한국지능시스템학회 논문지 2009, Vol. 19, No. 6, pp. 802-808

유전자 온톨로지를 활용한 클러스터링 성능 향상 기법

Improving Clustering Performance Using Gene Ontology

고송* · 강보영** · 김대원*

Song Ko*, Bo-Yeong Kang** and Dae-Won Kim*

* 중앙대학교 컴퓨터공학부

** 경북대학교 기계공학부

요 약

마이크로어레이 데이터의 클러스터링 성능을 향상시키기 위하여 유전자 온톨로지(GO)를 활용하는 연구가 최근 진행 중에 있다. 그 중 Biological Process(BP) GO를 활용한 Kustra et al.의 연구가 2006년에 소개된 바 있다. 본 연구는 Kustra et al.의 연구를 확장하여 일반적이고 실질적인 GO의 활용 방안을 위한 분석 결과를 제시하기 위하여 다양한 활용 방법을 적용한다. (1) GO의 거리를 측정하기 위하여 Lin et al, Resnik et al과 Jiang et al의 방법을 적용하였으며, (2) BP를 포함한 세 가지 GO 유형의 구조에 대해 적용하여 각 방법에 따른 성능 향상 정도를 분석한다. 각 방법에 대한 성능 분석 비교를 위하여 효모 유전자를 관측하여 형성한 데이터를 활용한다. 실험 결과를 통하여 GO 정보를 클러스터링에 적용하면 전반적으로 성능 향상을 유도하지만, 활용 방법에 따라서 성능 개선 정도의 차이가 발생한다. 그 중 Resnik의 거리 측정 척도와 BP GO를 활용하였을 때, 가장 개선된 성능을 유도함을 볼 수 있다.

키워드 : 반지도 클러스터링, 유전자 온톨로지, 마이크로어레이 데이터, 유전자 기능 예측, 의미론적 거리

Abstract

Recently many researches have been presented to improve the clustering performance of gene expression data by incorporating Gene Ontology into the process of clustering. In particular, Kustra et al. showed higher performance improvement by exploiting Biological Process Ontology compared to the typical expression-based clustering. This paper extends the work of Kustra et al. by performing extensive experiments on the way of incorporating GO structures. To this end, we used three ontological distance measures (Lin's, Resnik's, Jiang's) and three GO structures (BP, CC, MF) for the yeast expression data. From all test cases, We found that clustering performances were remarkably improved by incorporating GO; especially, Resnik's distance measure based on Biological Process Ontology was the best.

Key Words : Semi-supervised Clustering, GO, Microarray, Gene Function Prediction, Semantic Distance

1. 서 론

마이크로어레이 데이터를 활용한 유전자 기능 분석을 위한 연구에서 유사한 기능의 유전자는 유사한 발현 패턴을 갖는다는 특성에 기반하여 클러스터링 분석을 적용하는 다양한 방법이 소개되었다. 그 중 K-means 클러스터링, 계층적(Hierarchical) 클러스터링, SOM(Self Organizing Map), 그래프 이론과 통계학을 이용한 다양한 비지도 클러스터링 방법들이 소개되었다[1-7]. 그러나 수천 개 이상의 유전자를 동시에 관측하여 형성한 마이크로어레이 데이터는 다음과 같은 문제점을 갖기 때문에 분석 성능 향상의 한계점에 직면한다. 첫째, 실험 도구의 흠집 및 먼지 등에 의한 노이즈가 포함될 수 있으며, 둘째, 일부 유전자는 유사한 기능이지만 유사하지 않은 발현 패턴을 보일 수 있다. 이와 같은 문제로 인하여 클러스터링 방법론의 성능 향상에 한계점으로 작용하게 된다.

이러한 문제를 해결하기 위한 방법으로 유전자 온톨로지(Gene Ontology; GO)를 활용하려는 연구가 소개되고 있다. Dikla et al.이 제시한 방법은 계층적 클러스터링 방법에 기반하여, 순차적으로 각 그룹을 병합하여 최종 목적 클러스터 그룹 수로 형성될때까지 각 그룹을 병합할 때 기준으로 GO를 사용하는 방법을 제시하였다[12]. 병합 시 같은 레이블을 갖는 유전자가 최대가 되도록 병합 기준으로 활용하였다. 그러나 비정상적인 발현 패턴을 가지는 일부 유전자는 유사한 기능의 다른 유전자와 발현 패턴에 의한 거리가 멀기 때문에 계층적 구조에서 유사한 기능의 유전자 그룹으로 형성하기 어렵다. Fang et al.은 GO에 유전자가 포함된 모든 노드를 초기 그룹으로 형성하여, 각 그룹 간 거리를 측정하여 미리 선정한 임계치 값보다 작으면 해당하는 두 노드를 병합하는 방법을 제시하였다[13]. 그러나 임계치 값의 선택에 대한 문제를 가지고 있으며, 비정상적인 관측 값을 갖는 유전자에 의해 대표 값 계산에 영향을 받아 해당 그룹에 속하는 유전자의 발현 패턴에 대한 대표 값의 의미를 상

접수일자 : 2009년 7월 16일
완료일자 : 2009년 12월 2일
이 논문은 2009년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 20090259).

