Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

사전정보 활용을 위한 관련 규칙 기반의 Ensemble 클러스터링

Association-rule based ensemble clustering for adopting a prior knowledge

고송, 김대원

서울시 동작구 중앙대학교 컴퓨터공학과
E-mail: ssyong20@wm.cau.ac.kr

요 약

본 논문은 클러스터링 문제에서 사전 정보에 대한 활용의 효율을 개선시킬 수 있는 방법을 제안한다. 클러스터링에서 사전 정보의 존재 시 이의 활용은 성능을 개선시킬 수 있는 계기가 될 수 있으므로 그의 활용 폭을 늘리기 위한 방법으로 다양한 사용 방법의 적용인 semi-supervised 클러스터링 앙상블을 제안한다. 사전 정보의 활용 방법의 방안으로써 association-rule의 개념을 접목하였다. 클러스터 수를 다르게 적용하더라도 패턴간의 유사도가 높으면 같은 그룹에 속할 확률은 높아진다. 다양한 초기화에 따른 클러스터의 동작은 사전 정보의 활용을 다양화 시키게 되며, 사전 정보에 충족하는 각각의 클러스터 결과를 제시한다. 결과를 총 취합하여 association-matrix를 형성하면 패턴간의 유사도를 얻을 수 있으며 결국 association-matrix를 통해 클러스터링 할 수 있는 방법을 제시한다.

Key Words : 클러스터링, 클러스터링 앙상블, semi-supervised, association-rule

1. 서 론

데이터를 카테고리(category)로 구분하는 것은 오래 전부터 다루어져 오던 문제이며 여러 종류의 분류 문제가 있다[5]. 본 논문에서는 비지도 학습(unsupervised) 문제인 클러스터링에 대해서 다룬다. 클러스터 분석은 사전 정보 없이 데이터의 분석을 통한 정보를 취득해야 하는 문제이므로 탐험적인 데이터 분석(exploratory data analysis)으로 불리기도 한다. 클러스터링 문제는 데이터의 패턴을 분석하여 패턴간의 유사도를 측정하고 그룹을 형성하는 것이다. 실세계의 문제를 다루는 데이터에서 가지는 고차원의 특징은 동일한 그룹의 특징을 가지지 않는다. 유사도 측정과 클러스터 방법의 다양한 적용이 같은 결과를 제시하지 못하는 이유다.

클러스터링의 결과는 다음 단계의 연구 진행을 위한 기본 자료로서 이용이 되기 때문에 보다 의미 있고 안정적인 정보를 얻을 수 있는 방법에 대한 연구가 진행되고 있다. 이에 본 논문에서는 다양한 방법의 결과를 조합할 수 있는 앙상블 방법을 적용하였으며, 사전 정보의 활용 폭을 넓힐 수 있는 방법을 제안한다.

2. 관련 연구

2.1 클러스터링 앙상블

클러스터링의 앙상블은 다양한 방법론이 가지는 장·단점에 대한 특징을 극복할 수 있는 방법을 제시한다. 앙상블은 다양한 방법의 적용과 그 결과의 조합을 통한 상승효과(boosting)를 기대할 수 있다.

클러스터링 앙상블에는 2가지 방법이 제시되어졌고, EM(Expectation-Maximization)알고리즘[2]과 association-rule[1] 방법이다.

· EM의 동작은 아래 식과 같다.

$$E[z_{im}] = \frac{a'_m \prod_{j=1}^{H} \prod_{k=1}^{K(j)} (v'_{jm}(k))^{\delta(y_{ij},k)}}{\sum_{n=1}^{M} a'_n \prod_{j=1}^{H} \prod_{k=1}^{K(j)} (v'_{jm}(k))^{\delta(y_{ij},k)}}$$

$$\alpha_m = \frac{\sum_{i=1}^{N} E[z_{im}]}{\sum_{i=1}^{N} \sum_{m=1}^{M} E[z_{im}]}$$

$$v_{jm}(k) = \frac{\sum_{i=1}^{N} \delta(y_{ij},k) E[z_{im}]}{\sum_{i=1}^{N} \sum_{k=1}^{K(j)} \delta(y_{ij},k) E[z_{im}]}$$

클러스터 방법을 H개 적용하였을 때의 EM 알고리즘의 동작은 E의 기대치와 기대치(E)를 최대화 해주는 M의 단계의 연속이다.

