마이크로어레이자료분석에서의 최신 분류방법들의 비교연구

(Comparison of recently developed classification tools in microarray data analysis)

이재원[1] · 이정복[2] · 박미라[3]

요 약

cDNA 마이크로어레이자료를 이용한 분류방법은 수많은 유전자의 발현을 동시에 모니터링 할 수 있으므로 특정 질병간의 분자생물학적 변이를 이해하는데 있어 기존의 분류방법보다 신뢰성이 훨씬 높을 것으로 기대되고 있다. 최근에 Dudoit et al.(2001)은 cDNA 마이크로어레이를 이용한 유전자발현자료의 분석에 있어 분류를 위한 여러 고전적인 판별분류기법 및 최근에 개발된 기법들을 비교, 평가하였다. 본 논문에서는 Dudoit et al.(2001)에서 다루지 않았던 많은 최신 기법들을 포함하여 인간의 종양 자료뿐만이 아니라 농작물을 포함한 동식물 자료에 적용하여 보다 폭넓은 비교연구를 하였다.

1. 서론

인간 게놈 프로젝트(Human Genome Project)를 통해 인간 유전체의 분석이 완료되면서 방대한 분량의 유전정보가 사용 가능하게 되었다. 이러한 연구성과에 뒤이어 선진 각국에서는 최근 기능유전체학(functional genomics)에 관심을 집중하고 이에 대한 연구에 막대한 인적, 물적 투자를 시행하고 있다. 이러한 방대한 정보를 짧은 시간에 대량으로 처리할 수 있는 분석법으로서 cDNA microarray 기술이 개발되었다. cDNA microarray 기술은 수천 개의 유전자에 대한 발현양상을 동시에 관찰할 수 있도록 하는 방법으로, 유전자의 규칙(regulation)과 상호작용의 이해를 돕는데 큰 기여를 할 뿐 아니라, 정상인 세포와 질병세포를 비교함으로써 질병유전자(disease gene)를 식별하고 치료약을 개발하기 위해 사용될 수 있는 방법으로 크게 각광받고 있다(Chen et al., 1997; DeRisi et al., 1996; Khan, 1999; Scherf et al., 2000). 예컨대 현재 형태학적, 임상적, 분자생물학적 변수 등을 고려하여 악성종양을 분류하고 있으나 아직까지 진단의 불확실성이 남아있을 뿐 아니라, 분류된 것들이 실제로는 이질적이거나 서로 다른 임상과정을 거치는 질병들로 이루어진 경우도 많다. 그러나 cDNA microarray를 이용하면 수많은 유전자를 동시에 관찰할 수 있게 되어 질병간의 변이에 대한 보다 완벽한 이해를 끌어낼 수 있으므로 보다 정제되고 신뢰성 있는 분류를 가능하게 한다(Eisen et al., 1998).

유전자발현자료가 많아짐에 따라 microarray 실험에서 생성된 이미지의 분석 및 관측된 유전자 발현수준의 변이성 등에서부터 생화학적 경로의 해석에 이르기까지, 자료처리 및 분석에 걸쳐 각 과정에서 많은 통계적인 과제가 발생하고 있다. 특히 최근 생명체에서 어떤 질병의 발생 또는 진행은 특정 유전자의 이상으로 이해되는 것이 아니라 서로 연관된 여러 유전자들간의 상호관계 하에서 발현될 것이라는 예상이 일반적인 중론이므로, 유전자발현자료 특유의 복

1) 고려대학교 통계학과 교수, (136-701) 서울특별시 성북구 안암동 5가 1번지
2) 고려대학교 통계학과 박사과정, (136-701) 서울특별시 성북구 안암동 5가 1번지
3) 을지의과대학교 의예과 조교수, (301-112) 대전직할시 중구 용두 2동 143-5

