Support Vector Machine을 이용한 Microarray Gene Expression Data의 분류

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
구 경 민

Support Vector Machine을 이용한 Microarray Gene Expression Data의 분류

지도 변 혜 란 교수

이 논문을 석사 학위논문으로 제출함

2001년 12월 일

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
구 경 민

구경민의 석사학위 논문을 인준함

심사위원 ______________ 인

심사위원 ______________ 인

심사위원 ______________ 인

연세대학교 대학원

2001년 12월 일

감사의 글

찬바람에 가슴속까지 추워지는 이 시기에 조금은 따뜻한 마음을 가질 수 있었던 것은 미숙한 와중에 완성된 논문 때문인 것 같습니다.

천방지축인 저를 묵묵한 인내와 지속적인 가르침으로 지도해주신 김동기 교수님께 감사드립니다. 또한, 바쁘신 와중에도 친절한 지도로 이끌어주신 변혜란 교수님, 모자란 제자를 위해 아낌없는 도움을 주신 김동건 교수님, 조진남 교수님께도 감사의 말씀을 드립니다.

철없는 장녀인 저를 항상 깊은 사랑으로 감싸주시고 키워주셨던 아버님, 어머님께 다시 한 번 감사의 말씀을 드리며, 이제 저를 돌보아주신 은혜에 대한 보답을 시작해 드리고 싶습니다. 그리고 지금은 미국에 있지만, 항상 잘 챙겨주고 걱정해주는 나의 하나 뿐인 동생 동훈이에게도 고맙게 생각합니다. 또한, 논문 학기 내내 나의 투정과 괴롭힘을 다 받아준 사촌동생 정연에게도 고마운 마음을 정합니다.

아무 것도 모르던 서투른 시절, 저를 이끌어 주신 임종건 선생님, 희중오빠, 시내언니, 희철오빠, 우선오빠, 영진씨, 나의 물음에 항상 진지하게 답변해주었던 기준오빠, 모자람을 친절함으로 가르쳐주었던 윤주, 하나뿐인 동기이자 2년 내내 뒤치닥거리에 힘들었을 성민오빠, 밝은 웃음으로 반겨주던 현지엄마 효미, 같은 층에서 함께 추위와 씨름하며 우울함을 간식으로 달랬던 민지, 또한 한 연구실에서 지냈던 찬미, 은정, 무영, 원열에게도 고맙다는 말을 전합니다.

유학 중이지만 어려서부터 함께 한 자매 같은 지은, 친구보다 더 가까운 나영, 항상 논문을 걱정해주던 친구 문정, 은경, 요즘은 자주 만나지 못

했지만 언제나 주말을 함께 보냈던 서희, 지윤, 희정, 나의 졸업을 누구보다도 바랬고 걱정해주던 철없는 나의 동갑 친구 혜원, 나의 구박을 항상 웃음으로 지나쳤던 혜숙, 논문쓰고 일하느라 바빴던 재희, 새로운 전공에 열심인 지현, 효진, 일하느라 바쁜 경희, 유부녀 지현, 지영에게도 고마움을 전합니다.

2002년 겨울 끝자락에...

구경민 드림

제 목 차 례

표 차 례

그 림 차 례

국문요약

Support Vector Machine을 이용한 Microarray Gene Expression Data의 분류

.

본 논문에서는 microarray 자료에서 SVM(Support Vector Machine) 기법을 적용한 이원분류기법을 소개하고, 또한 모의실험을 통하여 얻은 microarray 자료를 이용하여 각 커널함수(kernel-function)에 따른 SVM에 대하여 평가하였다. 본 논문의 주 관심인 이원 분류기로서 SVM 성능을 평가하기 위해 microarray 자료를 이용하였는데, 이 microarray 자료는 통계프로그램(S-PLUS)을 이용하여 생성하였다. 또한 SVM 이원 분류기의 성능을 평가하기 위해서 Chen이 제시한 R-Package에 있는 SVM 라이브러리를 이용하였다.

결론적으로, 로그비가 높을수록 각각의 평가항목들(민감도, 특이도, 양성예측도, 음성예측도, 정분류율, 오분류율)이 더 좋게 나타났는데, 이것은 microarray 실험에서 Cy5의 강도가 높게 나타날수록 분류를 더욱 정확하게 한다는 것을 말하는 것이다. 그러나 커널함수와 분류방법에 따라서는 별 차이가 없다는 것을 알 수 있다. 또한, 표준편차가 높을수록 각각의 평가항목들이 더 낮게 나타났는데, 이것은 microarray 실험에서 오차가 높게 나타날수록 분류율이 더 낮아진다는 사실을 말하는 것이다.

핵심되는 말 : Support Vector Machine, Microarry, Kernel-function, SVM classifier

제 1 장 서 론

Human Genome Project의 10년간의 연구 결과로 우리는 인간이 지니고 있는 30억 개의 DNA 염기서열을 모두 해독하게 되었으며 이는 생명공학의 급속한 발전과 함께 한 결과이다.

종래의 생명과학은 이러한 유전자 발현의 변화를 일대일의 관계로부터 찾는 것이 관례였다. 즉 A 단백질의 발현 변화는 B 단백질의 발현 변화를 초래하고 이는 또 다시 C 단백질의 발현 변화를 유도한다는 것을 가정하고 이를 증명함으로서(hypothesis-driven) 유전자 발현 경로와 병리기전을 찾아내는 것이다. 그러나 특정 단백질의 작용이 하나의 단백질에 국한되는 것만은 아니며 다양한 단백질의 발현에 관여한다고 생각할 때 이러한 접근은 매우 제한적일 수밖에 없다. 이러한 이유에서 최근에는 수천 개, 수만 개의 유전자 발현을 일시에 검증하고 그 결과를 토대로 생물학적 의미를 찾아내는(discovery-driven) 효율적인 유전자 발현 검색 시스템이 개발되었으며, 이중 가장 보편적인 것이 DNA microarray나 SAGE(serial analysis of gene expression)이다.

질병의 진단에 있어 세포의 형태학적 소견이 중요한 근거이나 실제 세포의 형태학적 소견 변화에 앞서 이를 결정짓는 유전자 발현의 변화가 선행된다. 따라서 DNA microarray를 이용하여 유전자 발현을 대규모로 검증할 경우 질환의 발생에 선행되어 나타나는 특정 유전자의 발현 변화, 또는 특정 표현형(phenotype)의 발달과 관련된 유전자 발현 변화 등을 찾아낼 수 있을 것으로 기대하고 있다. 또한 DNA microarray 실험결과를 통하여 특정 유전자 발현 경로가 임상적으로 중요한 의미를 지니고 있는 것으로 판

단될 경우, 이를 target으로 하는 새로운 약물의 개발이나 평가가 이루어질 수 있다.

SVM은 1995년에 Vladmir Vapnik에 의해 이원 패턴 인식 문제를 해결하기 위해 제안된 학습방법으로 부정예제로부터 긍정예제를 분류해 낼 수 있는 결정면을 찾아내는 분류모형이다(Vapnik 1999). 최근에는 SVM을 응용한 연구영역이 분자 인식, 얼굴 인식 등의 패턴 인식 분야뿐만 아니라 문서 범주화나 microarray 데이터 등으로 확장되고 있다.

SVM은 이원 패턴 분리를 위한 알고리즘으로 개발되었기 때문에 k개의 범주를 학습시키기 위해서는 여러 개의 SVM 분류기를 조합해서 분류 업무를 수행하여 다원 분류기를 확장하게 된다. 패턴 인식 분야에서는 SVM 이원 분류기를 다원 분류기로 확장하는 것과 관련된 다양한 연구가 수행되고 있으나, Microarray 데이터 분야에서는 SVM 다원 분류기를 중심으로 고찰하는 연구가 거의 수행된 적이 없기 때문에 그 성능을 분석해 볼 필요가 있다.

SVM에서는 학습과정이 이루어지기 전에 사용자가 직접 모수(parameter)값을 결정해야 하며 모수 값에 따라 SVM의 성능은 달라지게 된다. 이렇게 미리 결정해야 하는 모수는 학습과정에서 마진폭과 분류 오류 사이의 타협점(trade-off)을 찾아주는 오류 패널티(penalty) 변수 C 값과 비선형 SVM에 적용되는 커널함수의 모수이다. 본 연구에서는 실험에 적합한 C 값과 ν 값을 결정하고, 이를 바탕으로 다항식 커널함수 및 RBF 커널함수를 적용한 비선형 SVM이 분류 성능을 비교 분석하였다.

제 2 장 SVM의 기본 원리

2.1 통계적 학습이론

$x_i \in R^n \quad i = 1, \cdots, l$

$(x_i, y_i), \cdots, (x_l, y_l) \qquad y \in \{+1, -1\}$

x_i는 n차원의 벡터이고, y_i는 각 클래스에 대한 인덱스이다. 이러한 과정의 학습을 통해 얻어진 f 는 새로운 데이터 x를 $\{+1, -1\}$중의 하나로 분류하는 것을 VC(Vapnik Chernonenkis) 이론이라 한다.

$$R(\alpha) = \int \frac{1}{2} |y - f(x, \alpha)| dP(x, y)$$

여기서 $P(x,y)$는 x와 y의 확률을 의미하고, $R(\alpha)$는 검증 집단상에서 나타날 수 있는 검증오류(test error)로서 리스크 함수(risk function) 또는 예상 리스크(expected risk)라고 부른다.

학습과정에서 얻어진 함수 f 에 대하여 어떤 제약조건을 지니게 하지 않는다면, 학습되어지지 않은 데이터에 대한 일반화를 보장할 수 없다. 그러므로 학습기계(trained machine)의 성능 오차(test error)를 측정하기 위해서 $R(\alpha)$를 최소로 하는 함수 α를 찾는다.

통계적 학습이론의 목적은 확률분포 $P(x,y)$와 학습 집단에 포함된 정보를 이용하여 예상 리스크를 최소화하는 함수 $f(x)$를 찾아내는 것이다 (Vapnik 1999).

2. 1. 1 경험적 리스크(empirical risk) 최소화 원칙

일반적으로 예상 리스크 $R(\alpha)$를 추정하기 위해서 필요한 확률분포 $P(x,y)$는 미지의 분포이기 때문에 $R(\alpha)$를 계산하는 것은 어려운 일이다. 따라서, 경험적 리스크(empirical risk) $R_{emp}(\alpha)$은 학습집단(training set)에서의 오차율(error rate)의 계산을 위해 사용된다.

$$R_{emp}(\alpha) = \int \frac{1}{2l} \sum_{i=1}^{l} |y_i - f(x_{i,} \alpha)|$$

경험적 리스크인 $R_{emp}(\alpha)$은 확률분포 값인 $P(x,y)$가 포함되어 있지 않기때문에 손실값을 $\frac{1}{2}|y_i - f(x_i, \alpha)|$을 계산할 수 있으며, 예상 리스크를 최소화시킬 수 있다(Vapnik 1999).

