다층퍼셉트론 기반 리 샘플링 방법 비교를 위한 마이크로어레이 분류 예측 에러 추정 시스템

박수영* · 정채영**

Classification Prediction Error Estimation System of Microarray for a Comparison of Resampling Methods Based on Multi-Layer Perceptron

Su-Young Park* · Chai-Yeoung Jung**

이 논문은 2009년도 조선대학교 학술연구비의 지원을 받아 연구되었음

요 약

게놈 연구에서 수천 개의 특징들은 비교적 작은 샘플들로부터 모아진다. 게놈 연구의 목적은 미래 관찰들의 결과를 예측하는 분류기를 만드는 것이다. 분류기를 만들기 위해서는 특징 선택, 모델 선택 그리고 예측 평가 등의 3단계 과정을 거친다. 본 논문은 예측 평가에 초점을 맞추고 모든 슬라이드의 사분위수를 똑같게 맞추는 quantile-normalization 적용하여 마이크로어레이 데이터를 표준화 한 후 특징 선택에 앞서 예측 모델의 '진짜' 예측 에러를 평가하기 위해 몇 개의 방법들을 비교하는 시스템을 고안하고 방법들의 예측 에러를 비교 분석 하였다. LOOCV는 전체적으로 작은 MSE와 bias를 나타내었고, 크기가 작은 샘플에서 split 방법과 2-fold CV는 매우 좋지 않는 결과를 보였다. 계산적으로 번거로운 분석에 대해서는 10-fold CV가 LOOCV보다 오히려 더 낮은 경향을 보였다.

ABSTRACT

In genomic studies, thousands of features are collected on relatively few samples. One of the goals of these studies is to build classifiers to predict the outcome of future observations. There are three inherent steps to build classifiers: a significant gene selection, model selection and prediction assessment. In the paper, with a focus on prediction assessment, we normalize microarray data with quantile-normalization methods that adjust quartile of all slide equally and then design a system comparing several methods to estimate 'true' prediction error of a prediction model in the presence of feature selection and compare and analyze a prediction error of them. LOOCV generally performs very well with small MSE and bias, the split sample method and 2-fold CV perform with small sample size very pooly. For computationally burdensome analyses, 10-fold CV may be preferable to LOOCV.

키워드

마이크로어레이, 리샘플링 방법, 다층 퍼셉트론

Key word

Microarray, re-sampling method(LOOCV, split, 2-, 10-fold CV), MLP

* 조선대학교 컴퓨터통계학과
** 교신저자

접수일자 : 2009. 08. 28
심사완료일자 : 2009. 09. 21

Ⅰ. 서 론

게놈 실험의 특징은 높은 차수의 데이터와 작은 샘플 크기 그리고 실험에서 발생하는 많은 잡음이다. 이러한 잡음들은 예측 정확도를 감소시킨다. 마이크로어레이 자료를 분석하는 초기 단계에서 잡음을 제거하는 과정을 거친다. 이런 과정을 표준화(normalization)라고 한다. 실험에서 관찰된 유전자들은 미리 정해진 클래스에 속한다고 알려져 있고 작업은 클래스가 알려져 있지 않은 새로운 유전자 데이터에 대해 예측자 또는 분류자를 만드는 것이다[1].

어떤 유전자가 예측에 포함될 지 결정하는 것을 특징 선택이라고 하며 클래스 예측자를 발달시키는데 중요한 단계이다.예측 정확도를 평가하기 위한 공통적인 접근은 원자료에 대해 몇몇의 리 샘플링방법을 실행하는 것이다. 예측 에러를 평가하기 위한 리 샘플링 기술에 대한 최근의 두 평가에서 유의한 유전자 선택은 리 샘플링 절차에 포함되지 않았고 그래서 높은 차수의 셋팅에는 부적절하다는 결론을 야기시켰다[2].

