비분류표시 데이터의 초기예측을 통한 제약기반 부분-지도 군집분석

A Constraint-based Semi-supervised Clustering Through Initial Prediction of Unlabeled Data

김응구, 전치혁*
포항공과대학교 산업경영공학과
{whiteg, chjun}@postech.ac.kr

Abstract

Traditional clustering is regarded as an unsupervised learning to analyze unlabeled data. Semi-supervised clustering uses a small amount of labeled data to predict labels of unlabeled data as well as to improve clustering performance. Previous methods use constraints generated from available labeled data in clustering process. We propose a new constraint-based semi-supervised clustering method by reflecting initial predicted labels of unlabeled data. We evaluate and compare the performance of the proposed method in terms of classification errors through numerical experiments with blinded labeled data.

1. 서론

전통적인 기계학습 분야는 지도학습(Supervised Learning), 비지도학습(Unsupervised Learning), 부분-지도 학습(Semi-supervised Learning)의 세 영역으로 나누어 진다. 지도 학습은 입력패턴과 그에 대응하는 출력값을 가진 학습 데이터를 통해 함수를 도출하고 이로부터 새로운 입력패턴에 대한 예측을 실시하는 방법이다. 반면에 비지도학습은 입력패턴만 가지고 있는 데이터를 다룬다. 부분-지도 학습은 지도학습과 비지도학습의 장점을 동시에 사용하는 방법으로 본 연구는 분류(Classification) 문제에 중점을 두고 있다. 전통적인 분류방법들은 분류표시 데이터(Labeled Data)만을 사용하여 분류기(Classifier)를 생성하는데 반해 부분-지도 학습은 소수의 분류표시 데이터뿐만 아니라 많은 양의 비분류표시 데이터(Unlabeled Data)를 함께 사용함으로써 보다 좋은 성능의 분류기를 생성한다. 따라서 부분-지도 학습은 문서 분류(Nigam et al., 2000), 필기체 숫자 인식(Chapelle and Zien, 2005), 의학 진단(Bouchachia and pedrycz, 2006)등과 같이 비분류표시 데이터는 얻기 쉬우나 분류표시 데이터를 얻기 위해서 추가적으로 많은 시간과 비용, 인력이 투입되어야 하는 분야에서 효과적으로 사용된다.

부분-지도 군집분석은 소수의 분류표시 데이터를 사용하여 군집분석의 성능을 향상 시킬 뿐만 아니라 비분류표시 데이터에 대한 범주 예측에도 사용한다. 2절에서는 기존의 부분-지도 군집분석 방법에 대한 연구 결과들을 정리할 것이다. 3절에서는 기존의 부분-지도 군집분석 방법중 제약기반 접근법들이 분류표시 데이터로부터 제약식을 생성하고 이를 군집화 과정에 반영한 것과 달리 비분류표시 데이터에 대한 초기 예측치를 추가적으로 반영한 제약기반 부분-지도 군집분석 방법을 제안한다. 4절에서는 다양한 실험 데이터를 통해 본 연구에서 제안한 방법과 기존 연구들의 분류성능을 비교하고 마지막으로 5절에선 이에 대한 결론을 내릴 것이다.

2. 관련 연구

기존의 부분-지도 군집분석을 위한 접근 방법은 분류표시 데이터를 사용하는 방식에 따라 척도기반 접근법(Metric-based approaches)과 제약기반 접근법(Constraint-based approaches)으로 구분된다.(Bilenko et al., 2004)

2.1 척도기반 접근법

척도기반 접근법은 분류표시 데이터의 학습을 통해 얻은 정보를 군집화 과정에서 사용되는 거리척도에 반영하는 방법이다. Klein et al.(2002)는 최단경로 알고리즘(Shortest-path algorithm)에 의해 훈련된 유클리드 거리를 사용하였으며, Xing et al.(2003)은 볼록 최적화(Convex Optimization)를 사용하여 훈련된 마할라노비스 거리를 사용 하였다. 척도 기반 접근법에서 거리척도의 학습 과정은 비분류표시 데이터를 배제하고 분류표시 데이터만을 이용한다.