잡한 상호작용를 고려한 질병의 분류는 연구할 가치가 매우 높다. 유전자발현자료를 이용한 암 분류에 있어서 근래의 논문들은 대부분 종양표본과 유전자에 대한 군집분석(cluster analysis)에 촛점을 맞추고 있으며, 여기에는 계층적 군집분석(hierarchical clustering methods)의 응용이나 self-organizing map 같은 분리기법 등이 포함되어 있다(Tibshirani et al. 1999; Tamayo et. al., 1999; Toronen et. al., 1999). 판별분석에 대해서는 최근 Golub et al.(1999)이 선형판별분석과 유사한 가중유전자투표방식(weighted gene voting)을 유전자발현자료에 적용한 바 있다. 본 연구에서 우리가 관심있는 분야는 이 가운데에서 판별 및 분류방법에 관한 것으로 군집분석에 비해 상대적으로 기존의 연구결과가 부족한 부분이다.

전통적인 판별분류기법으로서 피셔의 선형판별 분석(Fisher linear discriminant analysis), 최대우도판별분석(maximum likelihood discriminant analysis), 근접 분류자(nearest neighbor)등의 방법이 있으며, 최근에는 분류나무(classification tree)나 신경망(neural network)과 SVM(support vector machine)등을 이용한 분류기법도 선보이고 있고 현재까지 여러 연구자들이 이들을 이용한 분석기법을 유전자발현자료에 적용하여 왔다(Golub et al., 1999; Tamayo et al., 1999; Ross et al., 2000). 그러나 유전자발현자료를 이용하여 종양군들을 성공적으로 구분할 수 있는 능력이 암의 분류에 있어서 매우 중요한 부분임에도 불구하고 지금까지는 대부분의 종양분류에 대한 연구가 하나의 유전자발현자료세트에 대하여 하나의 방법을 이용해서만 이루어져왔다. 그러나 종합적인 비교연구가 없이는 각 방법의 장단점을 파악하기 어렵게 된다. 또한 해당되는 연구 자료에 적절하지 못한 방법이 남용되는 것을 막기 위해서라도 각 분류기법들이 신뢰성 있고 일치성이 있는 결과를 제공하고 있는지에 대한 연구가 필요하다.

최근에 Dudoit et al. (2000)은 cDNA microarray를 이용한 유전자발현자료의 분석에 있어 분류를 위한 여러 고전적인 판별분류기법 및 최신 기법들을 비교, 평가하였으나, 적용한 자료가 모두 인간의 암 자료이므로 충분하고 공정한 비교가 이루어졌다고 하기에 부족하다는 판단이 든다. 따라서 본 논문에서는 Dudoit et al.(2000)에서 다루지 않았던 많은 최신 기법들을 포함하여 인간의 종양 자료뿐만이 아니라 농작물을 포함한 동식물 자료에 적용하여 보다 폭넓은 비교를 해 보았다. 자료형태별로 여러 가지의 오분류율을 사용하여 비교를 실시하고 모의실험을 통해 민감도와 특이도를 구하였다. 이러한 연구의 결과로서 각 분야의 연구자들에게 해당되는 자료분석에 가장 적절한 분류방법을 선택할 수 있는 가이드라인을 제공하여 보다 신뢰성 있는 연구결과를 도출할 수 있도록 유도할 것이다.

2. 유전자발현자료의 분류방법

cDNA microarray 자료를 이용한 분류방법은 수많은 유전자의 발현을 동시에 모니터링 할 수 있으므로 특정 질병간의 분자생물학적 변이를 이해하는데 있어 기존의 분류방법보다 신뢰성이 훨씬 높을 것으로 기대되고 있다. 하지만 cDNA microarray 자료를 이용해서 질병을 분류하고자 하는 연구는 최근에 몇몇 연구자들이 관심을 가지고 몇 가지 기존의 군집/판별분석 방법들을 시도해보거나 유전자발현자료에 맞게 개량된 방법을 제안하기 시작하고 있는 수준이다. 몇 가지 방법을 살펴보기로 한다.

