사전정보를 활용한 앙상블 클러스터링 알고리즘

(An Ensemble Clustering Algorithm based on a Prior Knowledge)

고 송 † 김 대 원 ††
(Song Ko) (Dae-Won Kim)

요 약 사전정보는 클러스터링 성능을 유도할 수 있는 요인이지만, 활용 방법에 따라 차이는 발생한다. 특히, 사전정보를 초기 중심으로 활용할 때, 사전정보 간 유사도에 대해 고려하는 것이 필요하다. 레이블이 같더라도 낮은 유사도를 갖는 사전정보로 인해 초기 중심 설정 시 문제가 발생할 수 있기 때문에, 이들을 구분하여 활용하는 방법이 필요하다. 따라서 본 논문은 낮은 유사도를 갖는 사전정보를 구분하여 문제를 해결하는 방법을 제시한다. 또한 유사도에 의해 구분된 사전정보는 다양하게 활용함으로써 생성되는 다양한 클러스터링 결과를 연관규칙에 기반하여 앙상블 함으로써 통합된 하나의 분석 결과를 도출하여 클러스터링 분석 성능을 더욱 개선시킬 수 있다.

키워드 : 클러스터링, 반지도 학습, 앙상블, 연관규칙

Abstract Although a prior knowledge is a factor to improve the clustering performance, it is dependant on how to use of them. Especially, when the prior knowledge is employed in constructing initial centroids of cluster groups, there should be concerned of similarities of a prior knowledge. Despite labels of some objects of a prior knowledge are identical, the objects whose similarities are low should be separated. By separating them, centroids of initial group were not fallen in a problem which is collision of objects with low similarities. There can use the separated prior knowledge by various methods such as various initializations. To apply association rule, proposed method makes enough cluster group number, then the centroids of initial groups could constructed by separated prior knowledge. Then ensemble of the various results outperforms what can not be separated.

Key words : clustering, semi-supervised, ensemble method, association rule

1. 서 론

클러스터링은 오브젝트간의 유사도 측정으로 전체 데이타를 유사도가 높은 오브젝트 그룹으로 구분하는 비지도 학습 방법론이며 훈련 데이타를 통한 예측기를 설계함으로써 테스트 데이타의 예측 적중률로 평가하는 지도 학습 방법과 구분 된다[1,2]. 클러스터링은 데이타에 대한 정보 없이 분석이 가능하여 탐험적인 데이타 분석의 의미를 갖는다. 일례로, 바이오 정보학에서 마이크로어레이기술로 인간의 수천 개에 달하는 유전자에 대한 검사가 가능하게 되었지만, 각 유전자의 특성 및 관계에 대한 정립이 되어 있지 않고 그 관계 정보를 얻기 위해 완전탐사적인(exhaustive) 방법으로 분석하기에 적합하지 않다[3]. 클러스터링은 유전자간의 유사도 측정을 통하여 밀접한 관계를 갖는 유전자로 그룹을 형성할 수 있으며, 유전자 그룹의 공통된 특징을 정의할 수 있다. 또한 유전자의 전체 상황을 통해 사람의 질병 유무의 상태를 예측할 수 있게 된다[4]. 그래서 클러스터링은 유전자에 대한 정보를 추출함으로써 생물학자는 추출된 정보를 기반으로 구체적인 연구를 진행 할 수

† 학생회원 : 중앙대학교 컴퓨터공학과
ssyong20@wm.cau.ac.kr
†† 정 회 원 : 중앙대학교 컴퓨터공학과 교수
dwkim@cau.ac.kr
논문접수 : 2007년 12월 21일
심사완료 : 2008년 12월 16일

정보과학회논문지: 소프트웨어 및 응용 제36권 제2호(2009.2)

있다. 클러스터링의 성능이 개선됨은 의미 있는 정보를 생물학자에게 제공함을 뜻하며, 보다 의미 있는 정보 분석을 위한 클러스터링의 개선이 필요하다.

본 논문은 비지도 학습과 지도 학습 사이의 구분되는 분야로써, 사전정보의 활용을 다루는 반지도 클러스터링을 다룬다. 사전정보는 일부 오브젝트의 레이블을 알고 있는 것으로 가정하며, 사전정보를 클러스터링에 활용함으로써 성능의 개선됨을 여러 논문에서 보이고 있다[5-8].

사전정보의 활용을 통해 성능은 개선되었지만, k-means 계열의 반지도 클러스터링에서 분석 데이타의 형태에 따라 사전정보의 활용 시 발생 가능한 문제점이 있다. k-means는 정규 분포를 가정하고 있기 때문에 한 그룹의 평균으로 중심을 설정하여 그 그룹을 대표하므로, 정규분포가 아닌 형태의 데이타는 분석결과가 좋지 못하며 안정적이지 않다. 이 특성은 사전정보를 활용하더라도 개선되기 힘들다. 즉, 다양한 형태의 데이타 분석이 가능하지 않음을 보이며, 실세계 문제를 다루는 데이타의 형태를 정의할 수 없을 때, 이 문제는 방법론의 신뢰도를 얻기 힘들게 한다.

제안하는 방법은 사전정보의 유사도 측정으로 충돌 발생의 위험을 해소할 수 있도록 사전정보를 구분 사용하는 것으로, 단일 방법론과 그의 앙상블 방법론 두 방법을 제안하며, 이 방법론은 다양한 데이타 형태에 적용이 가능함을 실험 결과를 통해 제시한다. 본 논문의 구성은 2장에서 관련 연구에 대한 내용과 사전정보 활용 시 발생 가능한 문제점에 대한 내용을 지적한다. 3장에서는 문제 발생에 대한 해결 방법을 제안하며, 4장에서는 실험 결과와 분석을 다루고, 5장을 끝으로 마무리 짓는다.

2. 관련 연구

본 연구는 반지도 클러스터링의 사전정보간의 충돌 가능성을 회피 가능한 방법으로 사전정보의 구분 사용과 연관규칙을 기반한 앙상블을 다룬다. 2.1절과 2.2절은 본 연구 소개에 필요한 주제들을 관련 연구를 통해 제시하며, 2.3절에서는 사전정보의 활용 시 발생 가능한 문제에 대한 상황을 다룬다.

2.1 반지도 클러스터링

반지도 클러스터링은 전체 데이타 중 일부 오브젝트의 레이블을 알고 있을 때, 그 정보를 활용하는 방법을 다루는 분야이며, 미리 알고 있는 오브젝트의 레이블은 같은 레이블의 오브젝트 목록인 ML(Must-Link)과 다른 레이블의 오브젝트 목록인 CL(Cannot-Link)로 구분 가능하다[5-8]. 사전정보를 두 목록으로 구분하는 것은 여러 연구에서 공통적으로 제시하고 있으며, 이후의

1. Initialize

$c_h^{(0)} \rightarrow \frac{1}{|M_h|}\sum_{x \in M_h} x$, for $h = 1,...,k$; $t \leftarrow 0$;

c_h: 사전정보 오브젝트 M_h: 레이블이 h인 사전정보 목록

- 사전정보를 활용한 초기화 과정을 위해 사전정보 목록을 레이블(h)에 의해 구분,
- 초기화는 사전정보 목록을 통한 초기 중심을 형성

2. Repeat until convergence

2-1. assign_cluster

$h^* = \arg\min_{h \in 1,...,k} \|x - \mu_h^{(t)}\|^2$; μ_h: 그룹 h의 중심

2-2. estimate_means

$\mu_h^{(t+1)} \leftarrow \frac{1}{|X_h^{(t+1)}|}\sum_{x \in X_h^{(t+1)}} x$, X_h: h에 속하는 그룹의 개수

2-3. t ← (t+1)

3. end

그림 1 seed-means : 사전정보를 활용하여 레이블별 평균으로 초기 중심 설정 후, 유사도 측정에 의해 사전정보 레이블 변화 가능

활용 방법은 다양한 주제로 소개되어 있고 본 논문에서는 그 중 대표적인 방법론 3 가지를 소개하겠다.