경험적 리스크 최소화의 원칙은 $R_{emp}(\alpha)$이 $R(\alpha)$로 수렴하면, $R_{emp}(\alpha)$의 최소 값도 $R(\alpha)$의 최소 값으로 수렴한다는 것을 내용으로 한다.

$$\lim_{l \to \infty} R_{emp}(\alpha) = R(\alpha)$$

$$\lim_{l \to \infty} \min R_{emp}(\alpha) = \min R(\alpha)$$

경험적 리스크 최소화 원칙은 학습집단을 대상으로 경험적 리스크를 최소화시킴으로서 미지의 확률분포를 추정하기 때문에 학습집단의 크기와 정확도에 영향을 받게 된다. 또한 수렴의 경계범위를 나타내기 위해서 VC 차원(Vapnik-Chernonenkis dimension)을 도입하였다.

2. 1. 2 VC 차원(Vapnik-Chernonenkis dimension)

VC(Vapnik-Chernonenkis) 이론에서는 학습집단을 적합한 용량(capacity)으로 학습할 수 있는 함수의 클래스에 대한 제한을 하는 것이 필수적이다. 이를 위해 경험적 리스크와 함수 클래스의 용량을 모두 고려한 검증오류에 대한 경계범위(bounds)를 도입하였으며, 함수의 다양한 클래스를 측정하기 위해 VC 차원을 이용하여 경계범위를 나타내게 된다.

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\left(\frac{h(\log(2l/h) + 1) - \log(\eta/4)}{l}\right)}$$

여기서 h는 양수(non-negative)인 VC 차원이고, l은 학습 데이터의 수, η는 신뢰도(confidence)를 의미하며 $0 \leq \eta \leq 1$의 범위를 갖는다.
위의 식에서 우변은 리스크의 경계범위(risk bound)를 나타내며,
$\sqrt{\left(\frac{h(\log(2l/h) + 1) - \log(\eta/4)}{l}\right)}$는 VC 신뢰도를 나타낸다.

만약에 학습데이터의 주어진 양을 위해 자료가 복잡한 디자인을 하면, VC 신뢰도는 커지고 $R_{emp}(\alpha)$가 0까지 가서 적어진다 하더라도 검증집합(test set)에서의 오차(error)는 여전히 크다. 따라서 $R(\alpha)$을 최소화하기

위해서는 $R_{emp}(\alpha)$과 $\frac{h}{l}$를 동시에 줄임으로써, 우변의 항을 동시에 최저로 만드는 최적의 VC 차원의 학습모형을 얻어야 한다(Vapnik 1999).

2. 1. 3 구조적 리스크(structural risk) 최소화 원칙

기존의 학습 알고리즘은 학습집단을 이용하여 경험적 오차(empirical error)를 최소화하는 경험적 리스크 최소화 원칙(empirical risk minimizat-ion : ERM)을 구현하는 것을 목적으로 했다. 그러나 이러한 학습 알고리즘은 학습집단의 크기와 정확도에 많은 영향을 받게 되어 수집할 수 있는 데이터의 범위가 큰 경우에는 적합하지 않다.

Vapnik과 Chervonenkis(Vapnik and Chervonenkis, 1974)에 의해 경험적 리스크 최소화의 원칙의 단점을 보완하면서 최적의 VC 차원을 선택하기 위해 특정 데이터의 근사치에 대한 정확도(quality)와 경험적 리스크 함수의 복잡도(complexity) 사이의 타협점(trade-off)을 찾아주는 구조적 리스크 최소화의 원칙(structural risk minimization : SRM)이 제안되었다.

SRM은 전체 집단을 하위 집단으로 세분화 한 뒤 이 집단들에 대한 $R_{emp}(\alpha)$을 계산하여 이를 최소화하는 함수를 선택하는 것을 원칙으로 한다.

$$S_1 \subset S_2 \subset \cdots \subset S_n \cdots$$

VC 차원은 아래와 같다.

$$h_1 \le h_2 \le \cdots \le h_n \cdots$$

각 모형에 대한 경험적 리스크를 계산한 후 경계범위의 상한값을 최소로 하는 모형을 선택하게 된다.

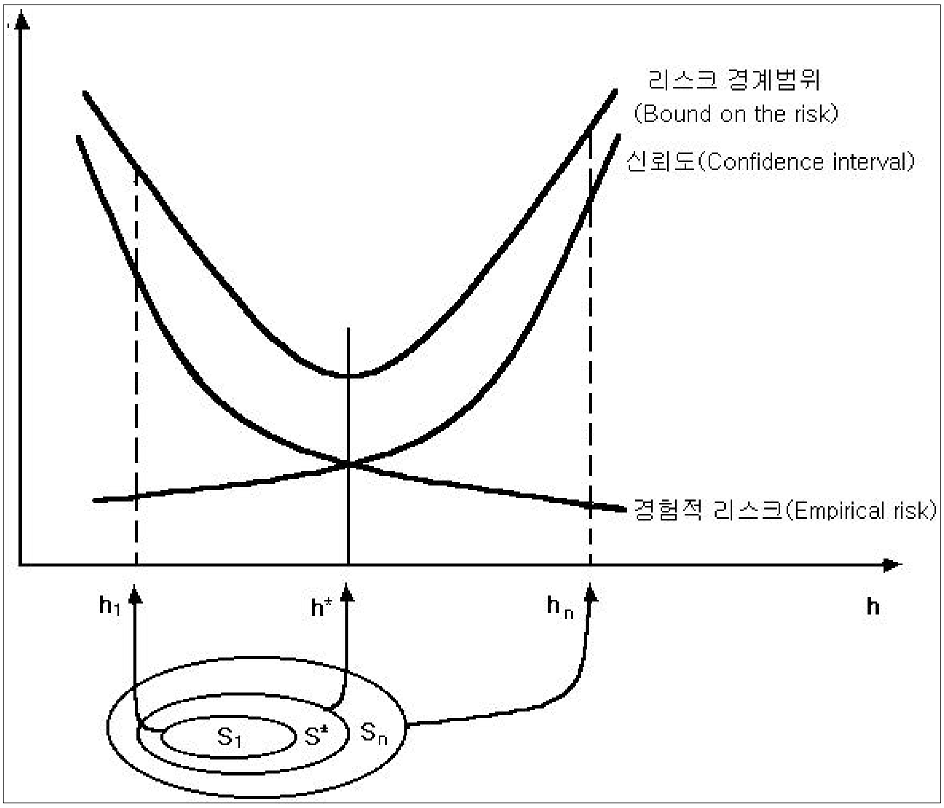


그림 2-1. 리스크의 경계범위

그림 2-1은 리스크에 대한 경계범위를 보여주고 있으며, 이때 경계범위는 경험적 리스크와 신뢰도의 합으로 나타낸다. 즉, VC 차원이 증가함에 따라 경험적 리스크는 계속 감소하고 반대로 신뢰도는 증가하게 된다. 그러나 실제 오류는 최적의 VC 차원 지점까지 감소하다가 다시 증가하게 되는데, 이 최소 지점이 최적의 모형에 해당하는 VC 차원(그림에서 h^*)이 된다(Vapnik 1999).

2. 2 SVM의 기본개념

SVM 분류기 구축을 위해서는 입력데이터 x와 지지벡터(support vector) x_i는 비선형 함수 Φ를 이용해서 $\Phi(x)$, $\Phi(x_i)$로 대체된 후, 특징공간(feature space)으로 변형된다 ($\Phi(\cdot) : R^n \rightarrow R^p$ ($p \gg n$)). 이 때, 커널함수 $K(x, x_i)$를 이용하면 입력공간(input space)과 특징공간에서의 역할이 한번에 처리되어 계산상의 어려움을 덜 수 있다(Scholkopf et al. 1999).

SVM은 선형적으로 분리할 수 있는 학습집단에 대해서는 최대 마진 분류기를 구축하는 선형 SVM을 이용하고, 선형적으로 분리할 수 없는 경우에는 커널함수를 이용한 비선형 SVM으로 분류한다.

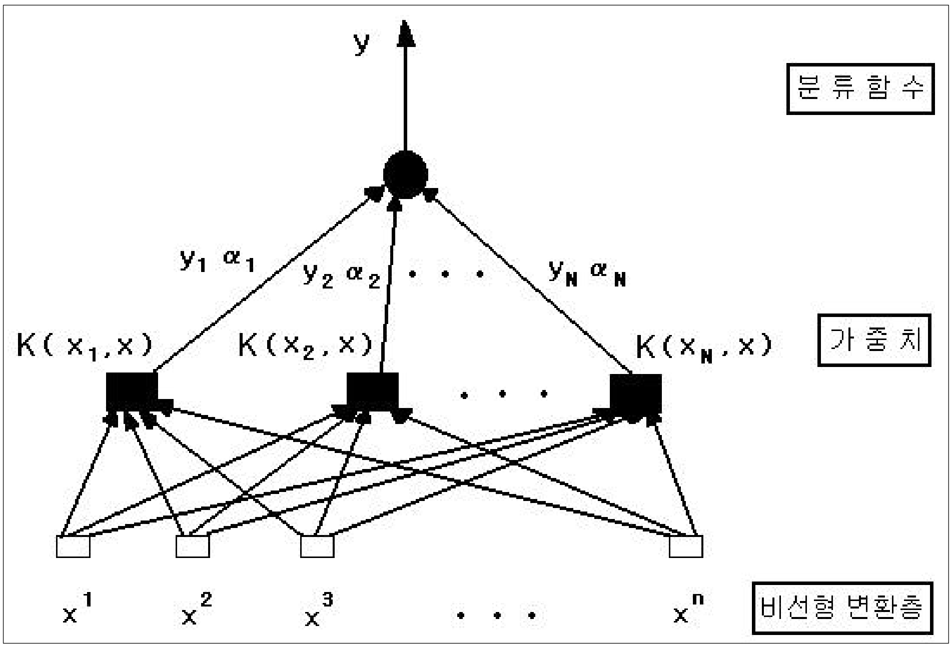


그림 2-2. SVM의 구조

2. 2. 1 선형 SVM (Linear Support Vector Machines)

SVM은 Vladimir Vapnik에 의해 처음 고안되었는데 그림 [2-3]처럼 선형 분리가 가능한 두 클래스를 구분지으며 마진(margin)을 최대로 하는 초평면 $w \cdot x + b = 0$을 찾는 문제다.
여기서 w는 가중치 벡터이고, x는 입력벡터, b는 기준값이다.

학습문서 집합을 $D = \{(x_i, y_i)\}$라고 할 때, 입력데이터 x_i가 범주(class)에 속하면 y_i는 +1의 값을, 그렇지 않으면 -1의 값을 갖는다. 결국 SVM은 다음을 만족하는 아래의 식에서 최적의 w와 b를 찾는 문제이다.

