다변량 통계 기법을 이용한 DNA microarray 자료 분석

권 성우, 한 종훈
포항 공과 대학교 화학 공학과

DNA microarray data analysis using multivariate statistical technique

Sungwoo Kwon,Chonghun Han
Dept. of Chemical Engineering, Postech

Introducton

암은 종류에 따라서 치료법이 크게 달라지며 사용하는 치료 약물이 다르기 때문에 환자가 어떤 암에 걸린 지를 예측하는 것은 매우 중요하다. 최근까지 암 진단은 현미경을 통한 암세포의 형태, 모양, 염색 정도와 조직의 출처에 따라 진단했다. 그러나 이러한 형태학적 암 진단 방법은 정확하지 못한 단점이 있다 (Marx, J et al., 2000).

최근 개발된 DNA microarray 는 세포의 표현형과 유전자 발현 양상간의 관계를 연구할 수 있도록 해 준다. 이러한 생물학적 이론을 바탕으로 암 진단에 DNA microarray 를 적용하는 연구가 현재 활발히 수행되고 있다. DNA microarray 를 이용한 방법은 형태학적 암 진단 방법보다 더 정확히 진단할 수 있다 (Golub.T.R et al., 1999; Marx, J et al., 2000).

DNA microarray 를 이용한 암 진단 방법은 변수로 사용되는 유전자의 개수가 2000 개에서 250000 개 정도로 60 개에서 100 개 정도인 암 환자의 표본 개수 보다 많으며 유전자간의 상관 관계가 높은 특징을 가지고 있다. 유전자간의 상관 관계를 무시하고 각각의 유전자에 대해 독립적으로 암을 진단하게 되는 경우, 진단 결과 이상이 있는 환자를 감지 하지 못하므로 잘못된 진단을 하는 문제가 있다 (Leping et al., 2001). 더욱이 표본 개수가 적기 때문에 암 진단 모델은 항상 overfitting 의 문제를 갖고 있다. 이런 문제점들을 극복하기 위해 데이터 전 처리로서 현재 PCA(principal component analysis)를 많이 사용한다 (Richard, O et al., 2000).

PCA 는 서로 상관관계가 있는 변수들 사이의 복잡한 구조를 좀더 간편하고, 이해하기 쉽게 설명하기 위해 사용되는 다변량 분석 기법이다. 하지만 PCA 는 암 종류에 상관없이 유전자의 발현 양상을 가장 잘 설명하도록 주성분을 정하며 이상 표본(outlier 와 noise)에 대하여도 가장 잘 설명하도록 주성분을 정하기 때문에 강건성(robust)이 떨어지는 단점을 가지고 있다. 따라서 암 진단에 PCA 를 사용하면 정확한 분류가 어렵다 (Richard, O et al., 2000).

본 연구에서는 DNA microarray 를 이용한 암 진단에 주로 사용하는 전처리 방법인 PCA 의 문제점을 해결하는 방법으로, DNA microarray 데이터를 PCA 로 전 처리하고 주성분들 중에서 암 진단에 중요한 의미를 가지는 주성분을 SDA(stepwise discriminant analysis)를 통하여 선별한다 (Subhash et al., 1996). 이렇게 선별된 주성분의 주성분 점수 값(score)으로 암 분류 모델을 만들고 이를 검증하도록 하는 방법을 제안한다. 한편 제안하는 방법과 비교하기 위하여 DNA microarray 데이터를 PCA 로 전 처리하여 유전자의 발현 양상을 잘 설명하는 주성분을 가지고 암을 진단하는 방법과

결과를 비교 한다.

The proposed method

표본 개수가 변수 개수보다 적은 데이터에 통계 방법을 적용하는 것은 쉽지 않다. 더욱이 DNA microarray 데이터와 같이 변수로 사용되는 유전자의 개수가 표본 개수 보다 매우 많은 경우에는 분류 방법을 적용하기에 부적합하다. 따라서 변수 개수를 줄이는 방법이 반드시 필요하다. 현재 DNA microarray 데이터 분석 방법은 우선 변수 개수를 줄인 후 여러 가지 분류 방법을 사용하고 있다. 현재 변수 개수를 줄이는 방법에는 PCA 를 많이 사용하고, 분류하는 방법은 neural network 와 K-NN (K-nearest neighbor)를 사용한다 (Khan, J et al., 2001).

본 연구에서 제안하는 방법은 세 단계로 구성된다. 첫 번째 단계는 변수 개수를 표본 개수 이하로 줄이는 전처리 단계이며 이때 PCA 를 사용한다. 두 번째 단계는 첫 번째 단계를 거쳐 구한 주성분들 중 암 진단에 중요한 주성분만을 선별하기 위한 단계이며 사용하는 방법은 SDA 이다. 세 번째 단계는 암 진단 모델을 만들 때 많이 사용하는 분류 방법인 neural network 와 K-NN 을 사용한다. 또한 제안하는 방법과 이전 방법의 결과를 비교한다. 이를 위해서 첫 번째 단계와 세 번째 단계만 거치는 원래 방법을 적용하여 얻은 결과를 제안하는 방법과 비교한다.

