한국지능시스템학회 논문지 2011, Vol. 21, No. 5, pp. 575-580
http://dx.doi.org/10.5391/JKIIS.2011.21.5.575

속성값 이산화 및 부정값 허용을 하는 의사결정트리 기반의 유전자 발현 데이터의 마커 후보 식별

Candidate Marker Identification from Gene Expression Data with Attribute Value Discretization and Negation

이경미, 이건명*
Kyung Mi Lee, Keon Myung Lee

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

요 약

맞춤형 의료에 대한 기대가 커지면서 분자생물학적인 의료정보의 분석이 중요해지고 있다. 유전자 발현 데이터는 생명현상의 분자생물학적 동태을 보여주는 대표적인 데이터이다. 유전자 발현 데이터의 분석을 통해서 유전자 발현 수준에서의 특정 질병의 발병, 전이, 재발 등을 예측하기 위한 마커에 대한 관심이 많다. 두 개의 대조적인 관심 집단을 식별하는 유전자를 찾기 위해 통계적인 방법 등이 이용되어 왔다. 이 논문에서는 여러 유전자의 조합을 통해서 집단을 식별할 수 있는 후보 마커를 찾는 의사결정트리 기반 방법을 제안한다. 제안한 방법에서는 수치적인 유전자의 발현값을 세 개의 범주값으로 이산화시키고, 유전자 발현값을 해당 범주값 뿐만 아니라 범주값의 부정값을 허용할 수 있도록 한다. 한편, 마커로 활용하기 위해서는 소수의 유전자만을 사용하는 것이 바람직하기 때문에, 마커에 소속할 유전자의 개수를 제한하여 마커를 찾도록 한다.

키워드 : 마이크로어레이, 데이터분석, 바이오인포매틱스

Abstract

With the increasing expectation on personalized medicine, it is getting importance to analyze medical information in molecular biology perspective. Gene expression data are one of representative ones to show the microscopic phenomena of biological activities. In gene expression data analysis, one of major concerns is to identify markers which can be used to predict disease occurrence, progression or recurrence in the molecular level. Existing markers candidate identification methods mainly depend on statistical hypothesis test methods. This paper proposes a search method based decision tree induction to identify candidate markers which consist of multiple genes. The propose method discretizes numeric expression level into three categorical values and allows candidate markers' genes to be expressed by their negation as well as categorical values. It is desirable to have some number of genes to be included in markers. Hence the method is devised to try to find candidate markers with restricted number of genes.

Key Words : microarray, data analysis, marker, bioinformatics

1. 서 론[1)]

통계적으로 유의한 의료시술을 기반으로 하는 증거기반 의료분야에서 인간 게놈의 확보와 유전자의 질환에 대한 역할이 규명되어 감에 따라 개인의 특성에 맞춘 맞춤의료(personalized medicine)에 대한 기대가 증가하고 있다. 맞춤의료에서는 임상적인 정보뿐만 아니라 개인의 DNA 서열에서의 차이, 세포 내의 유전자의 발현에서의 차이 등이 질병의 발생, 전이, 재발, 치료의 효과 등 예후에 어떤 영향을 주는지 파악하여 최선의 의료 서비스를 제공하려고 한다.[1] 예방적 의료 및 치료적 의료를 위해 건강 상태나 질환에 대한 예측 및 예후를 추정하기 위한 바이오마커(biomarker)에 대한 많은 연구 개발이 진행되어 왔다.[2] 맞춤의료에서는 질환에서 여러 상황에 따라 차별적인 발현정도를 보이는 유전자를 발굴하여 마커로서 활용하려는 연구가 많이 진행되어왔다.[3] 유전자의 발현정도를 측정하는 효과적인 도구의 하나로 마이크로어레이가 활용되고 있다. 마이크로어레이는 수만 가지 유전자의 발현정도를 동시에 측정할 수 있는 높은 처리율을 갖는 기술로서, 수치적인 데이터를 포함하는 고차원 특성을 갖는 유전자 발현 데이터를 생성하게 된다. 질환군 및 정상군과 같이 대조적인 집단에 대해서 차별적인 특성을 보이는 유전자는 질환에 대한 진단 또는 예측을 위한 마커로 활용될 수 있다. 기존의 유전자 발현 마커를 탐색하기 위한 방법으로 전형적으로 사용되어 온 방법은 t-검정 등의