$x_i \cdot w + b \geq +1$ ($y_i = +1$**인 경우**)

$x_i \cdot w + b \leq -1$ ($y_i = -1$**인 경우**)

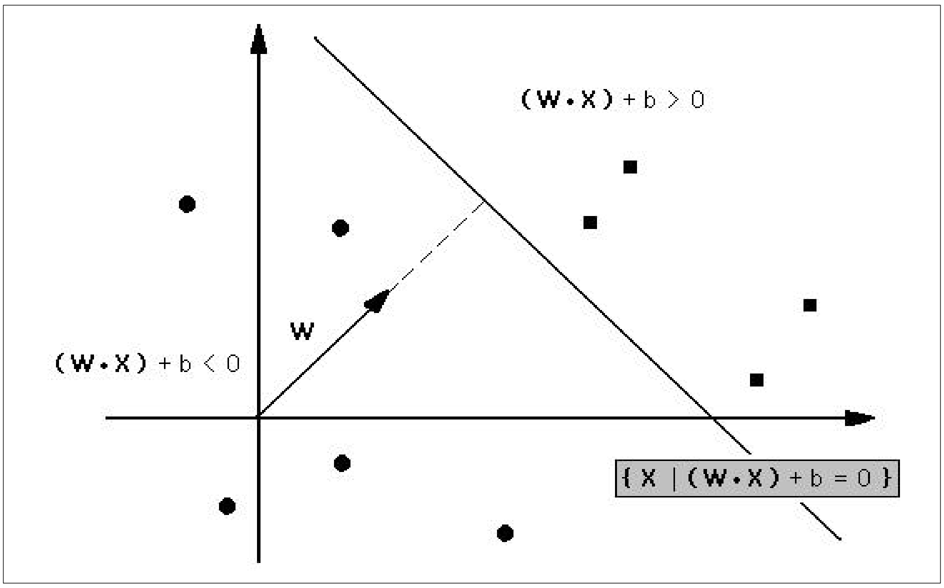


그림 2-3. **최적 경계면의 개념도**

선형적으로 분리가 가능한 학습집단에서는 최적의 초평면을 이용하여 최대 마진 분류기를 구축하며, **선형적으로 분리할 수 없는 경우에는 최적의 초평면 분류기 개념을 일반화시킨 소프트 마진 분류기를 구축하게 된다**.

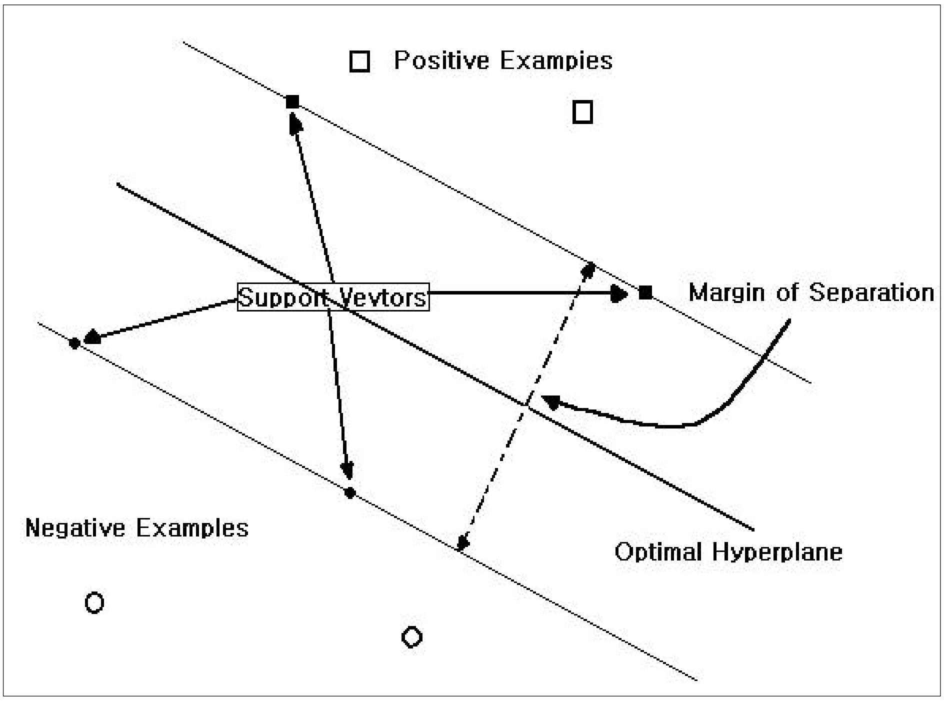


그림 2-4. 선형 SVM

2.2.1.1 최대 마진 분류기 (maximum margin classifier)

$$y_i(x_i \cdot w + b) \geq 1\ , \quad i = 1, \cdots, i$$

여기서 $\Psi(w) = ||w||^2$이므로 $||w||$가 최소가 되는 최적의 값을 결정해야 한다.

위의 식과 같은 제약조건하의 최적화(Constrained Optimization) 문제는 1차 영역(primal)에서의 해석과 2차 영역(dual)으로 나눌 수 있다.

1차 영역에서는 1차 변수 w와 b에 따라 최소화되어야 하며, 이를 라그랑즈 승수(Lagrange multipliers) α_i를 이용해서 표현하면 다음과 같다.

$$L(w, b, \alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{l} \alpha_i \{[(x_i \cdot w) - b] y_i - 1\}$$

$$(\alpha_i \geq 0, \forall_i)$$

이 식의 안장점(saddle point)을 찾기 위해서 다음과 같은 조건식을 구할 수 있다.

$$\frac{\partial L(w_0, b_0, \alpha^0)}{\partial b} = \sum_{i=1}^{l} \alpha_i^0 y_i = 0$$

$$\frac{\partial L(w_0, b_0, \alpha^0)}{\partial w} = w_0 - \sum_{i=1}^{l} y_i \alpha_i^0 x_i = 0$$

위의 결과를 이용해서 다음과 같은 식을 만들 수 있다.

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

최적 초평면을 설계하기 위해서는 이식을 최대화하는 α를 구하는데 이는 2차 계획법(Quadratic Optimization)을 통해서 이루어진다. 이 과정을 통해서 최종적으로 얻어진 해는 다음과 같다.

$$w_0 = \sum_{support\ vectors} y_i \alpha_i^0 x_i$$

이 해를 이용해서 얻어지는 분류법칙은 다음과 같다.

$$f(x) = sign\left(\sum_{support\ vectors} y_i \alpha_i^0 (x_i \cdot x) - b_o\right)$$

여기서 x_i를 지지벡터(support vector)라고 하는데 이것은 최적화 과정을 통해서 얻어진 α값 중에서 “0”이 아닌 값을 이용해서 결정되고 기준치 b_0는 다음 공식을 이용해서 구할 수 있다.

$$b_0 = \frac{1}{2}\left[(w_0 \cdot x^*(1)) + (w_0 \cdot x^*(-1))\right]$$

여기서 $x^*(1)$는 긍정 범주의 지지벡터, $x^*(-1)$는 부정 범주의 지지벡터를 의미한다(Vapnik 1999).

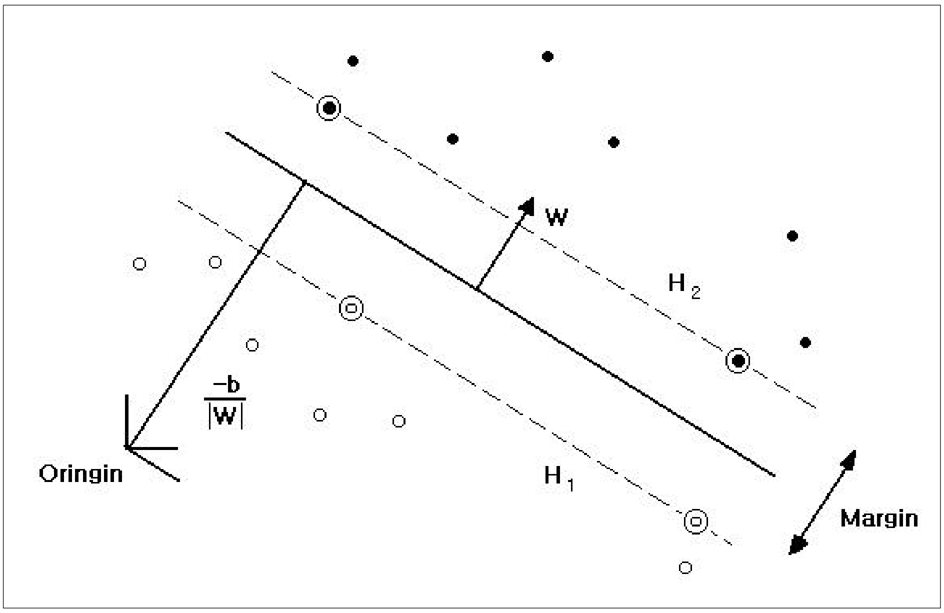


그림 2-5. 최대 마진 분류기

2. 2. 1. 2 소프트 마진 분류기 (soft margin classifier)

SVM에서는 동시에 여러 범주에 할당될 수 있는 데이터로 인한 오류 발생 등의 이유로 선형적으로 분리할 수 없는 경우에는 잘못 분류된 지점을 허용하는 완화(slack) 변수 ξ_i를 도입해서 일반화된 초평면(generalized hyperplane)을 구축하는 소프트 마진 분류기를 이용하게 된다. 이 때 초평면 (w, b)과 마진 r 상에서 데이터 $(x_{i,} y_i)$에 대한 완화 변수 ξ_i는 다음과 같이 정의된다.

$$\xi_i((x_{i,} y_i), (w, b), r) = \xi_i = \max(0, r - y_i(w \cdot x_i + b))$$

즉, $\xi_i > r$가 될 경우 데이터 (x_i, y_i)는 잘못 분류된 것을 의미하며, 완화 변수 ξ_i는 얼마나 많은 데이터들이 초평면에서 마진 r의 범위를 벗어나 있는지를 측정하는 척도가 된다(Cristianini et al. 2000).