유전자발현자료의 분석을 위해서 n개의 mRNA 표본에서 p개의 유전자를 측정하였을 때,

자료는 $n \times p$ 행렬 $$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

로 표현할 수 있다. 여기서 x_{ij}는 i-번째 mRNA 표본에서 얻은 j-번째 유전자의 발현 수준으로 추정된 형광강도를 나타낸다. 각 표본에서 얻은 자료(예측변수)를 $\boldsymbol{x}_i=(x_{i1},\cdots,x_{ip})$라고 표현하고, 이 자료가 포함되어 있는 분류집단(반응변수)을 y_i($y_i=1,\cdots,K$)라고 하면 분류자(classifier)는 유전자 발현 자료를 K 개의 서로 배반인 부분집합 $A_1,\cdots,A_K$로 나누게 될 것이고, 결과적으로 예측변수 $\boldsymbol{x}_i$를 통하여 반응변수 y_i를 예측하는 문제가 될 것이다.

피셔의 선형판별분석(FLDA)과 대각선형판별분석(DLDA)은 유전자 발현 자료에서 그룹간 제곱합과 그룹내 제곱합의 비 $\boldsymbol{a}'B\boldsymbol{a}/\boldsymbol{a}'W\boldsymbol{a}$를 이용하여 선형결합식 $\boldsymbol{xa}$를 찾는 과정을 거치게 된다. 근접 분류자(nearest neighbor classifiers)를 이용한 방법은 두 mRNA 표본의 유전자 발현 관측치쌍에 대해서 유클리드 거리 또는 1-상관계수를 기초하여 구해진다. 분류집단이 두 개인 경우 두 확률밀도함수의 비에 대한 로그값을 판별변수벡터와 회귀계수의 선형식으로 모형화하는 로지스틱판별분석은 개별적인 확률밀도 함수를 구체화할 필요가 없는 장점이 있다.

분류나무(Classification Trees)를 이용한 방법에 대해 평가하기 위해서는 이산형 나무구조의 분류자를 이용한다. 본 연구에서는 CART를 포함한 다양한 나무 구조의 분류자와 10-fold cross-validation방법을 적용하여 보았다. 예측의 정확도를 높이기 위해서 perturbed learning set에서 만들어진 예측자(predictor)를 구하여 통합하는 방법을 사용할 수 있다(Breiman, 1996). perturbed learning set을 만드는 방법에는 bagging(Breiman, 1996), boosting (Freund and Schapire, 1996) 과 arching(Breiman, 1996)의 세 가지가 있다. Bagging방법은 중복을 허락하여 랜덤하게 동일한 표본크기의 perturbed learning set을 만들고 중복투표(plurality vote)를 통하여 예측자를 구하고, boosting 방법은 resampling을 할 때 자주 오분류되는 자료에 더 많은 가중치를 주도록 bagging방법을 확장한 것으로 각 예측자는 가중투표(weighted voting)를 통하여 통합된다. arching 방법은 bagging에 비하여 반복 계산이 많이 요구되고 있으나 예측 오류를 줄여주는 장점을 가지고 있다.

유전자발현자료를 이용해서 암을 분류하고자 하는 연구에 self-organizing map (SOM) 방법이 많이 적용되고 있는데 이 방법은 계산이 빠르고 쉬워서 대량의 자료를 분석하는데 적합하며, 반복적인 계산과정에서 기저 클러스터 패턴을 탐색하고, 최적의 구분으로 수렴하는 장점이 있고 시각적인 분석을 용이하게 하는 장점을 가지고 있다 (Tamayo et. al., 1999; Toronen et. al., 1999; Golub et al., 1999). 또한 최근에는 신경망(Neural Network)과 Support Vector Machine(SVM)을 이용한 방법을 이용한 유전자 발현자료의 군집 및 판별분석이 제안되었다 (Brown, 2000). $\boldsymbol{x}_i$ 가 형광강도의 추정치이고 y가 반응변수일 때, 하나 이상의 hidden layer를 가진 신경망의 함수식은

$$y_k=f_0\Big(\sum_{i\to k}\boldsymbol{w}_{ik}\boldsymbol{x}_i+\sum_{j\to k}\boldsymbol{w}_{jk}\boldsymbol{x}_i f_h\Big(\sum_{i\to j}\boldsymbol{w}_{ij}\boldsymbol{x}_i\Big)\Big)$$

