977), pp.1-38.

[9] Klein, D., S.D. Kamvar, and C. Manning, From instance-level constraints to space-level constraints : Making the most of prior knowledge in data clustering. *Proceedings of the 19th International Conference on Machine Learning*, 2002, pp.307-314.

[10] Lee, D. and J. Lee, Equilibrium-based support vector machine for semi-supervised classification, *IEEE Trans. on Neural Networks*, Vol.18, No.2(2007), pp.578-583.

[11] Nigam, K., A. McCallum, S. Thrun, and T. Mitchell, Text classification from labeled and unlabeled documents using EM, *Machine Learning*, Vol.39(2000), pp.103-134.

[12] Tan, P.N., M. Steinbach, and V.Kumar, *Introduction to Data Mining*, Pearson Education, Boston, 2006.

[13] Wagstaff, K., C. Cardie, S. Rogers, and S. Schroedl, Constrained K-means clustering with background knowledge. *Proceedings of the 18th International Conference on Machine Learning*, Massachusetts, USA, 2001, pp.577-584.

[14] Xing, E.P., A.Y. Ng, M.I. Jordan, and S. Russell, Distance metric learning, with application to clustering with side information. *Advances in Neural Information Processing Systems*, Vol. 15(2003), pp.505-512.

[15] Zhu, X.Semi-supervised learning literature survey, Computer Sciences TR 1530, *University of Wisconsin-Madison.* http://www.cs.wisc.edu/~jerryzhu/pub/s sl_survey.pdf, 2007.

[16] UCI repository : http://www.ics.uci.edu/~mlearn/MLRepository.html.