2.1.1 Seed-means 클러스터링

k-means의 임의로 초기 중심을 설정하는 방법과 달리 seed-means 클러스터링은 초기 중심을 사전정보의 평균으로 중심을 설정하는 점에서 구분이 되고 이후 동작은 k-means와 동일하며, seed-means 클러스터링에 의한 결과가 사전정보에 의해 알고 있던 레이블과 다를 경우가 발생한다(그림 1). seed-means 클러스터링은 노이즈에 대해서 constraints-means 클러스터링보다 적합한 방법을 제공하며, 노이즈의 성격을 가지는 사전정보의 레이블을 순환 과정을 통해서 변경이 가능하다[5].

2.1.2 Constraints-means 클러스터링

constraints-means 클러스터링은 seed-means 클러스터링의 초기 중심 설정하는 방법과 동일하지만, 사전정보를 통해 알고 있는 오브젝트의 레이블에 대한 유지를 강제한다. 사전정보임을 알고 있다면 유사도 측정을 생략하고 레이블을 유지하게 함으로써 유사도 측정 방법에 상관없이 레이블의 변화를 가져오지 않는다(그림 2). constraints-means 방법은 노이즈를 포함하지 않았을 때 적합하며, 사전정보의 준수를 목표로 하는 방법이므로 노이즈가 섞인 사전정보에 대해서도 변화는 발생하지 않는다[8].

위 두 종류의 k-means 계열의 차이는 그림 2의 2-1-1)과 2-1-2)와 같이 사전정보에 포함되는지에 따른

1. Initialize

$c_h^{(0)} \rightarrow \frac{1}{|M_h|}\sum_{x \in M_h} x$, for $h = 1,...,k$; $t \leftarrow 0$

c_h: 사전정보오브젝트 M_h: 레이블이 h인사전정보목록

- 사전정보를 활용한 초기화 과정을 위해 사전정보 목록을 레이블(h)에 의해 구분,
- 초기화는 사전정보 목록을 통한 초기 중심을 형성

2. Repeat until convergence

2-1. assign_cluster

2-1-1) if $x \notin M_h \rightarrow h^* = \arg\min_{h \in 1,...,k} \|x - \mu_h^{(t)}\|^2$

2-1-2) if $x \in M_h \rightarrow h = \arg h$

2-2. estimate_means

$\mu_h^{(t+1)} \leftarrow \frac{1}{|X_h^{(t+1)}|}\sum_{x \in X_h^{(t+1)}} x$, X_h: h에 속하는 그룹의 개수

2-3. $t \leftarrow (t+1)$

3. end

그림 2 constraints-means : 사전정보를 활용하여 레이블별 평균으로 초기 중심 설정 후, 유사도 측정과 상관없이 사전정보의 레이블은 유지

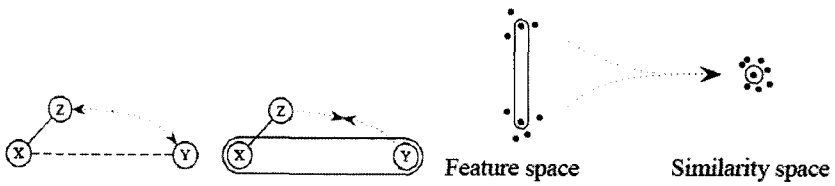
거리 측정 유무에 있다. 이로 인해 사전정보의 레이블의 변화 여부를 결정지운다.

2.1.3 Spatial-level 클러스터링

spatial-level 클러스터링은 계층적(hierarchical) 클러스터링과 유사하며, 사전정보를 활용한다는 점에서 구분이 된다. spatial-level 클러스터링은 모든 오브젝트간의 유사도를 저장하는 매트릭스를 계산하는 과정에서 사전정보에 따른 유사도를 변경한다[7]. 그림 3(a)에 세 오브젝트가 있고, X와 Y가 CL일 때, X와 ML인 Z의 Y에 대한 거리는 최대로 설정하고, (b)에서 X와 Y가 ML일 때, X와 ML인 Z의 Y에 대한 거리는 가장 가깝게 한다. 사전정보와 다른 오브젝트와의 유사도는 ML의 경우 ML 목록 오브젝트와의 거리 중 가장 가까운 거리로 대체한다(그림 3(c)).

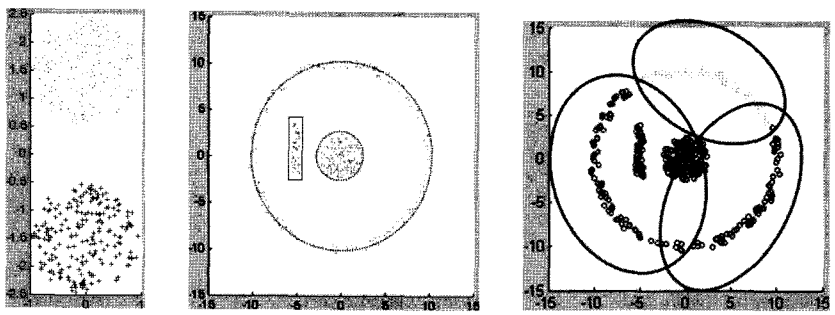
2.2 앙상블 클러스터링

단일 클러스터링 방법론은 특정한 가정을 가지고 설계되어 있으므로, 가정하고 있는 바에 적합한 데이타 셋의 분석은 우수한 성능을 보이지만, 가정하는 것과 다른 형태의 데이타 분석 시 결과는 만족스럽지 못하고 안정적이지도 못하다[9]. 일례로 그림 4(b)는 데이타의 실제 클러스터를 점선으로 표시한 것이며, 그림 4(c)는 k-means를 적용한 결과로써 실선으로 그룹을 표시한 것인데, 원래의 클러스터 그룹과 다름을 알 수 있다. 실세계 정보를 다루는 데이타의 형태에 대한 정보가 있으면



(a) (b) (c)

그림 3 spatial-level 클러스터링 : (a) X와 Y가 다른 레이블일 때, X와 유사도가 높은 Z는 Y와 거리를 멀게 함, (b) X와 Y가 같은 레이블일 때, X와 유사도가 높은 Z는 Y와의 거리를 작게 함, (c) 같은 레이블을 가지는 오브젝트를 동일한 오브젝트(거리 0)으로 설정하며, 다른 오브젝트들과의 거리는 가장 가까운 거리로 설정함



(a) (b) (c)

그림 4 (a) : 정규분포의 데이타로써 k-means의 가정에 적합한 데이타 형태, (b) : 복합적인 형태(3_cluster)의 실제 그룹을 점선으로 표현, (c) : (b)를 k-means 적용한 결과(k:3)

데이타 형태에 적합한 단일 방법론을 적용할 수 있지만, 클러스터링 문제는 비지도 학습문제로서 데이타 형태 정보가 없음을 가정한다. 형태에 대한 정의가 없는 상황에서의 특정한 가정을 가지는 단일 방법론의 적용은 좋은 결과를 제시하기 힘들다.

앙상블 클러스터링은 여러 가지의 단일 방법론을 적용하고 통합함으로써 단일 방법론의 가정을 극복할 수 있도록 하며, 단일 방법론의 사용 방법에 따라 두 가지로 나뉜다[10-12]. 다양한 방법론과 다양한 초기화를 적용한 방법인 이종(heterogeneous) 적용 앙상블과 동일한 방법론과 다양한 초기화를 적용한 동종(homogeneous) 적용 앙상블이 있다. 이종 적용 앙상블은 단일 방법론이 가지는 가정의 약점을 다른 단일 방법론으로 극복할 수 있도록 하는 방법이며, 동종 적용 앙상블은 단일 방법론을 적용하되 초기화만 변경하여 제시되는 각 결과를 통합하는 방법이다.

앙상블 방법론에서 중요한 과정은 통합 방법이며, 클러스터 수를 알고 있을 경우, 각 방법론의 결과를 통합시키기 위한 EM(Expectation-Maximization) 방법론과 클러스터 수를 모르는 경우, 오브젝트간의 관계를 이용하여 통합하는 연관규칙(association rule)등이 있다. 본 논문은 클러스터 수를 알지 못한 상황에서의 접근을 다루므로 연관규칙을 적용하였다[10].