우수한 성능을 위해서는 마진폭과 학습오류의 수를 조정하여 $\xi_i \geq 0$ $(i = 1, \cdots, l)$일 때 $y_i \cdot ((w, x_i) + b) \geq 1 - \xi_i$를 만족하는 범위에서 다음 수식에 C값을 최소화해야 한다(Cortes, and Vapnik 1995).

$$\tau(w, \xi) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i$$

이를 라그랑즈 승수 α_i와 커널함수 $k(x_{i,} x_j)$를 이용하여 비선형 함수에까지 일반화하기 위해서는 다음 수식을 최대화하는 값을 찾는 최적화 문제가 된다. 아래의 식은 $0 \leq \alpha_i \leq C (i = 1, \cdots, l)$과 $\sum_{i=1}^{l}\alpha_i y_i = 0$를 만족할 경우이다.

$$\max\ W(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j)$$

기준치 b는 $\alpha_i < C$인 모든 지지벡터에 대해서 완화 변수 ξ_i 값이 “0”이 되고 다음과 같이 표현할 수 있다(Scholkopf et al. 1999).

$$f(x) = sign\left(\sum_{i=1}^{l} y_i \alpha_i \cdot k(x, x_i) + b\right)$$

결국 완화 변수 ξ_i와 패널티 변수 C 값의 개념을 도입한 소프트 마진 분류기를 이용하게 되면 선형으로 분리할 수 없는 분류 문제의 경우도 어느 정도 해결할 수 있다(Coretes, and Vapnik 1995). 여기서 C 값은 마진 폭과 분류 오류 사이의 타협점(trade-off)을 찾아주는 역할을 담당하며, 분리할 수 없는 데이터에 대한 오류 패널티 값을 의미한다.

선형적으로 분리할 수 없는 데이터에 대해서 소프트 마진 분류기를 이용하여 분류를 수행하게 되면 초평면에서 마진 r의 범위를 벗어나서 잘못 분류된 두 개의 데이터는 완화 변수 ξ_i와 ξ_j 값이 마진 r보다 큰 값을 갖게 되고, 제대로 분류된 그 밖의 다른 데이터들은 완화 변수 값으로 "0"을 갖게 된다.

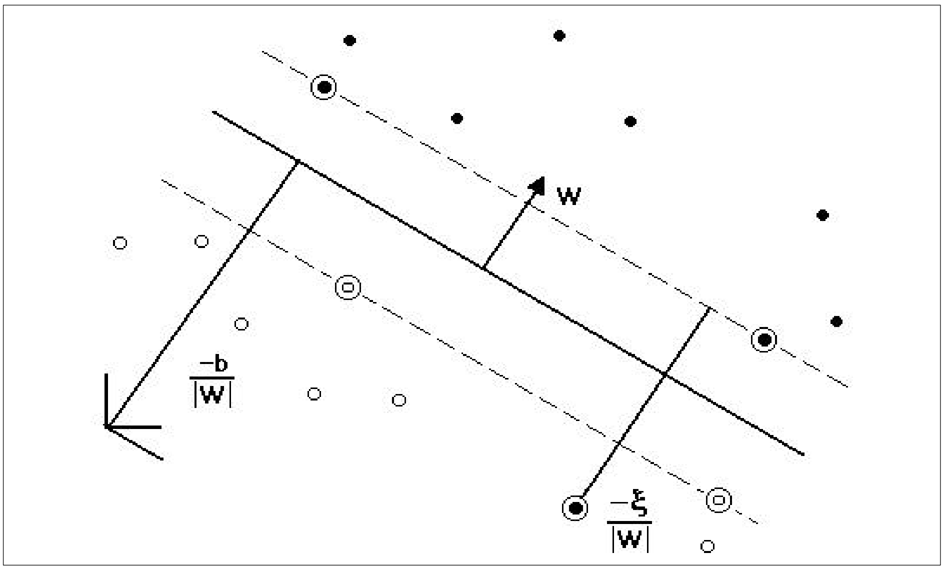


그림 2-6. 소프트 마진 분류기

2. 2. 2 비선형 SVM

입력데이터의 선형 분리가 불가능할 경우 입력공간을 분리하는 비선형 결정면(nonlinear decision surface)을 이용하게 된다.

함수 $K(\cdot,\cdot)$는 두 벡터의 내적을 나타내기 위한 커널 함수로 입력공간의 내적을 특징공간의 내적으로 변형시키는 역할을 한다. 이것은 특징공간으로 입력 벡터를 투영시킴으로써 내적에 대한 계산만을 필요로 하게 되어 고차원일 경우 계산상의 어려움을 덜 수 있다.

$$K(x, y) = (\Phi(x) \cdot \Phi(y))$$

즉, 벡터 x는 $\Phi(\cdot)$를 통하여 특징공간으로 변형시켜 두 벡터의 내적으로만 계산이 이루어지게 된다.

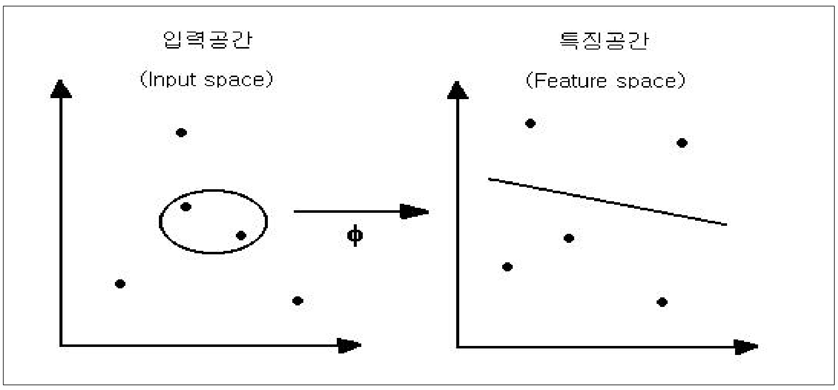


그림 2-7. 입력 공간의 특징 공간 투영

일반적으로 사용되는 커널함수는 다항식(polynomial) 커널함수, RBF (Radial Basis Function) 커널함수, 다층 퍼셉트론(multi-layer perceptron) 커널함수 등이며, 본 연구에서는 아래의 커널함수를 이용하여 실험을 수행하였다.

2. 2. 2. 1 다항식 커널함수

다항식 커널함수는 스칼라 곱(dot product)으로 처리되기 때문에 두 벡터간의 방향에 의존적이며, 결국 동일한 방향을 갖는 벡터들은 다항식 커널 함수를 이용했을 때 높은 결과 값을 갖게 된다.

$$K(x, x_i) = ((x \cdot x_i) + 1)^d$$

여기서 d는 다항식의 차수를 나타낸다.

2. 2. 2. 2 RBF 커널함수

$$f(x) = sign\left(\sum_{i=1}^{N} a_i K_\gamma(|x - x_i|) - b\right)$$

여기서 $K_\gamma(|x - x_i|)$은 두 벡터 사이의 거리인 $|x - x_i|$에 영향을 받게되고, 가장 일반적인 함수모형은 아래와 같다.

$$K(x, x_i) = \exp\left(\frac{||x - x_i||^2}{2\sigma^2}\right)$$

일반적으로 RBF 커널함수는 지지벡터와 검증 데이터간의 유클리드 거리와 연관되며, 여기서 지지벡터는 RBF의 중심점이 된다. σ 값이 커질수록 완만한(smoother) 결정면을 갖게된다.

제 3 장 Microarray 개요

3. 1 DNA Microarray 의 기본개념

Array에는 크게 3가지 종류가 있는데 filter array, Oligonucleotide array 및 cDNA microarray가 그것이다. 이 중 filter array와 Oligonucleotide array는 cDNA microarray와 기본 개념에서 많은 차이가 있으므로 여기서는 cDNA microarray만을 다루고자 한다. cDNA microarray는 분자생물학적 기법으로 볼 때 Northern blot의 역실험이라 할 수 있다(Reverse Northern). 즉, filter 위에 mRNA를 고정시키고 특정 유전자 탐침(probe)를 가하는 Northern 실험 대신 유전자 탐침을 먼저 고정시키고 조직에서 얻은 mRNA를 가하여 발현 양상을 조사하는 것과도 같다. 이러한 방법의 이점은 Northern과 달리 수많은 유전자 탐침을 동시에 사용할 수 있다는 점인데 이를 위해서는 먼저 몇 가지 트릭에 대한 기본적인 이해가 필요하다.

첫째, 수많은 탐침 유전자를 고정시키기 위해 filter paper 대신에 특별히 코팅된 유리 슬라이드를 사용한다는 점이다. Filter paper에서는 유전자 용액이 번지게 되므로 고정시킬 수 있는 유전자의 개수가 제한된다. 이때 유전자를 유리 슬라이드 위에 찍는 기술은 반도체 생산에 사용되는 미세 기술을 사용하며, 고정시키려는 유전자 간격을 마이크론 단위까지도 조절할 수 있다.

둘째, 유전자 탐침이 mRNA와 결합했는지 여부를 알기 위해서 일반적으로 Northern blot에서는 탐침 유전자를 방사성 동위원소로 tagging 하게

되는데, cDNA microarray에서도 마찬가지로 mRNA를 labeling 하는 절차가 필요하다. 그러나 방사성 동위원소는 신호가 예민한 대신, 신호가 강할 경우 인접한 유전자 영역까지 침범하게 되므로, 유전자가 고집적된 cDNA microarray에서는 일반적으로 형광물질을 사용하여 labeling 하게 된다. 형광물질은 Cy3와 Cy5나 Alexa dye 등을 일반적으로 사용하는데, 방사성 동위원소에 비하여 신호가 상대적으로 약하기 때문에 end labeling 대신에 Cy3-dUTP, Cy5-dUTP를 사용한 역전사(reverse transcription)를 통해 cDNA를 합성하고 이를 실험에 사용하게 된다.

셋째, 슬라이드 위에 고정된 유전자 탐침은 서로 농도와 DNA 크기가 다르기 때문에 화학반응(hybridization)에 의해 부착된 형광 신호로 절대적인 부착량을 계산할 수 없다. 따라서 cDNA microarray에서는 대조군(treatment group)과 실험군(control group) 2가지의 mRNA를 각기 Cy3와 Cy5로 labeling 한 후 동시에 동량을 사용하여 화학반응을 실시하며 결과적으로 녹색과 적색의 상대적인 비교치 만을 얻을 수 있다. 이는 실험군에서 대조군에 비하여 유전자 발현이 몇 배 증가하였다 또는 몇 분의 일로 감소하였다 하는 양상으로 실험 결과를 제시한다는 것이다.

