응용통계연구 제 **20**권 *3*호, *2007*년, *pp. 531–540*

붓스트랩 방법을 활용한 SVM 기반 유전자 선택 기법*

송석헌[1] 김경희[2] 박창이[3] 구자용[4]

요 약

본 연구에서는 유전자 선택 방법으로 최근 이용되는 SVM-RFE 알고리즘은 단순히 가중치의 절대값을 유전자 선택 기준으로 사용하여 유전자 값의 변동성을 고려하지 못하므로 가중치의 절대값을 그것의 표준오차로 나눈 보완된 통계량, B-RFE 알고리즘을 새로운 기준으로 제안하였다. 두 방법을 모의실험을 통해서 비교한 결과 본 연구에서 제안한 B-RFE 알고리즘이 더 의미 있는 순위를 도출하였다.

주요용어: 분류, 유전자 선택, RFE (Recursive Feature Elimination).

1. 서론

암을 정확하고 세밀히 분류하는 것은 암의 진단과 치료를 위해 매우 중요하다. 이에 최근 DNA 마이크로어레이 기술을 이용한 암 분류 연구가 활발히 진행 중이다. DNA칩 기술로 얻어지는 유전자 발현 자료는 생체 조직이나 세포의 수천 개에 달하는 유전자의 발현량을 측정한 것으로, 유전자 발현 양상에 기반을 둔 암 종류의 분류 등에 유용한 것으로 알려져 있다.

최근 많은 판별 방법들이 유전자 발현 자료에 적용되어져 왔다. 전통적인 방법으로 판별분석 (Dudoit 등, 2002), k-근방분류 (Koutsoukos 등, 1994), 의사결정 나무 등이 있고, 보다 새로운 방법으로 서포트 벡터 기계 (Support Vector Machines), Boosting (Philip과 Vinsensius, 2003)과 신경망 (Khan 등, 2001) 등이 있다. 또한 판별에 유용한 정보를 갖는 유전자의 선택 방법으로 전형적으로 이표본 t-검정이나 분산분석에서의 유의확률을 이용해 왔다. 이와 유사한 기준으로 Golub 등 (1999)은 유전자 각각의 집단별 변동성 대비 평균의 차이를 유전자 순위 기준으로 제안하였다. 그리고 Furey 등 (2000)은 Golub의 통계

* 송석헌의 연구는 정부재원으로 한국학술진흥재단 (R14-2003-002-01002-0)의 지원을 받아 수행되었으며 구자용 및 박창이의 연구는 2005년 정부재원으로 (교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음 (KRF-2005-070-C00020).

1) (136-701) 서울특별시 성북구 안암동 5-1, 고려대학교 통계학과, 교수
E-mail: ssong@korea.ac.kr
2) (156-756) 서울특별시 성북구 안암동 5-1, 고려대학교 통계학과, 대학원
E-mail: arlenent@hanmail.net
3) (136-701) 서울특별시 성북구 안암동 5-1, 고려대학교 통계연구소, 연구조교수
E-mail: park463@korea.ac.kr
4) (교신저자)(136-701) 서울특별시 성북구 안암동 5-1, 고려대학교 통계학과, 교수
E-mail: jykoo@korea.ac.kr

량을 수정하여 그것의 절대값을 순위의 기준으로 사용하였으며 Pavlidis 등 (2001)은 피셔의 판별기준과 유사한 계수를 제안하였다. 그러나 Golub의 통계량을 포함한 여러 유전자 선택 방법들은 개별적인 유전자 하나하나의 평균과 표준편차만을 고려하므로, 유전자 사이에 내재되어 있는 상관성을 고려하지 못한다. 즉, 위의 방법들은 각각의 유전자들이 서로 직교한다는 가정을 하고 있어, 판별에 기여하는 유전자를 찾지 못한다는 단점이 있다. 최근 Guyon 등 (2002)은 유전자들이 서로 상관되어 있는 경우에도 적용 가능한 Recursive Feature Elimination (RFE) 알고리즘을 제안하였다. 그러나 Guyon의 방법은 단순히 훈련표본의 선형결합인 가중치의 절대값을 기준으로 유전자의 중요도를 평가하고 있으므로 유전자 하나하나의 변동성을 고려할 수 없다.