비교 실험 대상 중 하나인 앙상블 클러스터링은 [10]의 것과 동일한 연관규칙을 이용한 앙상블을 적용하며, 구성 방법론은 k-means만을 적용하였으며 k는 데이타 수의 제곱근 보다 조금 큰 상태를 유지하며 임의로 선택한다.

2.2.1 관련 규칙을 이용한 앙상블

본 논문에서 제안하는 반지도 클러스터링 앙상블은 오브젝트간의 유사도를 측정함으로써 앙상블 하기 위하여 연관규칙을 접목하였으며, 연관규칙은 오브젝트나 특징 등이 독립적이지 않은 관계를 가지며 관계 정도의 크기가 다양함을 가정한다[10,13]. 오브젝트간의 관계는 가까운 오브젝트 일수록 강한 유사도를 갖게 되며 클러스터링 결과에서 같은 그룹에 속할 확률이 높음을 가정한다. 그림 5는 그림 4(b) 데이타를 클러스터 수 11개인 k-means로 적용한 것인데, 인접한 데이타는 유사도가 높아서 같은 그룹에 속할 확률이 높다는 것을 볼 수 있다.

오브젝트간의 관계로 클러스터링하기 위해서는 오브젝트 관계 정보의 누적이 필요하므로 수회의 반복적인 단일 방법론 적용이 필요하다. 식 (1)은 누적된 정보($C(i,j)$, association-matrix)를 생성하기 위해 단일 방법론을 N회 적용한 것이며, n_{ij}는 N회의 클러스터링 결과에서 같은 그룹에 속한 횟수를 의미한다. 그림 6은 누적된 정보이며 모든 오브젝트간의 같은 그룹에 속한 횟수가 저장되어 있으며 높은 값을 갖는 것은 그 오브젝트간의 유사도가 높음을 의미한다. C는 모든 오브젝트간의 유사도 정보로써 계층적 클러스터링을 적용하며, HAC 동작 중 발생하는 덴드로그램의 분석을 통해 클러스터 개수를 추정할 수 있다.

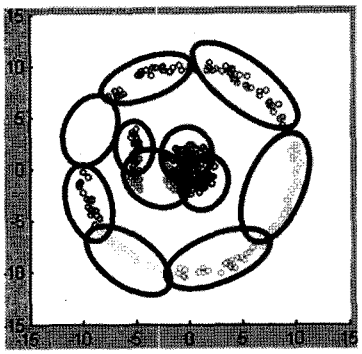


그림 5 k-means(k:11) : k를 크게 적용함으로써 서로 다른 그룹의 오브젝트가 같은 그룹에 속할 확률이 낮아짐

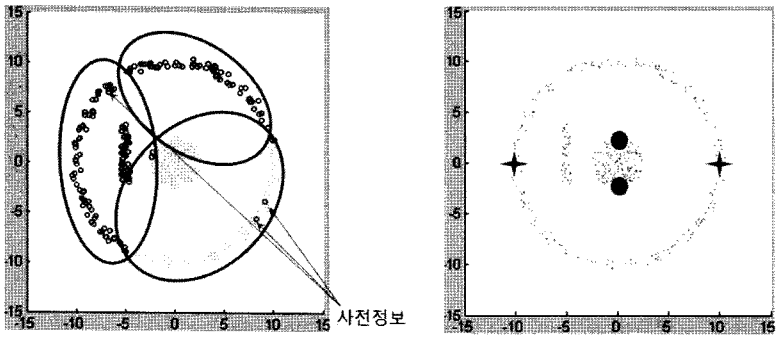
	1	2	3	4	5	...	N
	×	39	47	0	14	...	0
	39	×	40	0	8	...	0
	47	40	×	0	13	...	0
	0	0	0	×	0	...	0
	14	8	13	0	×	...	0
	...	...	...	...	...	×	...
	0	0	0	0	0	...	×

그림 6 C(association-matrix) : 모든 오브젝트(N개)간의 같은 레이블을 갖은 횟수를 저장하고 있으며, 숫자가 높은 것은 그 오브젝트간의 유사도가 높음을 의미하며, 반대의 경우는 유사도가 낮음을 의미함

$$C(i,j) = \frac{n_{ij}}{N} \tag{1}$$

2.3 사전정보 활용 시 발생하는 문제

k-means 계열의 반지도 클러스터링인 seed-means와 constraints-means는 데이타가 정규분포를 가정하고 있으므로(그림 4(a)), 그룹에 속하는 오브젝트들의 평균으로 중심을 설정한다. 정규분포의 데이타는 평균과 분산을 통해 분석이 가능하고, 이러한 형태의 데이타에는 k-means가 좋은 성능을 보이고 있지만, 그림 4(b)와 같이 복합적인 형태의 데이타에는 k-means의 클러스터링 결과가 좋지 않음을 보이고 있다. 즉, 분석하려는 데이타의 형태가 방법론이 가정하는 데이타의 형태와 비슷하지 않을 경우 분석 결과가 좋지 않다.



(a) (b)

그림 7 (a) : 사전정보의 구분 없이 constraints-means을 적용하면 서로 다른 레이블을 가지는 오브젝트가 뒤섞여 있는 경우가 발생함, (b) k-means 계열의 반지도 클러스터링에서 문제가 발생하는 경우

또한 같은 레이블을 갖는 모든 오브젝트의 평균으로 중심을 설정함으로써, 데이타 형태에 따른 사전정보 활용시 사전정보간의 충돌이 발생할 문제점을 가지게 된다. 그림 7(a)은 constraint-means 클러스터링을 적용하여 얻은 결과로써, 화살표가 가리키는 것은 서로 다른 레이블을 가지는 오브젝트가 뒤섞여 있는 것을 보이고 있으며, 문제가 발생하는 원인은 그림 7(b)와 같은 상황이 발생하기 때문이다. 쉘 형태 그룹의 사전정보 2개 (+)와, 정규분포 형태 그룹의 사전정보 2개(o)가 있을 때, k-means 계열의 반지도 클러스터링은 그룹의 평균을 적용함으로써 2개의 서로 다른 그룹의 중심이 같거나 비슷해지게 되므로, 그룹 간의 구분 할 수 있는 방법이 사라지게 된다. 이러한 문제에 대한 해결책 없이 constraints-means을 적용하면 그림 7(a)와 같은 결과가 나오게 된다.

3. 제안하는 연관규칙 기반 앙상블 클러스터링

제안하는 알고리즘은 2가지로 나눠서 소개한다. 3.1절은 클러스티 수를 알고 있는 것을 가정하고 있으며, 3.2는 클러스터 수를 모르는 것으로 가정하고 클러스터의 수는 데이타 분석을 통해 추출할 수 있도록 연관규칙을 적용하고 있다.

3.1 ECM(Expanded Constraints Means: proposed 1)

사전정보를 활용함에 있어서 유사도가 낮은 데이타 간에는 충돌이 발생 가능함을 2.3절에서 지적하였다. k-menas 방법의 특징은 그룹의 중심을 전체 데이타의 평균으로 설정하는 것이기 때문에 충돌의 문제에 대한 해결 방법이 필요하며, 특히 데이타의 양이 적은 사전정보의 활용에 있어서 해결책이 더욱 필요하다.

본 절에서의 충돌에 대한 문제의 해결책으로 사전정보 데이타 간 유사도를 측정하여 유사도가 낮은 데이타는 식 (2)와 같이 분리하는 것을 제시한다. 그림 8(a)에서 쉘 형태의 데이타 중 8개의 사전정보가 있을 때, 8개의 데이타에 대해 HAC를 적용하면 그림 8(b)와 같이 유사도가 높은 데이타 순으로 병합되는 것을 볼 수 있다. 병합되는 데이타 간의 유사도가 낮으면 그래프의 세로에 해당하는 거리가 커짐을 의미한다. HAC의 결과인 덴드로그램을 통해 유사도 추출이 가능하며 낮은 유사도를 갖는 사전정보 데이타에 대한 분리가 가능하다. 유사도가 낮아 분리되었더라도 사전정보를 통해 레이블을 알고 있기 때문에, 이의 준수를 위해 데이타의 레이블을 변함없이 서로 동일하게 유지한다.