2.2 제약기반 접근법

제약기반 접근법은 분류표시 데이터로부터 생성된 제약식을 통하여 군집화 과정에서 보다 적합한 자료 분할이 일어나도록 하는 방법이다. Demiriz et al.(1999)는 군집분석 시에 식(1)의 목적식을 사용할 것을 제안 하였다. (단, $\alpha>0, \beta>0$) 군집산포(Cluster Dispersion)는 군집의 퍼짐 정도를 의미하며, 군집불순도(Cluster Impurity)는 각 군집의 불순도의 척도로 지니 지수(Gini Index)를 사용한다. 군집 분석은 유사한 성질을 지닌 자료들을 같은 군집에 속하도록 하지만 같은 범주의 자료가 동일한 군집에 속하는 것을 보장하진 않는다. 따라서 군집분석의 목적식에 지니 지수를 반영함으로써 같은 범주의 분류표시 데이터들이 같은 군집에 속할 수 있도록 하는 것이다.

$$\min_{m_k, k=1,\ldots,K} \beta \times dispersion + \alpha \times impurity \quad (1)$$

Wagstaff et al.(2001)은 분류표시 데이터로부터 획득한 정보를 이용하여 제약이 있는 COP-KMeans 방법을 제안하였다. 우선 분류표시 데이터로부터 다음과 같은 두 종류의 제약 조건을 생성한다.

- Must-link : 두개의 관측치는 서로 같은 군집에 속해야 한다.
- Cannot-link : 두개의 관측치는 서로 같은 군집에 속할 수 없다.

그리고 EM(Expectation Maximization) 알고리즘을 통한 군집화 과정에서 각 관측치를 군집에 할당할 때 위 제약 조건을 적용한다. Demiriz et al.(1999)와의 차이점은 Wagstaff et al.(2001)이 제안한 제약식이 각 관측치 단위로 계산되는데 반해 지니 지수는 군집단위로 계산된다는 점이다.

Basu et al.(2002)는 분류표시 데이터를 제약식뿐만 아니라 초기 군집중심을 설정하는 과정에서도 사용할 것을 제안하였다. EM 알고리즘을 사용하는 군집분석은 초기 군집중심을 어떻게 설정하느냐에 따라 국소최적해(Local Optimal Solution)를 갖는다.(Dempster et al., 1977) 부분지도 군집분석의 경우 분류표시 데이터는 사전에 각 관측치들의 범주가 알려져 있는 상태이므로 각 범주별로 분류표시 데이터의 평균을 취하여 이를 초기 군집중심으로 사용한다. 그리고 이를 앞서의 COP-KMeans에 적용하여 EM 알고리즘 수행시 군집 중심은 매 반복시 마다 갱신하지만 분류표시 데이터의 소속 군집은 변경하지 않도록 하였다.

3. 제안 방법

3.1 비분류표시 데이터의 초기추정

제안된 방법은 부분지도 학습에 앞서 전통적인 지도-학습 방법인 분류방법을 통해 비분류표시 데이터의 예측을 실시한다. 이를 위하여 기존에 연구된 다양한 분류방법들이 사용 가능하나 단순히 비분류표시 데이터의 범주 예측이 아닌 사후확률(Posterior Probability)을 부여할 수 있는 방법을 사용하도록 한다. 그리고 분류표시 데이터와 비분류표시 데이터의 범주 정보는 각각 [그림 1]과 같이 벡터 형태로 변환한다. p_i는 각 관측치가 i번째 범주에 속할 사후확률을 나타내며 분류표시 데이터의 경우 0 또는 1의 값을 갖는다.

3.2 제약기반 부분-군집 분석

기존의 제약기반 부분-지도 군집분석은 군집형성 과정에서 사용되는 목적식에 분류표시 데이터로부터 생성된 제약식을 추가한다. 반면에 제안된 방법은 비분류표시 데이터로부터 생성된 제약식을 추가적으로 사용한다.

제약식의 생성

전통적인 K-Means 군집분석은 각 군집중심에서 관측치까지의 거리 합이 최소가 되도록 데이터를 분할하는 방법으로 거리척도로 유클리드 거리를 사용하는 경우 식(2)의 목적식(J)를 최소화하는 방법이다.