본 연구에서는 유전자의 상관성과 유전자 각각이 갖는 변동성 모두를 유전자의 순위에 반영하고자 한다. 우선, 자료의 판별 방법으로는 계산상의 이점을 갖는다고 알려져 있는 서포트 벡터 기계를 이용하려한다. 이를 통해 가중치 벡터를 계산하였고 가중치의 표준오차를 붓스트랩을 이용하여 추정하고 추정된 표준오차와 가중치의 비를 유전자 순위의 기준으로 사용하여 각 유전자의 변동성을 고려하고, RFE 알고리즘을 이용하여 유전자가 직교하지 않더라도 의미 있는 유전자를 찾을 수 있는 방법을 제안하고자 한다.

2. 서포트 벡터 기계 (Support Vector Machines)

본 연구에서는 선형 초평면에 의한 이항분류만을 다루기로 한다. $\{(x_i, y_i)\}_{i=1}^n$은 훈련표본으로서 $x_i \in R^l$이고 $y_i \in \{-1, +1\}$이다. 여기서 x_i와 y_i는 각각 입력변수와 출력변수라 불린다. $\langle \cdot, \cdot \rangle$을 R^l상의 내적이라 하자. 분류는 훈련표본을 이용하여 $f(x) = \langle w, x \rangle + b$로 정의되는 선형 판별함수를 얻은 후, 그 부호를 이용하여 새로운 입력 x에 대한 출력값을 예측하는 것으로 볼 수 있다. w의 해를 얻기 위하여 서포트 벡터 기계는 다음과 같은 이차함수를 최적화한다.

$$\min_{w,b} \frac{1}{2}\langle w, w \rangle + C\sum_{i=1}^{n} \xi_i, \tag{2.1}$$

$$\text{subject to } y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, n.$$

여기서 $C > 0$는 훈련오류와 벌점간의 상대적 크기를 조절하는 벌점항 모수이고 ξ_i는 slack 변수라 불린다. 흔히 (2.1)를 직접 최적화하는 대신 이에 대한 다음과 같은 쌍대 (dual) 목적함수를 최적화한다.

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle - \sum_{i=1}^{n} \alpha_i, \tag{2.2}$$

$$\text{subject to } \sum_{i=1}^{n} y_i \alpha_i = 0, 0 \leq \alpha_i \leq C, i = 1, \ldots, n.$$

(2.2)의 해를 $\hat{\alpha}_i, i = 1, \ldots, n$이라 하면 해가 되는 선형 판별함수는

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i y_i \langle x_i, x \rangle + \hat{b} \tag{2.3}$$

로 주어진다. 이 해가 0이 아닌 $\hat{\alpha}_i$값으로만 표현되므로 이러한 0이 아닌 $\hat{\alpha}_i$들을 대응되는 x_i들을 소위 서포트 벡터라 부른다. 여기서 $\hat{b}$는 Karush-Kuhn-Tucker 경계 조건들을 이용하여 결정된다. 비선형 분류 문제는 내적 $\langle \cdot, \cdot \rangle$ 대신 비선형 커널 $k(\cdot, \cdot)$을 이용하여 다룰 수 있다. 흔히 사용되는 비선형 커널로는 $k(x, y) = \exp(-\gamma \|x - y\|^2)$로 정의된 radial basis function (RBF) 커널이 있다. 여기서 $\gamma > 0$는 스케일 모수이다. 그러면 내적을 커널로 대체한 후 (2.2)의 쌍대 최적화 문제를 풀면 비선형 분류에서의 해는 (2.3)과 동일한 형태를 갖는다. 자세한 사항은 Vapnik (1998)을 참조하기 바란다.