Implementation and results

제안하는 방법과 원래 방법을 DNA microarray 데이터에 대해서 적용하였다. breast cancer 로서 Rosetta Impharmatics 의 Laura J. van’t Veer 가 97 개 lymph-node-negative Breast Cancer 표본을 25000 개 cDNA 가 고정화되어 있는 cDNA microarray 를 가지고 실험한 데이터이다 (http://www.rii.com/publications/2002/vantveer.htm). 데이터는 발표된 데이터로서 인터넷으로 이용 가능하다.

원래 데이터는 분석 표본이 78 개로 5 년내 전이(metastases)가 진행된 것은 34 개, 5 년 이내에 진행되지 않은 것은 44 개이다 (Laura J et al., 2002). 그러나 분석 표본 중 54 번이 결측치(missing value)가 많아 제외하여 총 77 개의 분석 표본을 사용했다. 검증 표본은 19 개로 5 년 이내에 전이가 진행된 것이 12 개이고 5 년 이내에 진행되지 않은 것이 9 개 이다. 데이터 값들 중 log ratio 을 사용했고 결측치의 경우 평균 값으로 보정했다. 그 외에 평균이 0 이고 표준 편차가 1 이 되도록 표준화시키는 것을 제외하고 다른 전 처리는 하지 않았다.

77 개 분석 표본에 대해 PCA 를 수행하고 구한 고유 벡터를 이용하여 19 개 검증 표본의 주성분 점수 값을 구했다. 분석 표본의 주성분 점수로 SDA 를 수행하고 암 전이 진단에 있어 중요한 주성분들을 순서대로 정열 했다. 분석 표본으로 주성분 개수를 증가시키면서 암 분류 모델에 사용되는 neural network 와 K-NN 을 각각 수행했다. 또한 모델 검증을 위해 leave-on-out 방식의 교차 타당성 방법을 분서 표본에 대해 수행하여 구했다. 사용되는 주성분 개수를 증가시키면서 분석 표본의 오분류율 값이 0.1%이하가 되거나 모든 값이 0.1% 보다 큰 경우에는 최소 오분류율일 때 모델을 최종 모델로 정했고 이것으로 검증 표본의 오분류율을 구했다.

제안한 방법과 원래 사용하는 방법을 비교하기 위해 77 개 분석 표본에 대해 PCA 만을 수행하고

고유값에 따라 정렬했다. 분석 표본으로 주성분 개수를 하나씩 증가시키면서 암 분류 모델에 사용되는 neural network 와 K-NN 을 각각 수행했다. 모델 검증을 위해 leave-on-out 방식의 교차 타당성 방법을 수행하여 분석 표본의 오분류율을 구했다. 사용되는 주성분 개수를 증가시키면서 분석 표본의 오분류율 값이 0.1%이하가 되거나 모든 오분류율 값이 0.1% 보다 큰 경우에는 최소 오분류율 일 때를 최종 모델로 정했고 이것으로 검증 표본의 오분류율을 구했다.

분류 모델로 첫 번째는 노드가 5 개인 시그모이드 뉴런의 단층으로 구성한 neural network 를 learning rate 는 0.01, momentum 은 0.71 로 적용했다. 원래 방법의 경우는 오분류율이 0.065%일 때 변수 개수를 16 개로, 제안 방법은 0.078%일 때 11 개로 결정했다. 최종 모델로 검증 표본을 예측하면 제안하는 방법의 오분류율은 0.11%이고 원래 방법은 0.53%이다. 분석 표본과 검증 표본의 오분류율을 비교하여 보면 제안하는 방법이 낮다. 또한 원래 방법의 경우 검증 표본의 오분류율이 분석 표본의 오분류율과 큰 차이를 보이며 overfitting 되었다. 두 번째로 K-NN 을 적용했을 때 제안 방법은 오분류율이 0.076%일 때 변수 개수를 29 개로, 원래 방법은 0.35%일 때 14 개로 결정했다. 최종 모델로 검증 표본을 예측하면 제안하는 방법의 오분류율이 0.09%이고 원래 방법은 0.47%이다.

분류 모델을 neural network 과 K-NN 으로 적용한 경우는 제안한 방법이 원래 방법보다 적은 수의 변수를 사용함에도 오분류율이 낮으며 정확한 암 진단을 한다. 더욱이 원래 방법으로 neural network 로 분석하면 overfitting 이 되었다.

Discussion and conclusion

DNA microarray 을 이용하여 암을 진단하는 새로운 방법을 제안하였다. 그 핵심은 원래 사용되고 있는 PCA 의 단점을 보완하기 위해서 SDA 를 같이 수행하는 것이다. 원래 많이 사용되는 분류 방법인 neural network 와 K-NN 에서 보다 정확히 암을 진단한다.