1. 분류표시 데이터 : $(X_l, Y_l) = \{(x_1,y_1),...,(x_l,y_l)\}$

범주	$\hat{Y}_l = (v_1,...,v_k)$
1	(1 , 0 , ... , 0)
2	(0 , 1 , ... , 0)
...	...
k	(0 , 0 , ... , 1)

2. 비분류표시 데이터 : $(X_u, \hat{Y}_u) = \{(x_{l+1},\hat{y}_{l+1}),...,(x_N,\hat{y}_N)\}$

$$\hat{Y}_u = (p_1,...,p_k)$$

where, $0 < p_i < 1, \sum_{i=1}^{k} p_i = 1$

[그림 1] 분류표시 데이터와 비분류표시 데이터의 벡터 변환

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

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

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j (\|X_n - \mu_j\|^2) \\ 0 & \text{otherwise} \end{cases}$$

본 연구에서는 분류표시 데이터뿐만 아니라 비분류표시 데이터의 초기 예측결과로부터 생성된 제약식을 사용하는 방법을 [그림 2]와 같이 제안한다. 부분-지도 군집분석에 앞서 분류표시 데이터와 비분류표시 데이터의 범주 정보를 벡터 형태로 변환하였으므로 입력패턴과 마찬가지로 이들 간의 유클리드 거리를 계산할 수 있다. 제안된 방법은 이를 목적식에 반영함으로써 동일 군집내의 관측치들 간의 입력패턴뿐만 아니라 범주의 동질성을 동시에 추구하고 있다. 그리고 이 둘간의 가중치를 부여하기 위하여 파라미터(λ)를 사용하였으며 λ는 0에서 1사이의 값을 갖는다. 특히, λ가 1일 경우 제안된 방법의 목적식은 식(2)의 목적식과 같아지게 되며, 이 경우 제안된 방법은 K-Means와 동일한 결과를 제공한다. 반대로 λ가 0인 경우 제안된 방법 은 이전 단계에서 사용한 분류방법과 동일한 결과를 보여준다.

초기 군집중심 설정

제안된 방법은 초기 군집중심 설정 과정에서 모든 분류표시 데이터를 잠재적 군집 중심으로 사용한다. 그리고 군집화 과정에서 각각의 군집들은 특별한 조건을 만족 시킬 경우 삭제된다. 따라서 알고리즘 초기 단계에서 군집 수는 분류표시 데이터의 수와 동일하다. 그리고 최종 군집 수는 전체 범주의 수보다 크거나 같고, 분류표시 데이터의 수보다 작거나 같다. 따라서 제안된 방법은 하나의 범주에 대하여 다수의 군집 형성을 가능하게 하는 특징이 있다. 텍스트 자료(Textual Data)와 같이 하나의 주제가 여러 개의 하위 주제들로 구성되어 있는 경우 하나의 범주에 대하여 다수의 혼합 성분(Mixture Components)을 사용하는

1. 비분류표시 데이터의 초기 예측 및 벡터변환

2. 제약기반 부분-지도 군집분석

2-1. 초기군집설정 : 모든 분류표시 데이터 사용

2-2. 목적식(J) 수렴시까지 a, b 반복

- 단, $0 \le \lambda \le 1$

a. E step.

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j (\lambda \|X_n - \mu_{xj}\|^2 + (1-\lambda)\|Y_n - \mu_{yj}\|^2) \\ 0 & \text{otherwise} \end{cases}$$

b. M step.

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

- 분류표시 데이터가 없는 군집 삭제

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

3. 비분류표시 데이터의 예측

[그림 2] 제안 방법

것이 좋다.(Nigam et al., 2000) 또한 제안된 방법은 초기 군집중심으로 항상 모든 분류표시 데이터를 사용하므로 λ가 고정되어 있을 경우 항상 동일한 군집 형성 결과를 제공한다.

군집 삭제

제안된 방법에 따른 군집화 과정은 초기에 각 군집별로 하나의 분류표시 데이터를 가지고 있다. 군집화 과정에서 서로 다른 군집에 속했던 분류표시 데이터가 같은 군집에 속하게 될 경우 분류표시 데이터가 존재하지 않는 군집이 발생하게 된다. 이 경우 해당(분류표시 데이터가 존재하지 않는) 군집을 삭제하고 이에 속한 비분류표시 데이터 들은 각각 가까운 군집으로 재할당 하는 과정을 거치도록 한다.