접수일자 : 2011년 9월 2일
완료일자 : 2011년 10월 15일
* 교신 저자
* 이 논문은 2011년 정부(교육인적자원부)의 재원으로 PT-ERC의 지원을 받아 수행된 연구임

통계적인 추정방법이다. 이 방법은 특정 유전자의 발현정도가 통계적으로 유의하게 두 집단에서 차이를 보인다면 마커로서 의미있다고 판정한다. 그런데 생명현상은 단순하지 않기 때문에 비교 집단에 대해서 분명하게 차이를 보이는 유전자를 찾는 것이 쉽지 않다. 따라서 복수개의 유전자의 조합을 통해서 두 비교집단을 차별화하는 시도를 하게 된다. 이를 위해서 분석자의 직관과 시행착오적인 방법이 사용되고 있어 체계적이고 노력이 절약되는 방법이 필요하다.

이 논문에서는 복수 개의 유전자를 사용하는 두 개의 대조집단을 식별하는 후보 마커를 탐색하는 방법을 제안한다. 기존의 대조집단 식별을 위한 유전자발현 데이터 분석 방법은 의미있게 차이가 나는 발현정도를 나타내는 개별 유전자를 식별하는 것에 관심을 갖는다.[4] 반면에 제안된 방법은 단일 유전자로는 대조집단을 식별하기 어려운 상황에서, 복수 개의 유전자를 동시에 고려하여 대조 집단을 식별하도록 한다. 또한 유전자의 수치적인 발현값을 세 개의 범주값으로 이산화시키고, 유전자 발현값을 해당 범주값 뿐만 아니라 범주값의 부정값(negated value)을 허용할 수 있도록 한다. 한편, 마커로 활용하기 위해서는 소수의 유전자만을 사용하는 것이 바람직하기 때문에, 마커에 소속할 유전자의 개수를 제한하여 마커를 찾도록 한다. 이 논문은 다음과 같이 구성된다. 2장에서는 기존 유전자발현 데이터분석 방법으로 마이크로어레이의 데이터 분석에서 사용된 대조집단간에 발현의 차이가 있는 유전자를 식별하는 방법에 대해서 소개한다. 3장에서는 제안된 결정트리 기반의 후보 마커 식별 방법을 소개하고, 기존의 방법과의 차이점에 대해서 기술한다. 4장에서는 제안된 방법의 적용 사례를 소개하고, 5장에서 결론을 맺는다.

2. 유전자발현 데이터 분석

2.1 유전자발현 데이터의 획득

생물학의 중심원리(central dogma)에서는 모든 유전정보가 DNA에 코딩되어 있고, DNA의 유전자 서열이 RNA 서열로 전사된 다음, RNA 서열에 따라 아미노산 서열이 합성된 다음 단백질로 접혀서 생체내의 생명현상에 개입한다고 한다.[5] 생명현상은 단백질들의 상호작용에 의해서 일어나기 때문에, 특정 상황의 생명현상을 이해하기 위해서는 어떤 유전자가 어떤 상황에서 발현되는지 관찰하는 것이 중요하다. 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.[4] 마이크로어레이는 유전자와 프로우브 간의 혼성화(Hybridization)의 정도를 정량적으로 측정하기 때문에, 기술의 발전에도 불구하고 재현성에 대한 검증이 분석에서 아직 주요 관건이 되고 있다. 최근 차세대 시퀀싱(Next Generation Sequencing) 기술이 정확도가 개선된 유전자발현 정보를 얻는데 활용되고 있다.[9] 이 논문에서는 마이크로어레이나 차세대 시퀀싱 기술을 통해서 얻어지는 유전자 발현 정보에 대해서 적용할 수 있는 후보 마커 식별방법을 제안한다.

마이크로어레이 실험 등을 통해서 측정된 유전자발현 데이터는 각 샘플에 대한 유전자별 발현정도에 대한 수치적인 측정값으로 2차원 배열과 같은 열지도(heatmap) 형태로 표현이 될 수 있고, 그림 1과 같이 발현정도가 클수록 빨간색이 진하게, 낮을수록 녹색이 진하게 기사화하여 데이터를 제공하기도 한다.