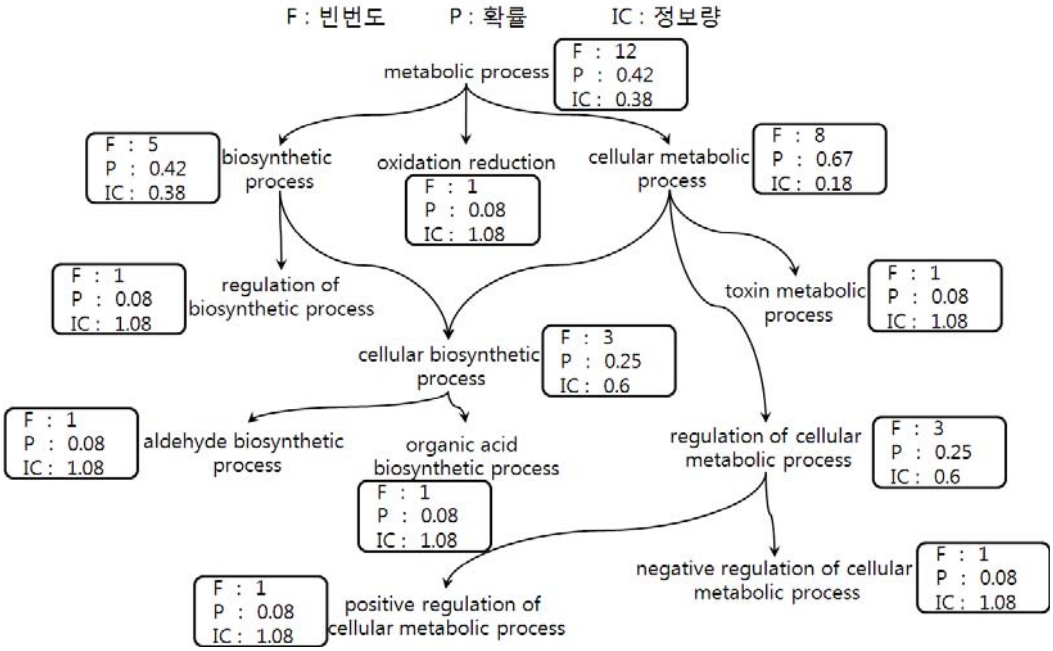


그림 1. IC를 반영한 유전자 온톨로지 구조
Fig. 1. Gene Ontology with IC

실할 수 있다.

Huang et al.의 연구에서는 GO를 바탕으로 유전자의 유사도 값을 추출하여 클러스터링에 적용하는 방법을 제시하였다[14]. 그러나 유전자의 기능 관계는 다양하지만 이 연구에서 사용한 방법은 유전자 간 관계를 유사한 기능과 유사하지 않은 기능 관계인 두 개의 관계로 한정시켰다는데 있다. Cheng et al.은 GO에서 유전자 간 거리를 에지 기반으로 측정하여 계층적 클러스터링에 적용하는 방법을 제시하였다[15]. 그러나 에지 기반 거리 측정 척도는 GO의 특정 구역에 많은 노드가 밀집하여 균형이 잡혀있지 않은 경우 그 구역에서의 노드 간 거리는 다른 지역에 비해 상대적으로 멀게 측정된다. 또한 생물학적인 현상에 대한 기능을 정의하는 노드는 기능의 상세함의 정도가 각기 다르므로 정의하는 상세함에 대한 관계를 해당하는 가중치로 거리를 계산할 수 있어야 한다. Kustra et al.의 연구에서는 정보량(Information Content; IC)에 기반한 Lin의 방법을 수정·적용하여 GO에서 노드 간 거리를 측정한 후 클러스터링에 조합하는 기법을 제안하였고 성능이 향상됨을 제시하였다[16]. 그러나 제안된 기법은 GO에 등록된 유전자만을 대상으로 실험함으로써 알려지지 않은 유전자 기능 예측을 위한 클러스터링 성능 향상에 대한 실질적인 결과를 제시하지는 못한다. 또한 다양한 정보량 기반의 유전자간 거리 측정 척도 중 Lin의 방법만을 제한적으로 적용하여 어떤 기법이 보다 효과적인 거리 측정 방법인지에 대한 비교 분석 결과를 제시하지 않았다.

따라서 본 논문에서는 Kustra et al.의 연구를 확장하여, 정보량 기반의 유전자간 거리 측정 기법을 활용한 반지도 클러스터링 기법을 제안함으로써 GO에서 유전자간 거리 정보 활용 방법을 위한 일반적이고 실질적인 결과를 제시하고자 한다. 제안된 기법은 (1) 정보량에 기반한 유전자간 거리 측정 기법에 따른 클러스터링 성능 결과를 비교 분석한다. (2) GO가 제공하는 세 가지 유형의 GO를 대상으로 클러스터링을 수행함으로써, GO에 따라 발생하는 차이가 클러스터링 분석에 어떠한 영향을 미치는지에 대해 분석한다. 다양한 적용 방법에 따른 성능 변화 정도를 세 종류의 효모 유전자 데이터를 활용하여 실험함으로써, GO의 적합한 활용 방법 연구를 위하여 다양한 실험 결과의 분석을 하였다.

