한국지능시스템학회 논문지 2011, Vol. 21, No. 5, pp. 569-574
http://dx.doi.org/10.5391/JKIIS.2011.21.5.569

마이크로어레이 데이터의 부공간 대조 샘플집단 마이닝

Mining of Subspace Contrasting Sample Groups in Microarray Data

이경미, 이건명*
Kyung Mi Lee, Keon Myung Lee

충북대학교 컴퓨터과학과, PT-ERC
E-mail: kmlee@cbnu.ac.kr

요 약

이 논문에서는 마이크로어레이 데이터에 대한 분석 문제로서 부공간 대조집단 식별 문제를 소개하고, 이를 해결하는 방법을 제안한다. 제안한 방법은 부공간에서 속성값이 대조적인 집단의 쌍들을 식별하기 위해, 먼저 각 속성에 대해서 분석자가 지정한 대조영역의 값을 갖는 두 개의 샘플집단을 선택한 다음, 연관규칙 마이닝과 유사한 형태의 방법으로 부공간의 차원을 점진적으로 확대해 가면서 대조집단을 추출한다. 마이크로어레이 데이터는 수천개 이상의 유전자에 대한 발현정보를 포함할 수 있는 다차원 데이터이기 때문에, 대조적인 발현특성을 갖는 유전자집합에 대한 샘플집단의 쌍을 모두 부차원에 대해서 질의를 통해 식별하는 것은 부담이 되지만, 제안한 방법을 사용하면 분석자가 지정한 대조영역 값의 범위를 기준으로 하여 모든 가능한 부공간에서의 대조집단을 효과적으로 추출할 수 있다.

키워드: 부공간 대조집단, 점진적 데이터마이닝, 마이크로어레이, 데이터분석, 군집화

Abstract

In this paper, we introduce the subspace contrasting group identification problem and propose an algorithm to solve the problem. In order to identify contrasting groups, the algorithm first determines two groups of which attribute values are in one of the contrasting ranges specified by the analyst, and searches for the contrasting groups while increasing the dimension of subspaces with an association rule mining strategy. Because the dimension of microarray data is likely to be tens of thousands, it is burdensome to find all contrasting groups over all possible subspaces by query generation. It is very useful in the sense that the proposed method allows to find those contrasting groups without analyst's involvement.

Key Words : subspace contrasting group, incremental data mining, microarray, data analysis, clustering

1. 서 론[1)]

20세기 후반의 분자생물학적 서열분석 기술의 비약적인 발전으로 인간 게놈을 비롯하여 다양한 생물종에 대한 분석이 이루어져 왔다.[1] 생물학의 중심원리(central dogma)에 따르면 모든 유전정보는 DNA에 코딩되어 있고, DNA의 유전자 서열이 RNA 서열로 전사된 다음, RNA 서열에 따라 아미노산 서열이 합성되어 단백질로 접혀서 생체내의 생명현상에 개입한다고 한다.[2] 생명현상은 단백질들의 상호작용에 의해서 일어나기 때문에, 특정 상황의 생명현상을 이해하기 위해서는 어떤 유전자가 어떤 상황에서 발현되는지 관찰하는 것이 중요하다. 인간의 경우에는 수만 종의 유전자가 있는 것으로 알려져 있는데, 이들 유전자의 발현정도를 효과적으로 측정하기 위한 대표적인 기술의 하나로 마이크로어레이가 사용되고 있다. 마이크로어레이 분석으로 나오는 데이터는 수만종의 유전자에 대한 측정값을 포함할 수 있기 때문에 기본적으로 다차원 데이터의 특성을 갖는다.

마이크로어레이 데이터는 속성의 값이 수치로 표현되는 전형적인 다차원 데이터이다.[3] 마이크로어레이로 측정된 유전자의 발현정도 데이터로부터 의미있는 정보나 지식을 추출하기 위해 다양한 데이터 분석 방법이 활용되고 있다. 데이터 분석이 성공적으로 이루어지면, 데이터를 생성한 시스템 자체 즉, 생체 내의 기작을 추정하기 위한 실마리를 얻을 수 있다. 데이터 군집화 분석은 데이터 간의 유사성에 따라 데이터들을 분할하는 것으로, 데이터 집합 고유의 구조를 파악하거나 추가적인 처리를 위한 일부 대상을 선택하기 위해 사용된다. 일반적으로 다차원 데이터는 차원의 증가에 따라 데이터간의 거리 차이가 적어지는 차원저주(the curse of dimensionality) 문제를 갖게 된다. 이 때문에 유사도나 거리 기반의 군집화 알고리즘이 다차원 데이터에서는 의미있는 패턴을 찾지 못하게

접수일자 : 2011년 9월 2일
완료일자 : 2011년 10월 12일
* 교신 저자
* 이 논문은 2011년 정부(교육인적자원부)의 재원으로 PT-ERC의 지원을 받아 수행된 연구임

