Journal of the Korean Data & Information Science Society
2009, **20**(2), 311–319

한국데이터정보과학회지

마이크로어레이 자료에서 서포트벡터머신과 데이터 뎁스를 이용한 분류방법의 비교연구[†]

황진수[1] · 김지연[2]

[1]인하대학교 통계학과 · [2]인하대학교 입학처

접수 2009년 1월 19일, 수정 2009년 3월 5일, 게재확정 2009년 3월 12일

요 약

군집과 분류분석에서 L1 데이터 뎁스를 이용한 DDclust와 DDclass라고 불리는 로버스트한 방법이 Jornsten (2004)에 의하여 제안되었다. SVM-기반방법이 많이 사용되나 이상치가 있는 경우에는 약간의 문제가 있다. 유전자 자료에서는 유전자 수가 많기 때문에 적절한 유전자 선택과정이 필요하다. 따라서 적절한 유전자 또는 유전자 군집을 선택하여 분류에 이용하면 분류의 성능을 향상시킬 수 있다. 이러한 관점에서 뎁스 기반 분류방법과 SVM-기반 분류방법을 비교 연구하여 그 성능을 비교하였다.

주요용어: 데이터 뎁스, 분류, 유전자 군집, 유전자 선택.

1. 서론

마이크로어레이 자료에서 중요한 유전자를 찾아내거나 조직 등의 표본을 질병의 유무나 질병의 종류별로 나누기 위하여 군집분석이나 분류분석 등을 이용하는 연구는 여러 가지로 활발하게 진행되고 있다. 전통적인 군집 또는 분류분석 방법으로는 K-평균 (K-means)법, 계층적군집법 (hierarchical clustering), SOM (self organizing map), kNN (k-nearest neighbor) 방법, 서포트벡터머신 (support vector machine, SVM), 신경망 (neural network) 방법 (Bishop, 2006)등이 있다. 이러한 방법들은 주로 이상관측치가 존재하면 영향을 받는 단점이 있어서 로버스트한 분류 및 군집분석의 방법의 개발이 요구되어 왔다. 최근에는 PAM (partitioning around medoids)이나 K-median 방법 등의 로버스트한 방법이 사용되고 있다. 위의 방법 중 PAM은 관측된 자료가 많은 변수를 가지는 마이크로어레이 자료나 잡음이 많은 자료에서는 문제점이 있다고 알려져 있어 이를 보완하기 위하여 관측값 간의 거리를 이용하는 분류(군집)방법인 *sil* Class 방법이 제안되었으나 이 방법은 군집들 간의 분산이 다른 경우에 분류에서 오류를 범하는 문제가 있었다.

로버스트한 자료분석 도구로는 데이터 뎁스 (data depth, Liu 등, 1999)의 개념이 많은 연구되고 있다. 그러나 다차원의 경우에 계산과정의 어려움 때문에 실제 자료 분석에서는 많이 사용되지 못하였다. Vardi와 Zhang (2000)은 이러한 문제를 해결하기위하여 여러 가지 데이터 뎁스 중에서 계산이 편리하고 우수한 로버스트성을 가지는 L1 데이터 뎁스를 제안하였다. 또한 Jornsten 등 (2002)은 L1 데이터

[†] 이 논문은 한국학술진흥재단 기초과학연구지원사업에서 지원되었음(KRF-2004-015-C00075).
[1] 교신저자: (402-751) 인천광역시 남구 용현동 253, 인하대학교 통계학과, 교수.
E-mail: jshwang@inha.ac.kr
[2] (402-751) 인천광역시 남구 용현동 253, 인하대학교 입학처, 전문위원.

뎁스의 개념을 이용하여 여러 군집간의 거리를 나타내는 상대적인 데이터 뎁스 (relative data depth, ReD)를 제안하였다. 이 측도는 군집들 간의 분산차이 때문에 발생하는 오류를 극복하는 척도가 될 수 있다.