· association-rule 방법은 다양한 방법으로 클러스터링을 적용하더라도 유사도가 높은 패턴은 같은 그룹에 속할 확률이 높다는 것을 이용하였다. 클러스터링의 적용을 여러 차례 반복하고 매 회마다 패턴간의 같은 그룹에 형성된 횟수를 누적시킨 것이 association matrix이며 이것은 similarity-matrix와도 동일하다.

2.2 semi-supervised 클러스터링

클러스터링은 본질적으로 비지도 학습 문제를 다루

게 되지만, 사전 정보를 가지고 있을 때 이용할 수 있는 방법에 대해서 다루는 분야가 semi-supervised 클러스터링이다. 이 문제는 비지도 학습과 지도 학습 문제와 구별되는 semi-supervised 문제로 분류가 되며, 사전 정보의 활용률을 높이기 위한 연구가 진행된다. semi-supervised 문제에서는 사전정보의 활용 방법에 대한 내용을 다룬다. 사전 정보라는 것은 일부의 패턴에 대한 라벨을 알고 있는 것으로 가정한다. 사전 정보를 통해 파악된 패턴 라벨의 이용은 pair-wise 제약조건(constraint)으로 정의한다[4].

- must_link_set : M
- cannot_link_set : C

M과 C의 활용에 대한 방법은 클러스터링 방법론마다 고유하며, pckmeans와 spatial-level 클러스터링이 소개되어졌다[3, 4].

pckmeans의 동작은 기본적으로 kmeans와 동일하다. 사전 정보의 처리를 다루는 식의 유무에 따라 구분이 된다.

· pckmeans의 목적함수는 (1)식과 같다.

$$J_{pckm} = \frac{1}{2}\sum_{x_i \in X} \|x_i - \mu_i\|^2 + \sum_{(x_i,x_j)\in M} w_{ij}[l_i \neq l_j] + \sum_{(x_i,x_j)\in C} w_{ij}[l_i = l_j] \quad (1)$$

식 (1)에서 앞의 식은 kmeans와 동일하며 뒤의 식은 사전 정보를 위배할 수 있는 확률을 낮출 수 있도록 페널티(penalty)를 적용하게 된다.

· spatial-level 클러스터링의 기본적인 동작원리는 아래 그림과 같다.

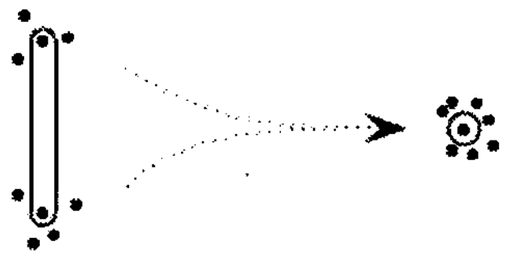

그림 1. 긴 타원 끝의 두 점을 동일한 점으로 처리함

[그림 1]의 왼쪽 그림을 통해 긴 타원 끝의 두 패턴이 같은 라벨임을 사전 정보로 받았을 때, 그 두 패턴의 거리를 0으로 하며, 각 패턴이 가지던 그 주변과의 패턴의 거리도 오른쪽 그림과 같이 짧은 패턴의 거리로 대체가 된다. 이러한 식으로 패턴간의 거리 matrix가 형성이 되면 HCA(Hierarchical Clustering Algorithm)으로 클러스터링 한다.

2.3 association-rule

semi-supervised 클러스터링 앙상블에서 패턴간의 유사도를 측정함으로써 클러스터링 앙상블의 진행을 위해서 association-rule을 접목하였다. association-rule은 패턴이 독립적이지 않은 관계(relationship)를 가지며 그 정도의 크기가 다양함을 가정한다. 패턴간의 관계의 정도가 크면 클러스터링의 방법에 따른 클러스터링의 결에서도 같은 그룹에 속할 확률이 높다.

association-rule을 최대한으로 이용하기 위한 비교적 큰 그룹 수를 가지는 클러스터링을 적용한다. 즉, 초기화시 중심의 개수(K)를 일반적으로 많이 설정한 후 그룹을 나누더라도 각 그룹의 패턴의 관계가 깊을수록 같은 그룹에 속할 확률이 높게 된다. 패턴간의 같은 그룹 유무에 따른 association-matrix를 누적시키게 된다.