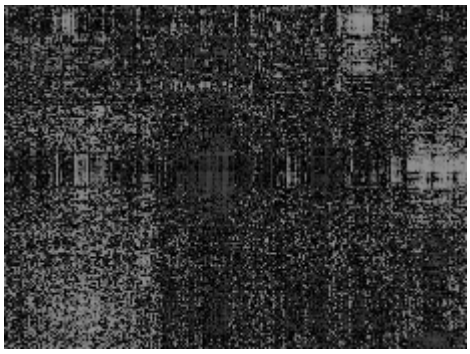

그림 1. 마이크로어레이 데이터의 열지도
Fig.1 Heatmap of Microarray Data

2.2 유전자발현 데이터 처리

마이크로어레이 등을 통해서 측정된 유전자 발현 데이터는 시료의 준비과정, 분석장치의 고유 특성 등으로 인해서 시스템적 왜곡이 개입될 수 있는 여지가 있다. 따라서 데이터 분석 이전에 데이터의 특성을 반영하여 왜곡을 완화시키기 위해 다음과 같은 전처리를 하게 된다. 극단적인 값에 대해서는 지정한 값으로 변경하고, 평균값이 0 값이 되도록 데이터값을 변환한다. 측정값에 대한 신뢰도 문제 등으로 결손치가 발생하면, 주변값 정보를 활용하여 결손치에 값을 채워넣는 처리를 한다. 한편, 어떤 샘플에서나 거의 일정하게 발현되는 참조 유전자(housekeeping gene)의 발현정도를 비교하여 샘플간의 값을 보정하는 방법이 전처리 과정에서 활용되기도 한다. 마이크로어레이 실험 등의 유전자 발현 정보를 측정하는 실험에 실험 대상(case)과 기준 대상(control)을 서로 비교하여 유의적인 차이를 보이는 유전자를 찾는데 주로 관심을 갖는다. 이러한 유전자를 식별하기 위해 여러 가지 방법이 도입되어 사용되고 있다.

배변화(fold change) 방법은 대조되는 것 간의 발현량의 값이 상대에 대해서 몇배인지를 계산하여, 일정 임계값 이상이면 차이가 있는 유전자로 선택하도록 한다. 기본적으로 이 방법은 집단간의 비교가 아니라 두 대조 샘플간의 실험으로부터 차이가 있는 유전자를 찾는 것이다. 집단간의 비교를 하기 위해서는 집단의 평균을 계산한 다음, 배변화를 계산하여야 하기 때문에 집단간의 비교에서는 구별력이 떨어진다. 한편, 임계값의 선택이 작위적이라는 점 때문에 유전자 선택의 정당성을 확보하기가 쉽지 않다. 이러한 점을 보안하기 위한 방법으로 비정상 비(unusual ratio)방법이 사용되기도 한다. 이 방법에서는 각 유전자에 대해서 실험대상과 비교대상 간의 발현량의 비를 모두

구한 다음, 이들 비의 평균을 계산하고, 평균이 0이 되도록 발현량의 비를 z-변환하여 분포를 구성한다. 구성된 분포에서 평균인 0으로부터 거리가 $\pm 2\sigma$이상 떨어진 유전자를 선택하는 것과 같이, 평균에서 일정 거리 이상 떨어진 것을 선택한다. 일반적으로 비정상 비 방법이 배변환 방법보다 나은 특성을 보이는 것으로 알려져 있다.[4]

통계적 가설 검정(hypothesis test)이 차별화된 발현특성을 보이는 유전자를 선택하기 위해 사용되기도 한다. 이 방법에서는 발현량의 비를 데이터로 하여 확률 분포를 상정하고, 귀무가설(null hypothesis)을 기각하는 영역에 들어가는 유전자를 차별화된 특성을 보이는 유전자로 선택한다. Tushher 등[6]이 제안한 SAM(Significance analysis of microarrays) 방법은 두 대조집단 사이에서 차별적인 발현특성을 보이는 유전자를 선택할 수 있도록 하는 것으로, 어떤 유전자가 우연히 유의적 차이가 있는 것으로 식별될 확률인 FDR(false discovery rate)를 계산하기 위해 샘플 집단을 무작위로 리샘플링하여 확률분포를 생성하여 통계치를 계산한다. 일원분산분석(ANOVA) 기법을 이용하여 측정치에 대한 변량들을 기반으로한 모델을 구하여 유의한 차이를 보이는 유전자를 선택하는 방법도 있다. 이들 통계적인 방법은 단일 유전자가 대조집단에 대해서 차이를 보이는지 판정하는데에 관심이 있다.