본 논문에서는 마이크로어레이 데이터를 사용하여 리 샘플링 방법들의 예측 에러를 다층 퍼셉트론 신경망(Multi-Layer Perceptron 이하 MLP)으로 평가하는 시스템을 제안하고 예측 에러 결과를 비교 평가 하였다. 논문의 구성은 다음과 같다. 2장에서 실험 방법을 소개하고, 3장에서 리 샘플링 방법과 다층 퍼셉트론 신경망을 설명한다. 4장에서 본 논문이 수행한 시스템 설계 및 구현과정을 설명하고 결과를 비교・분석한다. 5장 결론을 도출한다.

Ⅱ. 실험 방법

클래스를 예측하는 문제에서 알려지지 않은 분포P를 갖는 n개의 독립적이고 이상적으로 분포된 랜덤 변수 O1, ..., On은 관찰된다. 마이크로어레이 실험에서 X는 유전자 발현 측정을 포함한다. X는 또한 환자의 나이 또는 조직병리학의 측정과 같은 공 변수를 포함한다. 결과 Y는 병을 앓은 달수 같은 계속적인 측정이거나 병의 상태 같은 카테고리적인 측정이될 수 있다.

클래스 예측 목표는 Y를 예측하기 위해 X로부터의 정보를 실행하기 위한 규칙을 만드는 것이다. 그 의도는 이러한 규칙을 만듦으로써 관찰 O1, ..., On를 기반으로 미래에 관찰되지 않은 결과 Yo에 부합하는 관찰된 유전자 X0를 기반으로 결과 Yo가 예측 하는 것이다.

규칙 ψ는 $\psi(\bullet|P_n)$으로 쓰여 질 수 있다. Pn은 O의 실험 분포를 나타내고 관찰된 데이터에 관해서 만들어진 규칙에 대한 의존성을 반영한다. Loss 함수들은 주어진 규칙의 성능를 수량화 하는데 사용된다. 계속적인 결과 Y에 대한 일반적인 loss 함수는 제곱된 에러 loss 즉, $L(Y,\psi)=[Y-\psi(X)]^2$ 이고 카테고리적인 결과 Y를 가지고 널리 보급되어 있는 Loss 함수는 지시자(indicator) loss 함수, $L(Y,\psi)=I[Y\neq\psi(X)]$이다. loss 함수는 또한 차이를 나타내는 분류오류 비용을 통합하는데 사용될 수 있다[3].

결과의 어느 한쪽 타입에 대해, 기대된 손해 또는 손실은 식 (1)처럼 정의될 수 있다.

$$\tilde{\theta}=R(\psi,P)$$
$$=E_p[L(Y,\psi)]=\int L(y,\psi(x))dP(x,y) \quad (1)$$

Ⅲ. 리 샘플링 방법

크고 독립적인 테스트 셋이 없는 관찰된 데이터에 대해 분할 또는 리 샘플링 방법을 실행함으로써 예측 에러를 계산하는 많은 기술들이 있다. 각각의 이러한 기술들은 데이터를 훈련 셋과 테스트 셋으로 분할하고, 훈련 셋은 다시 훈련 셋과 검증 셋으로 분할한다. 관찰들을 서브 셋으로 분할하는 이진 랜덤 n-벡터 Sn은 $S_n=\{0,1\}$로 그리고 테스트 셋에서의 관찰 비율을 p로 정의한다.

본 논문은 MLP를 사용하여 리 샘플링 방법들의 예측 에러 추정지를 비교하였다.

3.1 v-fold cross-validation

이 방법은 크기가 거의 동일한 v개의 분할 중에 하나로 n개의 관찰들을 임의로 할당한다. 훈련 셋은 테스트 셋으로 레벨된 분할 하나를 제외한 모든 훈련 셋을 포함한다. 일반화 오류는 각각 v개의 테스트 셋에서 평가된 다음 평균을 구한다.

비율 p는 대략 1/v과 같다. p와 오류 평균은 에러 추정치에 반대로 또는 긍정적으로 영향을 줄 수 있다[4].