3. Recursive Feature Elimination

전통적으로 연구자들은 개별적으로 가장 자료를 잘 판별하는 유전자를 선택하는 방법을 사용해왔다. 몇몇의 연구자들은 하나의 유전자가 제거되었을 때 비용함수에서의 변화량을 순위의 기준으로 사용하는 방법을 제안하였다 (Kohavi와 John, 1997). J를 비용함수라 하고 $DJ(i)$를 i번째 유전자의 가중치를 0으로 제거하여 발생하는 비용함수에서의 변화라 하자. 또한 Dw_i는 i번째 유전자를 제거함에 따른 w_i의 값이라 하자. LeCun 등 (1990)은 J를 테일러 전개하여 식 (3.1)을 도출하였다.

$$DJ(i) = (1/2)\frac{\partial^2 J}{\partial {w_i}^2}(Dw_i)^2. \tag{3.1}$$

선형 판별함수에서는 비용함수가 w_i의 2차식으로 표현되어 $DJ(i)$와 w_i는 동일한 기준이 된다. 선형 서포트 벡터 기계의 경우 비용함수는 제약조건하에서 가중벡터의 제곱합을 최소로 한 것이므로 유전자 순위의 기준으로 $(w_i)^2$을 쓰는 것을 가능하게 해준다 (Guyon 등, 2002). 그러나 좋은 유전자 순위의 기준이 꼭 좋은 유전자 부분집합의 순위에 대한 기준이 되는 것은 아니다. $DJ(i)$나 $(w_i)^2$의 기준은 목적함수에서 한 번에 한 유전자를 제거함에 따르는 효과를 측정한다. 만약 여러 유전자를 한 번에 제거하는 경우 이러한 방법은 최적의 방법이 아니다. 이러한 문제를 해결할 수 있는 방법으로 RFE 방법을 소개하면 다음과 같다.

(단계 1) 비용함수에 대하여 가중치 w_i를 최적화

(단계 2) 모든 유전자에 대해 순위의 기준을 계산

(단계 3) 가장 작은 기준값을 갖는 유전자를 제거

(단계 4) 위의 (단계 1)–(단계 3)을 반복

RFE의 반복적인 과정은 Kohavi와 John (1997)의 Backward Feature Elimination 방법의 하나로 볼 수 있다. 또한 이 과정은 순위가 가장 낮은 유전자를 찾은 후 그것을 제외한 자료를 이용하여 다시 순위가 가장 낮은 유전자를 찾기 때문에 개별적으로 판별력이 가장 좋은 유전자를 고르는 방법과는 다르다.

따라서 Guyon 등 (2002)의 연구에서는 RFE 알고리즘 내의 가중치 w_i를 계산함에 있어 서포트 벡터 기계를 사용하는 알고리즘을 개발하였다. Guyon의 RFE 알고리즘은 단순히

SVM의 해의 절대값을 가중치로서 유전자 선택의 기준으로 삼았다. 만약 두 유전자가 모두 판별에 중요한 영향을 미치지만 그 중 한 유전자의 가중치는 변동성이 매우 크고 다른 유전자는 항상 비슷한 가중치 값을 갖는다면, 변동성이 작은 유전자의 순위가 더 높아야 할 것이다. 그러나 Guyon의 알고리즘은 이러한 문제를 해결할 수 없다. 따라서 본 연구에서는 보다 더 중요한 유전자를 선택하기 위해 유전자 각각의 변동성을 고려한 Bootstrap RFE (B-RFE) 알고리즘을 제안한다.

B-RFE 알고리즘은 유전자 순위의 기준으로 Guyon의 SVM-RFE 알고리즘의 가중치를 그것의 표준오차로 나눈 것, 즉 $|w_i|/\hat{se}(w_i)$, 을 사용한다. 이때 표준오차는 붓스트랩으로 추정한다. 따라서 단순한 가중치 계수의 크기 대신 변동성을 고려하여 계수와 표준오차의 비를 사용하였다. 이는 t검정통계량과 유사한 것으로 직관적으로도 타당할 것임을 예측할 수 있다.

붓스트랩의 절차를 이용하여 표준오차를 추정한 B-RFE 알고리즘은 다음과 같다.

(단계 1) 다음 자료를 입력한다.

훈련표본 $(X_0 = [x_1, x_2, \ldots, x_k, \ldots, x_l]^T)$

집단의 레이블$(y = [y_1, y_2, \ldots, y_k, \ldots, y_l]^T)$

(단계 2) 다음의 s, r을 초기화한다.