DNA chip 기술은 기존 연구와 근본적인 차이를 보이는 획기적인 연구방법으로 다수 또는 전체 유전자 발현상황을 총체적으로 탐색할 수 있는 기반 기술을 제공하고 있다. 즉, 한 두 개의 유전자의 기능탐색이라는 종래의 한계를 벗어나 생명현상과 관련된 유전체 수준의 연구가 가능해졌다는 것을 뜻한다. 이러한 DNA chip 기술에는 cDNA chip 방식과 Affimatrix사의 oligochip 방식이 있다. 이 중 Affimatrix사의 oligochip 방식은 반도체 집적기술을 접목시켜 높은 집적도와 응용성뿐만 아니라 신뢰성 높은 결과물을 제공하고 있어 주목받고 있는 기술이며 현재 여러 회사에서 개발에 성공했거나 추진 중에 있다. 그리고, cDNA chip은 비교적 적은 비용과 쉬

운 제작방식으로 인해 현재 널리 사용되고 있다.
이러한 DNA chip에서 얻어진 자료를 DNA microarray 자료, 간단히 microarray 자료라고 한다. 이러한 자료는 보통의 실험자료에 비해 잡음이 많이 포함되어 있으며 또한 자료에 일정한 패턴을 보이는 경우가 많다.
잡음을 제거하거나 일정한 패턴을 제거하는 등의 과정을 표준화(normalization)라고 하며 그간 microarray를 이용한 연구에서 여러 가지 표준화 방법들이 제안되었다. 처음으로 cDNA microarray에서 통계적인 분석법을 제시한 Chen et al.(1997)은 Cy3(녹색형광)의 강도(intensity)와 Cy5(적색형광)의 강도가 일정한 비율을 하고 있을 것으로 가정한 후 표준화와 검정을 동시에 처리하였다. 또한 여러 이미지 분석 프로그램(image analysis program)에서는 회귀분석을 통한 표준화를 많이 사용하고 있다. 그리고, Yang et al.(2001)은 2001년 1월 SPIE BiOE에서 발표한 글에서 여러 표준화 방법들을 소개하였고 거기에 자신이 제시한 LOWESS 적합을 이용한 표준화 방법을 설명하였다. 또한 Dudoit(2000) 등은 Yang이 제시한 LOWESS 적합 방법을 통해 표준화한 자료로 분석하였다. 이에 반해 사용된 microarray 데이터가 이미 표준화되었다고 가정하거나(Newton 2001) 아예 이런 것에 대한 고려가 없었던 경우도 많다(Eisen 1998). 그러나, 이러한 경우의 사용된 자료를 보면 일정한 패턴이 남아 있었음을 확인할 수 있었다.

3. 2 DNA chip 제작과 표준화 방법

3. 2. 1 DNA chip 제작

cDNA chip은 비교하고자하는 두 종류의 세포의 핵에 들어있던 여러

mRNA의 양의 비를 측정하는 방법을 이용한다. 특정 유전자를 전사(transcription)하여 만들어진 mRNA는 추후 전사 과정을 통해 단백질을 생성하는데 사용되므로 mRNA의 양은 해당 유전자의 발현정도를 나타내는 척도로 삼을 수 있다. 즉, 측정된 mRNA의 양이 많은 때에는 해당 유전자가 활성화되었다는 것을 뜻하며 적을 시에는 유전자가 비활성화되었다는 것을 뜻한다. 그러나, 세포에서 추출된 mRNA는 소량이므로 실험에서 요구되는 양을 얻기 위해 cDNA로 변환한 후 RT-PCR 방법을 이용해 증폭한다. 이렇게 증폭된 cDNA에 하나는 Cy3를 부착하고 다른 하나는 Cy5를 부착한다. 그런 다음 동일 양을 섞은 후 미리 준비된 관련 cDNA와만 결합할 수 있는 탐침을 화학 반응하게 한 후 슬라이드에 촘촘히 배열한다. 이런 식으로 슬라이드에 기록되는 cDNA 수는 적게는 수백 개에서 많게는 몇 만개에 달한다. 즉, cDNA 실험을 통해서 동시에 수백 개에서 몇 만개에 달하는 유전자의 발현양상을 한번에 살펴볼 수 있음을 뜻한다.

3. 2. 2 Oligochip 제작

Oligochip은 우선 각 유전자의 고유한 20개 정도의 길이의 DNA 염기서열(DNA sequence)을 찾아낸 후 그 온전한 것과 이 염기서열 중 가운데 하나의 염기만 다른 것을 반도체 제조공정과 비슷한 과정을 통해 합성한다. 그러한 후, 세포에서 추출한 mRNA를 cDNA로 만들고 형광물질을 부착한 후 슬라이드에 화학 반응하게 한다.

이런 식으로 제작된 cDNA chip과 oligochip을 confocal microscope를 통해 각 형광물질의 강도를 측정하여 cDNA chip의 경우 두 장, oligochip의 경우 한 장의 이미지(image) 파일을 생성하고 이미지 분석 프로그램(image analysis program) 등을 통해 각 유전자의 발현량을 수치화 한다.

3. 2. 3 표준화(Normalization) 기준

위에서 얻은 microarray 자료를 얻은 후 실제의 분석에 이용하기 위해 표준화 과정을 거쳐야한다. 이 때, 표준화는 두 가지 방식의 표준화 기준으로 처리할 수 있는데, 첫 번째로는 표준화만을 위해 일부로 넣은 몇 개의 유전자만을 사용하여 표준화하는 방법과, 두 번째로 전체 유전자를 사용하여 표준화하는 방법이 있다.

3. 2. 3. 1 일부의 유전자만을 기준으로 삼는 방법

이 방법은 분석에 사용된 유전자 정보 외에 표준화만을 위해 삽입된 유전자로 표준화하겠다는 전형적인 방법이다. 여기에 쓰이는 유전자로는 세포 내에서 생명활동을 위해 항상 일정한 양이 발현되고 있다고 생각되는 housekeeping 유전자(gene)를 이용하거나 아니면 전혀 발현되지 않으리라는 가정 하에 spiked 유전자를 이용하여 표준화에 사용하고자 한다.

3. 2. 3. 2 전체 유전자만을 기준으로 삼는 방법

두 번째로 제시된 방법을 위의 방법이 앞과 같은 경우 같이 housekeeping 유전자가 제 역할을 못하는 경우나 또는 Cy3가 강하게 나타나는 것과 같이 슬라이드에서의 일정한 변이 등이 나타나는 경우 등과 같이 일부의 유전자만으로는 보정하기 힘든 경우가 실제 실험에서 나타나기 시작하자 새로운 대안으로 제시된 방법이다. 이 방법은 수천, 수만 개의 유전자를 실험하는 경우에 전체 유전자를 모두 사용하여 보정하는 방법이다. 이 방법은 이런 실험에서 대다수의 유전자는 비슷하게 발현되고 몇몇 특정

의 유전자만 다르게 발현될 것이라는 가정을 전제로 한 것이다.

3. 2. 4 표준화(Normalization) 방법

이렇게 얻어진 표준화에 사용될 데이터에서 Cy3의 강도를 G_j라하고 Cy5의 강도를 R_j이라고 하자. 이 때, j는 각 유전자를 구분하는 id이다. 여기에 추가로

$$M_j = \log \frac{R_j}{G_j} = \log R_j - \log G_j$$

$$A_j = \log \sqrt{RG} = \frac{(\log R_j + \log G_j)}{2}$$

한 척도를 사용하도록 한다. 물론 여기에 사용되는 R_j과 G_j의 자료들은 위의 표준화 기준에서 선택된 유전자 자료이다.

3. 3 Microarray 분류기법

cDNA microarray 실험의 가장 큰 목적은 실험군과 대조군의 비교를 통하여 유전자들을 확인하는 것이 가장 큰 목적이라 할 수 있다.

Dudoit et al.(2000)에 따르면, 이러한 분류기법과 관련된 통계학적 문제에 대하여 크게 3가지로 제시하였다.

첫째로, 유전자 표현 측면을 이용하여 새로운 것이나 또는 알려지지 않은 집단들을 확인하는 방법이다. 이는 즉, 군집분석(cluster analysis) 또는 관리되지않은 학습(unsupervised learning)이라고 한다.

두 번째로, 첫 번째와는 반대로 알려진 어떤 질병에 관한 악성 집단들을 분류하는 방법이다. 이는 판별분석(discriminant analysis)으로 행할 수 있으며, 관리되는 학습(supervised learning)이라고 한다.

세 번째로, 서로 다른 질병의 class를 특징짓는 알려진, '표식유전자(marker gene)'의 확인하는 방법이다. 이는 변수선택(variable selection)방법으로 접근할 수 있다.

제 4 장 모의실험 자료를 이용한 SVM 분류기법

4. 1 모의 Microarray 자료의 생성

SVM 이원분류기의 성능을 모의실험을 통해 알아보기 위해서 자료를 이원 분리라 가정하고 모의 microarray 자료를 생성한다.

microarray 자료를 생성하기 위해서, 먼저 Cy3과 Cy5의 강도의 로그비를 구해야 한다. 또한, 환자군과 대조군을 각각 20명씩 생성하였고, 전체 유전자의 수는 100개로 하였다. 분류를 위해서 100개의 유전자 중, 환자군들에서 20%의 유전자는 유의한 유전자, 즉, Cy5의 강도가 더 강하다고 설정하고, 나머지 80%는 Cy3과 Cy5의 강도의 로그비는 0, 즉, 유의하지 않은 유전자라 설정한다.

환자군과 대조군에 대한 각각의 유전자(gene)들의 결과값, 즉, Cy3과 Cy5의 강도에 대한 로그비(y)는 각각 아래와 같은 분포에서 생성된다.

$$y = \log\left(\frac{R}{G}\right) + \varepsilon \quad , \ \varepsilon \sim MVN(\tilde{\mu}, \hat{\Sigma})$$

먼저, 유의하지 않은 80%의 유전자와 대조군의 모든 유전자를 생성하기 위해서는, $\left(\frac{R}{G}\right) = 1$ 이므로 아래와 같은 분포로 생성할 수 있다.

$$y = \varepsilon \quad , \ \varepsilon \sim MVN(\tilde{\mu}, \hat{\Sigma})$$

환자군들 중 유의한 집단에 속하는 20%의 유전자는 다음 분포에서 생성한다.

$$y = \log\left(\frac{R}{G}\right) + \varepsilon \quad , \quad \varepsilon \sim MVN(\tilde{\mu}, \tilde{\Sigma})$$

4.2 모의실험 자료를 이용한 SVM 기법 적용 방법

모의자료에 SVM 분류 기법을 적용시키기 위해서는 먼저 여러 가지 모수(parameter)들의 조절이 필요하다. 미리 결정해야 하는 모수는 학습과정에서 마진폭과 분류오류 사이의 타협점(trade-off)을 찾아주는 오류 패널티(penalty) 변수 C값이다.

SVM 분류 기법은 분류(classification)방법에 따라 c-분류와 ν-분류가 있다.

c-분류방법(c-classification)은 비용(cost)이라는 값을 미리 정해 줌에 따라 민감도(Sensitivity), 특이도(Specificity), 양성예측도(Positive Predicted Value), 음성예측도(Negative Predicted Value), 정분류율(Correct Proportion), 오분류율(Miss Correct Proportion)을 변화시킬 수 있다.
위와 같은 값들을 변화시키기 위해 사용하는 비용을 조절(tuninig)하는 방법에는 k-fold 교차타당도 검사(cross validation)와 LOO(Leave-One-Out) 방법이 있는데(Duan et al. 2001), 본 모의실험에서는 LOO방법을 고려하기로 한다.

k-fold 교차타당도 검사는 분석용 자료(training data)가 동일한 크기의 상호 배반적인(mutually exclusive) 부분집합으로 임의적인 k 개로 분리하

는 방법이다. LOO방법은 $l-1$ 개의 부분집합을 이용하여 얻어지고 나머지 1개로 검증(test)한다. 이러한 과정을 l 번 반복하여 기대되는 일반화 오류(expected generalization error)의 추정치를 얻는다. (Duan et al. 2001)

ν-classification은 ν 값을 정해줌에 따라서 값을 변화시킬 수 있는데 $0 < \nu < 1$사이의 값을 갖는다.