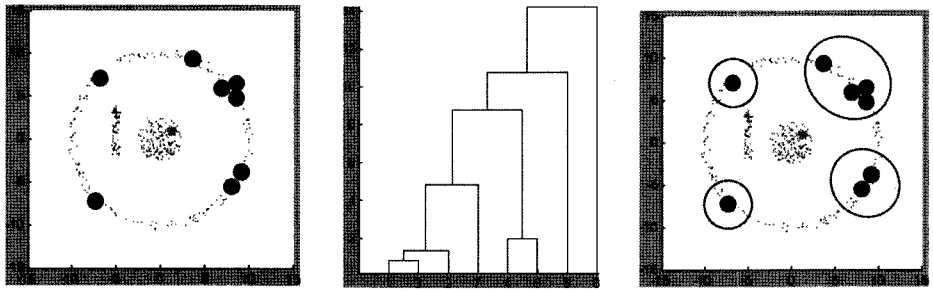
분리된 사전정보는 레이블을 동일하게 유지하면서 충돌을 해결하기 위해 그림 8(c)와 같이 독립적인 중심을 갖도록 한다. 그림 8(c)는 사전정보 데이타를 4개로 분리하여 독립적인 4개의 그룹을 형성하고 있음을 보인다.

ECM은 아래 식 (2), (3)을 통하여 사전정보 데이타를 구분하여 충돌 문제에 대한 해결이 가능한 클러스터링 방법이다. ECM에 대한 알고리즘의 구조는 알고리즘 1에서 보이고 있다. step 1은 식 (2), (3)과 같은 과정을 통해 사전정보를 구분하는 과정, step 2는 사전정보 데이타를 활용한 초기 중심 설정의 과정을 보이며, step 3은 거리 측정과 중심 갱신의 과정을 나타낸다.

그림 9는 알고리즘 1의 step 1과 2에서의 사전정보 데이타의 분리 정도에 따른 클러스터링 결과를 보이고 있다. 분리하지 않은 경우 그림 9(a)와 같이 서로 다른 레이블의 데이타가 뒤섞여 있음을 볼 수 있다. 분리하는 정도에 따란 그림 9(b), (c), (d) 순으로 뒤섞여 있는 경향이 줄어들고 있으며, 실제 클러스터 그룹의 형태와 비슷해지고 있음을 볼 수 있다.

사전정보 데이타 중 유사도가 낮은 데이타를 구분함으로써 그림 9와 같이 클러스터 결과가 향상됨을 볼 수 있다.

$$separated\_ML_j^M = HAC(ML_j) \quad (2)$$
$$j = 1,\ldots,k$$



(a) (b) (c)

그림 8 (a) 셀 형태의 사전정보 위치(●), (b) 셀 형태의 사전정보(8개)를 HAC 적용한 결과의 덴드로그램으로써 낮은 숫자의 세로축 값은 높은 유사도를 의미하며 높은 유사도를 가지는 오브젝트부터 병합됨, (c) HAC를 통해 4개의 그룹으로 구분하면 4개의 독립적인 중심을 형성함으로써, 충돌 발생 가능한 오브젝트는 분리함

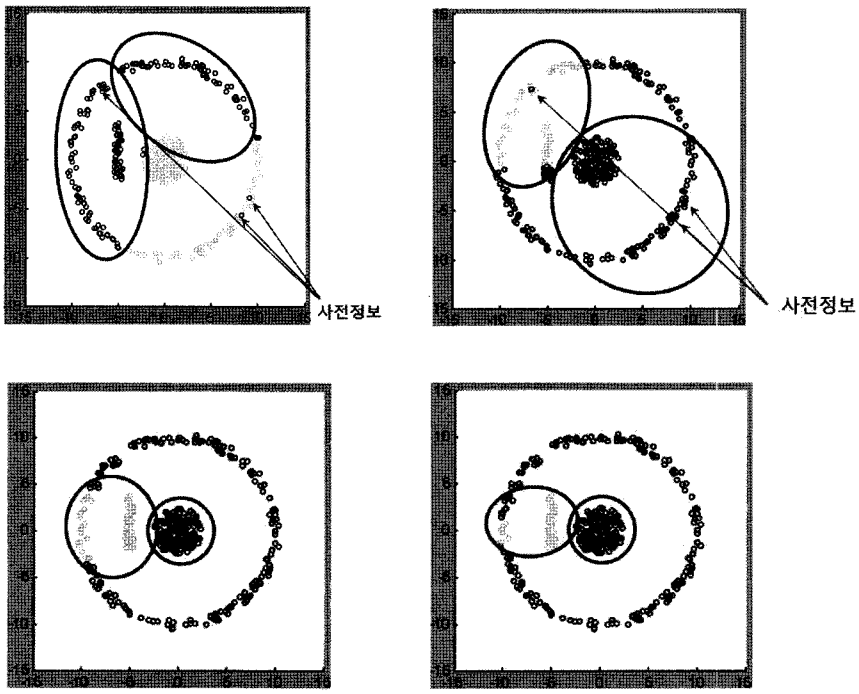


그림 9 3개의 그룹으로 클러스터링 한 결과, 두 개의 그룹은 원으로 표시하며 다른 그룹은 원 밖의 모든 오브젝트를 포함함. 사전정보를 유사도에 따라 구분하는 정도에 따라 서로 다른 레이블이 뒤섞여 있는 경우가 줄어들며, 클러스터링 결과 또한 실제 그룹의 형태와 비슷해짐을 볼 수 있다.

$$ML_j = \bigcup_{m=1}^{M} separated\_ML_j^m \quad (3)$$

3.2 ASEC(Association-rule based Semi_supervised Ensemble Clustering: proposed 2)

ASEC은 클러스터 수를 추정할 수 있는 연관규칙을 이용한 앙상블을 적용한다. 연관규칙을 이용한 앙상블은 오브젝트 간 관련정보의 누적에 의하며, 관련정보의 누적은 단일 알고리즘을 적용하여 앙상블 하는 동종 앙상블하며, 단일 알고리즘은 ECM을 적용한다. ECM는 그림 8에서와 같이 사전정보를 구분하여 활용함으로써, 다른 레이블이 뒤섞여 있는 경우가 없어지는 것을 볼 수 있고 결과가 개선됨을 볼 수 있다. 단일 알고리즘의 결과가 개선될수록 누적되는 관련정보는 개선된 오브젝트 간의 관계 정보를 저장하게 된다.

ASEC은 관련정보의 인접한 오브젝트는 높은 유사성을 갖는다는 가정을 이용하기 때문에 초기화 시 클러스

X : 전체 데이터, Y : 레이블을 알고 있는 일부 데이터로서 사전 정보
$X = \{x_1,...,x_n\}$, $x \in R^p$, 데이터 개수는 n개이며, p 차원의 데이터
$Y = \{x_m,...\}$, $Y \in X$, 사전정보 데이터 개수는 m개, 전체 데이터인 n개 보다 작음
종료 조건 ϵ

step 1) 사전정보 구분
step 1.1) 사전정보 초기 구분 ; 일부 오브젝트의 레이블을 알고 있기 때문에 같은 레이블을 가지는 오브젝트 목록(ML set)과 다른 레이블을 가지는 오브젝트 목록(CL set)으로 구분 가능
- ML set : Must Link set
- CL set : Cannot Link set
- prior_k : 사전정보에 있는 레이블 종류 수(클러스터 수)

step 1.2) 사전정보 세부 구분
- $separated\_ML_j^M = HAC(ML_j)_{j=1,...,k}$; 같은 레이블을 갖는 사전정보 목록이라도 유사도가 낮음으로써 충돌 유발 가능한 데이터는 구분하도록 함(그림 8.b를 통해 구분된 결과 그림 8.c)
- $ML_j = \bigcup_{m=1}^{M} separated\_ML_j^m$; 유사도가 낮은 오브젝트들은 구분하여 독립적인 중심을 형성(그림 8.c)하였지만, 사전정보를 통해 레이블이 같음을 알고 있기 때문에, 구분된 그룹이 독립적인 중심을 갖더라도 레이블은 동일하게 함
- $prior\_Y_j^M$; $separated_ML_j^M$에 포함된 사전정보 오브젝트 수 = $prior\_Y_j^M$