Jornsten (2004)은 L1 데이터 뎁스와 ReD를 이용하여 로버스트한 군집분석과 분류분석을 할 수 있는 DDclust와 DDclass라는 방법을 제안하였다. 이 방법은 군집의 수를 결정하거나 군집의 소속을 정할 때 사용하는 거리를 단순한 유클리드 거리 척도를 사용하지 않고 단위 방향벡터들을 이용하여 이러한 단위 방향벡터들의 평균을 이용하였다. 이러한 척도는 근본적으로 이상점의 영향으로부터 벗어날 수가 있게 된다. PAM에서 거리척도로 사용하는 실루엣너비 (silhouette width, sil)는 기본적으로 유클리드 거리 척도를 사용하므로 로버스트성이 약간 떨어진다. 따라서 DDclust에서는 거리척도와 ReD를 결합한 척도를 제시하여 그 평균값들을 최대로 해주는 군집을 생성하게 해준다. 즉 모든 관측값에서 계산한 $(1-\lambda)\cdot sil + \lambda \cdot ReD$ 값들의 평균을 최대로 해주는 방법으로 군집을 형성하는 것이다. 이때 λ 는 0과 1사이의 값을 갖는다. 따라서 위 값은 마치 실루엣너비와 ReD의 가중평균의 형태이다.

로버스트한 뎁스를 분류의 문제에 적용한 것으로는 Christmann (2002)이 Tukey에 의하여 제안된 하프스페이스 회귀뎁스를 이용하여 SVM 방법과 변수의 차원이 높지 않은 상황에서 비교 연구를 수행하였다. 그러나 일반적으로 마이크로어레이 자료와 같이 유전자 수가 많은 상황에서는 적용하기가 어렵다고 알려져 있다.

최근의 마이크로어레이 자료에 대한 분류분석이나 군집분석에 SVM을 이용한 방법이 아주 활발하게 연구가 진행되고 있다. 이 방법은 마이크로어레이 자료뿐 아니라 최근에는 원격 항공사진 이미지 (hyperspectral image) 자료 등에도 적용되어 좋은 성과를 내고 있다. 분류의 성과를 높이기 위한 방안으로 최근에는 변수 선택(유전자 선택 또는 feature 선택)을 이용하여 선택된 변수를 이용한 분류의 방법이 연구되고 있다. 더구나 마이크로어레이 자료는 변수(유전자)의 수가 자료의 수(표본)보다 너무 많기 때문에 변수 선택 또는 유전자 선택의 과정이 선행되는 것이 보통이다. 이러한 변수 선택에서도 SVM의 방법이 적용되고 있다. 최근에는 SVM보다 더 적은 수의 변수를 선택해주는 RVM (relevance vector machine)방법도 사용되고 있다. 변수 선택에서 사용되는 SVM방법으로 요즘 가장 많이 사용되는 것은 SVM-RFE (recursive feature elimination)라고 불리는 방법인데 이 방법은 Guyon (2002)에서 제안한 방법으로서 통계학적으로는 후진변수선택법에 해당한다. 이 방법에서 제거되는 변수의 기준은 SVM의 가중치 계수의 제곱값의 순위에 따른다. 이 방법은 좋은 분류성과를 보여주나 계산 시간이 많이 걸리거나 이상치가 있을 때 로버스트한 성능이 조금 떨어지는 것으로 알려져 왔다. 그 이후로 많은 후속 연구가 진행되어 SVM-RCE (recursive cluster elimination), R(recursive)-SVM 등과 RVM (relevance vector machine)-RFE방법과 Zhou와 Tuck (2007)은 SVM을 확장하여 다중 그룹으로 분류하는 MSVM-RFE 방법을 제안하는 등 이 방면으로 연구가 확장되고 있다. 또한 최근의 연구 결과로서 Shim 등 (2009)이 제안한 SWKC (supervised weighted kernel clustering)와 SVM을 이용한 분류방법도 유전자 선택과 분류를 결합한 새로운 연구로서 주목을 받고 있다.