3.2 Leave-one-out cross-validation (LOOCV)

이 방법은 *v-fold cross-validation*의 가장 극단적인 경우이다. 이 방법에서 각 관찰은 개별적으로 테스트 셋에 할당된다(즉, v = n, p = 1/n). LOOCV는 높은 분산을 가지고 작은 바이어스 갖는 경향 있다. LOOCV는 계산 부담 때문에 큰 샘플에 대해서는 선호하는 방법이 아닐 뿐만 아니라 일반화 에러를 평가하는데 있어 전적으로 연구되어 오지 않았다[4].

3.3 Split sample

훈련-테스트 분할로써 또한 알려져 있는 이 대중적인 리 샘플링 방법은 미리 결정된 p를 기반으로 훈련 셋과 테스트 셋으로 분할하는 단 하나의 테이터를 수반한다. 예를 들어, p = 1/3은 데이터의 2/3를 훈련 셋으로 1/3을 테스트 셋으로 할당한다. 이 방법의 장점은 계산이 쉽다는 것이다. 또한, 분류기가 단지 한번 개발되기 때문에 분류기 개발을 위해 완전하게 설명된 알고리즘은 유용할 필요가 없다[4].

3.4 알고리즘

인공 신경망의 대표적인 기계 학습 알고리즘인 다층 퍼셉트론 신경망은 대부분의 패턴 인식 문제에 대해 안정적인 성능을 보이며, 일단 학습이 끝나면 응용 단계에서는 매우 빠르게 결과를 출력한다. 다층퍼셉트론 신경망은 백프로퍼게이션(back propagation)알고리즘을사용하는데 이것은 출력 층의 오차 신호를 이용하여 은닉 층과 출력 층 사이의 연결 강도를 변경하고 출력 층의 오차 신호를 은닉 층에 역전파하여 입력 층과 은닉 층 사이의 연결 강도를 변경하는 학습법이다[5].

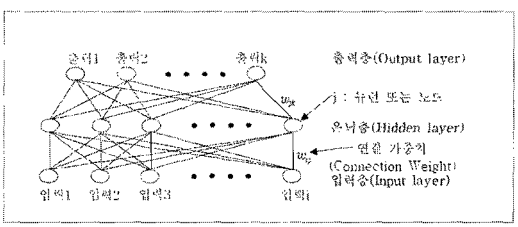

그림 1. MLP 신경망 구조
Fig. 1. Neural network structure of MLP

실제 분석은 기계학습(Machine Learning) Toolkit으로 Free open source 인 WEKA를 사용하였다. WEKA는 Waikato Environment for Knowledge Analysis의 약자로 프로그래밍 언어는 자바로 구성되어있고 모든 운영체제에서 실행 가능하다. Linux 환경에서 사용해 분석하였다[6].

Ⅳ. 실험 및 결과 고찰

4.1. 시스템 구성

본 논문에서의 시스템 구성은 데이터의 표준화, 특징 선택, 분류 알고리즘, 예측 평가 네 단계로 진행된다. 시스템 구성도는 그림 2와 같다.