남아있는 유전자의 부분집합 $s = [1, 2, \ldots, n]$

유전자 순위표 $r = [\ \ ]$

(단계 3) 다음과정을 반복한다. ($s = [\ \ ]$까지)

훈련표본들을 좋은 유전자들로 제한한다. $(X = X_0(s))$

입력변수(X, y)로 서포트 벡터 기계를 실행하여 차원이 s인 가중치 벡터를 계산한다.

$$w = \sum_k \alpha_k y_k x_k$$

붓스트랩을 이용하여 w_i의 표준오차를 추정한다.

$$\widehat{s.e.(w_i)} = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} (w^{*b} - \bar{w}^*)^2 \right\}^{1/2}$$

순위의 기준을 계산한다.

$$c_i = \frac{(w_i)^2}{\{\hat{s.e.}(w_i)\}^2}$$

가장 작은 기준을 갖는 유전자를 찾는다.

$$f = \arg\min(c)$$

유전자 순위표를 업데이트한다.

$$r = [s(f), r]$$

가장 작은 기준값을 갖는 유전자를 제거한다.

$$s = s(1 : f-1, f+1 : length(s))$$

(단계 4) 유전자 순위표 r를 출력한다.

4. 자료 분석

4.1. 모의실험

이 절에서는 SVM-RFE 알고리즘과 B-RFE 알고리즘을 비교하기 위해 모의실험을 실시하였다. 표본의 수는 100, 200, 500개이며 변수는 4개 $(X_1, \ldots, X_4)$이다. X_1, X_2의 절반은 이변량 정규분포 $N_2(-1, 1, 1^2, 2^2, 0)$에서 임의 표집하였고 이를 $+1$에 속하는 표본으로 정하였다. 다른 절반은 $N_2(1, -1, 1^2, 2^2, 0)$에서 임의 표집하였고 이를 -1에 속하는 표본으로 정하였다. 위의 두 변수가 판별에 의미 있는 영향을 미치게 되므로 중요 변수가 되며, 상대적으로 분산이 큰 X_2보다는 X_1의 순위가 더 높아야 할 것임을 짐작할 수 있다. 다른 두 변수 X_3, X_4는 집단의 레이블에 관계없이 $N_1(0, 1^2)$에서 임의 표집하였다. 즉 이 두 변수는 판별에 어떠한 영향도 미치지 못하는 변수로서 X_1, X_2보다는 항상 순위가 낮아야 할 것이다. 원자료와 이를 표준화한 자료에 각각의 알고리즘을 적용하였고 붓스트랩 표본의 수는 100, 500으로 정하였다. 이러한 자료에 SVM-RFE 방법과 B-RFE 방법으로 변수 순위를 구하였다. 순위로만 보면 어느 방법이 더 의미가 있는지 알 수 없으므로 자료를 훈련자료와 시험자료로 임의분할하였다. 우선 훈련자료로만 순위를 구하였다. 그리고 하나씩 변수를 제거해가면서 5-폴드 교차타당성 방법 (5-fold cross validation)을 통하여 가장 오분류율이 작은 변수의 수를 찾고, 이러한 변수와 가중치로 시험자료에서의 오분류율을 구하였다.

표 4.1은 모의실험 결과이다. 오분류율은 100번 실험을 반복한 뒤 평균을 구한 것이다. SVM-RFE를 적용했을때를 제외한 모든 경우에 1위와 2위 변수를 이용한 방법이 가장 작은 오분류율을 보인다. 따라서 B-RFE는 유의한 변수들의 순위를 잘 잡아주는 반면, SVM-RFE는 유의한 변수들간의 상대적 분산의 차이에 의해 순위가 뒤섞임으로써 더 많은 변수를 선택할 수록 오분류율이 줄어드는것으로 보인다. 그러나 SVM-RFE 역시 표본크기가 500으로 충분히 커지면 상대적인 분산차이에도 불구하고 1위와 2위를 사용한 경우 오분류율이 최소가 되었다. 중요한 변수가 2개로 정해져 있으므로, 1위와 2위의 변수를 이용한 판별식으로 구한 오분류율의 평균이 중요하다고 생각하여 구해보았다. 그 결과 표본이 200개인 경우 표준화 자료의 SVM-RFE 방법의 평균은 0.1372이나 표본의 수가 200개 이고 붓스트랩 반복의 수가 500일 때 B-RFE 방법의 평균은 0.1328였다. 오분류율들의 표준오차를 고려해 볼 때 SVM-RFE 방법과의 오분류율 측면에서의 B-RFE 방법의 차이는 그다지 크지는 않음을 알 수 있다.