되는 경우가 많다. 다차원 데이터에서 전체 차원이 아닌 일부 차원만을 고려하면 클러스터가 식별되고 이에 대한 의미있는 해석도 가능할 수도 있다. 다차원 데이터에 대해서 일부 차원에서 클러스터를 찾는 것을 부공간(subspace) 군집화라고 한다.[4]

마이크로어레이 데이터에서 대조적인 특성을 보이는 집단을 확인하는 것은 관심있는 분석대상이다. 질병의 유무, 전이 여부, 재발 여부, 생존 여부, 약물처리 전후 등에 따른 비교되는 샘플들의 집단이 있을 때, 대조적인 발현 특성을 보이는 유전자의 집단을 식별하는 것은 마커 개발 등에 유용하게 사용될 수 있다. 한편, 대조적인 발현특성을 보이는 유전자의 집단은 생물학적 경로의 예측 등에 활용될 수 있는 정보를 제공한다. 이러한 이유로 대조집단의 탐색에 대한 마이크로어레이 데이터 분석 연구가 진행되어 왔다.[5]

대조집단을 탐색하기 위해 군집화에 관련된 기법을 사용할 수 있는데, 먼저 군집화를 수행한 다음 대조적인 집단을 찾아보는 것이다. 부공간의 선택에 따라 즉, 유전자 집합의 선택에 따라 다양한 클러스터가 만들어질 수 있고, 유전자의 수가 수천에서 수만개까지 가능하기 때문에 부공간 군집화 기법을 직접 적용하는 것은 곤란할 수 있다. 따라서 이 논문에서는 점진적 데이터마이닝의 기법을 적용하여 모든 가능한 부공간에 대해서 대조집단을 추출하는 방법을 제안한다.

2. 관련연구

2.1 마이크로어레이 데이터 분석

생명현상을 이해하기 위한 분자생물학적 접근의 한 가지는 어떤 유전자가 발현되어 단백질로서 기능을 하는지 추정하는 것이다. 특정 생물 샘플에서의 유전자 발현 여부를 판정하기 위해서 DNA가 RNA로 전사될 때의 RNA 양을 측정하는 방법이 일반적으로 사용된다. 유전자의 발현량을 측정하는 방법으로 rtPCR 등이 사용되지만 개별 유전자별로 프라이머를 설계하여 실험을 하기 때문에, 많은 수의 유전자에 대해 적용하는 데는 제약이 있다. 이러한 제약을 보완하기 위한 방법으로 개발된 대용량 처리 기술인 마이크로어레이는 유리, 필터 또는 실리콘 판 위에 유전자를 검출할 수 있는 많은 수의 프로우브를 붙여 놓거나 합성하여 놓아서, 많은 유전자에 대한 발현량을 동시에 측정할 수 있도록 한 것이다.[3] 마이크로어레이 기술의 발전에 따라 현재 동시에 수만개 유전자의 발현을 동시에 측정할 수 있는 제품이 출현하는 등, 대규모 유전자의 발현정도 측정이 가능해지고 있다. 마이크로어레이는 동시에 많은 양의 데이터를 생성하기 때문에, 이에 대한 효과적인 분석기술이 필요하다. 마이크로어레이 데이터는 발현정도가 측정되는 유전자 각각이 하나의 차원에 해당하기 때문에 대표적인 다차원 데이터의 하나이다.

마이크로어레이 데이터는 시료의 준비과정, 분석장치의 고유 특성 등으로 인해서 시스템적 왜곡(systematic bias)이 개입될 수 있는 여지가 있기 때문에, 데이터 분석 이전에 정규화 등의 전처리를 하게 된다. 극단적인 값에 대해서는 지정한 값으로 변경하고, 평균값이 0값이 되도록 데이터값을 변환하고, 결손치(missing value)를 확인하여 주변값 정보를 활용하여 채워넣기 등을 한다. 한편, 어떤 샘플에서나 거의 일정하게 발현되는 참조 유전자(housekeeping gene)의 발현정도를 비교하여 샘플간의 값을 보정하는 방법이 전처리 과정에서 활용되기도 한다. 마이크로어레이 데이터 분석을 위해 t-test를 비롯한 다양한 통계검정 방법 뿐만아니라 여러 가지 군집화, 분류 등의 기법이 개발되어 적용되고 있다.[3,5,13]

2.2 부공간 군집화

부공간 군집화는 다차원 데이터에 대해서 모든 부공간에 있는 모든 클러스터를 찾아 내는 것이다. Parsons 등[4]은 사용된 탐색기법과 평가척도를 기준으로 부공간 군집화 알고리즘을 다음과 같이 분류하였다.