본 연구에서는 두 그룹으로 나누는 대표적인 SVM 기반 분류방법들과 데이터 뎁스를 기반으로 한 분류분석 방법들을 실제 자료 (Golub, 1999)를 통해서 비교분석을 하여 각자의 장단점을 알아보고자 한다. 2절에서는 비교에 사용되는 분류방법들에 대한 간단한 소개를 하며 3절에서는 비교실험의 결과에 대하여 논의를 하며 끝으로 4절에서는 결론과 추후 연구방향을 제시하고자 한다.

2. SVM과 뎁스 기반 분류방법의 소개

일반적으로 분류란 훈련자료를 이용하여 미리 정해진 여러 그룹으로 나누는 분류자 함수를 만드는 것이다. 분류분석에서는 셋 이상의 여러 그룹으로 분류하는 문제가 일반적이다. 그러나 두 집단으로 분류

하는 것이 기본이 되며 손쉽게 다중 집단으로 확장 할 수 있으므로 본 연구에서는 이진 분류의 상황만 다루고자 한다. 일반적으로 SVM 기반 변수 선택이나 분류 방법은 두 그룹으로 나누는 것에 적합하게 개발이 되었다. 그러나 이를 셋 이상으로 확장하는 방법 역시 많이 개발되어 이용되고 있다. MSVM-RFE 등이 그 예라고 할 수 있다.

2.1. SVM-RFE

이 방법은 Guyon (2002)에서 처음으로 제안하여 그 우수한 성능으로 현재까지 많은 후속 발전 연구가 진행되어온 방법이다. 기본적인 방법의 아이디어는 변수선택에서 후진소거법 (backward elimination)과 같은 것이라고 할 수 있다. 전체 유전자를 다 포함하여 분류자를 만들고 그 중에서 영향이 가장 적은 유전자를 차례로 제거해 나가는 것이다. 영향이 적음의 기준은 SVM 목적함수의 값의 변화를 가정적게 하는 것을 말한다.

<SVM-train 알고리즘>
Input : 훈련패턴 자료 벡터 $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_l}$
집단 소속 벡터 $y_1, y_2, \ldots, y_l$ 여기서 $y_k \in \{-1, +1\}$ 를 나타낸다.
Minimize over α_k :
목적함수 $J = (1/2)\sum_{hk} y_h y_k \alpha_h \alpha_k (\mathbf{x_h} \cdot \mathbf{x_k} + \lambda \delta_{hk}) - \sum_k \alpha_k$.
제약조건 $0 \le \alpha_k \le C$ and $\sum_k \alpha_k y_k = 0$.
Output : 목적함수 J 를 최소로 하는 모수 α_k 추정값

위 훈련 알고리즘의 결과로 얻어지는 결정함수, 즉 분류자는 $D(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + \mathbf{b}$ 이며 여기서 $\mathbf{w} = \sum \alpha_k y_k boldx_k$ 이고 $b = y_k - \mathbf{w} \cdot \mathbf{x}$ 로 주어진다. 이때 많은 α_k 값이 0이 되고 0이 아닌 훈련 벡터들이 서포트벡터가 된다. 즉 가중치 벡터 $\mathbf{w}$ 는 서포트벡터의 선형결합으로 주어지며 편의를 나타내는 b 는 $0 < \alpha_k < C$ 인 서포트벡터의 평균이다.

<SVM-RFE 알고리즘>

1. 전체 훈련자료에서 SVM-train 알고리즘으로 $\alpha = SVM - train(X, y)$ 를 구한다.
2. 가중치 벡터 $\mathbf{w} = \sum \alpha_k y_k \mathbf{x_k}$ 를 계산한다.
3. 가중치 벡터의 성분들의 제곱값 w_i^2 중 작은 유전자를 제거한다.
4. 나머지 자료를 바탕으로 SVM-train 알고리즘으로 새로운 α 를 구한다.
5. 위의 2-4 과정을 계속 반복한다.