순위를 살펴본 결과 원래 변수 1과 2가 중요한 변수로 뽑혀야 하나, 반복의 결과로, 혹은 변수 2의 분산이 다른 변수들 보다 상대적으로 크기 때문에 순위가 바뀌는 경우가 있다. 표 4.2는 표본의 수가 200개일 때 각 변수별로 1위와 2위로 뽑힌 횟수를 살펴보았다. SVM-RFE 알고리즘은 100번 중 17번 중요한 변수인 X_1과 X_2를 1, 2위 순위로 택하지 못했으나

표 4.1: SVM-RFE와 B-RFE 알고리즘의 오분류율의 평균값

표본수	변수	SVM-RFE	B-RFE(B=100)	B-RFE(B=500)
100	1위	0.1617 (0.0016)[1)]	0.1617 (0.0015)	0.1617 (0.0016)
	1~2위	0.1508 (0.0028)	0.1411 (0.0018)	0.1402 (0.0017)
	1~3위	0.1480 (0.0028)	0.1457 (0.0021)	0.1457 (0.0021)
	1~4위	0.1459 (0.0020)	0.1459 (0.0020)	0.1459 (0.0020)
200	1위	0.1623 (0.0014)	0.1623 (0.0014)	0.1623 (0.0014)
	1~2위	0.1372 (0.0024)	0.1328 (0.0016)	0.1328 (0.0016)
	1~3위	0.1357 (0.0016)	0.1357 (0.0016)	0.1357 (0.0016)
	1~4위	0.1368 (0.0016)	0.1368 (0.0016)	0.1368 (0.0016)
500	1위	0.1594 (0.0016)	0.1594 (0.0016)	0.1594 (0.0016)
	1~2위	0.1342 (0.0018)	0.1342 (0.0018)	0.1342 (0.0018)
	1~3위	0.1352 (0.0018)	0.1352 (0.0018)	0.1352 (0.0018)
	1~4위	0.1353 (0.0017)	0.1353 (0.0017)	0.1353 (0.0017)

1) 오분류율의 표준오차

B-RFE 알고리즘은 100번 모두 중요한 변수를 1, 2위로 택하였다. SVM-RFE와 B-RFE 모두 X_1을 100번 모두 1위로 선택했으나, SVM-RFE는 X_1에 비해 상대적으로 분산이 큰 X_2 대신 잡음변수인 X_3와 X_4를 각각 11번 6번을 2위로 선택하였다. 잡음변수인 X_3과 X_4 보다는 X_1과 X_2를 찾아내는 방법을 알아보는 것이 중요한 것이므로 17%를 잡음변수인 X_3과 X_4를 중요한 변수로 택한 SVM-RFE 알고리즘에 문제점이 있음을 알 수 있다. 따라서 비록 오분류율에는 큰 차이가 없지만 B-RFE 알고리즘이 SVM-RFE에 비해서 더 의미있는 변수들의 순위를 도출하였으므로 더 좋은 성능을 보여주었다고 할 수 있다.

4.2. 마이크로 어레이 자료 분석

분석에 사용된 Alon 등 (1999)의 직장암 자료는 사전 처리를 거친 후 DNA 마이크로어

표 4.2: SVM-RFE와 B-RFE 알고리즘의 오분류율과 선택된 유전자 개수

	SVM-RFE	B-RFE(B=100)
시험자료의	0.2370	0.2630
오분류율의 평균	(0.0080)[1)]	(0.0080)
교차타당성 오분류율을	5.8900	4.9900
최소로하는 유전자의 평균	(0.3110)[2)]	(0.1900)

1) 시험자료의 오분류율의 표준오차

2) 교차타당성 오분류율을 최소로 하는 유전자 수의 표준오차

레이 자료 결과로부터 추출된 유전자 발현 정보량이며 62개 표본과 2000개의 유전자 값으로 표현되어 있다. 62개 표본 중 40개는 정상이며 22개는 암 표본이다. 분석의 목표는 정상과 암 환자를 가장 잘 판별하는 유전자를 찾는 것이다. 우선 SVM-RFE 알고리즘과 B-RFE 알고리즘을 적용하여 유전자 순위를 구하고 이러한 순위를 적용하여 전체 자료에서 교차타당성 오분류율을 계산하였다.