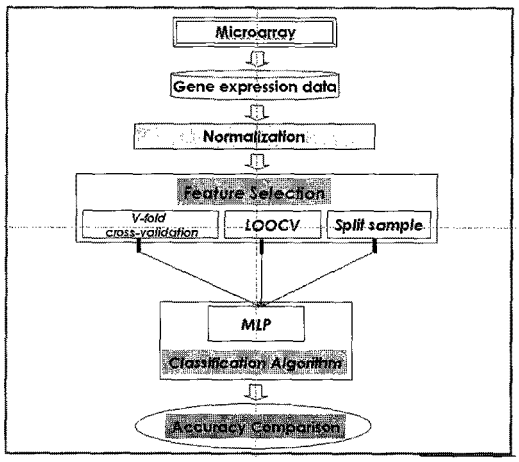

그림 2. 제안하는 분류 시스템
Fig. 2. porpoaing classification system

4.2. 실험 결과 및 고찰

실험용 데이터로 하버드대학교의 바이오인포메틱스 코어 그룹의 샘플데이터를 사용하였다. 데이터는 750마리의 흰쥐의 각 뇌신경조직 부위에서 획득한 유전체의 조정 인자를 각각 Cy5, Cy3로 염색한 다음, 2400개 이상의 알려진 유전체와 1700여개의 새로운 유전체가 찍힌 유리칩을 이용한 cDNA 마이크로어레이 실험에서 획득한 마이크로어레이 데이터를 사용하였다. 통계 컴퓨터 프로그램인 R의 stats 라이브러리의 IQR 함수를 이용하여 모든 슬라이드의 사분위수를 똑같게 맞추는 quantile-

normalization을 적용하였다. 그림 3은 원자료의 산점도 일부분이다.

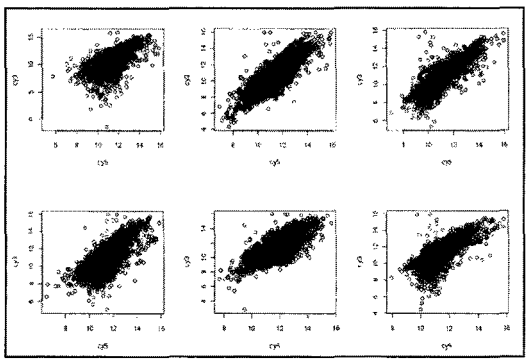

그림 3. 원자료의 산점도
Fig. 3. plot of original data

그림 4는 유전자의 발현 정도를 모든 슬라이드의 사분위수에 똑같게 맞추는 quantile-normalization을 적용한 결과의 일부분이다.

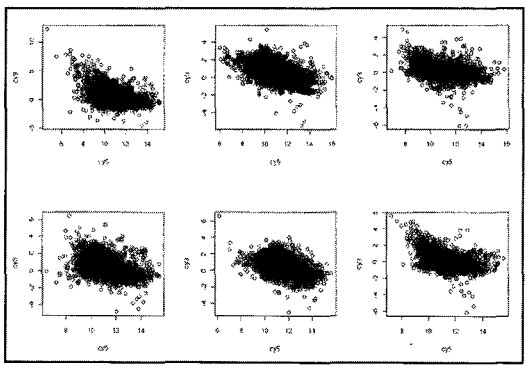

그림 4. quantile 표준화 후 산점도
Fig. 4. plot after quantile-normalization

4.3. 분석 결과

본 논문의 목적은 분류 문제에 제한된 일반화 에러를 추정하기 표준화 후 전체 마이크로어레이 데이터와 특징이 선택된 전체 마이크로어레이 데이터 그리고 샘플의 크기(40, 80, 120)에 각각 분류 알고리즘 MLP를 적용하여 리 샘플링 방법들의 예측 에러 차이를 조사하는 것이다.

WEKA를 이용하여 리 샘플링 방법들의 분류 예측 에러를 평가하기 위해 MLP 신경망을 구현하고 모멘텀은 0.09로, 총 레이어수는 3으로 고정한 후, 학습률을 0.01에서 0.05로 변화시켜가며 실험하여 리 샘플링 방법들의 예측 에러를 측정하였다.

리 샘플링 방법들을 비교하기 위해 각 방법들에 대한 조건 위험 추정치는 계산되었고 서로 그리고 실제 조건 위험과 비교되었다. 이것에 대한 평가는 평균제곱 오차를 나타내며, 실제 클래스와 예측한 클래스 차이를 제곱한 결과를 나타내는 MSE(the Mean Squared Error)와 치우침을 나타내는 bias로 수행되었고, 이 값이 작을수록 좋은 분류를 나타낸다. MSE와 bias는 다음 식 (2)와 (3)과 같다.

$$MSE = \frac{1}{r}\sum_{r=1}^{R}(\hat{\theta}_{n,r} - \tilde{\theta}_{n,r})^2 \quad (2)$$

$$Bias = \frac{1}{r}\sum_{r=1}^{R}(\hat{\theta}_{n,r} - \tilde{\theta}_{n,r}) \quad (3)$$

여기에서, $(\hat{\theta}_{n,r})$은 리 샘플링 조건 위험이고 $(\tilde{\theta}_{n,r})$는 r 번째 반복의 조건 위험이다. 모든 결과에 있어, 전체 반복 수 R = 100 으로 조정되었다.