보통 유전자 자료에서는 유전자 수가 많으므로 알고리즘에서 제거되는 유전자를 1 개 이상으로 설정할 수 있다.

2.2. SVM-RCE

이 방법은 유전자에 대한 군집을 K-평균법으로 한 다음에 군집들의 분류에 기여하는 점수에 따라서 군집들을 차례로 제거하는 방법이다. Yousef 등 (2007)에 의해서 제안된 방법으로 SVM-RFE에 비하여 계산시간 단축 및 정확성이 높다는 실험결과를 제시하고 있다. 군집 점수가 높은 군집들에 포함되는 유전자들을 이용하여 SVM 분류법을 이용한 분류를 한다.

<SVM-RCE 알고리즘>

1. 원 자료를 90%의 훈련자료와 10%의 검증자료 (test set)로 나눈다.
2. 훈련자료에서 K-평균법을 이용하여 유전자를 n 개의 군집 $S_1, S_2, \ldots, S_n$ 으로 나눈다.
3. 각 군집의 점수 산정(SVM-score) : 군집내의 자료를 가지고 SVM 분류자를 이용하여 f -fold cv를 r 회 반복하여 군집의 점수를 구한다.
4. SVM-score를 이용하여 군집의 순위를 정하고 가장 적은 d %의 군집을 제거한다.
5. 남은 군집의 유전자 자료를 바탕으로 10%의 검증자료에 대하여 SVM 분류자로 그 성능을 평가한다.
6. 남은 군집을 다시 통합하고 1-3 과정을 미리 정한 최적 군집 수가 될 때 까지 반복한다.

이 방법은 SVM-RFE 방법에 비하여 유전자들 간의 상관계수를 고려한다는 점이 다르다고 할 수 있다. 또한 군집을 K-평균법으로 나눌 때 상관계수 거리를 사용하였는데 다른 거리를 사용하거나 로버스트한 방법으로 군집을 하려면 다음에 소개되는 DDclust 방법을 활용하여 군집을 정하는 것도 고려할 수 있다.

2.3. R-SVM(Recursive SVM)

앞의 SVM 기반 방법들은 반복적으로 SVM을 이용하여 표본을 분류하고 그 분류자의 가중치에 따라서 유전자를 선택하는 방법이다. Zhang 등 (2006)에 의하여 제안된 R-SVM 방법은 SVM-RFE 방법과 유사하게 반복적으로 SVM을 이용하여 유전자를 선택하고 분류하지만 SVM-RFE 방법과는 달리 중요한 유전자를 선택하고 평가하는 기준이 다르다.

R-SVM 방법에서 유전자 j 가 두 집단으로 구별하는데 기여하는 점수는

$s_j = w_j(m_j^+ - m_j^-)$ 로 주어진다. 여기서 w_j 는 SVM 에서의 j 번째 가중치 벡터를 나타내며 m_j^+, m_j^- 는 해당 유전자의 + 집단과 - 집단 발현 값들의 평균을 나타낸다. 이 값은 SVM-RFE에서 사용하는 w_j^2 에 비하면 각 유전자가 두 집단으로 분별하는 정도의 추정치 $m_j^+ - m_j^-$ 가 추가된 것으로 생각할 수 있다.

<R-SVM 알고리즘>

1. 먼저 연속적인 선택 단계에서 선택할 유전자 개수의 리스트를 내림차순으로 정한다. 즉, $d_0 > d_1 > \cdots > d_k$ 이며 첫 단계인 $d_0 = d$ 로 전체 유전자 수를 의미한다.
2. i 번째 단계에서 d_i 개의 유전자를 이용하여 SVM 분류함수를 만든다. 단 이 때 선택하는 d_i 개의 유전자는 여러 가지 교차타당성(CV)을 측정할 수 있는 재표집 방법을 이용하여 선택된 유전자들 중에서 가장 많이 선택된 d_i 개를 택한다.
3. 각 유전자를 유전자별 기여점수 s_j 를 이용하여 순위를 정하고 앞에서부터 $d_i + 1$ 개의 유전자만을 선택한다.
4. 단계를 $i + 1$ 로 증가시키고 2-3 과정을 $i = k$ 까지 반복한다.