로 표현할 수 있다. 여기서 함수 f_0는 선형, 로지스틱, threshold등 다양한 연결함수(link function)를 가질 수 있다. training 자료인 $\boldsymbol{t}$ 에 대하여 가장 접근하기 쉬운 알고리즘으로 최소제곱법을 적용하면, $E(\boldsymbol{w})=\sum\|\boldsymbol{t}-f(\boldsymbol{x};\boldsymbol{w})\|^2$ 을 최소화하는 모수벡터 $\boldsymbol{w}$ 를 찾을 수 있다. 일반적으로 unsupervised learning의 형태를 지닌 방법들은 발현 패턴의 유사성을 정의하여 이를 기초로 자료를 구별하게 되나, Support Vector Machine(SVM)과 같은 supervised

learning 방법은 training set을 이용하여 어느 자료가 같이 분류되어야하는가를 미리 지정한다. SVM방법은 계층적 방법이나 self-organizing map(SOM) 방법보다 높은 차원에서 거리 함수를 적용할 수 있고, supervised learning 방법의 특징인 사전 정보를 이용할 수 있다는 장점이 있다. 유전자 발현 자료에 이용할 경우, SVM은 리보솜 단백질(ribosomal protein)과 같은 공통적인 기능을 하는 유전자 집합과 기능적으로 역할을 하지 않은 유전자의 집합을 구분한 training set을 이용하여 이 기준에 따라 새로운 유전자를 분류하게 된다. i-번째 실험($i=1,\ldots,m$) 에서 각 유전자의 발현 측정치를

$$X_i = \frac{\log(E_i/R_i)}{\sqrt{\sum_{j=1}^{m} \log^2(E_j/R_j)}}$$

를 표현할 수 있으며 (여기서 E_i, R_i 는 각각 i-번째 실험조건과 표준조건(reference state)에서의 유전자 발현 수준이다), 따라서 이들의 벡터 $\underline{X}=(X_1,\cdots,X_m)$ 은 m차원 발현공간에서의 한 점이 되고, 유클리드의 거리는 1이 된다. 벡터 $\underline{X}$ 를 두 그룹으로 분류하는 가장 간단한 방법은 구분하는 평면(separating hyperplane)을 구하는 것이나 실제로는 그 평면으로 구분할 수 없는 경우가 대부분이며, 이에 대한 해결책으로 더 높은 차원의 공간 (feature space로 정의함)으로 자료를 mapping해서 그 공간에서 평면을 구하는데 이러한 과정에서 커널 함수(kernel function)을 이용하면 복잡한 계산의 부담을 덜 수가 있다.

3. 분류방법의 비교를 위한 모의실험

본 연구에서는 이러한 여러 가지 고전적인 분류방법 및 최신의 방법을 모두 고려하여 cDNA microarray를 이용한 유전자발현자료의 분석에 있어 분류방법을 모의실험을 통하여 비교, 평가하였다. 고전적인 판별분석방법으로는 Dudoit et al.(2000)의 비교에서 우수한 결과를 보여준 피셔의 선형판별분석(FLDA)과 대각선형판별분석(DLDA)를 중심으로 대상방법을 선택하였으며 그 외에 통합분류자(aggregating classifiers)를 이용하는 세 가지 방법(boosting, bagging, arching)을 고려하였다. 여기에 self-organizing map (SOM) 방법과 support vector machine (SVM) 방법을 추가하였으며, 최신에 개발된 신경망(neural network)을 이용한 방법중 feed forward single layer방법과 multi layer방법 그리고 back propagation 방법을 적용하였다. 이외에 Dudoit et al(2000)에서 적용한 방법을 일부 보완수정한 유전자 알고리즘을 이용한 k-nearest neighbor방법(Li et al., 2001)과 partial least squares를 이용한 선형판별분석도 포함하였다.