사용자가 정의하는 테스트 샘플을 허락하는 v-fold cross-validation에 대해서는 2-, 5-, 10-fold가 조사되었고, split sample 평가 테스트 셋에 대해서는 p = 1/3과 p = 1/2의 비율이 bias/variance 교환을 평가하기 위해 조사되었다.

각각 흰쥐 750마리의 각 뇌신경조직 부위에서 획득한 유전체의 조정 인자를 공 변수를 갖는 전체 관찰에 대해 MLP 신경망으로 측정한 예측 에러 추정치 결과는 표 1과 같다. 이를 특징 선택에 따른 예측 에러와 샘플 크기에 대한 예측 에러를 비교하기 위한 실험의 대조군으로 하였다.

표 1. 전체 샘플에 대한 예측 에러 추정치
table 1. Prediction error estimate on total data

Resampling method	p	Estimate	Bias	MSE
LOOCV	0.025	0.146	0.025	0.018
v-fold CV	0.5	0.357	0.279	0.097
	0.2	0.284	0.163	0.055
	0.1	0.189	0.068	0.024
Split	0.333	0.371	0.25	0.121
	0.5	0.438	0.317	0.147

가장 작은 MSE와 bias는 LOOCV와 10-fold CV는 나타난 반면, 가장 큰 MSE, bias는 5-fold CV와 1/2 split에서 나타났다. 에러 추정치 LOOCV에서 0.146로 가장 낮게 나타났으며 1/2 split가 0.438로 가장 높게 나타났다.

표 2. 특징 선택을 갖는 전체 샘플에 대한 예측 에러 추정치
table 2. Prediction error estimate with feature selection on total data

Resampling method	p	Estimate	Bias	MSE
LOOCV	0.025	0.058	0.016	0.005
v-fold CV	0.5	0.277	0.235	0.077
	0.2	0.108	0.066	0.011
	0.1	0.078	0.036	0.005
Split	0.333	0.145	0.103	0.044
	0.5	0.265	0.223	0.086

특징을 선택한 후에 리 샘플링 방법을 테스트 한 결과 전반적으로 예측에러가 감소하는 경향을 보였다. LOOCV는 가장 작은 MSE와 bias를 보였으며 약 2배의 예측 향상을 보였다. 그러나 split 방법에서는 다소 큰 bias를 나타내었다.

각각의 샘플 크기 40, 80, 120에 대해 측정한 에러 추정치는 표 3과 같다. 샘플 크기 40, 80, 120으로 증가할 때 LOOCV의 bias는 커지고 MSE는 감소하였고 2-, 5-, 10-fold CV는 샘플이 크기가 증가함에 따라 MSE와 bias도 감소하였다. 5-fold CV와 10-fold CV에서 가장 작은 MSE와 bias를 나타내었고, 1/2 split과 2-fold CV는 가장 큰 MSE와 bias를 나타내었다.

표 3. 샘플 크기 대한 예측 에러 추정치
table 3. Prediction error estimate on each of sample size