제안하는 방법은 한 가지 특정 암 데이터뿐만 아니라 여러 가지 종류의 암에도 정확한 진단을 한다. 그리고 분류하고자 하는 암 종류가 두 가지 일 때뿐만 아니라 여러 가지 종류로 분류해야 하는 경우도 제안하는 방법이 PCA 만으로 전처리 하는 것보다 정확한 진단을 한다. 더욱이 DNA microarray 데이터와 같이 표본 수가 적고 변수가 많은 경우는 분류 방법을 적용하기가 어렵다. 특히 분석 표본이 적기 때문에 neural network 을 비롯한 모든 분류 모델을 만드는데 overfitting 되기 쉽다. 즉, breast cancer 데이터의 경우 Khan et al.,이 제안한 방법을 사용하면 overfitting 이 되며, 높은 오분류율 값을 가진다. 하지만 본 연구에서 제안하는 방법은 적은 수의 중요한 주성분만을 가지고 암 진단 모델을 만들기 때문에 overfitting 가능성이 적으면서 검증 표본의 오분류율도 감소 시킨다. 더욱이 제안하는 방법은 Laura J et al.,이 만든 암 전이 진단 모델보다 더 정확하다. 결론적으로 제안하는 방법은 PCA 만 사용하는 방법과 달리 이상치에 대해 강건성이 높으면서 오분류율도 감소 시킨다.

원래까지 대부분의 DNA microarray 데이터 분석시 암 진단에 유의미한 유전자를 선택하는 방법은 단변량 접근 방식이었다. 이러한 방법은 DNA microarray 데이터의 특징인 유전자간의 강한 상관관계를 고려하지 못하는 단점이 있다. 단변량적인 접근 방식은 유전자들간의 강한 상관 관계를 무

시하고 각각의 유전자에 대해 독립적으로 암을 진단하게 된다. 따라서 분석 결과도 이상 있는 사람도 이상으로 감지되지 않는다. 따라서 상관관계가 강한 변수들이 있는 데이터를 분석할 때는 PCA 와 같은 다변량 통계 방법을 많이 사용하여 왔다. 본 연구에서 제안하는 방법은 다변량 통계 방법에 기초하여 암 진단에 유의미한 주성분을 선택함으로써 DNA microarray 데이터의 특징인 유전자들간의 상관관계 문제를 해결할 수 있다. 더욱이, 각각의 유전자들이 암 진단에 유의미한 주성분에 대해서 얼마만큼의 기여도 값을 가지는지를 구할 수 있고 이를 바탕으로 각 암에서 특이적으로 과다 혹은 억제 발현하는 유전자들을 찾을 수 있다 (sungwoo et al., 2002). DNA microarray 데이터를 주성분 분석과 SDA 를 통해 유의미한 주성분을 선택하고 각각의 유전자들이 암 진단에 중요한 주성분들에 대해 얼마나 기여를 하는지 구하는 방법은 현재 많이 사용하는 단변량 접근의 단점을 해결하는 다변량 방식이며 최근 개발된 유전자 알고리즘을 통한 다변량 접근 방법보다는 분석시간이나 파라미터 결정에서 더 쉽다 (Leping et al., 2001).

제안하는 방법 중 데이터 전처리 과정인 PCA 를 nonlinear PCA, ICA(independent component analysis) 나 tree harvesting 으로 대체할 수 있다. Nonlinear PCA, ICA 나 tree harvesting 방법의 경우 여러 인공변수들 중 어느 것이 암 진단에 유의미한 것인지를 정확히 알 수가 없는 단점이 있어 여러 인공변수들에 대해 SDA 를 수행한다면 암 진단에 중요한 인공 변수들을 선별할 수가 있을 것이며 이들을 사용한다면 보다 정확한 암 진단을 할 수 있으리라 예상된다. 더욱이 nonlinear PCA 와 SDA 를 같이 사용한다면 DNA microarray 데이터에서 암 진단에 중요한 nonlinear 성질을 찾아낼 수 있다.

References

Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101-10106.

Golub.T.R., Slonim.D.K., Tamayo.P., Huard.C., Gaasenbeek.M., Mesirov.J.P., Coller.H., Loh,M.L., Douning,J.R., Casligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science,* **286**, 531-537.

Hastie,T., Tibshirani,R., Eisen,M.B., Alizadeh,A., Levy,R., Staudt,L., Chan,W.C., Botstein,D. and Brown,P. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Ge*nome Biol*, **1**, research0003.1-research0003.21.

Khan, J., Jun, S., Wei, M., Lao, R and Saal, H. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine,* **7**(6), 673-679.

Leping, Li., Clarice, R., Weinberg., Thomas, A., Darden and Lee G. Pedersen. (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.

Marx, J. (2000) DNA Arrays Reveral Cancer in Its Many Forms. *Science*. **5485**. 1670-1675.

Richard, O., Peter, E., and David, G., (2000) *Pattern Classification*. John Wiley & Sons, Inc.

Subhash Sharma. (1996) *Applied multivariate techniques*, John Wiley & Sons, Inc.

Sungwoo Kwon, Chonghun Han, (2002) Multivariate identification of maker gene for cancer classification using DNA chip data, preprocessing.