2. GO에서 유전자 간 거리

GO는 유전자의 기능 관계에 따라 계층적으로 구성된 의미론적 구조이므로 유전자 기능의 유사도 정도를 거리로 계산하기 위해서 의미론적 거리를 수치적으로 계산할 수 있어야 한다. 이를 위해 정보 이론의 개념인 정보량(Information Contents; IC)을 활용할 수 있다.

2.1 GO 특성 및 노드의 정보량(IC)

GO는 생명 현상을 보는 관점에 따라 세 가지의 독립적 구조인 Biological Process(BP), Cellular Compon- ent (CC)와 Molecular Function(MF)으로 구분할 수 있다[8]. 각 구조는 가장 일반적인 기능 정보를 최상위 노드에 배치하고 보다 상세한 기능 정보는 상대적으로 하위 노드로 배치하는 계층형 비방향성 그래프 구조를 갖는다. 기능에 의한 노드의 상대적 위치에 따라 기능 정보를 정량화 할 수 있는 정보량 기법이 정보이론에서 제시되고 있다[9]. 즉, 노드에서 정의하는 기능의 상세한 정도를 정보량의 의미를 갖는 수치를 통해 부여하는 방법이다. 식(1)～식(3)은 정보량을 구하기 위한 과정을 표현하고 있다. 보다 일반적인 기능을 정의하는 노드는 루트 노드에 가깝게 위치하여 상대적으로 자손 노드가 많을 것이고, 해당하는 자손 노드의 전체 개수를 식 1과 같이 세는 것이 빈도수(Frequency; F)이다. 따라서 그림 1의 최상위 노드인 “Metabolic Process” 노드의 빈도수는 12가 된다. 각 노드는 비슷한 방법으로 빈도수

를 계산할 수 있으며, 최상위 노드의 빈도수로 각 노드의 빈도수에 나눠준 것이 식(2)와 같이 확률(Probability; P)를 계산할 수 있다. 마지막으로 식(3)과 같은 과정을 통해 최종적으로 각 노드의 정보량을 수치형 값으로 표현할 수 있다.

$$F = \sum count(n) \tag{1}$$

$$P = \frac{F}{Root\_Node(F)} \tag{2}$$

$$IC = -\log_{10} P \tag{3}$$

2.2 정보량 기반 거리 측정 척도

본 논문은 정보량을 기반으로 하여 GO에서 노드 간 거리를 측정하는 방법으로 3 가지를 적용하였다.

Resnik이 제시하였던 유사도 측정 방법은 식 (4)와 같이 유사도를 측정하려는 두 노드로부터 가장 가까운 조상 노드의 정보량 값으로 계산하였다[9]. 예를 들어, 그림 1의 “aldehyde biosynthetic process” 노드와 ”organic acid biosynthetic process” 노드의 유사도 측정 시 각 노드는 c_1과 c_2에 해당한다. 위 두 노드의 가장 가까운 조상 노드는 (Parent, P) ”cellular biosynthetic process” 노드이므로 해당 노드의 정보량 값인 0.6이 두 노드의 유사도가 된다.

$$sim(c_1, c_2) = P(c_1, c_2)$$

Jiang et al.은 두 노드의 정보량 값 차이로 거리를 계산하는 방법을 식 (5)와 같이 제시하였다[11]. 예를 들어, 위에서 언급한 두 노드의 거리는 ”aldehyde...”와 “organic...”의 최단 경로인 ”aldehyde...”-”cellular...”- ”organic acid...”의 거리가 되며 식 (6)과 같은 과정으로 거리는 0.96이 된다.

$$dist(c_1, c_2) = [IC(c_1) - P(c_1, c_2)] + [IC(c_2) - P(c_1, c_2)] \tag{5}$$

$$dist(cellular..., organic...) = (1.08 - 0.6) + (1.08 - 0.6) = 0.96 \tag{6}$$

Lin이 제시하였던 노드 간 유사도는 식 (7)과 같이 두 노드의 정보량 값의 합을 분모로 하며, 공통 조상 노드의 정보량 값을 분자로 하여 계산하는 방법이다[10]. 예를 들어, 위에서 언급한 두 노드의 유사도는 식 (8)과 같은 과정을 통해, 두 노드의 정보량 값을 합한 2.16이 분모가 되고, 공통 조상 노드인 “cellular...”의 정보량 값인 0.6을 분자로 하면, 유사도는 0.28로 계산된다.

$$sim(c_1, c_2) = \frac{2 \times P(c_1, c_2)}{IC(c_1) + IC(c_2)}$$

$$sim(aaldehyde..., organic...) = \frac{2 \times 0.6}{1.08 + 1.08} = 0.28$$

위에서 제시한 방법은 클러스터링에 적용 시 거리의 개념을 가져야 하므로 유사도를 의미하는 결과 값은 (dist = 1 - sim)의 식을 통하여 거리를 의미하도록 처리한다.