유전자 순위를 비교해 보면, 판별에 가장 중요한 역할을 하는 1위, 2위 유전자가 두 알고리즘에서 다른 것으로 나타났다. 그러나 10위 중 6개의 유전자를 공유하는 것을 알 수 있었다. SVM-RFE 알고리즘의 경우 1위 유전자로만 판별했을 때 교차타당성 오분류율이 0.25였으며 1위부터 14위 유전자로 판별했을 때 이 값이 0으로 떨어지는 것을 볼 수 있었다. 반면, B-RFE 알고리즘의 경우 1위 유전자로만 판별했을 때 교차타당성 오분류율이 0.27이나 1위에서 6위 유전자로 판별했을 경우 이 값이 0으로 떨어지는 것을 볼 수 있었다. 그러나 이러한 방법은 전체 자료로 구축한 모형을 다시 그 자료로 평가하므로 낙관성 편향을 갖게 된다. 이러한 편향을 감소시키고 각 알고리즘의 타당성을 살펴보기 위해 자료를 훈련자료와 시험자료로 나누어서 분석을 시도하였다. 훈련자료에서는 각 알고리즘을 적용하여 유전자의 순위를 구하였고 1위 유전자부터 각 순위의 유전자를 사용하여 교차타당성 오분류율을 구하였다. 이러한 오분류율이 가장 작게 나온 유전자들과 그것이 갖는 가중치를 이용하여 시험자료 내에서의 오분류율을 계산하였다. 분할된 자료는 각각 20개의 정상인 표본과 11개의 암 표본을 포함하게 하였고 한 번만 분할한다면 임의 분할에 따른 편의가 발생할 것이므로 100회 반복하여 분석하였다.

표 4.2는 시험자료에서의 오분류율의 평균과 그것들의 표준오차, 교차타당성 오분류율을 최소로 하는 유전자 개수의 평균과 그것들의 표준편차를 나타낸 것이다. 우선 SVM-RFE 알고리즘의 오분류율의 평균값이 B-RFE 알고리즘의 경우보다 조금 더 낮다는 것을 알 수 있다. 그러나 시험자료의 표본의 수가 31개임을 감안하면 이러한 차이는 하나의 유전자를 잘못 분류하는 비율도 되지 않는 값이라는 것을 알 수 있다. 이는 표본의 개수가 아주 작은 자료를 또다시 반분하였고 실제적인 해답이 없는 데이터라는 점에서 다른 실제 자료를 통해 이를 비교해 보는 것이 바람직할 것으로 여겨진다. SVM-RFE와 B-RFE의 오분류율의 평균은 각각 0.2370과 0.2630이고, 표준오차가 0.0080임을 고려하면 두 방법은 오분류율 측

면에서는 큰 차이가 없음을 알 수 있다. 각 방법에 의하여 선택된 유전자수는 SVM-RFE 알고리즘의 경우 5.89이고 B-RFE 알고리즘의 경우 4.99이다. 실제 자료에서는 유의한 유전자를 알 수 없으므로 유전자 선택에 대한 직접적인 질적 비교가 가능하지는 않지만, 적은 수의 유전자로 비슷한 오분류율을 주므로 B-RFE가 더 좋은 성능을 보여 준다고 할 수 있다. 또한 선택된 유전자 수의 평균에 대한 표준오차값을 보면 SVM-RFE 알고리즘은 유전자 선택에 있어서 B-RFE에 비해 상당한 변동성을 보이고 있음을 알 수 있다.

5. 결론

본 논문에서는 붓스트랩을 이용하여 유전자 선택기준을 조정한 B-RFE 알고리즘을 제안하고 기존의 SVM-RFE 알고리즘과 비교하였다. 모의실험을 통하여 판별에 중요하나 변동성이 큰 변수의 순위를 낮게 제시하는 SVM-RFE 알고리즘의 문제점을 B-RFE 알고리즘이 극복하고 있음을 알 수 있었다. 이는 시험자료에서의 선택된 변수에 대한 질적 비교를 통해서도 확인해 볼 수 있다. 마이크로 어레이 자료를 적용해 보았을 때 B-RFE 알고리즘의 오분류율이 SVM-RFE 알고리즘의 오분류율과 큰 차이를 보이지 않으나 더 적은 수의 유전자 더 안정적으로 선택함을 알 수 있다.