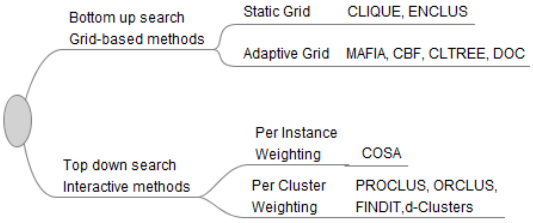


그림 1. 부공간 군집화 기법의 분류[4]
Fig.1 Classification of subspace clustering algorithms[4]

부공간 군집화를 하는 가장 단순한 방법은 모든 가능한 부공간에 대해서 군집화를 수행하고 클러스터를 잘 형성하는 부공간과 클러스터들을 선택하는 것이지만, 부공간의 개수가 기하급수적으로 증가하기 때문에 이러한 방법은 적용하기 곤란하다. 이러한 이유로 휴리스틱을 이용하는 부공간 탐색방법이 사용된다. 탐색방법 관점에서 부공간 탬색 알고리즘들은 상향식 방법과 하향식 방법으로 대별될 수 있다. 상향식 방법은 APRIORI[6]에서와 같이 밀도의 하향포함성질(downward closure property of density)을 이용하는데, 먼저 각 차원에 대해서 히스토그램을 작성하여 주어진 임계치 이상의 밀도를 갖는 구간을 탐색한 다음, 각 쌍의 차원에 대해서 구간을 결합하여 만들어지는 2차원 구간들에 대해서 임계치 이상 밀도를 갖는 것들을 선택한다. 하향포함성질을 이용하여 k-1차원의 구간들을 결합하여 k 차원의 구간을 찾는 방법을 반복하는 것이 상향식 알고리즘이 사용하는 방법이다. 상향식 방법을 채택한 알고리즘에는 CLIQUE[7], MAFIA[8] 등이 있다.

하향식 방법을 채택한 부공간 군집화 알고리즘은 전체 공간에서 시작하여 부공간의 클러스터를 찾아가게 되는데, 각 차원에 가중치를 부여하는 방식을 이용한다. 우선 공간을 구성하는 각 차원 즉, 속성에 대해서 동일한 가중치를 부여하여, 전체 공간에서 클러스터를 근사시키는 것에서부터 시작한다. 찾아진 각 클러스터에 대해서 차원별로 가중치가 부여되고, 이 가중치를 이용하여 클러스터가 다시 형성된다. 이 방법은 전체 차원에 대해서 군집화를 반복적으로 수행하기 때문에, 성능 개선을 위해 샘플링 기법을 사용하는 경우가 많다. 하향식 방법을 채택한 알고리즘에는 PROCUS[9], COSA[10] 등이 있다.

부공간 군집화 방법은 다차원 데이터를 대상으로 하기 때문에, 적용 대상 문제로 마이크로어레이 데이터를 다루

어왔다. 이들 방법은 부공간에서 클러스터를 찾는 것에 목적을 두고 있으며, 이 논문에서 대상으로 하는 대조적인 집단을 식별하는 데는 적합한 방법이 아니다.

2.3 반지도(semi-supervised) 학습기반 대조 클러스터 식별 방법

마이크로어레이 데이터가 다차원 데이터이고, 가시화가 가능하기 때문에 이러한 특성을 이용하는 군집화 방법이 연구되고 있다. 마이크로어레이 데이터 분석에서는 분석자가 특정 관심대상 유전자집단을 지정하여 분석하는 경우가 있다. 예를 들면, 전체 마이크로어레이 데이터에 대해서 군집화를 수행한 후에 이들 중에서 일부 영역을 대한 관심영역을 표현하고 이와 유사한 특성을 보이는 유전자집단 및 샘플집단을 추출하고자 한다. 이러한 분석을 지원하기 위해서 논문[11]에서는 설정된 관심영역에 대해서 클러스터를 추출하는 방법을 제안하고 있다. 위 방법은 하나의 클러스터만을 식별해 내는 것이지만, 이 논문에서는 관심영역을 지정하지 않고, 모든 부공간에 대해서 발견가능한 모든 대조집단을 찾는 방법을 제안한다.

3. 부공간 대조집단 마이닝

이 절에서는 마이크로어레이 데이터의 유전자 집합을 속성으로 하여 대조적인 특성을 보이는 샘플 집단의 쌍을 찾아내는 마이닝 방법을 제안한다. 논문에서 사용하는 개념을 분명하게 하기 위해 다음과 같이 용어를 정의한다.

정의. 대조집단(contrasting group)

두 집단 S_A, S_B가 속성집합 IA의 각 속성에 대해서 각 집단 내에서의 속성별 값들은 유사하면서, 집단간의 속성값은 분명한 차이를 보이면, 두 집단은 대조집단이라 한다.

위의 대조집단의 조건을 정형화하여 표현하면 아래와 같다.

(1) $\forall (a,b) \in S_A \times S_A$ or $S_B \times S_B, \forall g \in IA, g(a) \approx g(b)$.

(2) $\forall (a,b) \in S_A \times S_B$ or $S_B \times S_A, \forall g \in IA, g(a) \asymp g(b)$.

위 식에서 $g(a) \approx g(b)$는 객체 a의 속성 g의 값이 객체 b의 속성 g의 값과 유사한 것으로, $g(a) \asymp g(b)$는 객체 a의 속성 g의 값이 객체 b의 속성 g의 값과 상이한 것으로 간주할 수 있다는 것이다.