2.3 마이크로어레이 기반 클러스터링에 적용

발현 패턴에 의한 거리($dist_{exp}$)와 GO에서 유전자 간 거리($dist_{GO}$)를 통해 전체 거리($dist_{total}$)를 계산하는 경우 서로 다른 범위의 거리 값을 갖으므로, 각 방법에 의한 거리는 0 ~1이 되도록 정규화한다. 정규화 된 $dist_{exp}$와 $dist_{GO}$를 통하여 $dist_{total}$를 식 (9)와 같이 그룹 X와 유전자 x의 거리를 계산할 수 있다. 또한 두 거리의 가중치 α를 통하여 성능 결과를 분석할 수 있다.

$$dist_{total}(X,x) = \alpha \times dist_{\exp}(X,x) + (1-\alpha) \times dist_{GO}(X,x) \tag{9}$$

3. GO 활용 방법에 따른 비교 분석

3.1 실험 디자인

실험 데이터는 유전자 기능 분석을 위하여 활발히 사용되고 있는 유전자 데이터를 활용하였으며 세 데이터 모두 효모 유전자이다[3,4,5]. Cho et al.의 데이터는 세포 주기에 대한 과정을 시간의 흐름에 따라 17번 관측한 6,456개의 유전자 데이터이다. Eisen et al.의 데이터는 포자 형성, 세포 주기와 diauxic shift등의 환경과 alpha-factor 혼합 및 정화 등의 상태에서 시간의 흐름에 따른 유전자의 발현 패턴 변화를 80번 관측한 6,220개의 유전자 데이터이다. Spellman et al.의 데이터는 세포 주기의 환경에서 alpha-factor나 cdc15등을 혼합한 상태에서 시간의 흐름에 따른 유전자의 발현을 77번 관측한 6,178개의 유전자 데이터이다.

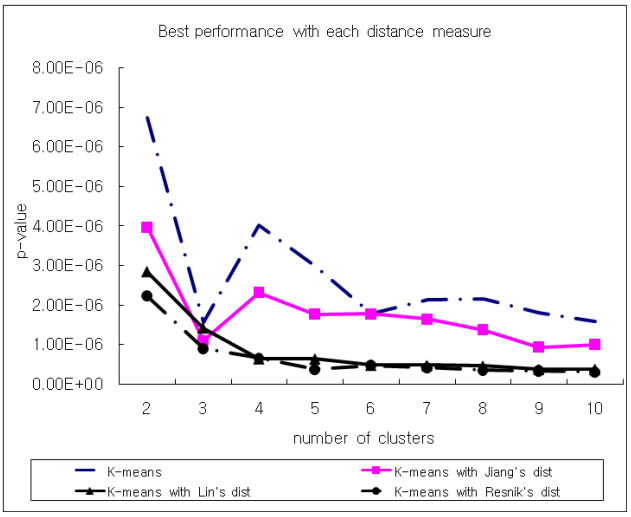
실험 방법은 기본 방법론으로 K-Means를 활용하며 그룹 수는 2개에서 10개 사이로 한 후 10회 반복하여 평균으로 성능을 평가한다. 클러스터링 성능 평가는 통계학적인 방법인 p-value를 이용하였으며, 유사한 기능의 유전자가 같은 그룹에 속한 것이 많아질수록 좋은 클러스터링 결과를 의미하고, 이 때 p-value는 낮은 값을 갖는다[2].

3.2 세 가지 효모 데이터에 대한 클러스터링 결과

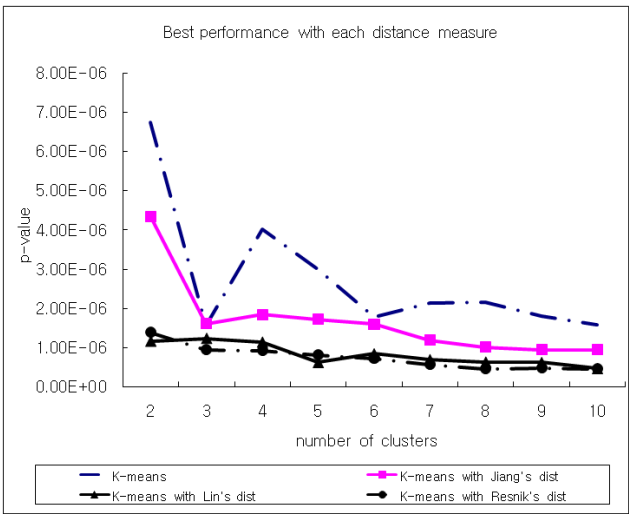
Cho et al.의 유전자 데이터에 대한 실험 결과를 그림 2에서 보이고 있으며 범례와 같이 각 실험 결과는 실선 및 점선 그리고 각각의 표식으로 구분하였다. 실험 결과를 통해 GO를 활용한 경우 클러스터링 성능의 개선이 유도됨을 볼 수 있다. 성능 개선의 정도는 거리 측정 척도에 따라 변화가 있으며, Resnik과 Lin의 거리 측정 척도를 적용하였을 때 가장 좋은 성능을 유도하며, 나타내는 성능의 정도가 대체적으로 비슷한 수준임을 보인다. 다만 Jiang et al.의 방법론을 적용하였을 때, 성능 개선을 유도하나 다른 두 거리 측정 척도에 비해 저조하다.