본 연구에서 군집화 과정에 추가한 제약식은 같은 범주에 속한 분류표시 데이터간의 거리가 항상 0 이다. 반면에 분류표시 데이터와 비분류표시 데이터 간의 거리 또는 비분류표시 데이터들간의 거리는 항상 이보다 멀다. 이 경우 입력패턴으로부터 계산되는 거리가 동일할 지라도 같은 범주에 속한 분류표시 데이터 간의 거리는 다른 데이터들에 비해 상대적으로 작다. 따라서 군집화 과정에서 같은 범주에 속한 분류표시 데이터 간에는 비분류표시 데이터와 비교하여 상대적으로 같은 군집으로 뭉치려는 경향이 강하게 나타난다. 그리고 이는 λ가 0에 가까울수록 그 경향이 강하다.

파라미터(λ) 최적화

전통적인 군집분석이 초기 군집중심에 따라 다른 군집형성 결과를 제공하는데 반해 제안된 방법은 λ가 고정되어 있을 경우 항상 동일한 결과를 제공한다. 따라서 5-fold 교차타당성(Cross-validation) 검증과 베이즈정보기준(Bayesian Information Criterion)등을 통해 최적 λ를 설정할 수 있다. 5-fold 교차타당성 검증은 우선 분류표시 데이터를 5개의 집단으로 나누고 각 집단을 비분류표시 데이터와 동일하게 취급하여 제안된 방법을 통해 예측을 실시한다. 이때 예측오차를 최소화 하는 λ값을 최적값으로 선정하도록 한다. 베이즈 정보기준에 의한 모형선택은 식(3)의 BIC를 최대화 하는 λ를 선정한다. (Hastie, 2001)

$$\text{BIC} = \text{log-likelihood} - \frac{1}{2} p \log(N) \tag{3}$$

단, p : the number of free parameters

N : the number of data points

3.3 비분류표시 데이터의 예측

부분-지도 군집분석 과정을 통해 최종적으로 분류표시 데이터와 비분류표시 데이터의 혼합물에 대한 군집 형성 결과를 얻을 수 있다. 이때 최종 군집의 수는 전체 자료의 범주 수 보다 크거나 같으며 각각의 군집은 최소한 한 개 이상의 분류표시 데이터를 포함하고 있다. 따라서 비분류표시 데이터에 대한 예측은 동일한 군집에 속한 분류표시 데이터를 이용할 수 있다. 간단한 방법으로 각 군집에 속한 분류표시 데이터 중 가장 다수가 속한 범주를 같은 군집내의 비분류표시 데이터에 할당할 수 있다.

4. 실험 결과

4.1 실험 데이터 설명

제안된 방법을 통한 비분류표시 데이터의 성능을 평가하기 위하여 Lee and Lee(2007)의 실험에서 사용한 데이터 일부를 사용하였으며 이를 [표 1]에 정리하였다. 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. tae와 g50c는 실험을 위하여 인공적으로 생성된 데이터들이며 sonar, segment는 분류 문제에서 자주 사용되는 데이터들이다.(UCI repository) sonar의 경우 변수간 척도의 차이가 심하기 때문에 입력패턴에 대한 정규화를 하였으며 나머지 데이터는 그대로 사용 하였다. 테스트(test) 데이터는 비분류표시 데이터에 대한 예측 완료 후 새로운 입력패턴에 대한 예측 성능을 평가하기 위한 데이터로 사용 되었다.

[표 1] 실험데이터 설명

data set	data set description				
	dims	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