이 방법은 SVM-RFE 방법에 비하여 이상치가 있는 경우에 좋은 경향을 보인다고 모의실험을 통하여 밝히고 있다. SVM 기반 방법들과 같은 다변량의 분류기법이 일반적으로 오분류의 확률을 최적으로 해주지만 중요한 유전자를 선택하는 점에서는 일변량 방법인 가중 투표법 (weighted voting)에 비하여 약간 뒤지는 결과를 준다. 유전자 별 점수를 계산하는데 있어서 두 그룹의 평균값을 사용하는 문제는 로버스트한 관점에서 재고가 필요하다.

2.4. Depth Based Classification(DDclass)

Vardi와 Zhang (2000)에서 제안된 L1 데이터 뎁스와 군집간의 상대적인 거리 측도인 ReD를 활용한 DDclass는 유전자 자료에서도 좋은 성능을 보임이 Jornsten (2004)에 의하여 밝혀졌다. 이 방법은 로버스트한 방법 중의 하나인 *sil* Class와의 비교실험을 하였으나 본 논문에서는 SVM 기반 방법들과의 폭 넓은 비교를 하였다.

유전자수 p , 표본수 N 인 자료 $x_1, \ldots, x_N in R^p$ 이 주어졌을 때 $\bar{d}(x_i|k)$ 는 x_i 로부터 군집 k 내의 모든 관측값까지 거리의 평균을 나타낸다고 하자. 군집수가 K 일 때 임의의 점 $z \in R^p$ 의 군집 k 와의 L1 데이터 뎁스는

$$D(z|k) = 1 - \max[0, ||\bar{e}(z|k)|| - f(z|k)]$$

로 정의한다. $D(z|k)$ 값이 1에 가까우면 z 는 군집 k 의 중앙에 있는 것이며 0에 가까우면 군집 k 의 중앙에서 멀리 떨어져 있음을 나타낸다. $\bar{e}$ 는 z 로부터 군집 k 에 속하는 모든 자료와의 단위방향벡터의 평균을 나타낸다. 따라서 $\bar{e}$ 가 0에 가까우면 z 가 군집의 중앙부분에 있음을 나타내는 것이며 1에 가까우면 군집에서 떨어져있음을 나타낸다. 그리고 f 는 자료가 겹치는 정도를 나타낸다. 자세한 정의는 Jornsten (2004)의 논문이나 백수진 등 (2006)의 논문을 참고하면 된다.

<DDclass 알고리즘>

1. 훈련자료 TR 에서 새로운 leave-one-out(LOO) 훈련자료 TR_b , $b = 1, \ldots, N$ 생성한다.
2. 각 b 에서
   (i) 관측치 b 에서 각 소속 군집 $k(= 1, \ldots, K)$ 별 L1 뎁스 $D(b|TR_b^k)$ 를 계산하고
   (ii) b 의 소속을 $\hat{k}_b = \arg\max_k D(b|TR_b^k)$ 로 예측한다.
3. 집합 $T = \left\{ b : b \in TR, k(b) \neq \hat{k}_b \right\}$ 를 구한다. 여기서 $k(b)$ 는 b 의 정확한 소속 군집을 나타낸다.
4. 축소된 훈련자료 생성한다. $TR^* = TR\_T$
5. 검증자료 TE에 포함된 자료 j 에 대하여 $\hat{k}_j = \arg\max_k D(j|TR^{k*})$ 로 예측한다.
6. 자료 j 에 대하여 상대 데이터 뎁스 ReD_j^{test} 를 계산한다.