본 연구에서는 선형인 경우의 서포트 벡터 기계를 사용하였다. 그러나 실제적으로 자료가 선형식으로 분리 가능하다고 단정지을 수 없으므로 여러 커널함수를 이용하여 보다 일반적인 경우로 확장하여 분석해 볼 수 있을 것이다. 이 경우 선형 서포트 벡터 기계를 통한 분석보다는 시간이 오래 걸리나 방법의 특성 상 적은 수의 서포트 벡터만 계산하면 되므로 비선형인 경우에도 충분히 적용할 수 있을 것으로 보인다. 또한, 후진적인 제거 방법을 취하는 본 논문의 방식과는 달리 전진적으로 가장 중요한 유전자를 선택하고, 나머지 자료로 중요한 유전자를 선택하는 방법이나, 전진적 방법을 취하다가 오분류율이 증가할 때 다시 후진적으로 유전자를 제거해 나가는 방법을 추후 연구 과제로 고려해 볼 수 있을 것이다.

참고문헌

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra. S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6745–6750.

Dudoit, S., Fridlyand, J. and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, **97**, 77–87.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, **16**, 906–914.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh M. L., Downing, J. R., Caligiuri, 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.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines, *Machine Learning*, **46**, 389–422.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine*, **7**, 673–679.

Kohavi, R. and John, G. (1997). Wrappers for feature subset selection, *Artificial Intelligence*, **97**, 273–324.

Koutsoukos, A. D., Rubinstein, L. V., Faraggi, D., Simon, R. M., Kalyandrug, S., Weinstein, J. N., Kohn, K. W. and Paull, K. D. (1994). Discrimination techniques applied to the NCI in vitro anti-tumour drug screen: predicting biochemical mechanism of action, *Statistics in Medicine*, **13**, 719–730.

LeCun, Y., Denker, J. S. and Solla, S. A. (1990). Optimum brain damage, *Advances in neural information processing systems*, **2**, 598–605.

Pavlidis, P., Weston, J., Cai, J. and Grundy, W. N. (2001). Gene functional classification from heterogeneous data, *Annual Conference on Research in Computational Molecular Biology Proceedings of the fifth annual international conference on Computational biology.*

Philip, M. L. and Vinsensius, B. V. (2003). Boosting and microarray data, *Machine Learning*, **52**, 31–44.

Vapnik, V. N. (1998). *Statistical Learning Theory*, John Wiley & Sons, New York.

[2007년 2월 접수, 2007년 5월 채택]

Gene Selection Based on Support Vector Machine using Bootstrap*

Seuck Heun Song[1] Kyoung Hee Kim[2] Changyi Park[3] Ja-Yong Koo[4]

ABSTRACT

The recursive feature elimination for support vector machine is known to be useful in selecting relevant genes. Since the criterion for choosing relevant genes is the absolute value of a coefficient, the recursive feature elimination may suffer from a scaling problem. We propose a modified version of the recursive feature elimination algorithm using bootstrap. In our method, the criterion for determining relevant genes is the absolute value of a coefficient divided by its standard error, which accounts for statistical variability of the coefficient. Through numerical examples, we illustrate that our method is effective in gene selection.

Keywords: Classification, gene selection, recursive feature elimination.

* This research of Seuck Heun Song was supported by the Korea Research Foundation Grant founded by Korean Goverment (MOEHRD)(R14-2003-002-01002-0), and the research of Ja-Yong Koo and Changyi Park was supported by the Korea Research Foundation Grant funded by Korea Government (MOEHRD, Basic Research Promotion Fund) (KRF-2005-070-C00020).

1) Professor, Department of Statistics, Korea University, Seoul 136-701, Korea
E-mail: ssong@korea.ac.kr

2) Graduate student, Department of Statistics, Korea University, Seoul 136-701, Korea
E-mail: arlenent@hanmail.net

3) Research Assistant Professor, Institute of Statistics, Korea University, Seoul 136-701, Korea
E-mail: park463@korea.ac.kr

4) (Corresponding author) Professor, Department of Statistics, Korea University, Seoul 136-701, Korea
E-mail: jykoo@korea.ac.kr