발현량이 적은 것과 많은 것이 두 개의 가우시안 분포로부터 각각 나온 것이라는 가정을 하여 가우시안 혼합모델(mixture model)을 최대 우도법(maximum likelihood estimate)을 통해서 모델링하는 방법도 있다.[4] 구성된 모델을 기준으로 각 유전자의 발현정도값에 따라 발현량이 많은 것에 대한 가우시안 모델에서 나왔을 확률값과 반대 모델에서 나왔을 확률값을 계산하여, 해당 유전자의 발현량이 작은 것인지 많은 것인지를 판정한다. 이를 기준으로 대조집단에서의 발현량의 판정이 상이한지 비교하여, 대조집단 간에 유의적인 발현량의 차이를 보이는 유전자를 결정한다.

앞에서 살펴본 방법들은 대조집단에서 유의적인 차이를 보이는 개별 유전자를 식별하는 것으로, 두 집단을 구별하기 위한 것을 목표한 것이 아니다. 실제로 단일 유전자 만으로는 대조집단을 구별하기 어렵기 때문에, 복수개의 유전자를 활용하는 것이 요구된다. 한편, 집단의 식별 문제는 전형적인 분류(classification) 문제이기 때문에, k-최근방(k- nearest neighbor), 의사결정트리, 신경망, SVM 등이 집단의 식별에 사용되어 왔다.[7] 이들 분류 방법은 집단의 식별에 주된 관심이 있고, 식별을 위한 최소한 유전자가 어떤 것이 되어야 하는지에 대해서는 고려를 충분히 하지 않고 있다.

3. 의사결정트리 기반의 집단 식별 후보 마커 탐색 방법

여기에서는 유전자발현 데이터에 대해서 두 개의 대조집단을 식별하기 위한 최소 개수의 유전자와 이들의 특성을 찾는 방법을 제안한다. 먼저 3.1절에서는 후보 마커로서 고려해야할 성질에 대해서 기술하고, 3.2절에서는 후보 유전자를 찾기 위해 변형된 결정트리 구성하는 방법을 소개한다.

3.1 집단 식별을 위한 후보마커

의료에서 마커는 건강 또는 질병에 대한 특정 상태의 유무를 판정하기 위해 사용되는 것으로, 마커에 대한 측정치는 비용을 요구하는 검사를 통해서 얻어진다. 따라서 유전자 기반 마커라면 최소한의 유전자 만을 대상으로 하는 것이 경제적이다. 또한 마커로서 적용범위 및 정확도가 큰 것이 바람직하다. 유전자 발현 정도의 정성적인 특성 때문에, 정확한 수치값을 마커의 판단 기준으로 활용하는 것보다는 분석자의 관점에서 이해하기 쉬운 형태로 제공하는 것이 데이터의 전반적인 특성을 파악하도록 하는데 도움이 될 수 있다. 마커로서의 유효성은 후보로 추천된 것들에 대해서 엄밀한 분석을 해서 판정하게 됨으로, 후보 마커의 기술 자체가 수치적으로 엄밀할 필요는 없다. 이러한 관점에서 제안한 방법은 집단의 식별을 위한 유전자의 특성값을 정성적인 것으로 표현한다. 발현정도를 정성적으로 언급할 때는, 발현정도가 증가한 것(up- regulation), 감소한 것(down-regulation), 큰 변동이 없는 것(neutral)인 것으로 구별한다. 제안한 방법에서도 발현정도의 값을 이에 대응하여 U(up- regulation), N(neutral), D(down-regulation) 등으로 나타낸다. 이를 위해 연속구간의 수치적인 발현정도 값을 $\{D, N, U\}$로 변환하는 매핑 M이 필요하다.

$$M: R \rightarrow \{D, N, U\} \quad (1)$$

매핑을 어떻게 하는 가에 따라 두 대조집단을 식별하는 후보 마커가 찾아질 수도 있고 만족스러운 것을 찾지 못할 수도 있다. 마이크로어레이 데이터 등의 유전자 발현 데이터는 분석을 위해 평균값이 0이 되도록 정규화시키는 것이 일반적이다. 따라서 0보다 작아질수록 D이 되고, 0 부근에서는 N이 되고, 0보다 커질수록 U의 성향이 강해진다. 이러한 특성에 맞게 매핑을 하도록 하게 위해 퍼지이론의 소속함수를 사용한다. 소속함수는 그림 2와 같이 치역이 구간 [0,1]인 함수로서, 지정된 제약조건이 만족되면 1, 만족정도가 낮을수록 0에 가까워지는 함수이다.