상대 데이터 뎁스 ReD_j^{test} 는 다음과 같이 구한다.

$$\mathrm{ReD}_j^{test} = D(j|\hat{k}_j) - \max_{l \neq \hat{k}_i} D(j|l).$$

단 $\hat{k}_j = \arg\max_i D(i|l),\ l \in \{1, \ldots, K\}$ 이다.

백수진 등 (2006)에서는 L1 데이터 뎁스의 장점과 일반적인 L1 거리를 조합한 새로운 분류방법인 DnDClass를 제안하고 그 상대적인 장단점을 연구하였다. 즉, 검증자료 x_j 를 DnDClass에서는 다음의 집단

$$\hat{k}_j = \arg\min_k \left\{ (1-\lambda)\bar{d}(j|TR^k) - \lambda D(j|TR^k) \right\}, \lambda \in [0,1],$$

로 분류한다. 이때 λ 는 CV 방법을 통하여 적당한 값을 찾는다.

3. 실험 결과

본 연구에서는 크게 SVM 기반 분류법 (SVM-RFE, R-SVM, SVM-RCE)과 데이터 뎁스 기반 분류법 (DDClass, DnDClass) 두 가지의 분류기법을 비교하였다. SVM 기반 분류 기법은 유전자를 선택하는 방법에 따라서 구별이 되며 뎁스 기반 분류법에서는 분류시 L1 뎁스만을 고려한 DDClass 방법과 L1 뎁스와 거리를 함께 고려한 DnDClass 방법 두 가지를 고려하였다. 뎁스 기반 방법에서 유전자의 선택은 여러 방법이 있지만 비교 편의상 SVM-RFE 방법에서 선정된 유전자를 바탕으로 하였다.

3.1. 실험 자료

비교에 쓰인 실제 자료는 Golub (1999) 등이 사용한 백혈병 자료로 일반적인 분류분석에서 보편적으로 널리 쓰이는 자료 중의 하나이다. 이 자료는 유전자 분석을 통하여 새로운 그룹을 찾아내거나 임의의 샘플에 대하여 어떤 그룹에 속하는지를 예측하려는데 사용되었다. 백혈병의 두 가지 종류로 알려진 ALL 타입과 AML 타입이 있는데 주어진 자료는 72명의 환자에게서 채취한 62개의 골수 샘플과 10개의 혈액 샘플들로 이루어져 있다. 원 자료에는 6817개의 유전자 자료가 있으나 표준화 과정 등을 통하여 3571개가 되었다. 따라서 자료의 형태는 행을 유전자 열을 샘플이라 한다면 3571×72 행렬 자료가 된다.

3.2. 결과 정리

방법별로 훈련자료와 검증자료로 나누어 각 분류자들의 예측 정확도를 반복 측정하여 정리한 자료가 표 3.1 과 표 3.2에 있다. 여러 방법 중에 SVM-RCE 방법은 군집의 수를 선택하는 것이므로 유전자 수를 동일하게 할 수 없다. 따라서 동일한 유전자 수를 이용한 방법이 표 3.1에 정리하였으며 가능한 한 다른 방법들과 유사한 수의 유전자를 가지는 군집을 선택하여 별도로 표 3.2에 정리하였다. 물론 유전자 수는 동일하지만 유전자 종류는 일반적으로 달라진다. 즉 RFE 방법과 R-SVM 방법에서 유전자를 선택하는 방법은 서로 다르다.

일반적으로 모든 유전자를 다 사용하지는 않고 t-test 방법이나 분산이 큰 유전자의 순서대로 적당한 개수만큼만 이용하여 분류에 사용한다. 여기서는 일단 RFE 방법으로 선정된 유전자를 사용하여 뎁스 기반 분류법에 사용하였다. 이러한 부분에 대해서는 추후 여러 방향으로 논의가 진행되어야 할 것으로 생각한다.