Eisen et al.의 유전자 데이터에 대한 실험 결과에서 Jiang et al.의 방법이 BP 활용 시 그룹 수 2~5개에서 성능 개선이 미미하거나 오히려 K-means 보다 저조한 성능을 보이지만, 나머지 그룹에서는 좋은 성능을 보인다(그림 3). Lin과 Resnik의 방법론은 모든 GO 구조와 그룹 수에서 월등한 성능을 유도하고 있다. 다만 CC 활용 시 Resnik의 방법을 적용하였을 때, 그룹 3개에서 오히려 저조한 성능을 유도하지만, 이를 제외한 전체 성능 평가에서 Resnik의 방법이 가장 좋은 성능을 유도하고 있다.

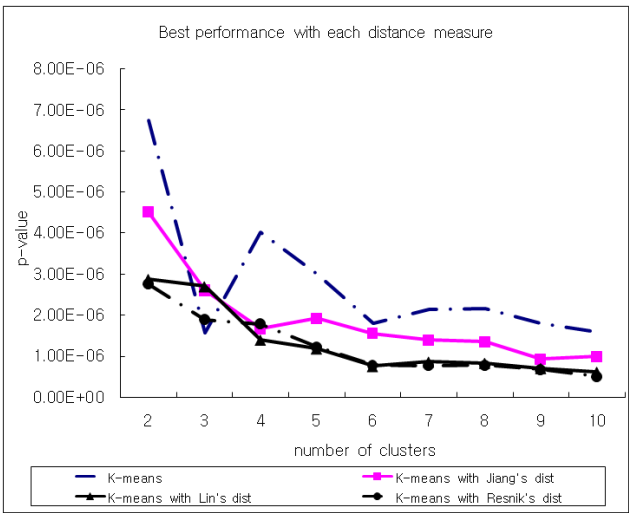
Spellman et al.의 유전자 데이터에 대한 실험 결과에서도 GO를 활용함으로써 성능 향상이 이뤄짐을 볼 수 있다(그림 4). Resnik과 Lin의 방법은 그룹 수가 2~4개일 때 약간의 성능 차이가 있으나, 그룹 수가 5개 이상부터는 비슷한 수준에서 가장 좋은 성능을 유도한다. Jiang et al.의 방법은 MF 활용 시 그룹 수가 4개일 때 K-means 보다 저조한 성능을 보이지만 전반적으로 성능 개선을 유도하고 있다.



(a) BP 정보를 활용한 결과

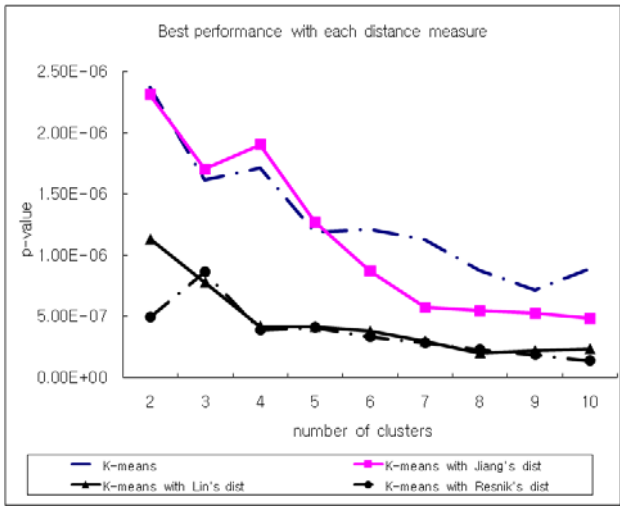


(b) CC 정보를 활용한 결과

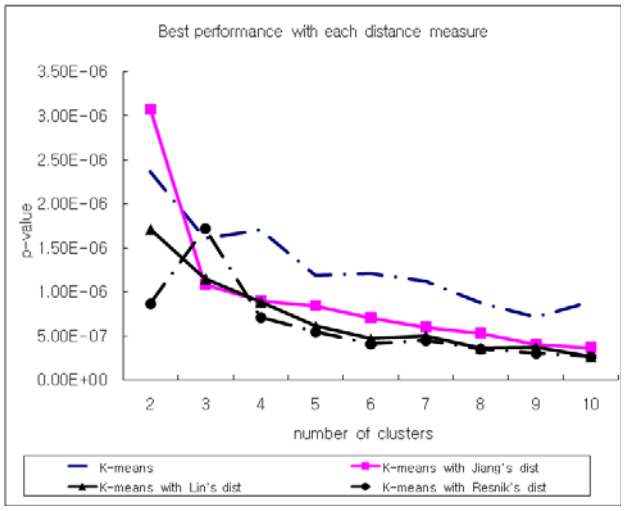


(c) MF 정보를 활용한 결과

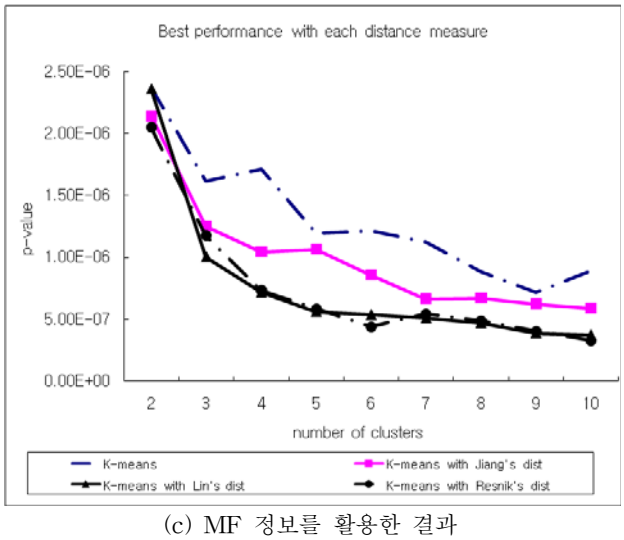
그림 2. Cho et al.에서의 유전자에 대한 분석 결과
Fig. 2. Results of gene data of Cho et al.