그림 2. 이산화를 위한 소속함수의 형태
Fig.2 Membership functions for discretization

분석자에 의해 이산화 구간에 대한 소속함수들이 주어지면, 발현정도의 값 e의 각 소속함수에 대한 소속정도 $\mu_D(e), \mu_N(e), \mu_U(e)$를 계산하고, 소속정도가 특정임계값 θ_m이상인 것의 이산값 기호 $label(e)$를 선택한다.

$$label(e) = \arg max_{l \in \{D,N,U\}} \{\mu_l(e) | \mu_l(e) \geq \theta_m\} \quad (2)$$

하나의 발현정도 값이 두 가지 이상의 이산화 기호로 변환되는 것을 피하기 위해, 두 개 이상의 소속함수의 소속정도가 임계값 θ_m보다 커지지 않게 소속함수를 정의하는 것이 필요하다. 한편, 세 개의 소속함수가 발현정도 값의 전체 범위를 모두 포함할 필요도 없다. 즉, 그림 2에서 인접한 소속함수가 중첩되지 않고, 일부 영역은 소속함수가 정의되지 않을 수 도 있다는 것을 의미한다.

편의상 분석 대상이 되는 유전자발현 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 나타내도록 한다. 유전자발현 데이터는 다음과 같이 2차원 행렬 형태로 나타낼 수 있고, 각 행렬의 원소는 행에 해당하는 유전자의 열에 해당하는 샘플에서의 발현정도값을 나타낸다. 유전자발현 데이터 E에서, 원소 e_{ij}는 i번째 유전자 g_i의 j번째 샘플 s_j에서 발현정도를 나타낸다.

$$E=\begin{pmatrix} e_{11} & e_{12} \cdots & e_{1m} \\ e_{21} & e_{22} \cdots & e_{2m} \\ \vdots & \vdots \ddots & \vdots \\ e_{n1} & e_{n2} \cdots & e_{nm} \end{pmatrix} \quad (3)$$

후보마커를 $\{D, N, U\}$의 정성적인 값을 사용하여 표현하기 때문에, 발현정도값을 이들 정성적인 값으로 변환한 다음과 같은 2차원 행렬 L을 생성한다.

$$L=\begin{pmatrix} l_{11} & l_{12} \cdots & l_{1m} \\ l_{21} & l_{22} \cdots & l_{2m} \\ \vdots & \vdots \ddots & \vdots \\ l_{n1} & l_{n2} \cdots & l_{nm} \end{pmatrix} \quad (4)$$

여기에서 $l_{ij}=label(e_{ij})$로서, 발현정도값 e_{ij}를 기호로 변환한 값을 나타낸다.

제안한 방법에서 후보 마커는 다음과 같이 목표집단에 대해 만족되는 선택된 유전자 이름들과 이들에 대한 기호값으로 표현된다. 여기에서 후보마커에 대한 표현식은 논리곱의 합 형태인 DNF(disjunctive normal form)로 나타낸다.

$$(g_i = L_i \wedge g_j = L_j) \vee (g_i = L_p \wedge g_k = L_q) \quad (5)$$

위의 후보마커 표현 예에서 g_i, g_j, g_k등은 마커로 선택된 유전자들을 나타내고, 대상집단에 대한 해당 유전자의 발현 특성을 나타내는 L_i, L_j, L_p, L_q 등은 집합 $\{D, N, U\}$의 하나를 나타낸다.

3.2 후보 마커 식별 방법

여기에서는 유전자 발현 데이터에 대해서 두 집단을 구별할 수 있는 최소한의 유전자를 이용한 후보마커 표현식을 추출하는 방법을 소개한다. 제안한 방법은 기본적으로 의사결정트리(decision tree)를 생성하는 과정을 통해서 후보마커를 생성한다. 의사결정트리는 점진적으로 트리를 확장해 하는 방법으로 구성되는데, 모든 데이터를 포함한 루트노드에서 시작하여, 노드에 포함된 데이터를 이질도(impurity)를 최소화시킬 수 있도록 분할하는 속성을 선택하여 이에 대응하는 자식노드들을 만들고 이들 노드에 부모노드의 데이터를 분할하도록 한다. 이러한 과정을 미리 지정된 조건이 만족될 때 까지 반복하여 의사결정트리를 구성한다. 데이터의 이질도를 측정하는 전형적인 방법으로 엔트로피(entropy)가 사용된다. 어떤 데이터 집합 S에 두 개의 집단 A, B가 있고, 집단의 비율이 각각 p_A와 p_B $(1-p_A)$라고 할 때, 이 집합 S에 대한 엔트로피 $E(S)$는 다음과 같이 계산된다.