[표 3] 오분류율(%)의 평균과 표준편차

Data sets	wKNN	LDS	MEA-EM		semi-SVC		Proposed(CV)		Proposed(BIC)	
	unlabeled	unlabeled	unlabeled	test	unlabeled	test	unlabeled	test	unlabeled	test
coil20	27.8±2.0	13.9±1.8	35.3±5.1	34.0±5.9	23.9±1.2	22.0±2.1	26.6±3.6	27.0±3.6	26.6±3.6	26.9±3.6
uspst	28.0±2.8	15.4±1.9	35.7±2.7	35.7±3.1	29.6±4.4	28.2±3.1	22.1±3.2	20.5±3.5	25.0±2.8	22.4±2.8
sonar	40.1±5.8	38.8±5.6	47.7±5.3	48.4±7.7	34.9±4.9	27.6±5.1	40.6±5.2	31.1±5.7	41.4±5.6	34.6±6.1
segment	8.1±0.9	28.9±2.2	33.5±4.1	37.0±2.9	13.0±1.4	17.0±1.9	7.8±0.9	10.2±1.1	7.8±1.5	10.5±2.4
g50c	12.6±2.0	8.2±1.6	15.2±5.1	15.4±5.2	8.7±3.4	6.9±3.0	10.8±1.7	11.8±2.4	9.9±1.6	6.1±1.2
tae	2.9±1.4	3.1±1.6	5.2±0.2	5.4±0.5	3.7±2.1	3.5±1.9	2.4±1.3	2.9±1.3	2.7±1.6	3.2±1.5

4.2 실험 결과

제안된 방법은 5-fold 교차타당성 검증과 베이즈 정보기준을 사용하여 최적 파라미터를 설정하였으며 그 결과를 [표 2]에 정리 하였다.

[표 2] 최적 파라미터 설정 결과

data set	Proposed(CV)	Proposed(BIC)
	(λ)	(λ)
coil20	0.2	0.16
uspst	0.38	0.06
sonar	0.67	0.3
segment	0.06	0.4
tae	0.35	0.67
g50c	0.33	0.18

제안된 방법과 기존의 부분-지도 학습 방법들과의 성능 평가를 위해 LDS(Chapelle and Zien, 2005), semi-SVC(Lee and Lee, 2007), MEA-EM (Dimitriadou et al., 2002) 3개의 방법을 선택 하였다. LDS와 semi-SVC는 Lee and Lee(2007)의 연구에서 가장 분류 성능이 좋은 것으로 보고된 방법이며, MEA-EM은 상대적으로 분류성능이 떨어지는 것으로 보고 되었으나 군집분석을 이용한 접근방법을 사용하였고 제안된 방법과 밀접한 관련이 있다. 제안된 방법을 사용하여 100번의 반복실험을 통해 계산된 오분류율의 평균과 표준편차를 [표 3]에 정리 하였다.

wKNN(weighted K Nearest Neighbor)은 지도학습 방법의 하나로 비분류표시 데이터와 거리가 가장 가까운 K개의 이웃(분류표시 데이터)를 이용하여 예측하는 방법이다. 이때 각 이웃과의 거리를 이용하여 가중치를 부여함으로써 예측결과가 K에 덜 민감하게 할 수 있다.(Tan et al., 2006) 본 연구에선 식(4)를 만족하는 최대의 정수를 K로 사용하였으며 유클리드 거리를 이용한 가중치를 사용하여 비분류표시 데이터의 초기 예측에 사용 하였다.

$$K \le \min(10,\ l/c)$$

단, l : the number of labeled data (4)

c : the number of classes

Proposed(CV)와 Proposed(BIC)는 [표 2]의 최적 파라미터 설정 결과를 사용한 제안된 방법의 분류성능을 보여준다. LDS, semi-SVC, MEA-EM의 결과는 Lee and Lee(2007)의 실험 결과이다.

λ 최적화 방법에 따른 제안된 방법의 성능은 5-fold 교차검정을 이용하는 경우가 베이즈정보기준을 사용하는 경우보다 근소하게 좋은 성능을 보여주고 있다. LDS는 다른 방법들과 비교할 때 6개 데이터 중 3개 데이터에서 가장 우수한 성능을 보이고 있으며 나머지 3개 데이터에 대해서도 상당히 좋은 성능을 보여준다. 하지만 LDS는 새로운 입력패턴([표 1]의 test)에 대한 예측을 위해서 전체 알고리즘을 새로 수행해야 한다는 점에서 한계가 있다. 반면에 semi-SVC는 알고리즘 수행 과정에서 평형점(Equilibrium)을 이용하여 전체 입력공간에 대한 분할을 실시하며, MEA-EM과 제안된 방법의 경우 최종 결과로부터 얻은 군집중심을 이용하여 새로운 입력패턴에 대한 예측을 실시할 수 있다는 점에서 장점이 있다.