(a) BP 정보를 활용한 결과

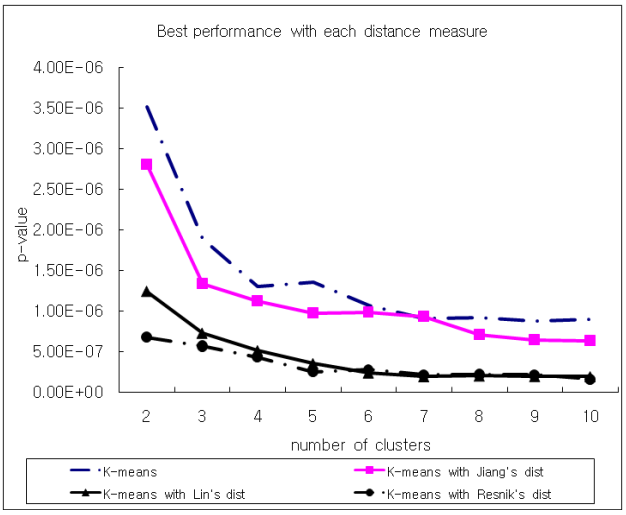


(b) CC 정보를 활용한 결과

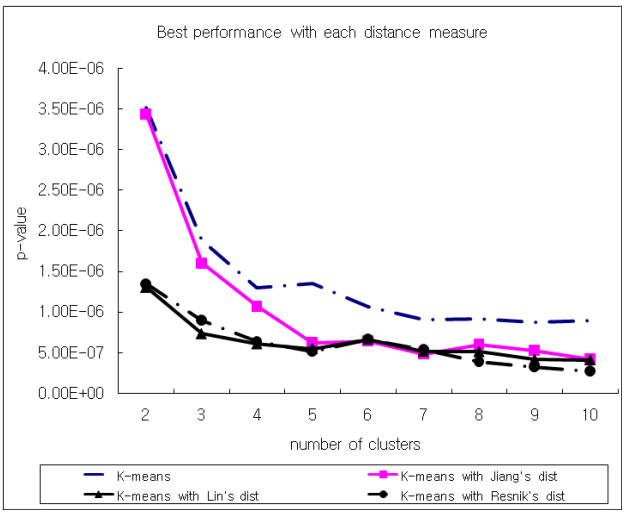


(c) MF 정보를 활용한 결과

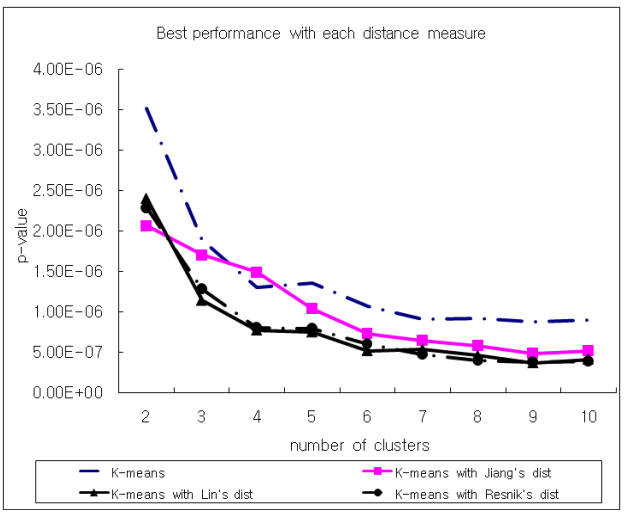
그림 3. Eisen et al.에서의 유전자에 대한 분석 결과
Fig. 3. Results of gene data of Eisen et al.



(a) BP 정보를 활용한 결과



(b) CC 정보를 활용한 결과



(c) MF 정보를 활용한 결과

그림 4. Spellman et al.에서의 유전자에 대한 분석 결과
Fig. 4. Results of gene data of Spellman et al.

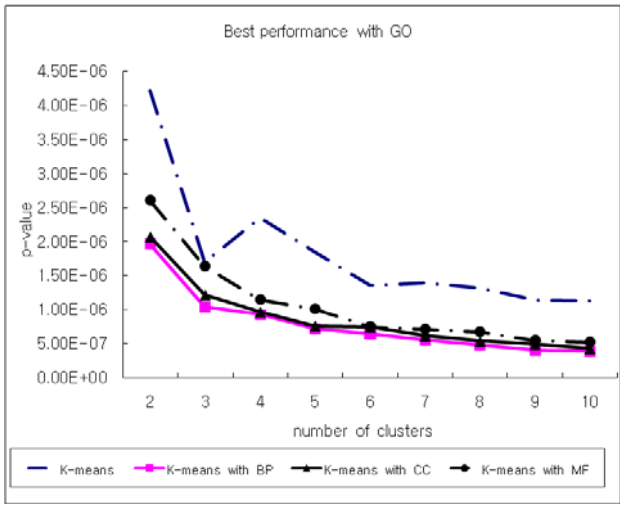


그림 5. GO 별 통계: 유전자 데이터 및 거리 측정 척도에 의한 실험 결과의 평균
Fig. 5. Performance according to GO