Resampling method	sample = 40			
	p	Estimate	Bias	MSE
LOOCV	0.025	0.072	-0.019	0.008
v-fold CV	0.2	0.085	0.038	0.01
	0.5	0.07	0.004	0.007
	0.1	0.078	-0.007	0.006
Split	0.333	0.119	0.001	0.017
	0.5	0.117	0.37	0.018
	sample = 80			
	p	Estimate	Bias	MSE
LOOCV	0.025	0.04	-0.013	0.004
v-fold CV	0.2	0.043	0.002	0.004
	0.5	0.045	-0.008	0.005
	0.1	0.036	-0.009	0.003
Split	0.033	0.071	0.0	0.007
	0.5	0.058	0.001	0.005
	sample = 120			
	p	Estimate	Bias	MSE
LOOCV	0.025	0.033	-0.004	0.003
v-fold CV	0.2	0.031	0.0	0.003
	0.5	0.032	-0.006	0.003
	0.1	0.031	-0.006	0.003
Split	0.033	0.059	-0.004	0.005
	0.5	0.046	-0.001	0.004

또한 split 방법이 가장 높은 예측 에러 추정치를 나타내었다.

V. 결 론

다수의 공 변수와 작은 샘플 크기에 직면 했을 때 예측 에러 추정은 비교적 새로운 문제이다. 유의한 유전자 선택, 샘플 크기 그리고 신호 대 잡음 비율은 리 샘플링 방법의 상대적인 성능에 중요한 영향을 끼친다. 본 논문에서는 MLP를 사용하여 리 샘플링 방법들의 예측 에러를 비교 평가하는 시스템을 고안하고 리 샘플링 방법들을 비교 분석하였다.

split 방법과 2-fold CV는 작은 샘플 크기에서 매우 좋지 않은 결과를 보였고 이러한 결과는 줄여진 훈련 셋 크기의 사용으로 발생하는 큰 bias에 기인하는 것으로 여기진다. LOOCV는 MSE와 bias에 관해서는 일반적으로 작은 에러 예측을 나타내었다. 10-fold CV 예측 에러 추정치는 거의 모든 셋팅에서 LOOCV의 예측 에러 추정치 비슷했으나 계산적으로 번거로운 분석에 대해서는 10-fold CV가 LOOCV보다 오히려 더 나은 경향을 보였다. 그러나 샘플 크기가 증가했을 때 리 샘플링 방법들 사이에 차이는 감소하였다.

향후 연구 과제로는 다양하고 체계적인 많은 데이터의 획득과 계속적인 결과에 대한 리 샘플링 방법의 비교와 bootstrap 추정치의 작용을 계속적으로 조사해야 할 것이다.

표준화 방법들이 유의한 유전자 선택에 미치는 영향에 대해 더 많은 연구를 진행하고자 한다.

참고문헌

[1] Ransohoff,D.F., "Rules of evidence for cancer molecular marker discovery and validation.", Nature Reviews/ Cancer, 4, 309-313, 2004.

[2] Breiman,L. and spector,P., "Submodel selection and evaluation in regression.", The X-random case. Int. Stat. Rev., 60, 291-391, 1992.

[3] S. Dudoit, "Comparison of discrimination methods for the classification of tumors using gene expression data", Journal of the American Statistical Association, vol. 97, pp. 77-87, 2002.

[4] Vreiman,L. Friedman,J.H., Olshen,R.A and Stone,C.J., "Classification and Regression Tress.", Wadsworth and Brooks/Cole, Monterey, CA., 1984.

[5] Golub, T.R., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, vol.286, no. 5439, pp. 531-537, 1999.

[6] WEKA, http://www.cs.waikato.ac.nz/ml/weka/

저자소개

박수영(Su-Young Park)

2001년 조선대학교 컴퓨터통계학과 이학사
2003년 조선대학교 컴퓨터통계학과 이학석사
2007년 조선대학교 컴퓨터통계학과 이학박사
※관심분야 : 신경망, 인공지능, 정보보호, 멀티미디어, 멀티미디어 컨텐츠, Bioinformatics

정채영(Chai-Yeoung Jung)

1987년 조선대학교 컴퓨터공학과 공학석사
1989년 조선대학교 컴퓨터공학과 공학박사
1986년~현재 조선대학교 컴퓨터통계학과 교수
※ 관심분야 : 신경망, 인공지능, 정보보호, 멀티미디어, 멀티미 디어 컨텐츠, Bioinformatics