$$C(i,j) = \frac{n_{ij}}{N} \quad (2)$$

N : 적용한 클러스터링 횟수
n_{ij} : i, j번째 패턴 라벨 비교
C(i, j) : association-matrix

[그림 2]을 통해 가까운 지점에 있는 패턴들은 서로 같은 그룹에 있을 확률이 높다는 것을 알 수 있다. 이 점을 이용하여 다양한 초기화를 통해 클러스터링을 적용하여 association-matrix를 갱신하는 것이 식 (2)와 같다

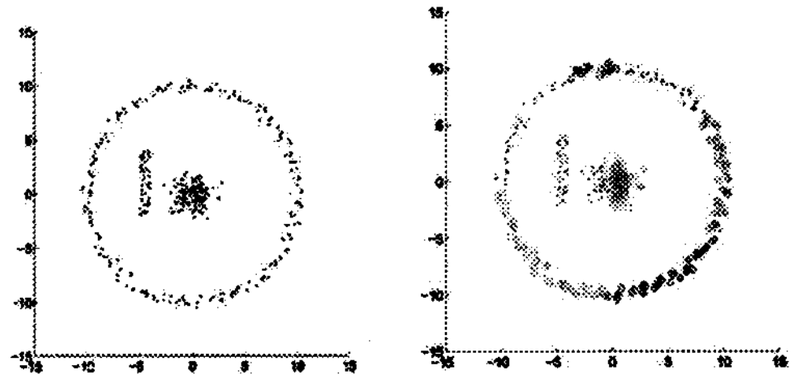

그림 2. 오른쪽 그림은 클러스터 개수를 11개로 설정한 후 kmeans 적용한 결과

[그림 2]에서 클러스터 수를 11개를 적용한 결과를 확인 할 수 있다. 거리가 가까운 패턴이 한 그룹에 속함을 확인할 수 있다. 여러 클러스터 수로 적용한 후 association-matrix에 누적시킴으로써 전체적인 모델을 형성할 수 있다.

3 관련 규칙을 활용한 앙상블 알고리즘

클러스터링에서 해결해야 할 문제는 다양하다. 실세계의 데이터가 보일 수 있는 중첩, 불균형(imbalanced data)의 문제는 아직도 연구가 진행 중이다. 데이터의 중첩의 발생은 그룹의 구분을 어렵게 만들며, 불균형의 발생은 거대 그룹에 과도한 영향을 받는 경우가 발생한다.

본 논문에서는 불균형의 발생에 대해 사전 정보에 대한 활용에 의해 개선됨을 보인다.

위에서 언급했던 semi-supervised 클러스터링은 단일 방법론만을 적용함으로 인한 방법론에 대한 치우침(bias)이 발생하게 된다. 단일 방법론은 전체적인 데이터 형태의 모습에 대해 강건(robust)하게 성능을 발휘하지 못한다. 데이터의 형태에 따라 성능의 차이가 심한 방법론은 실세계의 데이터에 적용하는 것에 대한 신뢰를 얻기 힘들다.

이에 대한 극복 방법으로 사전 정보를 묶어서 관련 규칙으로 설정한 후, 이 규칙을 준수하는 한도 내에서 다양한 초기화 및 방법론을 적용한다. 각 결과를 총 취합한 하나의 결과는 보다 강건한 성능을 보일 수 있음을 실험 결과를 통해 확인할 수 있었다.

앙상블 클러스터링의 적용은 크게 2가지 방법으로 적용하였다.

· homogeneous : pckmeans만을 통한 앙상블
· heterogeneous : pckmeans + spatial-level clustering

같은 방법론의 결과를 앙상블 하는 homogeneous과 다양한 방법론을 통한 앙상블인 heterogeneous로 분류되며, 본 논문에서는 homogeneous에 주안점을 두고 진행하였다.