step 2) 초기화
step 2.1) 초기 중심
- $u_j^M = \frac{1}{prior\_Y_j^M} \sum_{c=1}^{prior\_Y_j^M} separated\_ML_j^M$; step 1.2와 같이 M개의 그룹으로 구분된 사전정보는 각각의 중심으로 그룹을 형성(그림 8.c)
- 중심의 개수 = $u_1^M + ... + u_k^M$; 그룹 수는 k개로 주어지며, 그에 따른 중심 μ는 k개를 형성하지만, 위의 step 1.2)를 통해 같은 레이블을 갖더라도 유사도가 낮아 충돌을 유발할 수 있는 오브젝트들은 구분하는 과정을 거치므로 μ_k^M인 M개로 구분함

step 3) 클러스터링
step 3.1) 유사도 측정
(a) $Distance_j^M = \|u_j^M - x_n\|^2$; 모든 중심과 x_n의 유사도 측정
(b) $label_n^M = \arg\min \|Distance_j^M\|$; 가장 가까운 거리의 중심의 레이블 할당
step 3.2) 중심 갱신
(a) $count_j^M = \sum_{x_j \in label_n^M} 1$; 각 중심으로 그룹이 형성된 오브젝트의 수(평균을 위해 필요)
(b) $u_j^M = \frac{1}{count_j^M} \sum_{x_j \in j} x_j$; 각 중심으로 그룹이 형성된 오브젝트를 통해 새로운 중심 설정

step 4) 종료 조건 만족 때까지 (step 3) 반복
step 4.1) 종료 조건 : $(before\_update\, u_j^M) - (after\_update\, u_j^M) < \epsilon$

알고리즘 1. ECM(proposed 1) : 낮은 유사성의 사전정보는 구분함, 구분된 사전정보의 레이블은 동일하게 유지하며, 동작은 k-means와 동일하며 사전정보의 레이블은 고정

터 수를 일반적으로 크게 한다(알고리즘 2). 클러스터 수를 크게 설정하여 얻을 수 있는 장점은 2가지가 있다. 첫째, 연관규칙의 특성을 더욱 활용할 수 있다는 것으로 클러스터 수를 크게 함으로써 유사도가 높은 오브젝트가 같은 그룹에 속할 확률이 높아지게 된다. 둘째, 클러스터 수를 크게 함으로써 하나의 중심으로 형성하는 그룹의 형태는 정규분포에 가까워지게 됨으로써, ECM의 가정에 보다 적합한 데이타의 형태로 형성할 수 있게 한다(그림 5). 관련정보의 누적을 위해 N회의 단일 알고리즘을 적용하며, 관련정보는 오브젝트간의 같은 그룹

데이터 $X = \{x_1, \ldots x_n\}$, $x \in R^p$, 데이터 개수는 n개이며, p 차원의 데이터
사전정보 $Y = \{y_1, \ldots y_m\}$, $Y \in X$, 사전정보 개수는 m개이며, n개 보다 작음

step 1) 사전정보 구분
- ECM(algorithm 1)과 동일

step 2) 초기화
step 2.1) 클러스터 수
- temp_K = sqrt(n)+sqrt(n)*rand() ; 클러스터 수를 크게 함으로써 인접한 오브젝트는 유사도가 높아 같은 그룹에 속할 확률이 높다는 것을 이용함

step 2.2) 초기 중심
- temp_K의 수는 사전정보를 통해 알고 있는 레이블의 수보다 크므로 아래 식과 같이 구분하여 중심을 설정함
- $u_{1\ldots k} = \frac{1}{prior\_Y^M_{1\ldots k}} \sum_{c=1}^{prior\_Y^M_{1\ldots k}} separated\_ML_j^M$;
- $u_{k+1\ldots temp\_K} = rand() * n$; step 2.1을 통해 중심을 충분히 크게 하며, 초기 중심은 위의 식을 통해 사전정보를 활용하며 사전정보를 활용한 초기 중심의 개수가 temp_K보다 적으면, 적은 만큼 다른 오브젝트를 랜덤 선택하여 초기 중심 설정함

step 3) ECM(클러스터 수 : temp_K)
- 클러스터 수를 크게 적용하는 것만 제외하고 알고리즘 1과 동일함

4. $C(i,j)$ 매트릭스(association-matrix) 갱신 및 2-3 반복(N회)
- $C(i,j) = n_{ij}$; i번째 오브젝트와 j번째 오브젝트가 N회의 클러스터링 중 같은 그룹에 속한 횟수

5. association_matrix : $C(i,j)$
5.1 HAC($C(i,j)$) ; $C(i,j)$를 Hierarchical Agglomerative Clustering을 적용하여 유사도가 높은 오브젝트 순으로 클러스터링

알고리즘 2. ASEC : 일반적으로 크게 설정한 클러스터 수를 적용하여 ECM을 N회 적용한 후 앙상블 함에 속한 횟수로써 유사도를 의미하여 HAC를 통해 클러스터링 결과를 제시한다.

4. 실험 및 분석

데이타는 인공 데이타와 실세계 데이타로 구분 실험하였으며, 인공 데이타는 다양한 형태의 데이타를 적용하였고, 실세계 데이타는 마이크로어레이 데이타를 적용하였다. 반지도 학습은 일부의 사전정보를 이용하므로 사전정보의 양은 5, 10, 20, 30, 40개로 하였으며, 사전정보는 임의로 선택하여 실험하였다. 실험 비교 방법론은 표 1과 같이 적용하며, 실험 방법은 30회 적용하여 정확도의 평균으로 평가한다.

4.1 인공 데이타에 대한 실험 결과

인공 데이타 중 표 2 (a), (d), (e)의 모델은 [10]에서 사용했던 형태를 가져와 다시 만든 것이며, (b), (c)는 (a), (d), (e)에서 보이지 않는 형태의 데이타를 임의로 만들었다. 각 데이타는 다양한 형태를 보이고 있으며, 모든 데이타에 대해 전반적으로 좋은 결과가 제시된다면 실세계 데이타의 분석 결과도 좋을 것이라 예상된다. 표 2는 인공 데이타에 대한 설명과 특징에 대한 내용을 보이고 있다.

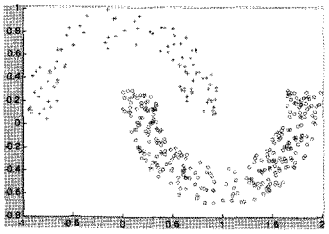
그림 12와 그림 14를 통해 k-means 계열인 seed-means와 constraints-means의 반지도 클러스터링 실험 결과는 데이타가 정규분포의 형태가 아니어서 전반적으로 좋은 결과를 제시하지 못함을 볼 수 있다. XOR 데이타의 경우 각 그룹은 정규 분포에 가깝지만, 같은 레이블의 그룹이 떨어져 있기 때문에 그림 11(c)와 같이

표 1 비교 방법론

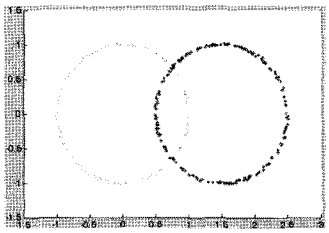
방 법 론	설 명	참고문헌
seed-means	사전정보를 초기화에만 사용	[4]
constraints-means	사전정보를 제약조건 규정	[7]
앙상블 클러스터링	연관규칙 기반 앙상블	[9]
ECM(proposed 1)	구분된 사전정보를 독립된 위치에서 초기화	알고리즘 1
ASEC(proposed 2)	ECM의 결과를 연관규칙 기반 앙상블	알고리즘 2

표 2 데이타 셋

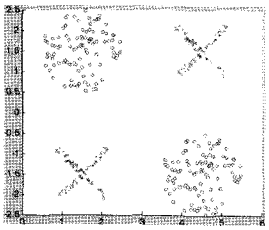
데이타 셋	데이타 수	차원	그룹 수	오브젝트 분포	k-means 정확도
인공 데이타(그림 10)					
(a) half_rings	400	2	2	25%, 75%	66.5 ± 0
(b) occlude_rings	400	2	2	각 50%	71.7 ± 5
(c) XOR data	400	2	2	각 50%	61.2 ± 11.2
(d) three_rings	450	2	3	10%, 45%, 45%	67.6 ± 9.3
(e) 3_clusters	400	2	3	50%, 12.5%, 37.5%	68.5 ± 16.5
실세계 데이타					
colon	62	2000	2	35%, 65%	62.5 ± 17.8
leukemia	72	7129	2	35%, 65%	66.7 ± 9.7
nci60	60	7129	9	3.3% ~ 15%	75.9 ± 11.7