$$E(S) = -p_A \log p_A - p_B \log p_B \quad (6)$$

어떤 집합 S가 임의의 속성 a의 값 $\{a_1, a_2, \cdots, a_n\}$에 따라 $S_1, S_2, ..., S_n$ 으로 분할되고, 각 집합에서의 A와 B의 비율이 p_{Ai}, p_{Bi}이라고 하자. 이때 속성 a를 이용한 분할에서 이질도 감소에 대한 척도인 정보이득(information gain) $I(S,a)$은 다음과 같이 정의된다.

$$I(S,a) = E(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|}(-p_{Ai}\log p_{Ai} - p_{Bi}\log p_{Bi}) \quad (7)$$

유전자발현 데이터에서 각 샘플이 하나의 데이터에 해당하고, 유전자 각각은 속성에 대응한다. 샘플에는 목표집단 또는 비교집단으로 분류 정보가 있다고 전제한다.

제안한 방법에서는 후보 마커를 기술할 때, 속성의 값을 3개의 기호값 즉, D, N, U로 나타낼 뿐만 아니라 이 값들의 부정 즉, $\neg D, \neg N, \neg U$도 허용한다. 이들 값에 따라 데이터를 분할하게 되면, 하나의 데이터가 여러 개의 분할에 포함되게 되는 경우가 발생한다. 예를 들어 어떤 유전자의 값이 L이라면, L에 해당하는 분할 뿐만아니라, $\neg N$과 $\neg U$의 분할에도 포함되게 된다. 하나의 노드가 확장될 때 6개의 자식 노드가 만들어질 수 있기 때문에, 트리 구조가 지나치게 커질 수 있다. 이러한 문제에 대응하기 위해 샘플을 포함하고 있는 동기(sibling) 노드들 중에서 다른 노드에 포함되는 노드는 가치치기를 한다. 예를 들면, D 또는 N에 대응하는 두 노드는 $\neg U$에 해당한다. 그러므로 D와 N에 대응하는 두 노드를 가지치기해서 제거한다. 이러한 가지치기 조건들을 기술하면 다음과 같다.

$$\begin{array}{ll} D \vee N \equiv \neg U & \rightarrow D\text{와 } N\text{노드 삭제} \\ N \vee U \equiv \neg D & \rightarrow N\text{와 } U\text{노드 삭제} \\ D \vee U \equiv \neg N & \rightarrow D\text{와 } U\text{노드 삭제} \end{array} \quad (8)$$

Algorithm: 후보 마커 식별

Input: 유전자발현 데이터 E
발현값의 기호변환 소속함수 μ_D, μ_N, μ_U
최소 소속정도 θ_m
노드 확장 종료 조건 T

Output: 후보마커 표현식

begin

1. 주어진 유전자발현 데이터 E에서 대해서, 소속함수 μ_D, μ_N, μ_U를 사용하여 식 (2)를 통해 식 (4)에 보인 바와 같이 발현정도를 기호로 변환한 데이터 L을 생성한다.
2. 전체 샘플을 포함한 루트 노드를 생성하고, 루트 노드의 샘플집합에 대한 엔트로피를 식 (6)을 사용하여 계산한다.
3. 각 유전자 g_i에 대해서 가장 깊은 레벨(level)의 단말 노드(들)의 샘플집합을 $g_i = D$ 또는 $g_i = N$ 또는 $g_i = U$ 또는 $g_i = \neg D$ 또는 $g_i = \neg N$ 또는 $g_i = \neg U$에 따라 분할하는 경우의 정보이득을 식 (7)을 사용하여 계산한다.
4. 정보이득을 가장 크게 하는 유전자를 선택하여, 이를 이용하여 샘플들을 분할하여 단말노드들의 자식노드를 만들어 분할된 샘플들을 저장한다.
5. 식 (8)의 조건들을 이용하여 생성된 노드들에 대해서 가지치기를 한다.
6. 종료조건 T를 만족하면 단계 7로 가고, 그렇지 않으면 단계 3으로 간다.
7. 각 단말노드 T_i에 대해서 과반을 차지하는 샘플의 집단을 해당 노드의 대표 집단 값 C_i으로 지정하고, 해당 집단의 노드에 있는 샘플들에 대한 비율을 정

확도 a_i로 설정하고, 샘플개수의 전체 샘플에 대한 개수를 가중치 w_i로 한다.

end.