클러스터링 적용시의 클러스터 수는 association-rule을 이용하기 때문에 비교적 크게 잡는 것이 적합하다.

$$\text{클러스터수}(K) = \sqrt{\text{데이터수}} + \sqrt{\text{데이터수}} * rand() \quad (3)$$

식 (3)의 앞의 식과 같이 $\sqrt{\text{데이터수}}$는 최소한의 클러스터 수를 보장한다. 이것은 유사성이 높을수록 같은 그룹에 속할 수 있는 확률이 높다는 점을 최대한 이용하기 위함이다. association-rule은 패턴과 패턴의 관계를 따지는 개념이므로, 전체적인 데이터 모델을 형성하기 위해서는 패턴 관계의 적당한 누적이 필요하다. 누적을 위한 횟수는 식 (2)와 같은 N회를 시행하는 데 본 논문에서는 50회를 시행하였다.

association-rule의 개념은 작은 부분의 패턴을 봄으로써 모델을 형성하는 것이다. 이 rule의 개념을 이용한 관련 규칙의 활용의 개선을 통해 [그림 3]의 (b)와 같은 불균형의 데이터에서도 큰 그룹에 쏠리는 현상이 개선됨을 실험 결과로 확인할 수 있었다.

표 1. 제안하는 방법의 알고리즘

알고리즘. 1

1.초기화
클러스터수$(K) = \sqrt{\text{데이터수}} + \sqrt{\text{데이터수}} * rand()$
2. pckmeans
2.1 초기화 : 중심(K개)
- must-link set **M**
- cannot link set **C**
- 사전 정보의 그룹 수 : temp_k
- ***1.초기화***에 의한 클러스터 K개로 조정
  · rest_K=K-temp
  · rest_K개의 초기 랜덤 중심 값

2.2 distance measure

$$J_{pckm} = \frac{1}{2}\sum_{x_i \in X} \|x_i - \mu_i\|^2 + \sum_{(x_i,x_j)\in M} w_{ij}[l_i \neq l_j] + \sum_{(x_i,x_j)\in C} w_{ij}[l_i = l_j]$$

2.3 assign_cluster
- data(i)=arg min(J_{pckm})

3. spatial_level clustering
3.1 distance matrix
- 데이터의 모든 패턴간의 거리 측정
- [그림 1]과 같이 사전 정보에 따라 거리 갱신

3.2 HCA
4. 1-3까지 N회 반복
- $C(i,j) = \frac{n_{ij}}{N}$
- **매 회마다 C(i,j) matrix 갱신**

5. association_matrix : C(i,j)
5.1 HCA

4. 실험 방법 및 결과

실험은 인공적인 데이터 셋과 실세계의 데이터 셋을 구분하였다. 인공적인 데이터 셋은 특정한 형태의 문제 해결을 위해 만든 것이다[1]. 다양한 데이터 셋의 적용은 우리가 궁극적으로 풀어야 할 실세계의 데이터 형태는 알 수 없지만 복잡할 것이라는 것을 알고 있기 때문이다. 다양한 인공적인 데이터 셋에 대한 적용에 대한 결과가 전반적으로 좋게 나온다면 실세계의 데이터의 분석 결과에 대해서 기대할 수 있게 된다.

본 논문에서의 인공적인 데이터 셋은 4개를 사용했다. 데이터의 형태는 [그림 3]과 같다.