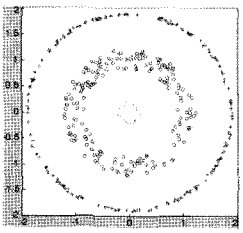
(a) half_rings



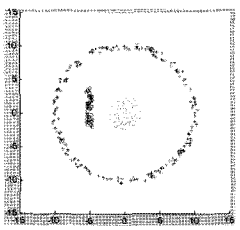
(b) occlude_rings



(c) XOR

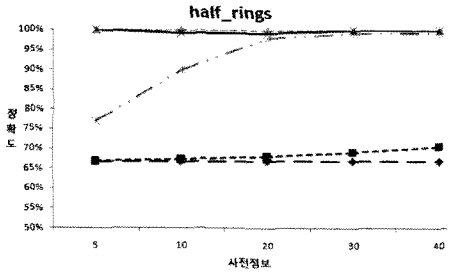


(d) three_rings

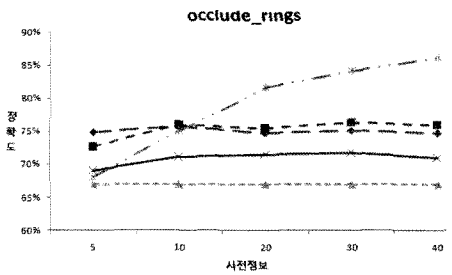


(e) 3_clusters

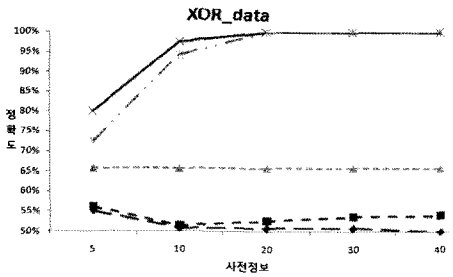
그림 10 인공 데이타 형태



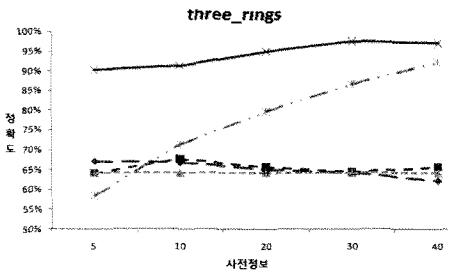
(a)



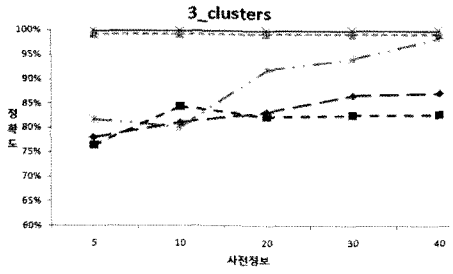
(b)



(c)



(d)



(e)

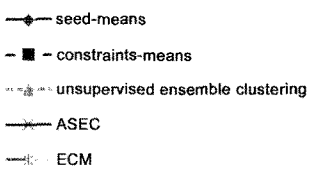


그림 11 인공 데이타 실험 결과

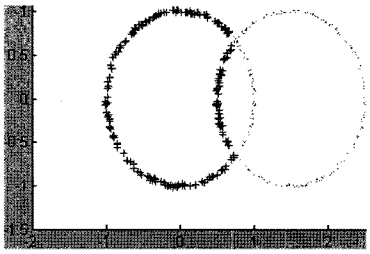


그림 12 seed-means : 중첩된 부분에 대한 분석은 전혀 할 수 없음

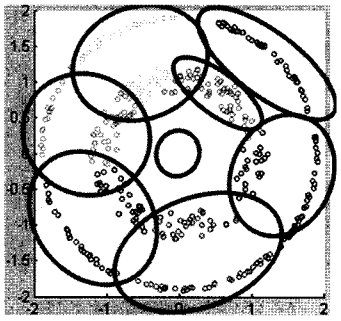


그림 13 k-means(k:8) : 클러스터 수를 작게 설정함으로써 서로 다른 그룹의 오브젝트와 같은 그룹에 속하는 경우 발생

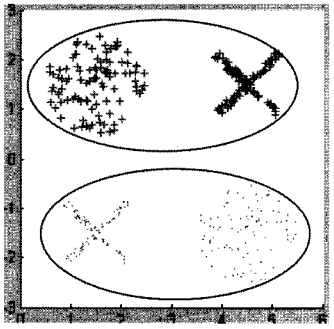


그림 14 앙상블 클러스터링 : XOR 형태의 분석은 어려움

좋지 못한 결과를 보이고 있으며, 같은 이유로 사전정보 양의 증가에 따른 정확도 상승도가 크지 않다. 앙상블 클러스터링은 그림 10(a),(d),(e)와 같이 같은 레이블의 데이타가 비슷한 간격으로 이어져 있는 형태를 보일 때 좋은 성능을 보임을 가정한다[10]. 그럼에도 그림 10(e)의 분석이 좋지 못한 것은 셀 형태인 3개의 동심원은 그림 10(a),(d)의 것보다 복잡한 형태의 것이어서, [10]에서 제시하는 초기화 방법에 의한 k-means의 적용은 잘못된 오브젝트간의 관계 정보를 추출할 경우가 발생하기 때문이다. 그림 13은 [10]의 앙상블 방법을 구성하는 기본 알고리즘으로 이용하는 k-means의 k를 8개로 적용하였을 때의 결과이다. 가장 외곽의 셀 형태의 오브젝트와 가운데 셀 형태의 오브젝트가 같은 그룹에 속하게 되는 경우가 많은 것을 알 수 있다. XOR 데이타는 같은 레이블의 그룹이 떨어져 있는 데이타 형태이므로 그림 11(c))와 같이 클러스터링 결과가 좋지 않으며, occlude_rings 데이타는 같은 레이블의 데이타 분포가 비슷한 간격으로 이어져 있지만 중첩이 발생함으로써 그림 12와 같이 클러스터링 분석 결과가 좋지 않다.

ECM은 정규분포를 가정하는 k-means 계열의 문제점을 해결함으로써 반지도 클러스터링을 개선시켜 다양한 데이타 형태를 분석할 수 있도록 한 것이다. half_rings 데이타의 경우 k-means 계열의 것과 비교하여 개선되었음을 알 수 있다. 사전정보가 늘어남에 따른 성능의 개선이 안정적으로 증가함을 볼 수 있다. occlude_rings의 데이타 분석 결과 사전정보가 늘어남에 따라 다른 알고리즘과 달리 분석 성능이 개선됨을 알 수 있다. 중첩이 발생하더라도 ECM을 이용한 분석의 가능성을 제시한다. XOR_data의 경우 모든 오브젝트로 중심을 설정함으로써 충돌의 가능성이 높은 데이타 형태이지만, 사전정보를 구분 사용함으로써 사전정보의 양이 적더라도 높은 성능을 보이며, 사전정보의 양이 증가함에 따른 성능의 개선은 안정적으로 증가하고 있다. three_rings 데이타의 경우 사전정보의 양이 적을 때에는 낮은 성능을 보이고 있고 한 그룹에 치우쳐진 사전정보를 활용할 경우 낮은 성능을 보일 수 있는 단점을 내포하고 있다. 하지만 사전정보의 양이 증가함에 따른 단점의 극복이 이뤄짐을 볼 수 있다. three_cluster의 결과 사전정보가 증가함에 따라 성능이 개선되고 있다.

단일 알고리즘이 가질 수 있는 한계에 대한 극복 가능성을 ECM을 통해 볼 수 있으며, 표 2에서 제시한 데이타 형태에 대한 분석이 모두 가능함을 보이고 있다. 하지만, 사전정보에 따른 분석 결과의 차이가 크며 한 그룹에 치우쳐진 사전정보를 활용 시 그 그룹에 치우쳐진 결과 분석을 제시할 수 있는 단점을 보이고 있다(그림 15).