semi-SVC와 제안된 방법은 서로 6개 데이터 중 3개 데이터에서 좋은 예측 성능을 보이고 있으며 전반적으로 비슷한 예측 성능을 보여준다. 하지만 제안된 방법이 성능이 뛰어난 uspst, segment, tae의 경우 지도학습 방법인 wKNN이 semi-SVC 보다 좋은 예측 성능을 보여준다. 특히, segment의 경우 제안된 방법을 제외한 어떤 부분-지도 학습방법 보다도 wKNN이 좋은 예측 성능을 보여준다. 이러한 현상은 비분류표시 데이터의 활용을 통한 분류성능 향상 이라는 부분-지도 학습의 근본 취지에서 벗어난다. 즉, 학습 과정에서 비분류표시 데이터를 추가적으로 사용하는 것이 분류표시 데이터만을 사용하는 것 보다 항상 좋은 결과를 보장하는 것은 아니다.(Cozman et al., 2003) 반면에 제안된 방법의 경우 모든 데이터에서 wKNN보다 뛰어나거나 비슷한 예측 성능을 보여준다.

5. 결 론

본 연구에서는 비분류표시 데이터의 초기 추정치를 반영하는 제약기반 부분-지도 군집분석 방법을 제안하였다. 기존의 제약기반 부분-지도 군집분석에서 제약식이 분류표시 데이터만을 이용하여 생성하는데 반해 제안된 방법은 분류표시 데이터와 비분류표시 데이터의 초기 예측치를 동시에 반영하는 방법을 제안하고 있다는 점에서 특징이 있다. 특히 제안된 방법은 부분-지도 학습에 의한 예측이 전통적인 지도학습 방법에 의한 예측보다 좋지 않은 결과를 가져오는 것을 방지한다. (λ =0 인 경우 지도 학습과 같은 결과를

제공함) 하지만 sonar의 경우와 같이 최적 λ 설정에 따라 낮은 성능을 제공할 수도 있다. 따라서 효과적인 최적 λ 값을 찾는 방법에 대한 연구가 추가되어야 한다. 또한 제안된 방법은 비분류표시 데이터의 초기 예측에 사용되는 방법에 따라 성능의 차이가 있을 수 있다. 따라서, 다양한 지도 학습 방법을 적용하는 비교연구가 필요하다.

참고문헌

[1] Basu, S., Banerjee, A. and Mooney, R. (2002). Semi-Supervised Clustering by Seeding. *Proceedings of the 19th International Conference on Machine Learning*, 19-26.

[2] Bilenko, M., Basu, S. and Mooney, R. (2004). Integrating Constraints and Metric Learning in Semi-Supervised Clustering. *Proceedings of the 21st International Conference on Machine Learning*, 81-88

[3] Bouchachia, A. and pedrycz, W. (2006). Data Clustering With Partial Supervision. *Data Mining and Knowledge Discovery*, vol. 12, no. 1, 47-78

[4] Chapelle, O. and Zien, A. (2005). Semi-supervised Classification by Low Density Separation, *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 57-64

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

[6] Demiriz, A., Bennett, K. And Embrechts, M. (1999). Semi-Supervised clustering using genetic algorithms. *Intelligent Engineering Systems*, 809-814

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

[8] Dimitriadou, E. Weingessel, A. and Homik, K. (2002). A Mixed Ensemble Approach for the Semi-supervised Problem. http://citeseer.ist.psu.edu/590958.html

[9] Hastie, T., Tibshirani, R. and Friedman, J. (2001). The Elements of Statistical Learning, *Springer*, New York

[10] Klein, D., Kamvar, S. D. and Manning, C. (2002). 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*, 307-314

[11] Lee, D. and Lee, J. (2007). Equilibrium-Based Support Vector Machine for Semi-supervised Classification, *IEEE Trans. on Neural Networks*, Vol. 18, No. 2, 578-583

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

[13] Tan, P. N., Steinbach, M. And Kumar, V. (2006). Introduction to Data Mining, *Pearson Education*, Boston

[14] UCI Repository of Machine Learning Database, http://www.ics.uci.edu/~mlearn/MLRepository.html

[15] Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge, *Proceedings of the 18th International Conference on Machine Learning*, 577-584

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