정의. 마이크로어레이 부공간 대조집단 탐색 문제

마이크로어레이 데이터에서 샘플들을 객체로, 유전자들을 속성으로 간주하여, 대조집단이 성립되도록 하는 모든 유전자 집합과 이에 대응하는 대조 샘플집단을 탐색하는 문제를 마이크로어레이 부공간 대조집단 탐색 문제라고 한다.

대조집단을 갖는 유전자의 집합을 결정하는 것은 부공간 클러스터링에서 부공간을 선택하는 것에 대응한다. 부공간 대조집단 탐색 문제는 부공간에 대해서 대조적인 집단의 쌍을 탐색대상으로 하는 점에서 부공간 클러스터링과 다르다.

3.1 단일 차원에서의 대조집단 탐색

제안한 방법에서는 마이크로어레이 데이터의 고유한 특성을 반영하여, 대조적인 집단을 판정한다. 마이크로어레이 데이터는 평균값이 0이 되도록 정규화시킨다. 정규화된 마이크로어레이 데이터에서 유전자의 발현량이 작은 경우에는 음수값을 갖게 되고, 발현량이 높은 경우에는 양수값을 갖게 된다. 대조집단을 표현하기 위해서 제안한 방법은 두 개의 소속함수(membership function)를 발현량의 정의구역에 대해서 정의한다. 편의상 낮은 발현량의 범위를 나타내는 소속함수는 μ_L로, 높은 발현량에 대응하는 소속함수는 μ_H로 나타낸다. 소속함수는 치역이 구간 [0,1]인 함수로서, 지정된 제약조건을 만족하면 1, 만족정도가 낮을수록 0에 가까워지는 함수이다. (그림 2)는 대조집단의 각 집단에 대한 대조영역을 지정하는 소속함수의 예를 보인 것이다. 대조집단 식별을 원하는 분석자는 자신이 관심있는 있는 부분에 대한 대조영역을 (그림 2)와 같은 소속함수를 사용하여 지정한다. 대조집단의 쌍에서 각 집단의 속성값들은 하나의 소속함수에 대해서 높은 소속정도를 보여야 대조집단이 된다.

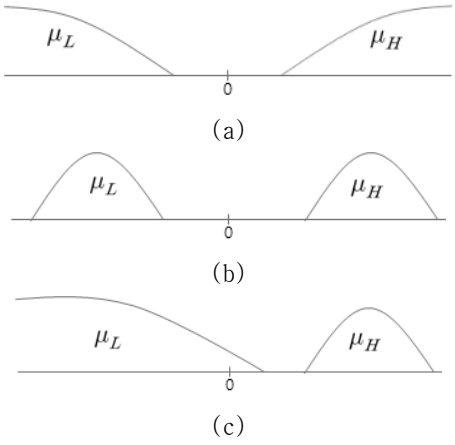


그림 2. 대조 영역 지정을 위한 소속함수의 형태
Fig.2 Shapes of Membership functions for specifying contrasting regions

기술의 편의상 분석 대상이 되는 마이크로어레이 데이터에 관련된 샘플 집합을 $S=\{s_1, s_2, \cdots, s_m\}$ 로 나타내고, 유전자 집합을 $G=\{g_1, g_2, \cdots, g_n\}$ 로 나타낸다. 마이크로어레이 데이터는 다음과 같이 2차원 행렬 형태 E로 나타낼 수 있고, 각 행렬의 원소 e_{ij}는 행에 해당하는 유전자 g_i가 열에 해당하는 샘플 s_j에서 발현정도값을 나타낸다.

$$E=\begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1m} \\ e_{21} & e_{22} & \cdots & e_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nm} \end{pmatrix} \quad (1)$$

지정된 소속함수 μ_L과 μ_H에 부합되는 샘플집단은 대조적인 것으로 간주한다. 다음은 유전자 g_i를 기준으로 해당 소속함수에 대한 소속정도를 임계값 θ_m이상 만족하는 대

조 샘플집단을 나타낸다.

$$S_L^1(\{g_i\}) = \{s_j | \mu_L(e_{ij}) \geq \theta_m\} \quad (2)$$

$$S_H^1(\{g_i\}) = \{s_j | \mu_H(e_{ij}) \geq \theta_m\} \quad (3)$$

즉, $S_L^1(g_i)$와 $S_H^1(g_i)$은 g_i에 대해서 대조집단이 된다. 예를 들면 다음과 같은 마이크로어레이 데이터가 있다고 하자.

$$E = \begin{array}{c} \\ g_1 \\ g_2 \end{array} \begin{array}{c} \begin{array}{cccc} s_1 & s_2 & s_3 & s_4 \end{array} \\ \begin{pmatrix} -0.5 & 0.5 & 0.4 & -0.1 \\ 0.8 & -0.7 & 0.6 & -0.5 \end{pmatrix} \end{array}$$