ASEC은 ECM을 기본 방법론으로 이용하고 클러스터 수를 알 수 없는 상황에서 데이타를 분석하며, 기본적인 동작은 [10]의 연관규칙을 이용한 앙상블 클러스터링과 동일하지만 구성 단일 알고리즘으로 사전정보를 이용하는 ECM을 적용하였다는 점에서 구분이 된다. ASEC은 같은 레이블이 비슷한 유사도로 이어져 있을 때 좋은 성능을 보이는 특징이 있어서 그림 10(a), (d), (e) 데이타는 좋은 성능을 보이고 있으며, 특히 [10]의 것과 달리 그림 10(e)의 데이타도 좋은 성능을 보임을 알 수 있었다. 클러스터 수를 크게 설정하여 유사도가 강한 오브젝트 간의 정보를 추출할 수 있기 때문에 가능하다. [10]에서 접근하기 힘들었던 XOR 데이타의 분석 또한

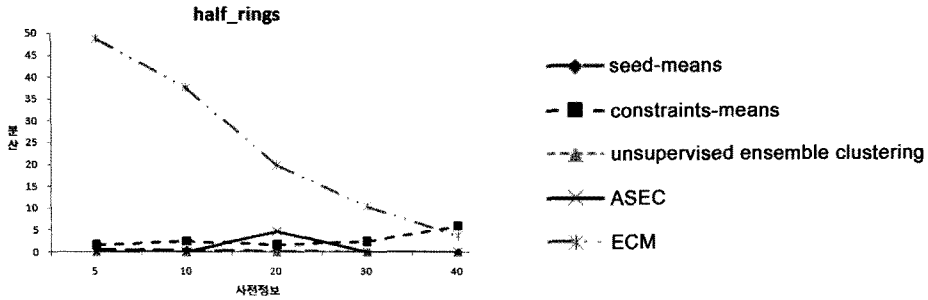


그림 15 half_rings의 평균 정확도의 분산 : 사전정보에 따라 분석 결과가 차이를 많이 보임은 ECM의 개선이 필요한 부분임

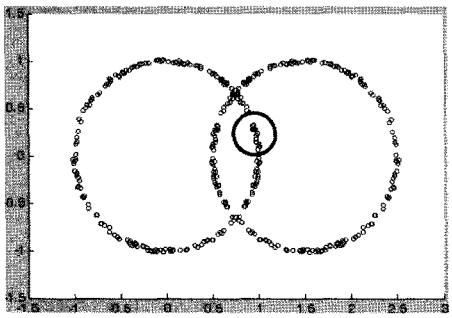


그림 16 occlude_rings 데이타에 ASEC을 적용한 결과 : 두 그룹(실선으로 표시된 원의 내부와 외부)으로 클러스터링

사전정보를 활용함으로써 같은 레이블을 가지는 오브젝트가 떨어져 있더라도 클러스터링이 가능함을 볼 수 있었다. occlude_rings의 중첩이 발생한 데이타에 대한 분석은 ECM보다 성능이 떨어지며, 사전정보가 증가함에도 정확도의 상승을 볼 수 없는 것은 중첩에 의해 다른 레이블이 연결됨으로써 한 그룹으로 치우쳐진 결과를 제시하게 되는데 이에 대한 해결 방법은 연구가 필요하다(그림 8).

4.2 실세계 데이타

실세계 데이타는 생체정보학의 문제를 다루는 데이타로써 유전자 정보 분석을 통해 암의 유무를 판별할 수 있는 colon, leukemia과 nci60을 적용하였으며, 인공 데이타의 실험 방법과 동일하게 진행하였다.

그림 17(a)을 보면 leukemia 데이타의 분석은 오히려 constraints-means가 가장 좋은 성능을 보이고 있고, ECM이 비슷한 성능을 보이고 있으며, ASEC은 안정적이지 않은 결과를 보이고 있다. seed-means의 결과를 통해 사전정보에 의한 중심을 설정 시 사전정보 양이 많아질수록 좋은 초기 중심점에서 시작하여 성능이 향상되고, constraints-means은 사전정보를 제약조건으로 규정함으로써 높은 정확도를 보임을 알 수 있었다. ASEC은 좋지 않은 결과를 보이며 특히 사전정보가 20개일 때 오히려 성능이 떨어짐을 볼 수 있는데, 그 이유는 그림 19와 같은 상황이 발생하기 때문이다. 그림 19의 세로축이 의미하는 것은 N회의 클러스터링을 적용 시 같은 그룹에 속한 횟수인 유사도를 의미하여 가로축은 환자를 의미하는데, ASEC을 적용하여 클러스터링 한 결과의 덴드로그램이다. C그룹의 33번 환자의 유전자 변화폭과 다른 그룹 오브젝트의 유전자 변화폭의 차이를 극명하게 내는 노이즈가 33번 환자의 유전자에 발생한 것으로 예상한다. 노이즈가 포함된 leukemia 데이타의 분석은 다른 방법론에 비해 우수한 성능을 보이고 있지 않다.

colon의 분석은 제안하는 두 가지 방법들이 사전정보 양의 상승에 따라 정확도가 상승하며, k-means 계열의 방법은 오히려 정확도가 떨어지는 것을 볼 수 있었다. ASEC이 ECM을 구성 알고리즘으로 사용하여 그 결과를 앙상블 하는 방법이지만 오히려 ASEC의 성능이 떨어짐을 볼 수 있는데, 그 이유는 그림 19와 같은 상황이 발생하기 때문으로 예상한다. 환자 그룹이 3개로 나옴을 알 수 있고 그 중 눈에 띄는 것은 그룹 C의 부분이다. C의 그룹에 속하는 환자(가로축 14)는 하나만이 존재한다는 것을 알 수 있으며, C에 속하는 환자의 유전자가 다른 환자의 것과 큰 차이를 보임으로써 같은 그룹에 속하는 확률을 낮추기 때문이다. 실제로 14번 환자가 다른 그룹과 구분 지어질 상태에 있을 수도 있겠으나, 실세계의 데이타는 노이즈를 배제할 수 없을 때 노이즈를 선별할 수 있는 분석이 필요하다.

nci60 데이타의 분석 결과는 ECM의 경우 전체적으로 개선된 성능을 보이고 있지만 사전정보가 10개일 때 성능이 오히려 감소하는 것을 볼 수 있는데, 그 이유는 9클래스의 데이타이며 각 그룹의 오브젝트의 수가 불균형의 형태를 띠고 있기 때문이다. 사전정보로 많은 오브젝트를 가지고 있는 그룹의 것이 뽑힐 확률이 높기 때문에 적은 개수의 사전정보를 활용 시 큰 그룹에 치우친 결과를 제시하므로 결과가 좋지 못하다. 하지만 사전