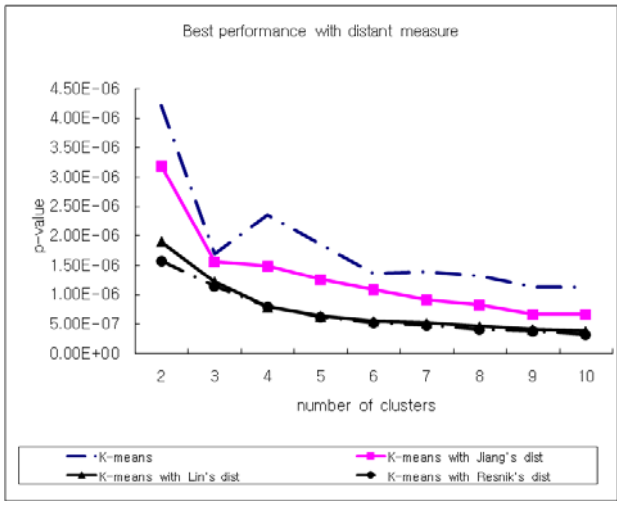


그림 6. 거리 측정 척도 별 통계 : 유전자 데이터 및 GO 구조에 따른 실험 결과의 평균
Fig. 6. Performance according to distance measure

전체적으로 GO를 활용함으로써 클러스터링 성능 개선에 직접적으로 영향을 주고 있으며, 거리 측정 척도에 따라 성능 개선의 차이가 발생하고 있음을 볼 수 있다.

3.3 GO 유형에 따른 클러스터링 결과

본 절에서는 GO 구조의 차이에 따라 발생하는 성능 변화를 비교한다. 따라서 데이터와 거리 측정 척도에 따른 모든 결과를 구조별로 평균을 적용함으로써 상대 비교 한다 (그림 5). 예를 들어, BP의 범례가 나타내는 결과 그래프는 세 가지 효모 데이터에 대하여 BP를 Lin, Resnik과 Jiang et al.에 의한 방법을 적용하여 클러스터링에 적용한 성능을 평균으로 나타낸 것이다. BP를 활용하였을 때, 월등한 성능 개선 요인이 되며, MF 또한 성능 개선을 유도하나 다른 구조를 적용한 결과보다 가장 저조한 개선 요인이 됨을 볼 수 있다. BP를 활용하는 경우 가장 좋은 성능을 보이는 이유는 생체 현상과 직접적으로 연관된 내용을 담는 구조이고 분석

유전자 데이터는 세포 주기 등에 대한 현상을 관측한 데이터이기 때문이라고 예상할 수 있다.

3.4 유전자 간 거리 계산 방법 별 클러스터링 결과

본 절은 GO의 활용 시 거리 측정 척도에 따른 성능 결과 분석을 다룬다. 각 GO 구조와 유전자 데이터에 따른 모든 결과의 평균치를 적용하여 비교 분석하였다. 예를 들어, 그림 6의 검정색 실선은 Jiang et al.의 방법론을 적용하여 얻어낸 모든 실험 결과에 대한 평균을 의미한다.

모든 거리 측정 척도에서 성능 개선을 유도하고 있으나, Jiang et al.의 방법은 다른 두 방법인 Lin과 Resnik의 방법보다 저조한 성능을 보이고 있다. 이에 반해 Lin과 Resnik의 방법을 활용하였을 때 월등한 성능 개선을 유도하며, 그 중 그룹 수 2개와 3개에서 Resnik의 방법이 Lin의 방법보다 성능 향상을 유도하고 있다. 그룹 수 4개 이상에서는 두 방법론의 성능이 비슷한 양상을 보이면서 가장 좋은 성능을 유도하고 있다.

가장 우수한 성능을 유도한 Resnik의 방법은 두 유전자에 해당하는 공통 부모 노드의 IC 값을 유사도로 추출하였다. 가장 저조한 성능을 유도한 Jiang et al.의 방법은 두 유전자의 기능 차이를 해당 노드의 IC 값 차이로 거리를 추출하였다. 결과적으로 GO를 통한 거리 추출 시 두 유전자의 기능이 비슷한 정도를 클러스터링에 반영하였을 때, 좋은 성능을 보인다. 또한 리프노드에 가까워질수록 IC 값의 크기 변화가 심해지는 문제가 있는데, Jiang et al.의 방법이 가장 큰 영향을 받아 저조한 성능을 보인다.