이에 대한 대조영역에 대한 소속함수가 사다리꼴 퍼지숫자 $trap(a,b,c,d)$로 아래와 같이 정해지고, 소속정도 임계값 θ_m이 0.5로 지정된다고 하자.

$$\mu_L = trap(-1.0, -1.0, -0.2, -0.1)$$

$$\mu_H = trap(0.1, 0.2, 1.0, 1.0)$$

여기에서 $trap(a,b,c,d)$는 다음과 같이 정의되는 소속함수이다.

$$trap(a,b,c,d)(x) = \begin{cases} 0 & \text{if } x < a \\ (x-a)/(b-a) & \text{if } a \leq x < b \\ 1 & \text{if } b \leq x < c \\ (x-c)/(d-c) & \text{if } c \leq x < d \\ 0 & \text{if } x > d \end{cases}$$

이때 두 유전자 각각에 대한 대조집단의 쌍은 다음과 같다.

$$S_L^1(\{g_1\}) = \{s_1\} \quad S_H^1(\{g_1\}) = \{s_2, s_3\}$$

$$S_L^1(\{g_2\}) = \{s_2, s_4\} \quad S_H^1(\{g_2\}) = \{s_1, s_3\}$$

3.2 부공간에서의 대조 샘플 집단 마이닝

제안한 방법은 마이크로어레이 데이터의 유전자 차원의 부공간에 대한 대조 샘플집단을 모두 찾아내기 위해, APRIORI 알고리즘[6] 방식과 유사한 하향포함성질을 이용한다. 대조집단을 선택하기 위해서 우선 식(2), (3)을 이용하여 각 유전자별로 대조집단을 찾아낸다. 대조집단으로서 의미를 갖기 위해서는 대조집단의 크기가 일정 수준이 되는 것이 바람직하다. 이를 위해 최소 집단의 크기를 나타내는 임계치로 θ_s를 지정하여, θ_s이상의 크기를 갖는 유전자들만을 차원 1인 대조 샘플집단 L_1의 원소로 한다. 유전자 2개인,즉 차원 2인 부공간에서의 대조집단의 유전자들은 L_1에 포함되어 있어야 한다. 따라서 대조집단을 탐색할 유전자 2개를 갖는 차원 2의 후보 부공간의 집합 C_2는 L_1에 포함된 유전자로 구성한다. 차원 2인 후보 부공간 $\{g_i, g_j\}$에 대한 대조집단은 $S_L^1(\{g_i\})$와 $S_H^1(\{g_i\})$, $S_L^1(\{g_j\})$와 $S_H^1(\{g_j\})$의 교집합으로 계산된다. 예를 들어 유전자 g_i와 g_j 각각에 대한 대조집단이 다음과 같다고 하자.

$$S_L^1(\{g_i\}) = \{s_a, s_b, s_c, s_d\}$$

$$S_H^1(\{g_i\}) = \{s_e, s_f, s_g, s_h\}$$

$$S_L^1(\{g_j\}) = \{s_a, s_b, s_c, s_k\}$$

$$S_H^1(\{g_j\}) = \{s_f, s_g, s_k, s_l\}$$

차원 1인 위의 부공간의 대조집단들로부터 차원 2인 후보 부공간 $\{g_i, g_j\}$이 만들어지고, 이 부공간에서의 대조집단은 차원 1인의 부공간에 대응되는 샘플집단의 교집합을 통해서 다음과 같이 구해진다.

$$S_L^2(\{g_i, g_j\}) = \{s_a, s_b, s_d\}$$

$$S_H^2(\{g_i, g_j\}) = \{s_f, s_g\}$$

차원 2인 부공간에서 다음과 같은 대조집단이 얻어졌다고 하자.

$$S_L^2(\{g_i, g_j\}) = \{s_a, s_b, s_c, s_d\}$$

$$S_H^2(\{g_i, g_j\}) = \{s_f, s_g, s_l\}$$

$$S_L^2(\{g_i, g_k\}) = \{s_a, s_b, s_e\}$$

$$S_H^2(\{g_i, g_k\}) = \{s_f, s_g, s_h\}$$

위의 대조집단으로부터 차원 3인 후보 유전자 부공간으로 $\{g_i, g_j, g_k\}$가 만들어지고, 이에 대한 대조집단은 대응하는 샘플집합 간의 교집합을 통해서 다음과 같이 결정된다.

$$S_L^3(\{g_i, g_j, g_k\}) = \{s_a, s_b\}$$

$$S_H^3(\{g_i, g_j, g_k\}) = \{s_f, s_g\}$$

대조집단의 최소 크기 θ_s가 3으로 설정되었다면, 유전자 부공간 $\{g_i, g_j, g_k\}$에 대해서는 만족스러운 대조집단이 없는 것으로 간주된다.