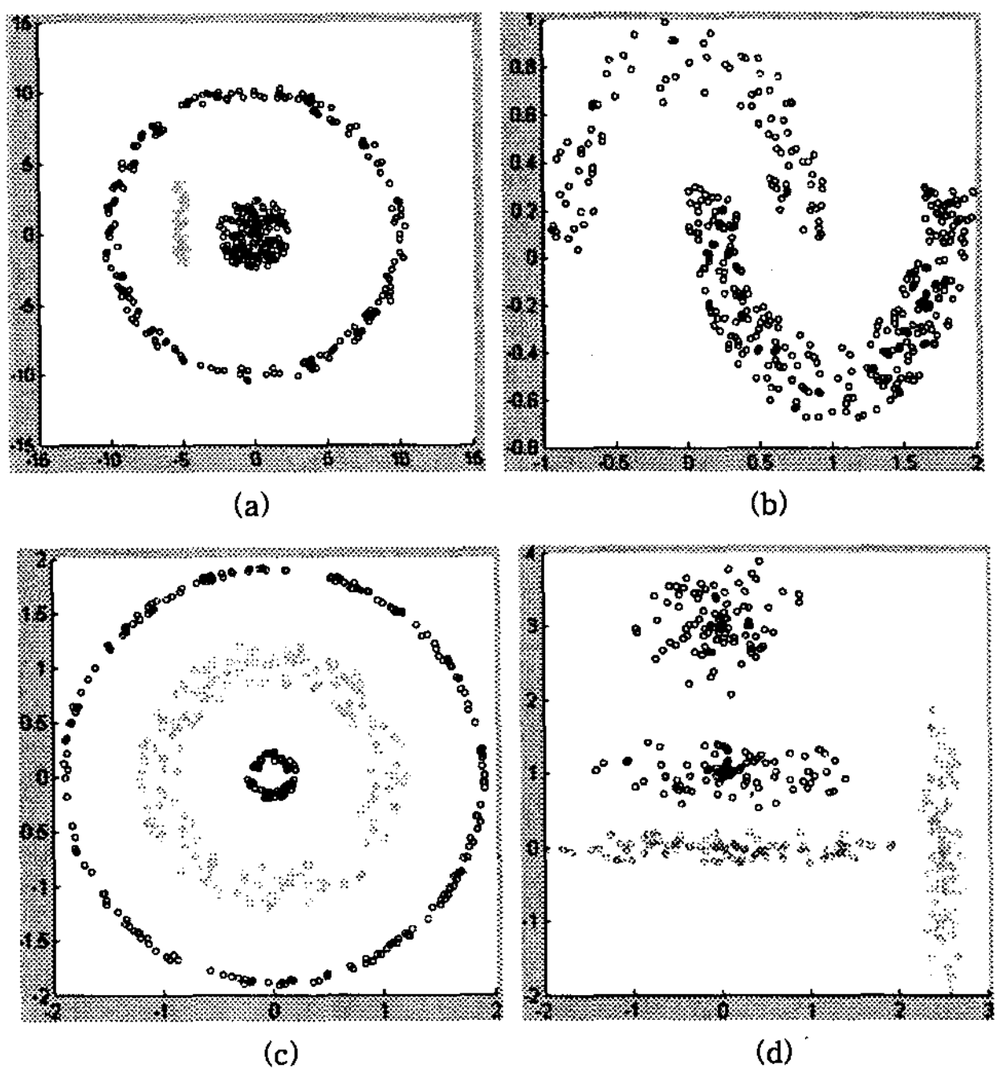

그림 3. 실험 데이터

[그림 3]의 각 데이터에 대한 세부 설명 :
(a) 3 클러스터
  바깥 원 : 200개, 사각 모양 : 50개, 가우시안 분포 : 150개
(b) half rings : 2차원의 2클래스 문제
  · 위의 반원 : 100개, 아래 반원 : 300개 패턴
(c) Three rings : 3클래스
  · 바깥 원부터 200, 200, 50개 패턴
(d) e5 : 4클래스 문제
  · 각 클래스 100개 패턴

실세계의 데이터 셋은 바이오 정보학에서 많이 다루고 있는 leukemia와 colon을 사용한다. 이 데이터는 인체의 질병 유무를 담고 있다.

비교 알고리즘은 pckmeans, spatial-level 클러스터링, 비지도 학습 앙상블 클러스터링을 사용하며, 비지도 앙상블 클러스터링은 association-rule을 적용하여 실험하였다.

실험 방법은 사전 정보의 양의 변화에 따른 정확도의 변화를 체크하며, 실험은 30회 시행하여 평균을 기입하고 결과는 [그림 4]를 통해 제시한다.

3_cluster를 제외한 나머지 3개의 인공의 데이터와 2개의 실세계 데이터를 적용한 결과에서 다른 방법론보다 좋은 성능을 보이고 있음을 볼 수 있었다.

일부 데이터 셋에서 사전정보의 양이 늘어남에 따라 정확도가 격감하는 상황이 발생하게 되는데, 이는 사전정보 활용이 효율적이지 못함을 뜻한다.