원래 자료는 47개의 ALL 타입과 25개의 AML 타입의 자료로 구성되어 있다. 각 방법별 예측력은 47개의 ALL 중 정확한 예측수(LL), 25개의 AML 중 정확한 예측수(MM), 그리고 72개 전체에서의 정확한 예측수(LL+MM)를 이용하여 나타내었다. 표 3.1에서는 LOO-CV 방법을 이용하여 전체 중에 정확하게 예측한 횟수를 나타내었고 표 3.2에서는 10-fold CV를 10회 반복한 평균 예측정확도와 표준편차를 표시하였다.

4. 토의 및 결론

본 연구에서는 뎁스 기반 분류방법과 선형 SVM 기반 방법들을 비교하였다. 훈련 자료가 선형함수로 분리 가능하다면 선형 SVM은 최대마진분류자가 된다. 비선형인 경우나 일반적인 커널 함수로의 확장은 약간의 계산이 추가되지만 어렵지 않게 확장될 수 있음을 Guyon (2002)에서 보여주고 있다.

앞의 결과를 보면 유전자 수를 감안한다면 모든 방법이 많은 유전자 1000개 정도를 가지면 거의 정확하게 예측을 하고 있다. 그러나 유전자 수가 적어질수록 특히 뎁스 기반 방법은 그 성능이 떨어짐을 알 수 있다. 표 3.1을 보면 R-SVM 방법이 다른 방법에 비하여 우수하지만 그 결과를 표 3.2의 RCE 방법

표 3.1 사용된 유전자 수에 따른 각 분류방법별 예측도

사용한 유전자수	방법별 예측도 (R-SVM, RFE, DD, DnD)		
	LL	MM	LL+MM
16	43,39,31,32	21,13,15,16	64,52,46,48
32	44,37,36,35	23,15,12,14	67,52,48,49
64	47,40,37,38	24,18,19,19	71,58,56,57
128	46,43,47,45	24,23,18,21	70,66,65,66
256	45,46,47,44	24,24,21,21	69,70,68,65
512	46,47,47,45	24,24,22,23	70,71,69,68
1024	46,47,47,46	24,24,23,23	70,71,70,69
2048	47,46,46,46	24,24,24,24	71,70,70,70
3571	47,47,46,46	24,24,24,24	71,71,70,70

표 3.2 SVM-RCE 방법에서 사용된 군집과 유전자 수에 따른 예측정확도

군집수	유전자수	평균	표준편차
2	29	0.975179	0.060039
3	42	0.961726	0.069371
6	68	0.954762	0.073974
16	128	0.960000	0.071057
30	200	0.971131	0.062683

과 비교하면 비슷한 유전자 수에서 RCE 방법이 더 정확한 결과를 줌을 알 수 있다. 보다 엄밀한 성능 비교를 위해서는 반복을 통한 신뢰구간 또는 통계적인 검정이 실시되어야 하지만 본 연구에서는 특정한 방법의 우수성을 알리려는 것 보다는 여러 방법들의 제시와 이에 대한 일차적인 분석 결과에 중점을 두었으므로 차후에 보다 엄격한 통계적인 비교 검증이 필요하다고 할 수 있다.

뎁스 기반 방법의 성능은 본 자료에서는 별로 인상적이지 않다고 할 수 있다. 그러나 유전자를 많이 사용하면 동일한 결과가 나오므로 유전자 선택의 문제를 좀 더 뎁스 기반 방법에 적합한 방법으로 연구한다면 결과가 향상될 수 있다고 생각한다. 또한 현 상태의 자료는 이상치 등이 없어서 뎁스 기반 방법의 특성이 나타나지 않았다.