각 방법을 적용함에 있어서 여러 가지 형태의 오류율(test set error rate, observation-wise error rate, individual misclassification rate)을 비교할 것이며, 이와 더불어 모의 실험을 통해 민감도(sensitivity)와 특이도(specificity)를 구해 볼 것이다. 위의 방법들의 공정한 비교를 위해서 모든 방법을 형태가 다른 각각의 자료에 적용할 것이다. 연구자료로는 Dudoit et al. (2000) 에서 사용한 세가지 암 자료인 임파구 자료, 백혈병 자료와 NCI 60 cell 자료외에, 보다 다양한 형태의 자료를 대상으로 폭 넓은 비교를 하기 위해 동식물 자료(Zebra Fish, Arabidopsis thaliana)와 농작물 자료(Soybean)등을 추가하였다.

공정한 비교를 위해서는 위의 세 가지 자료를 포함하여 여러 형태의 자료를 통일시켜야 하며, 이를 위해서 먼저 관측치(array)를 각 유전자(변수)에 따라 평균이 0 이고 분산이 1 이

되도록 표준화(normalization)하였으며, 유전자의 수(p)가 너무 많으면 유전자발현 수준을 구분하기 어렵기 때문에 Dudoit et al.(2000)에서와 같이 각 자료로부터 가장 관심거리가 될 만한 유전자를 약 50개 정도로만 선택하였다. 이러한 실제자료의 기초분석결과를 통해서 모의실험의 구도를 잡을 것이며 모의 실험을 통해서 각 방법의 민감도(sensitivity)와 특이도(specificity)를 비교함으로써 연구형태에 가장 적합한 가이드라인을 제공하였다.

4. 결론 및 토의

유전자발현자료를 분류함에 있어서 반드시 검증해야 할 부분 중 하나는 측정된 유전자발현의 변동과 불확실성의 효과를 고려한 신뢰성 계산이다. microarray 자료분석의 일치성(consistency)과 유의성(significancy)을 확보하기 위해서는 여러 분류방법의 비교연구가 필요할 것이다. 하지만 종양분류에 관한 대부분의 연구가 단지 하나의 유전자발현자료세트에 한 가지 분류방법을 적용시켜서 적용된 방법의 강점만을 부각시키고 있으므로, 한 자료에 근거해서 제안된 방법을 다른 자료에 널리 적용하는 것은 많은 무리가 있다. 따라서 본 연구에서와 같이 여러 가지 방법을 여러 가지 자료세트에 동시에 적용시키고 또한 모의실험을 통해서 서로 공정한 비교를 해보는 과정이 반드시 필요하다. 또한 support vector machine 방법이나 신경망을 이용한 방법등과 같이 다변량자료의 분석을 위해 최근에 개발된 분류기법들을 유전자발현자료에 적용할 수 있도록 개량하여 분석적용범위를 확대하는 것이 바람직하다.

유전체 연구에 있어서 선도적 역할을 하고 있는 미국의 경우에도 생물과 통계적 지식을 모두 갖춘 이른바 생물정보학 연구자의 수가 많지는 않다. 그러나 대학 및 여러 연구기관에 생물통계학자들이 다수 포진해있고 각 기관에서의 풍부한 지원 및 관련분야와의 물적, 인적 교류를 바탕으로 유전체 연구분야에 관한 통계적 방법의 연구가 활발히 진행되고 있다. 국내에서는 유전체 연구에서의 통계적 방법에 관한 경험과 조직이 거의 없고 연구 여건 또한 많은 차이가 있어 기술격차가 상당히 있는 실정이다.