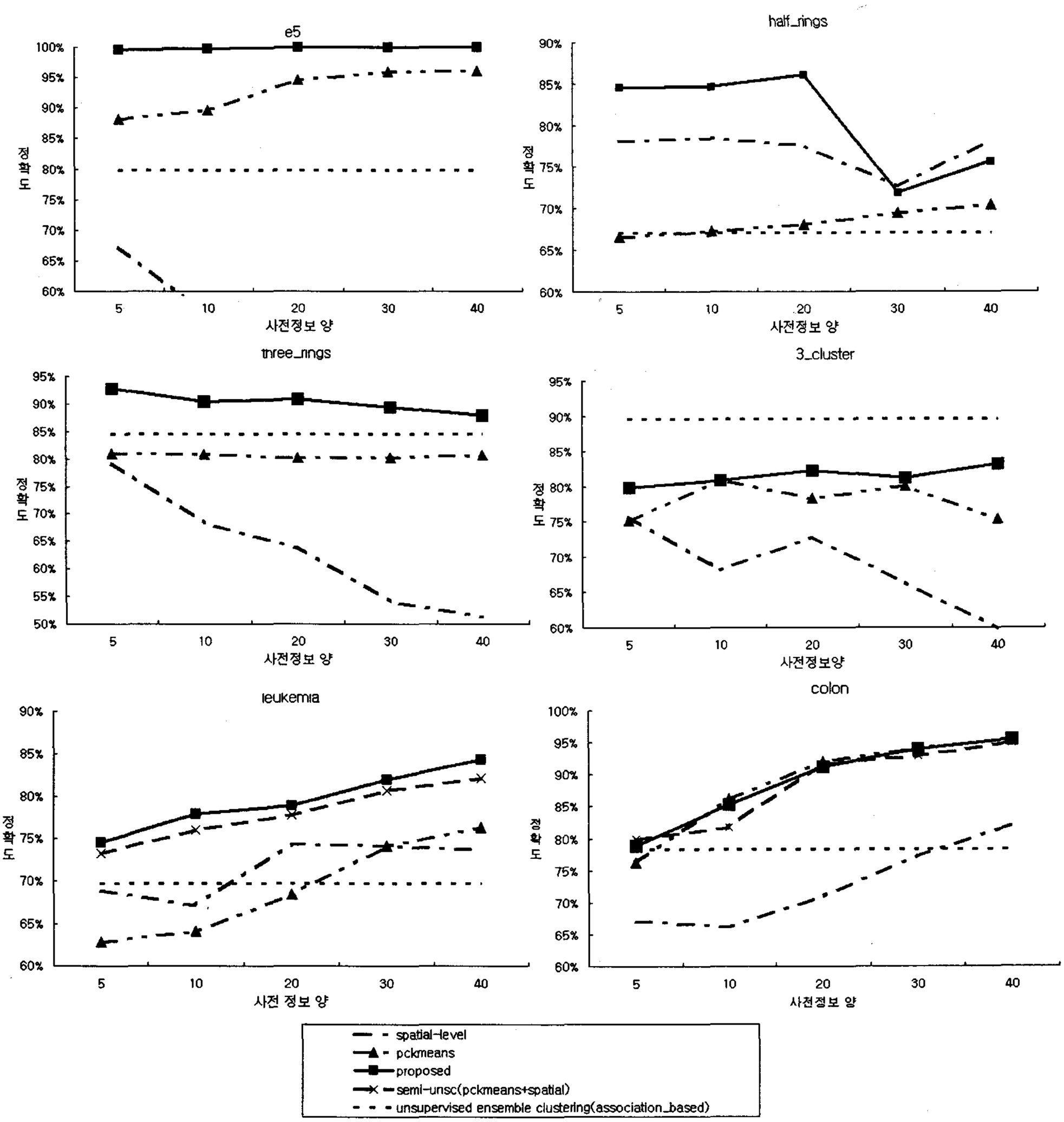

그림 4. 각 데이터 실험 결과

5. 결론 및 향후연구

[그림 4] 3_cluster의 결과에서 보이듯이 제안한 알고리즘이 항상 우월한 성능을 보이고 있지만은 않다는 것을 알 수 있었다. 사전 정보의 활용에 있어 효율적인 사용을 위한 개선책이 필요하게 됨으로써, 이에 대한 연구를 진행할 계획이다.

참 고 문 헌

[1] 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

[2] Alexander Topchy, Anil K.Jain Ailliam Punch, "Clustering Ensembles Models of Consensus and Weak Partitions" IEEE Trans, PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 27, NO. 12, DECEMBER2005

[3] 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"

[4] Sugato Basu, Arindam Banerjee abanerje, Raymond Mooney, "Active Semi-Supervision for Pairwise Constrained Clustering"

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