다음은 주어진 마이크로어레이 데이터에 대해서 모든 부공간에 대해서 대조집단을 추출하는 알고리즘을 기술한 것이다. 대조영역에 대한 소속함수 μ_L, μ_H는 분석자에 의해 지정되어야 하는데, 열지도(heatmap)을 제공하는 가시화 도구를 사용하면 대조 영역에 대한 소속함수를 정의하는데 도움을 받을 수 있다.

Algorithm: 점진적 부공간 대조 샘플집단 추출
Input: 마이크로어레이 데이터 E
대조 영역에 대한 소속함수 μ_L, μ_H
최소 소속정도 θ_m
최소 대조집단 크기 θ_s
Output: 모든 부공간의 대조집단 L
begin

1. 각 유전자 $g_i \in G$에 대해서, 대조집단을 식 (2)와 (3)을 통해서 결정한다.
2. 대조집단의 각 원소의 개수가 임계크기 θ_s이상인 유전자로 차원 1인 유전자 집합 L_1을 결정한다.

$$L_1 = \{g_i \mid |S_L^1(g_i)| \geq \theta_s, |S_H^1(g_i)| \geq \theta_s\}$$

3. $k \leftarrow 2$
4. L_{k-1}로부터 C_k를 생성한다.

$$C_k = \{g_{i1}g_{i2}\cdots g_{ik-2}g_pg_q |$$
$$g_{i1}g_{i2}\cdots g_{ik-2}g_p \in L_{k-1},$$
$$g_{i1}g_{i2}\cdots g_{ik-2}g_q \in L_{k-1},$$
$$S_L^k(\{g_{i1}, g_{i2}, \cdots, g_{ik-2}, g_p, g_q\}) =$$
$$S_L^k(\{g_{i1}, g_{i2}, \cdots, g_{ik-2}, g_p\})$$
$$\cup\, S_L^k(\{g_{i1}, g_{i2}, \cdots, g_{ik-2}, g_q\})$$
$$S_H^k(\{g_{i1}, g_{i2}, \cdots, g_{ik-2}, g_p, g_q\}) =$$

$$S_H^k(\{g_{i1}, g_{i2}, \cdots, g_{ik-2}, g_p\})$$
$$\cup\, S_H^k(\{g_{i1}, g_{i2}, \cdots, g_{ik-2}, g_q\})$$

여기에서 $k-2=0$인 경우에는 $g_{i1}, g_{i2}, \cdots, g_{ik-2}$ 부분이 없는 것으로 간주한다.

5. C_k로부터 L_k를 결정한다.

$$L_k = \{g_{i1}g_{i2}\cdots g_{ik} \mid g_{i1}g_{i2}\cdots g_{ik} \in C_k,$$
$$|S_L^k(\{g_{i1}g_{i2}\cdots, g_{ik}\})| \geq \theta_s,$$
$$|S_H^k(\{g_{i1}g_{i2}\cdots, g_{ik}\})| \geq \theta_s\}$$

6. $L_k \neq \phi$ 이면, $k \leftarrow k+1$를 한 다음 단계 4부터 반복한다.

7. 모든 부공간에 대한 대조집단의 결과로서, 모든 L_k에 대응하는 유전자 집합과 대조집단을 반환한다.

$$L = \cup_{i=1}^{k} \cup_{l_j \in L_i} \{(l_j, G_L^i(l_j), G_H^i(l_j))\}$$

여기에서 $G_L^i(l_i)$, $G_H^i(l_i)$는 L_k에 속하는 유전자 집합 l_j의 μ_L과 μ_H에 대응하는 샘플집합을 나타낸다.

end.

위 알고리즘은 Apriori 알고리즘에서와 같이 이전 단계의 결과를 다음 단계의 후보를 생성하는데 사용하는 점진적인 방법으로 대조집단을 찾아간다. 일반적으로 이러한 알고리즘은 차원 2인 부공간 즉, 2차원 부공간의 경우에 가장 많은 후보의 조합이 만들어지고, 크기가 커질수록 후보가 점점 축소되는 행태를 보인다. 제안된 알고리즘을 적용하기 위하여 분석자는 자신이 대조하고자 하는 영역을 지정하는 소속함수 μ_L과 μ_H를 정의하고, 대조 영역에 대한 소속정도의 최소값 θ_m과 대조집단의 최소크기 θ_s를 지정하는 것이 필요하다. 알고리즘의 출력인 L은 대조집단이 존재하는 부공간인 유전자의 부분집합 l_j과 이에 대응하는 샘플집합쌍 ($G_L^i(l_i)$, $G_H^i(l_i)$)의 튜플들$(l_j, G_L^i(l_j), G_H^i(l_j))$의 집합이다. 이들 각 튜플은 하나의 부공간 대조집단에 해당한다.