본 논문은 기능유전체 연구자들에게 각자의 실험상황이나 자료성격에 적합한 분류방법을 제시함으로써 부적절한 방법의 적용으로 인한 연구결과의 오류를 최소화시키는데 기여하고자 하였다. 특히 연구의 결과로서 사용자를 위한 가이드라인을 제공함으로써 각 분야의 연구자들이 쉽게 접근하여 활용할 수 있게 되었다. 뿐만 아니라 유전체 자료의 분석을 연구하는 수학, 전산학 및 통계학 전문가에게는 기존의 분류 방법들이 자료형태나 연구설계등에 따라 어떠한 장단점이 있는가를 정리, 요약하는 계기가 될 수 있고 그들로 하여금 새로운 분류기법을 개발하는 데에 필요한 유용한 정보와 모티브를 제공할 수 있을 것이다.

참고문헌

Breiman, L., Greidman, JH., Plshen, R., and Stone, CJ.(1984). Classification and regression trees. The Wadsworth statistics/probability series. Wadsworth International Group.

Brown, MP., Grundy, WN., Lin, D., Cristianini, N., Sugnet, CW., Durey, TS., Ares Jr, M., and Haussler, D.(2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS. 97-1;262-267.

Chen, Y., Dougherty, ER., Bittner, ML.(1997). Ratio-based decisions and the quantitative analysis of cDNA Microarray Images. Journal of Biomedical Optics. 2(4);

364-374.

DeRisi, J, Penland, L., Brown, PO, Bittner, ML, Meltzer, PS, Ray, M., Chen, Y., Su, YA, Trent, JM. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nature Genetics. 14; 457-460.

Dudoit, S., Fridlyand, J. Speed, TP.(2001). Comparison of Discrimination Mehtods for the Classification of Tumors Using Gene Expression Data. TR #576. UC Berkeley.

Eisen, MB., Spellman, PT., Brown, PO., and Botstein, D.(1998). Clustering analysis of display of genome-wide expression patterns. PNAS. 95; 14863-14868.

Freund, Y. and Schapire, RE.(1996). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 55;119-139.

Golub, TR., Slonim, DK., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, JP., Coller, H., Loh, ML., Downing, JR., Caligiuri, MA., Bloomfiled, CD., and Lander, ES.(1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Sclience. 286-5439; 531-537.

Khan, J., Saal, LH., Bittner, ML., Chen, Y., Trent, JM., Meltzer, PS.(1998). Expression profiling in cancer using cDNA microarrays. Electrophoresis. 20; 223-229.

Li, L., Weinberg, CR., Darden, TA., and Pedersen, LG.(2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics. 17; 1131-1142.

Nguyen, DV. and 깨찰, DM.(2002). Tumor classification by partial least squares using microarray gene expression data. Bioinformatics. 18; 39-50.

Ross, DT., Scherf, U., Eisen, MB., Perou, CM., Rees, C., Spellman, P., Iyer, V., Jeffrey, SS., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, JC., Lashkari, D., Shalon, D., Myers, TG., Weinstein, JN., Botstein, D., and Brown, PO.(2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet. Mar24-3; 227-235.

Scherf, U., Ross, DT., Waltham, M., Smith, LH., Lee, JK., Kohn, KW., Reinhold, WC., Myers, TG., Andrews, DT., Scudiero, DA., Eisen, MB., Sausvillie, EA., Pommier, Y., Botstein, D., Brown, PO., and Weinstein, JN.(2000). A cDNA microarray gene expression database for the molecular pharmacology of cancer. Nat. Genet. 24-3, 236-244.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, ES., and Golub, TR.(1999). Interpreting patterns of gene expression with self-organzing maps: methods and application to hematopoietic differentiation. PNAS. 96; 2907-2912.

Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., and Brown, P.(1999). Clustering methods for the analysis of dna microarray data. Technical report, Department of Health Research and Policy, Stanford Univ.

Toronen, P., Kolehmainen, M., Wong, G., and Castren, E.(1999). Anaysis of gene expression data using self-organizing maps. FEBS letters. 451; 142-146.