4.3 모의실험 자료를 이용한 사전실험

일반적으로 C가 작을수록 보다 더 간단한 모델을 선호하게되는 효과를 지니게 되며, C를 무한대(infinite)로 하게 될 경우 모든 분석용 자료(training data)가 정확하게 분류되어야 한다는 것을 의미한다(smola, 1997).

비선형 SVM에서 사용되는 C 값과 ν 값은 각각 다음과 같이 정했다. 모의실험에서 적절한 C 값을 구하기 위하여, LOO(Leave-One-Out) 방법을 이용하여 탐색한 결과 그림 [4-1], [4-2]에서 보는 것과 같이 $C = 1.4$의 값에서 가장 안정적인 값을 갖는 것으로 나타났다.

또한, ν 값을 찾기 위해 $0 < \nu < 1$사이에서 탐색한 결과 그림 [4-3], [4-4]에서와 같이 0.3부터 가장 안정적인 값을 갖는 것으로 나타났다.

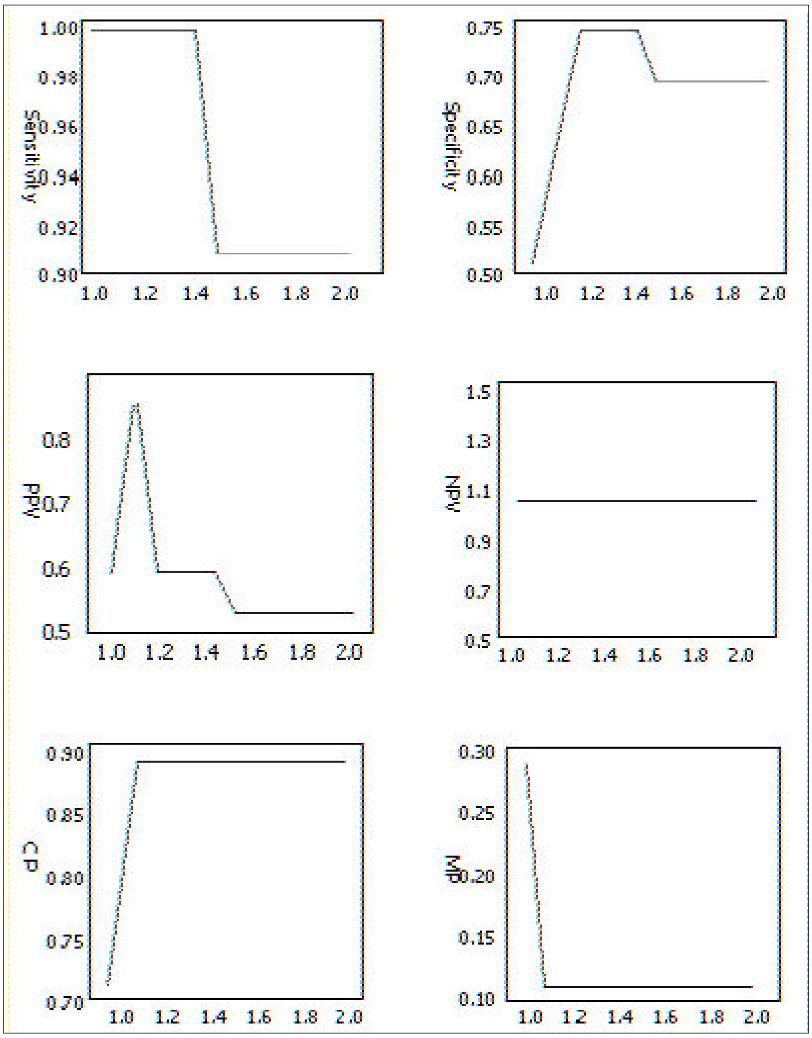


그림 4-1. **비용 값에 따른 각 항목들의 변화**(radial)

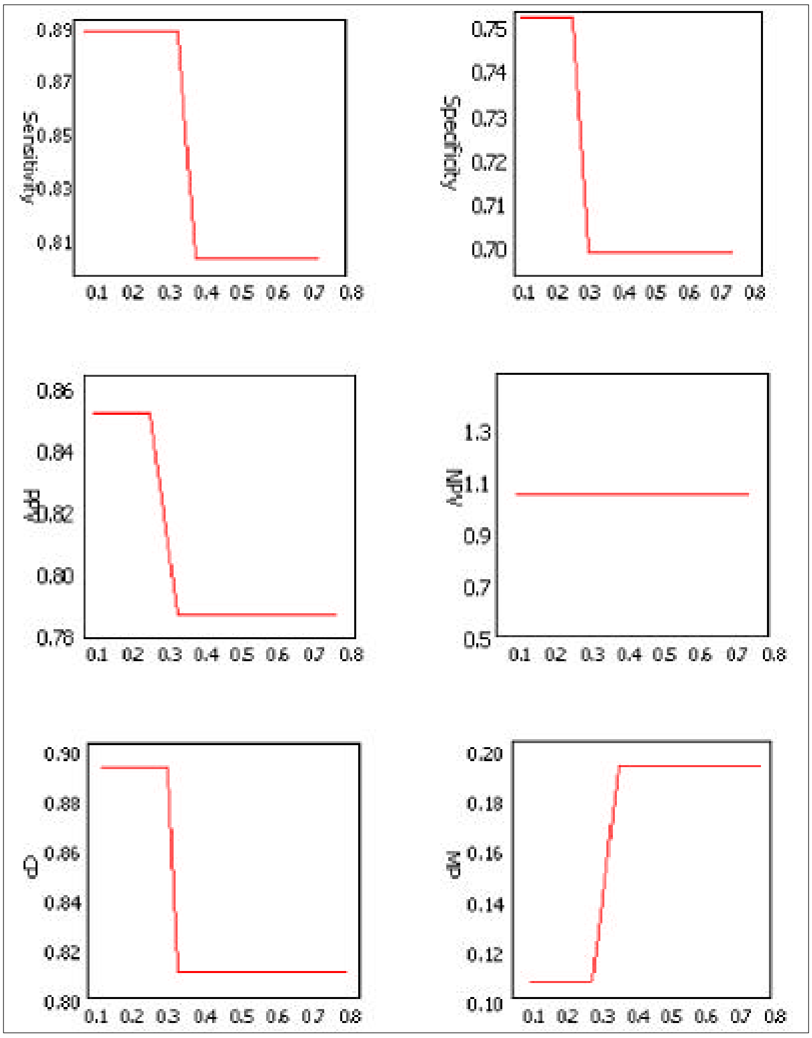


그림 4-2. **비용 값에 따른 각 항목들의 변화**(polynomial)

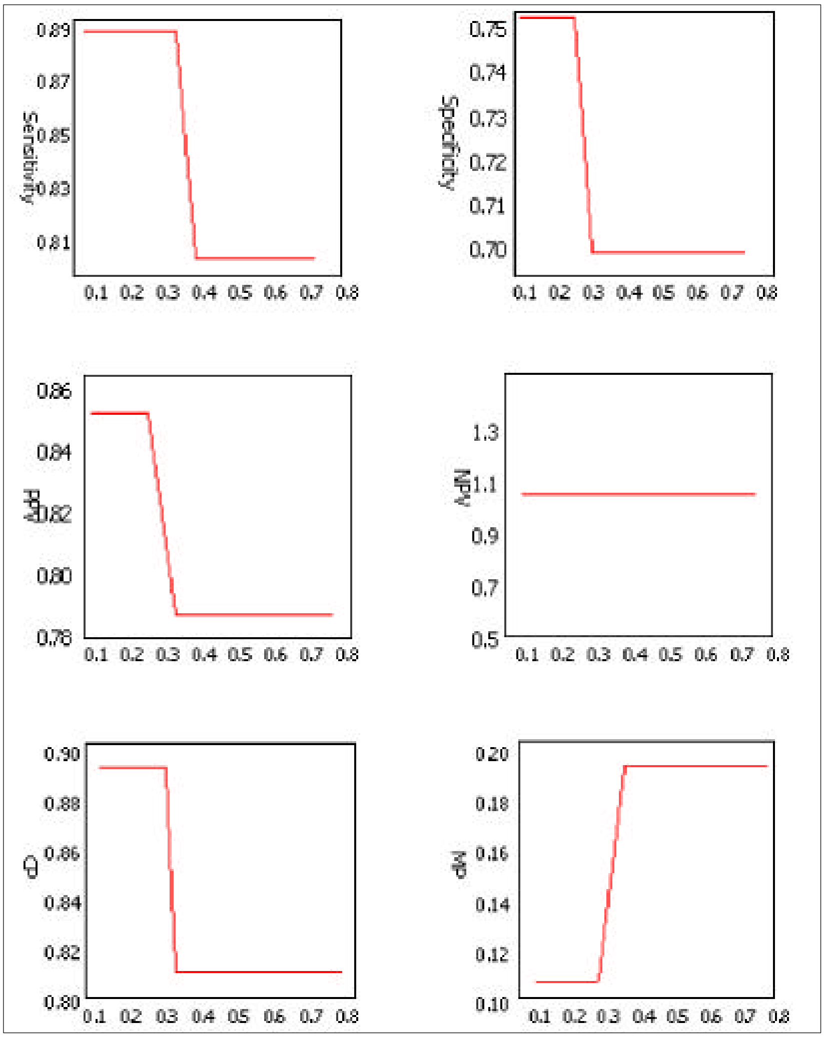


그림 4-3. ν 값에 따른 각 항목들의 변화(radial)