종료조건 T는 노드 확장시에는 해당 노드에 포함된 샘플의 개수가 미리 지정된 임계값이 이내이거나, 노드에 엔트로피가 지정된 임계값이 이하일 때 해당 노드의 확장을 종료하고, 한편 선택된 유전자의 개수가 미리 지정된 개수에 도달하면 종료조건이 만족된 것으로 간주한다.

3.3 구축된 후보마커 트리를 이용한 추론

앞에 소개한 후보 마커 식별 방법을 통해서 구성된 후보마커 트리를 이용하여 새로운 샘플에 대한 분류는 다음과 같은 방법으로 한다. 구성된 트리에서는 부정기호를 허용하기 때문에, 하나의 노드에서 여러 개의 자식노드로 샘플들이 중복해서 분할될 수 있다. 새로운 샘플의 분류를 위해서 루트 노드에서 시작을 하여, 노드에 대응된 유전자의 값에 따라 샘플을 복제하여 자식노드로 전달한다. 이와 같은 과정을 단말노드를 만날 때까지 반복한다. 이 경우 평가되는 샘플 데이터가 여러 개의 단말노드에 도달할 수 있고, 이들 단말노드에 부여된 분류 정보 C_i가 일치하지 않을 수도 있다. 주어진 샘플 d에 대해 도달한 단말노드들 $T_i\ (i=1,\cdots,k)$이 주어지면 이에 대한 집단 $c(d)$은 아래와 같이 결정된다. 단말노드의 샘플 분류의 분포가 다르고, 단말노드에 도달하는 데이터 개수가 많을 때 분류의 정확도가 높다고 할 수 있다. 이를 위해서 제안한 방법에서는 단말노드의 정확도 p_i와 샘플의 개수 정보 w_i를 함께 이용한다. 정확도 p_i는 노드 T_i에서 비율이 높은 집단의 비율이며, 가중치 w_i는 전체 단말노드들에 도달한 샘플의 개수에 대한 T_i에서의 노드의 비로 설정한다.

$$\kappa_1 = \frac{\sum_{C_i=1} w_i p_i}{\sum_{i=1}^{k} w_i} \qquad \kappa_2 = \frac{\sum_{C_i=2} w_i p_i}{\sum_{i=1}^{k} w_i} \qquad (8)$$

$$c(d) = \begin{cases} C_1 & \text{if } \kappa_1 \geq k_2 \\ C_2 & otherwise \end{cases} \qquad (9)$$

식 (8)의 κ_1, κ_2는 각각 샘플에 대한 가중치를 고려한 부류의 선호정도를 나타내고, 식 (9)는 이들 선호정도를 기준으로 샘플의 부류를 결정하는 것이다.

3.4 의사결정트리와 제안된 방법과의 비교

제안된 방법은 결정트리 생성 방법에 기반하여 후보마커에 대한 트리를 생성한다. 후보마커 트리는 기존의 의사결정트리 생성방법과 다음 측면에서 차이가 있다. 첫째, 기존 의사결정트리는 속성에 나타나는 값 자체나 값을 이산화시킨 것을 노드를 확장할 때 사용한다. 반면 제안된 방법은 수치적인 속성값을 D,N,U 세 가지 이산값으로 변환한 후에 이들 값에 대한 부정(negation)을 사용할 수 있도록 한다. 둘째, 의사결정트리는 매 노드를 확장할 때 현재 노드의 데이터를 최적으로 분할하는 속성을 선택한다. 반면, 제안된 방법은 만들어지는 트리의 레벨(level) 별로 동일한 속성 즉 유전자를 선택하도록 한다. 이러한 제약은 최소한의 유전자들로 후보마커를 생성하기 위한 것이다. 셋째, 제안된 방법은 노드 확장시에 트리 구조를 단순화시키기 위한 가지치기 과정을 수행한다.

4. 구현 및 적용

전형적인 유전자발현 데이터인 마이크로어레이 데이터에 대해서 적용하여 적용가능성을 확인하기 위하여 제안한 방법을 구현하였다. 실험 데이터는 평균값이 0이 되도록 정규화하고, 기호변환을 위해서 D,N,U에 대응하는 소속함수를 정의해서, 기호화된 형태로 원래 데이터를 변환하였다. 실험에서 방광암 환자의 마이크로어레이 데이터로 1000개의 유전자에 대한 235개의 샘플의 데이터를 사용하였으며, 샘플은 암세포 및 암주변 세포에 대한 정보가 부류정보를 주어진 것이다. 실험에서는 해당 데이터에 대해서 유전자의 개수를 2개부터 6개까지 확대하면서 제안된 방법에 따른 분류 정확도를 leave-one-out 교차검증(cross-validation)을 계산하였다.