본 연구에서는 동일한 자료를 가지고 가능하면 동일한 실험조건하에서 공정하게 여러 방법들을 비교하고자 하였다. 현재는 L1 데이터 뎁스를 이용한 로버스트한 분류 방법과 SVM 방법의 장점인 유전자 선택을 결합하는 여러 방법에 대한 연구를 진행하고 있다. 또한 로버스트한 군집 방법을 활용하여 SVM 기반의 분류 방법을 적용하는 것도 의미 있는 연구가 될 것으로 기대하고 있다.

참고문헌

백수진, 김진경, 황진수 (2006). L1 거리와 L1 데이터 뎁스를 이용한 분류방법의 비교연구. <응용통계연구>, **19**, 183-193.

Bishop, C. (2006). *Pattern recognition and machine learning*, Springer, New York.

Christmann, A. (2002). Classification based on the support vector machine and on regression depth. In *Statistical data analysis based on L1-norm and related methods,* Ed. Y. Dodge, 341-352, Birkhäuser, Boston.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J. and Caligiuri, M. (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.

Jornstern, R. (2004). Clustering and classification based on L1 data depth. *Journal of Multivariate Analysis*, **90**, 67-89.

Jornsten, R., Vardi, Y. and Zhang, C.-H (2002). A robust clustering method and visualization tool based on data depth, In *Statistical data analysis based on the L1-norm and related methods*, Ed. Y. Dodge, 313-366, Birkhäuser, Boston.

Liu, R., Parelius, J. and Singh, K. (1999). Multivariate analysis by data depth : descriptuve statistics, graphics and inference (with discussion). *The Annals of Statistics*, **27**, 783-858.

Seok, K. H. (2007). Semi-supervised learning using kernel estimation. *Journal of Korean Data & Information Science Society*, **18**, 629-636.

Seok, K. H., Hwang, C. H. and Cho, D. H. (2002). On approximate prediction intervals for support vector machine. *Journal of Korean Data & Information Science Society*, **13**, 65-75.

Shim, J., Sohn, I., Kim, S., Lee, J., Green, P. and Hwang, C. (2009). Selecting marker genes for cancer classification using supervised weighted kernel clustering and the support vector machine. *Computational Statistics and Data Analysis*, **53**, 1736-1742.

Vardi, Y. and Zhang, C. (2000). The multivariate L1 median and associated data depth. *Proceedings of the National Academy of Sciences*, **97**, 1423-1426.

Yousef, M., Jung S., Showe, L. and Showe, M. (2007). Recursive cluster elimination (RCE) for classification and feature selection from gene expression data. *BMC Bioinformatics*, **8**, 144.

Zhang, X., Lu X., Shi, Q., Xu, X., Leung, H., Harris, L., Iglehart, J., Miron, A., Liu, J. and Wong, W. (2006). Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, **7**, 197.

Zhou, X. and Tuck, D. P. (2007). MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics*, **23**, 1106-1114.

Journal of the Korean Data & Information Science Society
2009, **20**(2), 311–319

한국데이터정보과학회지

A comparison study of classification method based on SVM and data depth in microarray data[†]

Jinsoo Hwang[1] · Jeeyun Kim[2]

[1]Department of Statistics, Inha University
[2]Admission Officer, Inha University

Received 19 January 2009, revised 5 March 2009, accepted 12 March 2009

Abstract

A robust L1 data depth was used in clustering and classification, so called DDclust and DDclass by Jornsten (2004). SVM-based classification works well in most of the situation but show some weakness in the presence of outliers. Proper gene selection is important in classification since there are so many redundant genes. Either by selecting appropriate genes or by gene clustering combined with classification method enhance the overall performance of classification. The performance of depth based method are evaluated among several SVM-based classification methods.

Keywords: Classification, data depth, gene clustering, gene selection.

[†] This research was supported by Korean Research Foundation Grant(KRF-2004-015-C00075).
[1] Corresponding author: Professor, Department of Statistics, Inha University, Incheon 452-751, Korea. E-mail: jshwang@inha.ac.kr
[2] Admission Officer, Inha University, Incheon 452-751, Korea.