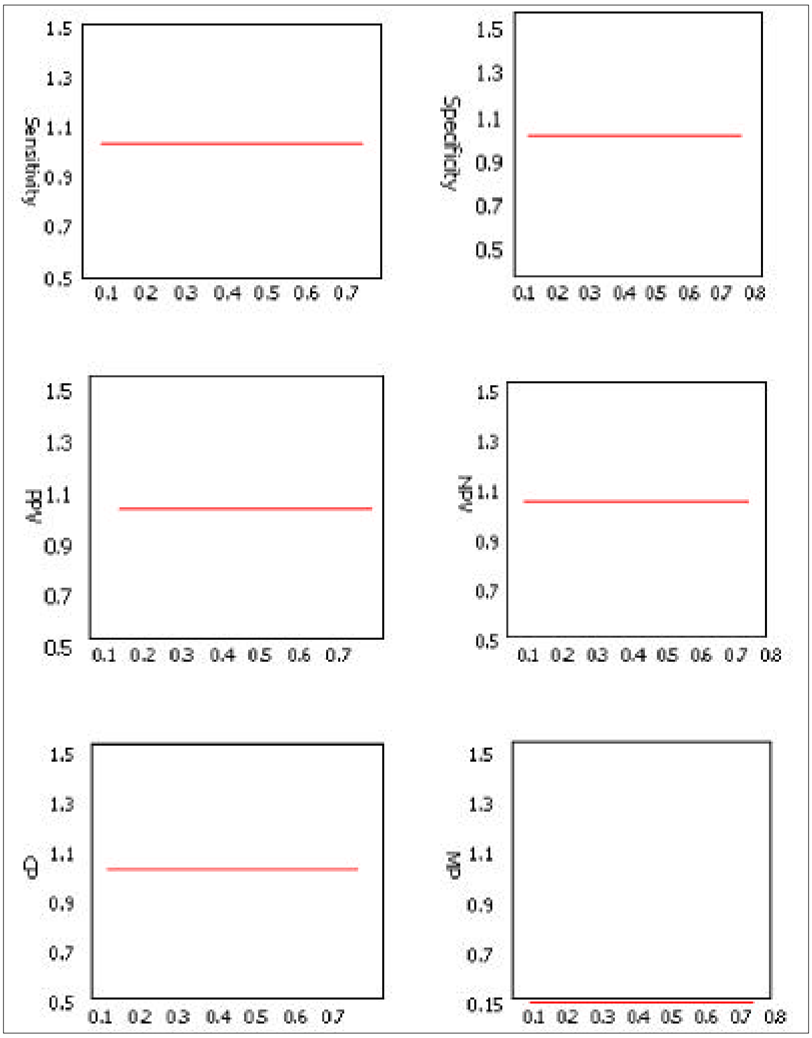


그림 4-4. ν **값에 따른 각 항목들의 변화**(polynomial)

표 1 . 비용과 편차에 따른 SV개수의 변화(c-classification)

커널함수	cost	σ=0.5	σ=1.0	σ=1.5
radial	1.0	30	30	30
	1.3	30	30	30
	1.5	30	30	30
	1.7	30	30	30
	1.9	30	30	30
	2.0	30	30	30
polynomial (degree=3)	1.0	24	25	26
	1.3	24	25	26
	1.5	24	25	26
	1.7	24	25	26
	1.9	24	25	26
	2.0	24	25	26

c-분류에서 비용과 표준편차에 따른 Support Vector 개수의 변화를 살펴본 결과는 위의 [표1]에 나타내었다.
위의 표에 나타난 결과로서 radial에서는 비용과 표준편차 값에 상관없이 Support Vector의 개수가 변화하지 않는 것으로 나타났으나, 다항식에서는 비용에 따라서는 변화가 없으나 표준편차가 커질수록 Support Vector의 개수가 증가함을 알 수 있다.

4.4 로그비 및 편차에 따른 커널 함수별 모형의 평가

c-분류의 비용을 1.4로, ν-분류의 ν 를 0.3으로 하고, 커널함수는 radial 과 3차 다항식으로 하여 Cy5가 높게 나타나는 로그비의 변화에 따른 모형의 평가를 모의 실험의 정확도를 위해 100번 반복한 민감도, 특이도, 양성예측도, 음성예측도, 정분류율, 오분류율 들의 평균 및 표준편차로서 [표2,3]에 나타내었다.

표 2 . 로그비에 따른 커널함수별 모형의 평가(c-classification)

커널함수	parameter mean±sd	log 1.5	log 2.0	log 2.5	log 3.0
radial	Sensitivity	0.729±0.218	0.841±0.188	0.896±0.163	0.968±0.083
	Specificity	0.761±0.220	0.812±0.194	0.882±0.193	0.942±0.128
	PPV	0.746±0.239	0.821±0.187	0.892±0.168	0.958±0.082
	NPV	0.726±0.231	0.845±0.170	0.917±0.128	0.970±0.082
	CP	0.700±0.152	0.804±0.130	0.883±0.121	0.957±0.062
	MP	0.300±0.152	0.196±0.130	0.117±0.121	0.043±0.062
polynomial (degree=3)	Sensitivity	0.765±0.206	0.883±0.166	0.949±0.118	0.992±0.043
	Specificity	0.727±0.207	0.771±0.199	0.837±0.201	0.921±0.145
	PPV	0.697±0.215	0.753±0.212	0.845±0.178	0.936±0.110
	NPV	0.788±0.201	0.896±0.148	0.962±0.089	0.994±0.031
	CP	0.718±0.140	0.800±0.138	0.885±0.118	0.958±0.069
	MP	0.282±0.140	0.200±0.138	0.115±0.118	0.043±0.069

* cost=1.4, number of iteration=100

표 3 . 로그비에 따른 커널함수별 모형의 평가(ν- classification)

커널함수	parameter mean±sd	log 1.5	log 2.0	log 2.5	log 3.0
radial	Sensitivity	0.727±0.214	0.842±0.188	0.898±0.159	0.968±0.083
	Specificity	0.762±0.219	0.817±0.189	0.883±0.191	0.942±0.128
	PPV	0.749±0.245	0.827±0.182	0.893±0.164	0.958±0.082
	NPV	0.721±0.226	0.844±0.171	0.919±0.127	0.970±0.082
	CP	0.700±0.144	0.808±0.129	0.885±0.118	0.957±0.062
	MP	0.300±0.144	0.192±0.129	0.115±0.118	0.043±0.062
polynomial (degree=3)	Sensitivity	0.765±0.206	0.883±0.166	0.949±0.118	0.992±0.043
	Specificity	0.727±0.207	0.771±0.199	0.837±0.201	0.921±0.145
	PPV	0.697±0.215	0.753±0.212	0.845±0.178	0.936±0.110
	NPV	0.788±0.201	0.896±0.148	0.962±0.089	0.994±0.031
	CP	0.718±0.140	0.800±0.138	0.885±0.118	0.956±0.069
	MP	0.282±0.140	0.200±0.138	0.115±0.118	0.043±0.069

* ν=0.3, number of iteration=100

위의 표에 나타난 결과로서 로그비가 높을수록 각각의 평가항목들(민감도, 특이도, 양성예측도, 음성예측도, 정분류율, 오분류율)이 더 좋게 나타났는데, 이것은 microarray 실험에서 Cy5의 강도가 높게 나타날수록 분류를 더 정확하게 한다는 것을 말하는 것이다. 그러나 커널함수와 분류방법에 따라서는 별 차이가 없다는 것을 알 수 있다. log 0.5일 때, c-분류에서 radial 일 경우의 민감도가 0.729±0.218인데, ν-분류에서는 0.727±0.214로서 평균과 표준편차에 별 차이가 없다는 것이다.

c-분류는 비용을 1.4로, ν-분류는 ν 를 0.3으로 하고, 커널함수는 radial 과 3차 다항식으로 하여 로그비의 변화를 고정시키고 표준편차에 따른 모형의 평가를 100번 반복한 민감도, 특이도, 양성예측도, 음성예측도, 정분류율, 오분류율 들의 평균 및 표준편차로서 [표4,5]에 나타내었다.

표 4 . 편차에 따른 커널함수별 모형의 평가 (c-classification)

커널함수	parameter mean±sd	σ=0.5	σ=1.0	σ=1.5
radial	Sensitivity	0.920±0.157	0.808±0.202	0.763±0.237
	Specificity	0.941±0.113	0.784±0.209	0.767±0.232
	PPV	0.941±0.109	0.766±0.241	0.749±0.264
	NPV	0.927±0.136	0.814±0.195	0.764±0.235
	CP	0.925±0.100	0.756±0.151	0.708±0.173
	MP	0.075±0.100	0.244±0.151	0.292±0.173
polynomial (degree=3)	Sensitivity	0.976±0.061	0.885±0.175	0.851±0.178
	Specificity	0.917±0.135	0.761±0.237	0.687±0.202
	PPV	0.928±0.114	0.760±0.238	0.644±0.235
	NPV	0.969±0.081	0.879±0.181	0.860±0.189
	CP	0.944±0.079	0.790±0.154	0.722±0.147
	MP	0.056±0.079	0.210±0.154	0.278±0.147

* log 2.0, cost=1.4, number of iteration=100

표 5 . 편차에 따른 커널함수별 모형의 평가(ν- classification)

커널함수	parameter mean±sd	σ=0.5	σ=1.0	σ=1.5
radial	Sensitivity	0.922±0.145	0.806±0.204	0.770±0.225
	Specificity	0.937±0.113	0.782±0.210	0.770±0.227
	PPV	0.932±0.116	0.765±0.241	0.753±0.259
	NPV	0.925±0.137	0.812±0.194	0.765±0.236
	CP	0.925±0.089	0.756±0.153	0.712±0.170
	MP	0.075±0.089	0.244±0.153	0.288±0.170
polynomial (degree=3)	Sensitivity	0.976±0.061	0.885±0.175	0.851±0.178
	Specificity	0.917±0.135	0.761±0.237	0.687±0.202
	PPV	0.928±0.114	0.760±0.238	0.644±0.235
	NPV	0.969±0.081	0.879±0.181	0.860±0.189
	CP	0.944±0.079	0.790±0.154	0.722±0.147
	MP	0.056±0.079	0.210±0.154	0.278±0.147

* log 2.0, ν=0.3, number of iteration=100

위의 표에 나타난 결과로서 표준편차가 높을수록 각각의 평가항목들 (sensitivity, specificity, PPV, NPV, CP, EP)이 더 낮게 나타났는데, 이것은 microarray 실험에서 오차(noise)가 높게 나타날수록 분류율이 더 낮아진다는 사실을 말하는 것이다.

또한, 분류방법과 커널함수에 따라서는 별 차이가 없었다는 것을 알 수 있다. 표준편차 (σ) 0.5일 때, c-분류에서 radial일 경우의 민감도가 0.920±0.157인데, ν-분류에서는 0.922±0.145로서 평균과 표준편차에 별 차이가 없다는 것이다. 그리고, 커널함수에서 c-분류에서 radial일 경우의 민감도가 0.920±0.157인데, 다항식에서는 0.976±0.661로서 평균과 표준편차에 차이가 없다는 것이다.

제 5 장 결 론 및 고찰

지금까지 microarray 자료에서 SVM 기법을 적용한 이원분류기법을 소개하였다. 또한 microarray 모의실험을 통한 자료생성(generating) 및 각 커널함수별 SVM 분류에 대하여 평가하였다.

SVM은 Vapnik에 의해 이원 패턴 인식 문제를 해결하기 위해 제안된 학습방법으로서, 부정예제로부터 긍정예제를 분류해 낼 수 있는 결정면을 찾아내는 분류모형을 의미한다.