3.3 부공간 대조 샘플 집단 마이닝의 구현

일반적으로 데이터마이닝 알고리즘은 데이터의 크기가 컴퓨터의 주기억장치에 탑재될 수 없는 크기이기 때문에, 성능평가에서 전체 데이터를 주기억장치에서 몇 번 읽어오는지가 관심이 대상이 된다. 마이크로어레이 데이터는 대개 주기억장치에 충분히 탑재될 수 있는 크기이기 때문에 일반적인 알고리즘의 시간복잡도를 성능평가에 이용한다. Hegland 등[12]에 따르면 Apriori 알고리즘은 $O(m^2n)$의 시간 복잡도를 갖는다. 여기에서 m과 n은 각각 항목의 수와 데이터의 수를 나타낸다. 제안한 알고리즘에서는 속성의 개수, 즉 유전자의 개수가 항목의 수에 대응하고, 샘플의 개수가 데이터의 개수에 대응한다. Apriori 알고리즘에서와 다르게 제안된 알고리즘에서는 대조집단의 계산을 위해 교집합을 하는 연산이 발생한다. 그런데 대조집단의 샘플들을 정렬된 순서에 따라 저장하면 샘플크기의 선형시간만에 교집합 연산이 가능하게 된다. 유전자별로 대조집단을 결정하는 L_1 관련 처리는 샘플이름과 유전자이름에 일련번호를 부여하여 래딕스 정렬(radix sort)[14]를 하면 $O(nm)$만에 가능하다. 이후 단계에서는 원소를 일련번호순으로 정렬하여 관리하면, 후보 부공간 각각에 대한 교집합 연산이 $O(n)$에 가능하다. 제안된 알고리즘에 의해 생성되는 후보 부공간의 개수의 복잡도는 하향포함성질을 갖기 때문에 Apriori와 같은 시간 복잡도를 갖게 된다. 따라서 제안 알고리즘의 시간 복잡도는 $O(m^2n)$가 된다.

4. 구현 및 적용

마이크로어레이 데이터는 일반적으로 발현정도를 색상으로 표현하는 2차원 이미지인 열지도(heatmap) 형태로 가시화하기 때문에, 제안된 방법은 그래픽 사용자 인터페이스(GUI)를 제공하는 도구로 개발하였다. 제안된 도구는 사용자가 대조집단의 발현영역을 소속함수로 지정할 수 있는 인터페이스를 제공하고, 마이닝 결과로 얻어진 모든 가능한 부공간에 대해 대조적인 특성을 보이는 샘플 집단을 2차원 열지도로 가시화하는 기능을 제공한다. (그림 3)은 제안한 방법을 구현한 도구의 인터페이스를 보인 것으로, 왼쪽의 열지도는 1000개의 유전자에 대한 235개의 샘플에 대한 마이크로어레이 데이터를 입력으로 읽어들인 것이다. GUI 환경의 메뉴를 통해서 대조영역을 지정하기 위한 소속함수를 정의하고, 마이닝을 지시하면, 대조집단의 최소크기 요건을 만족하는 대조집단을 모든 가능한 부공간, 즉 유전자의 조합에 대해서 찾을 수 있게 된다. 마이닝에 의해서 결정된 부공간 및 대응되는 대조집단은 파일로 저장될 수 있고, (그림 4)와 같이 메뉴를 통해서 가시화시켜 대조적인 열지도 이미지를 통해 대조집단의 유효성을 확인할 수 있도록 구현되었다. (그림 3)의 화면 왼쪽의 열지도는 입력으로 주어진 마이크로어레이를 보인 것인데, 빨간색에 가까울수록 값이 큰 것이고, 녹색에 가까울수록 값이 작은 것이다. 그런데 전체 데이터에서는 어떤 유전자의 집합에 대해서 어떤 샘플들이 대조적인지 확인하기 어렵다. (그림 4)는 제안된 알고리즘에 따라 추출된 대조집단의 예를 보인 것인데, 가로축의 라벨들은 유전자의 이름이고, 새로축의 라벨은 샘플이름이다. 그림에서 왼쪽의 녹색부분이 하나의 집단을 구성하고, 오른쪽의 빨간색이 대조적인 집단을 나타낸다.

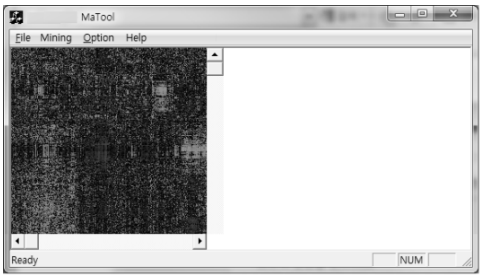
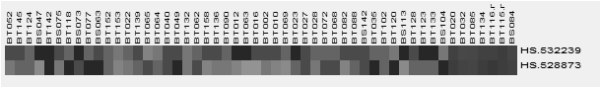
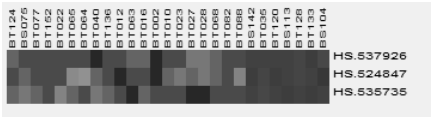