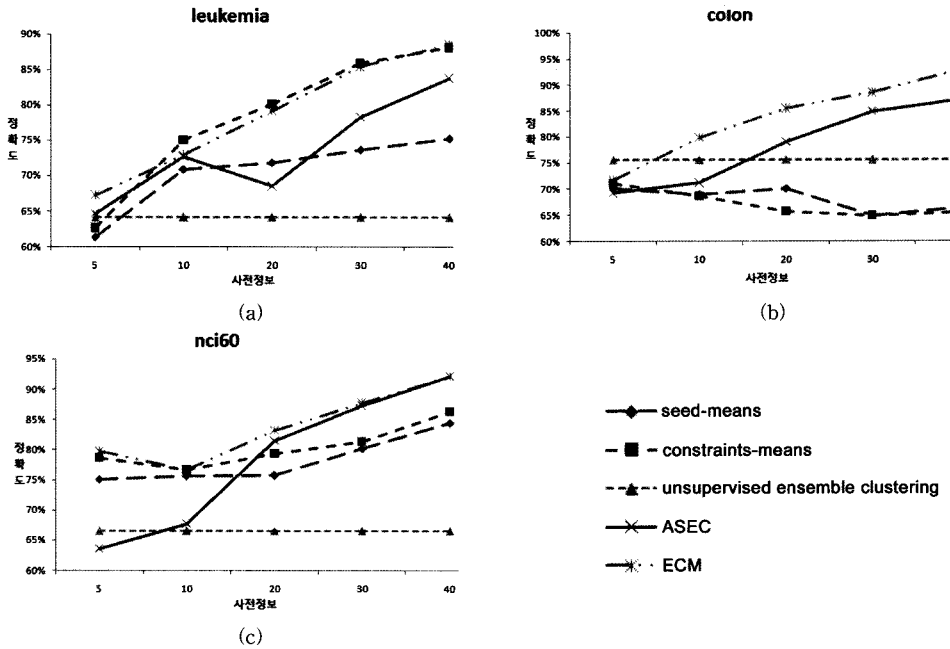


그림 17 마이크로어레이 실험 결과

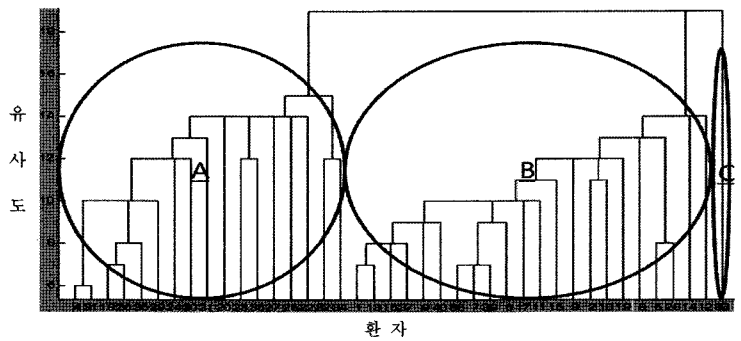


그림 18 leukemia 데이타의 ASEC 클러스터링 결과인 누적정보(association_matrix)에 HAC를 적용한 결과의 덴드로그램, C그룹에는 한 명의 환자만이 있는 것을 볼 수 있음

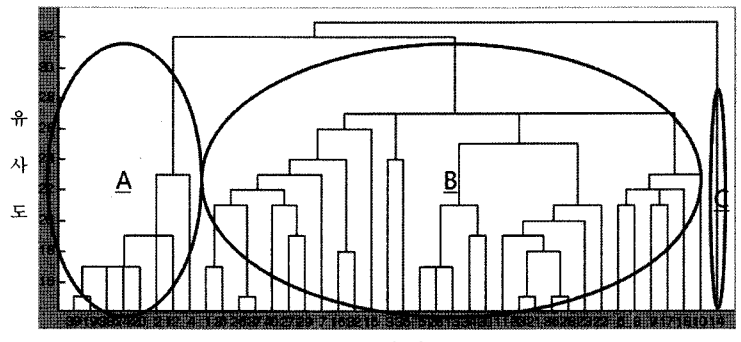


그림 19 colon 데이타의 ASEC 클러스터링 결과인 누적정보(association_matrix)에 HAC를 적용한 결과의 덴드로그램, C그룹에는 한 명의 환자만이 있는 것을 볼 수 있음

정보가 많아짐에 따라 불균형의 현상도 극복하고 있음을 알 수 있다. ASEC의 경우도 ECM의 이유와 마찬가지로 불균형의 발생으로 인해 한쪽 클래스에 극명하게 치우친 결과를 제시하므로 초기의 성능은 가장 좋지 않게 나타나지만, 활용하는 사전정보의 양이 증가함에 따라 성능의 개선정도가 커짐을 알 수 있다.

nci60의 데이타에 대한 실험 결과 제안하는 알고리즘들은 사전정보의 양이 증가함에 따라 성능이 개선됨을 보이지만, 초기 적은 양의 사전정보 활용 시 불균형에 의한 악영향으로 성능이 좋지 않은 점은 앞으로 연구가 진행해야 할 부분이다.

5. 결론 및 향후 과제

본 논문은 클러스터링 분석 성능을 개선시킬 수 있는 요인으로 활용할 수 있는 사전정보의 활용 시, 문제 발생 요인에 대하여 지적하고 이를 해결함으로써 분석 성능을 향상시킬 수 있도록 하였다. 즉, 사전정보 중에서 같은 레이블을 갖는 오브젝트라도 유사도는 다양할 수 있는데, 낮은 유사도 간 오브젝트를 클러스터링 분석 시 동시에 활용할 때 문제가 발생할 수 있기 때문에, 유사도가 낮은 오브젝트들을 구분한 후 사전정보를 다양하게 클러스터링 분석에 활용하는 방법을 통한 결과를 통합하는 방법을 제시하였다. 실험 결과에서 다양한 형태의 데이타인 인공 데이타에서 성능이 개선되고 있음을 볼 수 있다. 그러나 실세계 데이타인 바이오 데이타의 분석 성능은 눈에 띄게 향상되지 않고 있다. 이는 바이오 데이타가 가지고 있는 문제 중 하나인 데이타의 불균형이 제안하는 방법에서 제대로 해결하지 못한 것으로 보이며, 이에 대한 보완 방법이 추후 필요하다.

또한 사전정보의 획득 시 본 논문에서는 인공 데이타와 바이오 데이타의 구분 없이 데이타 중 일부를 랜덤으로 선택하여 사전정보로 활용하였다. 바이오 데이타는 유전자에 대한 것으로써 많은 연구가 진행되어 유전자 온톨로지 등을 통하여 그 동안 밝혀진 정보를 보관 및 공유하고 있다[14]. 추후 사전정보에 대해 다뤄야 할 부분은 유전자 온톨로지를 통해 유전자간 유사도를 측정하여 클러스터링에 활용하는 방법이 될 것이다.

참 고 문 헌

[1] A.K. Jain, M.N. Murty, P.J. Flynn, "Data Clustering : A Review," ACM Computing Surveys, Vol.31, No.3, September.

[2] Brian S.Everitt et al, "Cluster Analysis," ARNOLD

[3] Aidong zhang, "advanced analysis of gene expression microarray data," World Scientific, 2006.

[4] Danh V. Nguyen et al, "Tumor classification by partial least squares using microarray gene expression data," Bioinformatics, Vol.18, No.1, p. 39-50, Jun 2002.

[5] Sugato Basu, "Semi-supervised Clustering by Seeding," Proceedings of the 19th International Conference on Machine Learning, (ICML-2002), pp. 19-26, Sydney, Australia, July 2002.

[6] Akinori Fujino et al, "Semisupervised Learning for a Hybrid Generative/Discriminative Classifier Based on the Maximum Entropy Principle," IEEE Trans, Pattern Analysis and machine intelligence, Vol.30, No.3, MARCH 2008.

[7] Dan Klein, Sepandar D. Kamvar, Christopher D. Manning, "From Instance-level Constraints to Space-level Constraints : Making the Most of Prior Knowledge in Data Clustering".

[8] Kiri Wagsta, "Constrained K-means Clustering with Background Knowledge," Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577-584, 2001.

[9] M.A.T. Figueiredo et al, "Unsupervised Learning of Finite Mixture Models," IEEE Trans, Pattern Analysis and machine intelligence, March Vol.24, No.3, pp. 381-396, 2002.

[10] Ana L.N. Fred, Anil K. Jain, "Combining Multiple Clusterings Using Evidence Accumulation," IEEE Trans, Pattern Analysis and machine intelligence, Vol.27, No.6, JUNE 2005.

[11] Yi Hong, "Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm," Pattern Recognition, Vol.41, Issue. 9, SEPTEMBER 2008.

[12] Lawrence Hubert, "Comparing Partitions," journal of Classification, 2:193-218, 1985.

[13] David Hand et al, "principal of Data mining," A Bradford Book The MIT Press Cambridge, Massachusetts London, England, 2001.

[14] http://www.geneontology.org

고 송

2006년 전북대학교 컴퓨터공학과 학사 2006년~현재 중앙대학교 컴퓨터공학과 석사과정. 관심분야는 데이타마이닝, 바이오정보학, 유전자 온톨로지

김 대 원

1997년 경북대학교 컴퓨터공학과 학사 1999년 KAIST 전산학과 석사. 2004년 KAIST 전자전산학과 박사. 2005년~현재 중앙대학교 컴퓨터공학부 조교수. 관심분야는 바이오정보학, 의료정보학, 데이타마이닝