4. 결론 및 향후 과제

본 논문을 통해 마이크로어레이 데이터의 비지도 학습 방법론의 성능 향상을 위해서 GO를 활용할 수 있으며 GO의 활용 방법에 따라 성능의 차이가 발생함을 보였다. GO는 사전정보로 활용 가능한 정보 데이터베이스로써 최근 활용 기법들이 소개 되고 있다. 그 중 Kustra et al.에서 GO를 정보량 기반으로 하여 Lin의 거리 측정 척도를 적용하여 활용하는 방법이 최근 소개되었다. 본 논문은 Kutra et al.의 연구를 확장하여 (1) 동일한 유전자라도 GO의 유형에 따라 저장되는 위치가 달라짐으로써 발생하는 유전자 간 거리 차이와 (2) 동일한 GO라도 거리 측정하는 방법에 의한 유전자간 거리 차이 등이 클러스터링 분석에 미치는 영향을 분석하였다. 전체적으로 GO 정보를 클러스터링 분석에 적용함으로써 성능이 향상되고 있음을 볼 수 있지만, 방법에 따라 성능의 차이가 발생하고 있다. 그 중 효과적인 성능 향상은 GO의 유형으로 BP를 활용하고 거리 측정 척도로 Resnik의 방법을 적용하였을 때이며, 높은 성능은 유전자 기능의 예측에 신뢰도를 높이는 요인이 된다.

GO를 활용함으로써 클러스터링 성능이 향상됨을 볼 수 있었는데, 앞으로 진행해야할 연구 방향은 GO의 활용 방법의 차이에 따라 클러스터 그룹에서 형성되는 유전자 멤버가 변화해가는 과정을 분석함으로써, GO 활용의 적합성을 생물학적으로 검증할 계획이다.

참 고 문 헌

[1] J.Herrero et al. "A hierarchical unsupervised growing neural network for clustering gene expression patterns,", *Bioinformatics*, Vol. 17, no. 2, pp. 126-136, 2001.

[2] R. Sharan et al. "CLICK and EXPANDER : a system for clustering and visualizing gene expression data,", *Bioinformatics*, Vol. 19, no. 14, pp. 1787-1799, 2003.

[3] RJ. Cho et al. "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle,", *Molecular Cell*, Vol. 2, pp. 65-73, 1998.

[4] MB. Eisen et al. "Cluster analysis and display of genome-wide expression patterns," *Proc Natl Acad Sci,* Vol. 95, pp. 14863-14868, 1998.

[5] PT. Spellman et al. "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization,", *Molecular Biological of the Cell*, Vol. 9, pp. 3273-3297, 1998.

[6] P. Tamayo et al. "Interpreting patterns of gene expression with self-organizing maps : Methods and application to hematopoietic differentiation,", *Proc. Natl. Acad. Sci. USA,* Vol. 96, pp. 2907-2912, 1999.

[7] S. Tavazoie et al. "Systematic determination of genetic network architecture,", *Nature Genetics,* Vol. 22, pp. 281-285, 1999.

[8] The Gene Ontology Consortium, "Gene Ontology : tool for the unification of biology,", *Nature Genetics*, Vol. 25, 2000.

[9] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy,", *cmp-lg/9511007*, 1995.

[10] D. Lin, "An Information-Theoretic Definition of Similarity,", *In Proceedings of the 15th International Conference on Machine Learning*, 1998.

[11] JJ. Jiang and DW. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,", *ROCLING X*, 1997.

[12] D. Dotan-Cohen et al. "Hierarchical tree snipping : clustering guided by prior knowledge,", *Bioinformatics*, Vol. 23, no. 24, 3335-3342, 2007.

[13] Z. Fang et al. "Knowledge guided analysis of microarray data,", *Journal of Biomedical Informatics*, Vol. 39, pp. 401-411, 2006.

[14] D. Huang and W. Pan, "Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data,", *Bioinformatics*, Vol. 22, no. 10, 1259-1268, 2006.

[15] J. Cheng et al. "A Knowledge-Based Clustering Algorithm Driven by Gene Ontology,", *Journal of Biopharmaceutical Statistics*, Vol. 14, no. 3, pp. 687- 700, 2004.

[16] R. Kustra and A. Zagdanski, "Incorporating Gene

Ontology in Clustering Gene Expression Data,", *CBMS'06*, 2006.

저 자 소 개

고송(Song Ko)
2006년 : 전북대학교 컴퓨터공학과 학사
2009년 : 중앙대학교 컴퓨터공학과 석사
2009년 ~ 현재 : 중앙대학교 컴퓨터공학 박사과정

관심분야 : 데이터 마이닝, 베이지안 네트워크, 바이오정보학
Phone : +81-2-821-5304
Fax : +81-2-820-5301
E-mail : sko22.cau@gmail.com

강보영(Bo-Yeong Kang)
2002년 ~ 2004년 : 경북대학교 컴퓨터공학과 박사
2004년 ~ 2006년 : KAIST ICC/서울대학교 박사후연구원
2005년 ~ 2009년 : 서울대학교 치의학전문대학원 연구조교수
2009년 ~ 현재 : 경북대학교 기계공학부 조교수

관심분야 : 인공지능, 바이오메디컬 인포머틱스, 지능형 기기
Phone : +81-53-950-7542
E-mail : kby09@knu.ac.kr

김대원(Dae-Won Kim)
한국지능시스템학회 이사
현재 중앙대학교 컴퓨터공학부 부교수

Phone : +81-2-820-5304
Fax : +81-2-820-5301
E-mail : dwkim@cau.ac.kr