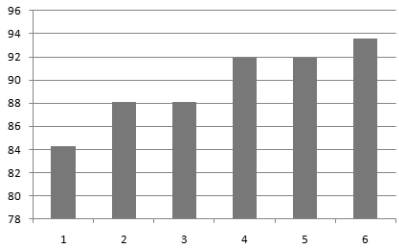

그림 3. 유전자 개수 증가에 따른 정확도 분포

Fig.3 Classification accuracy over the number of selected genes

그림 3은 선택되는 유전자 개수를 증가시키면서 분류 정확도를 측정한 결과를 보인 것이다. 동일한 데이터에 대해서 대표적인 의사결정트리 생성 알고리즘은 C4.5를 사용하는 경우, 6개 유전자를 사용하여 91.0%의 정확도를 보이는 규칙이 선택되었다. 제안된 방법이 C4.5의 경우보다 더 작은 개수의 유전자를 사용하여 보다 나은 결과를 낼 수도 있음을 실험을 통해서 확인하였다.

5. 결론

고성능 분자생물학적 측정 도구의 발달로 유전자 발현 데이터가 많이 생성되고 있으며, 유전자 발현 수준에서의 임상적인 유의성을 보이는 패턴을 추출하기 위한 다양한 데이터 분석이 시도되고 있다. 이 논문에서는 유전자의 발현정도를 세 단계의 기호값으로 변환한 다음, 이러한 정성적인 값을 제약조건으로 하는 비교되는 두 샘플집단을 식별할 수 있는 후보마커를 찾는 의사결정트리 기반의 방법을 제안하였다. 효과적인 표현을 위해서 세 가지 기호값에 대한 부정값을 허용하도록 하고, 부정값 허용에 따른 트리의 기하급수적 확대를 막기 위한 가지치기를 도입하였다. 최소의 유전자 집합만을 사용할 수 있도록 하기 위해 트리의 각 레벨에서 동일한 유전자를 선택하도록

트리가 구축되었다. 루트 노드에서 단말노드까지의 각 경로가 하나의 규칙에 대응하는데, 각 규칙의 정확도 및 가중치를 해당 단말 노드의 학습 샘플에 대한 대표 부류의 비율과 전체 학습데이터 대비 비율로 설정하여, 분류 정확도를 결정하도록 하였다. 실험 데이터에 대해서 적용가능성을 확인하였다. 분류정확도는 수치값을 세 개의 기호값으로 변환하는 소속함수에 영향을 많이 받았는 것을 실험적으로 확인하였다. 따라서 정확도 향상의 관점에서 바람직한 소속함수를 효과적으로 찾는 방법에 대한 추가적인 연구가 필요하다. 제안한 방법은 기호된 후보마커 규칙을 제공하기 때문에, 의미있는 패턴에 대한 직관적인 정보를 제공할 수 있게 하여, 추가적인 분석을 위한 실험 설계를 하는데 도움을 줄 수 있다.

참 고 문 헌

[1] G. McDougall, Personalized Medicine : The Time to Act and Collaborate Is Now, *Breakthroughs*, 7-9, Apr. 2010.

[2] E. Dalmasso, Planning for Success in Biomarker Discovery, *GEN*, Vol.28, No.12, Jun. 2008.

[3] R. Rosell, E. Felip, M. Taron, et al., Gene Expression as a Predictive Marker of Outcome in Stage IIB-IIIA-IIIB Non-Small Cell Lung Cancer After Induction Gemcitabine-Based Chemotherapy Followed By Resectional Surgery, *Clinical Cancer Research*, Vol.10, No.12, 2004.

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

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

[6] V. G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation reponse, P*roc. of Nat. Acad.Sci.*, Vol.98, No.9, 5116-5121, Apr. 2001.

[7] 강호일, 김야석 역, *DNA:마이크로어레이 데이터 해석*, 월드사이언스, 2005.

[8] K. M. Lee, K. S. Hwang, C. H. Lee, Fuzzy Set-based Microarray Data Analysis Techniques for Interesting Block Identification, *Proc. of FUZZ- IEEE2009*, 2009.

[9] M. L. Metzker, Sequencing Technologies - the next generation, *Nature Review Genetics,* Vol.11, pp.31-46, 2010.

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