그림 3. 대조 유전자 집단 마이닝 도구
Fig. 3 A snapshot of the developed tool

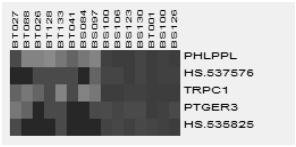


(a) 크기 2의 부공간에서의 대조집단 예



(b) 크기 3의 부공간에서의 대조집단 예

(c) 크기 4의 부공간에서의 대조집단 예



(d) 크기 5의 부공간에서의 대조집단 예

그림 4. 마이닝된 샘플대조집단 가시화의 예
Fig. 4 Displays of mined sample contrasting group pairs

5. 결 론

전형적인 다차원 데이터의 특성을 갖는 마이크로어레이 데이터는 가시화 도구의 도움을 받아 직관적인 분석을 하는 경향이 있다. 이 논문에서는 마이크로어레이 데이터의 분석 방법으로 모든 부공간의 대조집단을 마이닝하는 알고리즘을 제안하였다. 제안된 기법은 분석자가 대조영역의 값의 범위를 지정할 수 있도록 하기 위해 소속함수를 사용하여, 융통성있는 분석을 가능하게 하였다. 한편, 분석자의 개입없이 모든 부공간에 대해서 대조집단을 추출할 수 있도록 하여 분석자의 부담을 줄일 수 있게 하였다. 한편, 분석 및 분석결과의 확인을 용이하게 할 수 있도록 제안된 방법을 가시화 도구로 개발하여, 실제 분석에 효과적으로 활용할 수 있도록 하였다. 향후 단순히 대조적인 집단만을 식별하는 것뿐만 아니라 두 개의 비교 샘플집합을 지정하면, 이 두 집합을 식별할 수 있는 부공간의 대조 패턴조합을 찾을 수 있도록 확장하면 유전자 마커 설계에 유용할 것으로 기대된다.

참 고 문 헌

[1] ____. The Human Genome at Ten, *Nature*, Vol. 464, pp.649-650, Apr. 2010.

[2] T. A. Brown, *Genomes*, John Wiley & Sons, 1999.

[3] S. Draghici, *Data Analysis Tools for DNA Microarrays*, Chapman & Hall/CRC, (2003).

[4] L. Parsons, E. Haque, H. Liu, "Subspace Clsutering for High Dimensional Data: A Review," S*IGKDD Explorations*, Vol.6, No. 10, pp.90-105, 2004.

[5] K. M. Lee, K. S. Hwang, C. H. Lee, "Fuzzy Set-based Microarray Data Analysis Techniques for Interesting Block Identification," *Proc. of FUZZ-IEEE 2009*, 2009.

[6] R. Aggrawal, T. Imielinski, A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *SIGMOD*, Vol. 22, No. 2, pp.207-216, 1993.

[7] R. Aggrawal, J. Gehrke, D. Gunopulos, P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," *Proc. of the 1998 ACM SIGMOD*, pp.94-105, ACM Press, 1998.

[8] S. Goil, H. Nagesh, A. Choudhary, "Mafia: Efficient and scalable subspace clustering for very large data sets," *Technical Report CPDC-TR-9906-010*, Northwestern University, 1999.

[9] C. C. Aggrawal, J. L. Wolf, P. S. Yu, C. Procopiuc, J. S. Park, "Fast algorithms for projected clustering," *Proc. of the 1999 ACM SIGMOD*, pp.61-72, ACM Press, 1999.

[10] J. H. Friedman, J. J. Meulman. Clustering objects on subsets of attributes, http:// citeseer.nj.n ec.com/friedman02clustering.html, 2002.

[11] 황경순, 이건명, 이찬희, "마이크로어레이 데이터에 대한 퍼지 경계 지역 클러스터링," *한국지능시스템학회 춘계학술대회논문지*, 2009.

[12] M. Hegland, "The Apriori Algorithm - a Tutorial," *Mathematics and Computation in Imaging Science and Information Processing*, pp.209-262, World Scientific, 2007.

[13] 박대훈, 김연태, 김성신, 이춘환, "마이크로어레이 데이터에 적용된 2단계 K-means 클러스터링의 소개," *한국지능시스템학회 논문지*, 17권, 2호, pp.167-172, 2007.

[14] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, *Introduction to Algorithms*, The MIT Press, 2001.

저 자 소 개

이경미
1992 건국대학교 산업공학과 석사
1993~1994 일본 九州工業大學 研究生
1994~1996 일본 九州工業大學 정보공학과 박사과정수료
2011~현재 충북대학교 컴퓨터과학과 박사과정
관심분야: 소프트 컴퓨팅, 기계학습, 최적화

이건명
1990, 1992, 1995 KAIST 전산학과(학사, 석사, 박사)
1995~1996 : INSA de Lyon (Lyon, France), Post-Doc Fellow
1996 Park Scientific Instruments (Sunnyvale, USA) Staff Scientist
2001~2003 Univ. of Colorado at Denver, Visiting Professor
2008~2009 Indiana University, Visiting Scholar
1996~현재 충북대학교 전자정보대학 교수
관심분야: 기계학습, 바이오인포매틱스, 소프트컴퓨팅