본 논문의 주 관심인 이원 분류기로서 SVM 성능을 평가하기 위해 microarray 자료를 이용하였는데, 이 microarray 자료는 통계프로그램(S-PLUS)을 이용하여 생성하였다. 또한 SVM 이원 분류기의 성능을 평가하기 위해서 Chen이 제시한 R-Package에 있는 SVM 라이브러리를 이용하였다.

모의실험 결과, c-분류에서 radial에서는 비용과 표준편차 값에 상관없이 Support Vector의 개수가 변화하지 않는 것으로 나타났으나, 다항식에서는 비용에 따라서는 변화가 없으나 표준편차가 커질수록 Support Vector의 개수가 증가함을 알 수 있었다.

로그비가 높을수록 각각의 평가항목들(민감도, 특이도, 양성예측도, 음성예측도, 정분류율, 오분류율)이 더 좋게 나타났는데, 이것은 microarray 실험에서 Cy5의 강도가 높게 나타날수록 분류를 더욱 정확하게 한다는 것을 말하는 것이다. 그러나, 커널함수와 분류방법에 따라서는 별 차이가 없다는 것을 알 수 있다. 또한, 표준편차가 높을수록 각각의 평가항목들(민감도, 특이도, 양성예측도, 음성예측도, 정분류율, 오분류율)이 더 낮게 나타났는데, 이것은 microarray실험에서 오차가 높게 나타날수록 분류율이 더 낮아

진다는 사실을 말하는 것이다.

SVM 학습 알고리즘에서는 이용자가 직접 모수 값을 결정하여 실험을 수행해야 하며 모수 값에 따라 분류기의 성능이 달라지게 된다. 모수 값을 결정하기 위해서는 모수에 대한 이론적인 이해와 실험과정에서 시행 착오를 거치면서 모수 값을 조정해야 하기 때문에 SVM 이론에 익숙하지 않은 이용자가 직접 모수를 결정한다는 것은 어려운 일이다. 그러므로 이러한 모수 값들을 SV의 VC 차원과 관련하여 기계학습과정에서 결정할 수 있는 연구에 대한 접근이 필요하다고 생각된다.

본 연구에서는 microarray 모의실험에 대해서만 분류성능을 확인하였고, 실재 자료에 관해서는 고려를 하지 않았다. 이러한 분석을 더욱 심화하기 위해서는 실재자료를 통하여 구체적인 접근이 필요하다고 생각된다.

참고문헌

김진혁, **이용성**, cDNA microarray. ***천식 및 알레르기***, 2001, 21(3):467-476

나종화, **김정숙**, S-PLUS **사용법 및 프로그램밍**. **자유아카데미**, 2000

송영기, **강환일**, SVM**을 이용한 패턴 인식**. ***전기학회지***, 2000, 49(4): 44-48

유종영, **이승천**, **차경준**, **허문열**, S-PLUS**를 이용한 통계계산**. **박영사**, 1997

오장민, **장병탁**, **김영택**, SVM **학습을 이용한 다중 클래스 뉴스그룹 문서 분류**. ***한국정보과학회 가을 학술 발표 논문집(II)***, 1999, 26(2):60-62

Burges, Christopher. J. C., A Tutorial on Support Vector Machines for Pattern Recognition. Boston : Kluwer Academic Publisher, 1998

Chen, Ratio-Based Decision, and the Quantitative Analysis of cDNA Microarray Images. *Journal of Biomedical Optics*, 1997, 2(4):364-374

Chris J. C. Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, Vladimir Vapnik., Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers. *IEEE Transactions on signal processing*, 1997, 45(11):2758-2765

Cristianini, Nello, and John Shawe-Taylor., An introduction to support Vector Machines and other kernel-based learning method. Cambridge: Cambridge University Press, 2000

Cleveland, Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 1979, 74(368):829-836

Cortes, C., Vladimir Vapnik, Support Vector Networks. Machine Learning, 1995, 20:273-297

Kaibo Duan, S Sathiya Keerthi, Aun Neow Poo., Evaluation of Simple Performance measures for Tuning SVM Hyperparameters. *Control Division Technical Report*, 2001

Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Bostein, Cluster analysis and display of genome-wide expression patterns. *In Proceedings of the National Academy of Sciences*, 1998, 95:14863-14868

M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, K. W. Tsui, On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of the American Computation Biology*, 2001, 8(1):37-52

M. Kathleen Kerr, Mitcheel Martin, Gary A. Churchill, Analysis of Variance for Gene Expression Data. *Journal of the American Computation Biology*, 2001, 7:819-837

Michael P. S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Sugnet, Manuel Ares, Jr., David Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS*, 2000, 97:262-267

Michael P. S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Sugnet, Manuel Ares, Jr., David Haussler, Support Vector Machine Classification of Microarray Gene Expression Data. *Technical Report* UCSC-CRL-99-09, 1999

Sandrine Dudoit, Jane Fridlyand, and Terry Speed, Comparison of discrimination methods for the classification of tumors using gene expression data. *Technical Report*, 2000

,

Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, and Terence P. Speed, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Technical Report* 578, Department of Biochemistry, Stanford University School of Medicine, 2000

Scholkopf, Bernhard, Chris Burges, and Alex J. Smola, Advences In kernel methods:Support Vevtor Machines. Cambridge:MIT Press, 1999

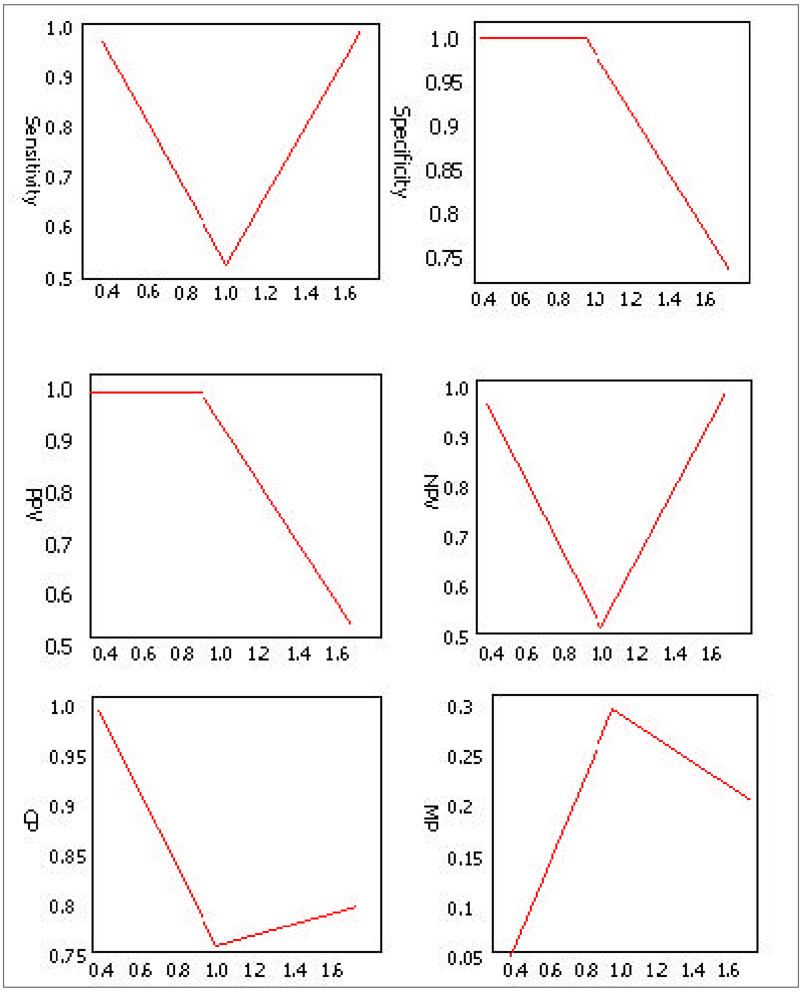
A. Smola, and B. Scholkopf, On a kernel-based Method for Pattern Recognition, Regression, Approximation and Operator Inversion. *GMD Technical Report*, 1997

T. R. Golub, D. K. Slonim, P.Tamayo, C. Huard, M. Gaasenbeek, J. P. Meirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander, Molecular Classification of Cancer: Class Discovery and Class Prediction by gene Expression Monitoring. *Scinece*, 1999, 286:531-537

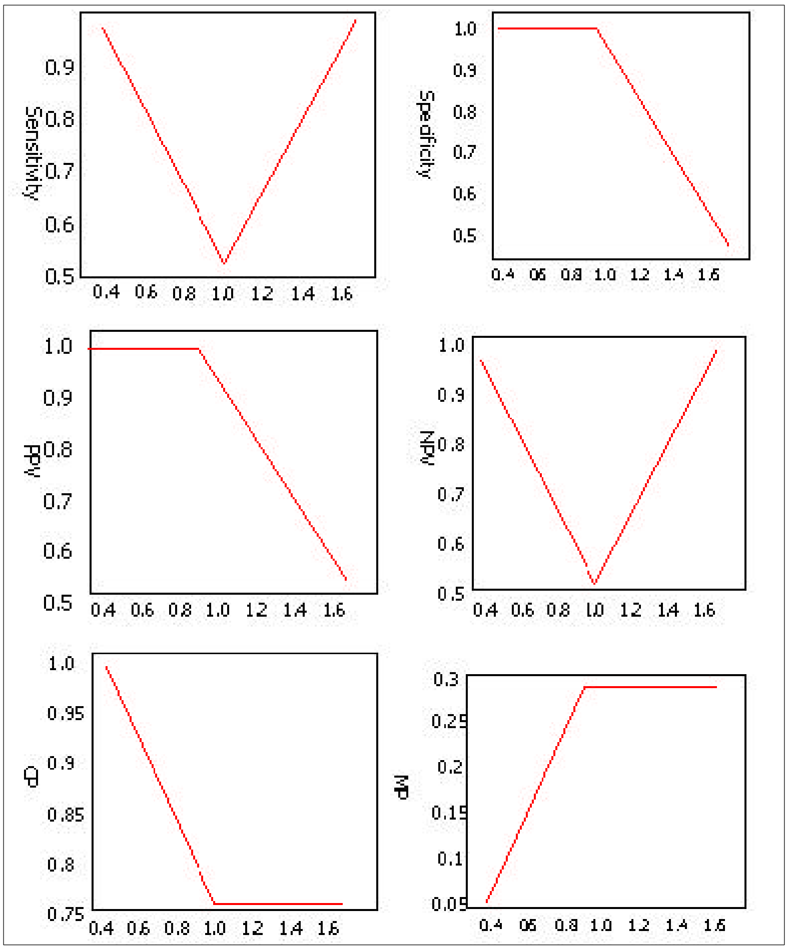
Vladimir N. Vapnik, The Nature of Statistical Learning Theory. Springer, 1999

W. N. Venables, B. D. Ripley, Modern Applied Statistics with S-PLUS(3rd edition). Springer, 1999

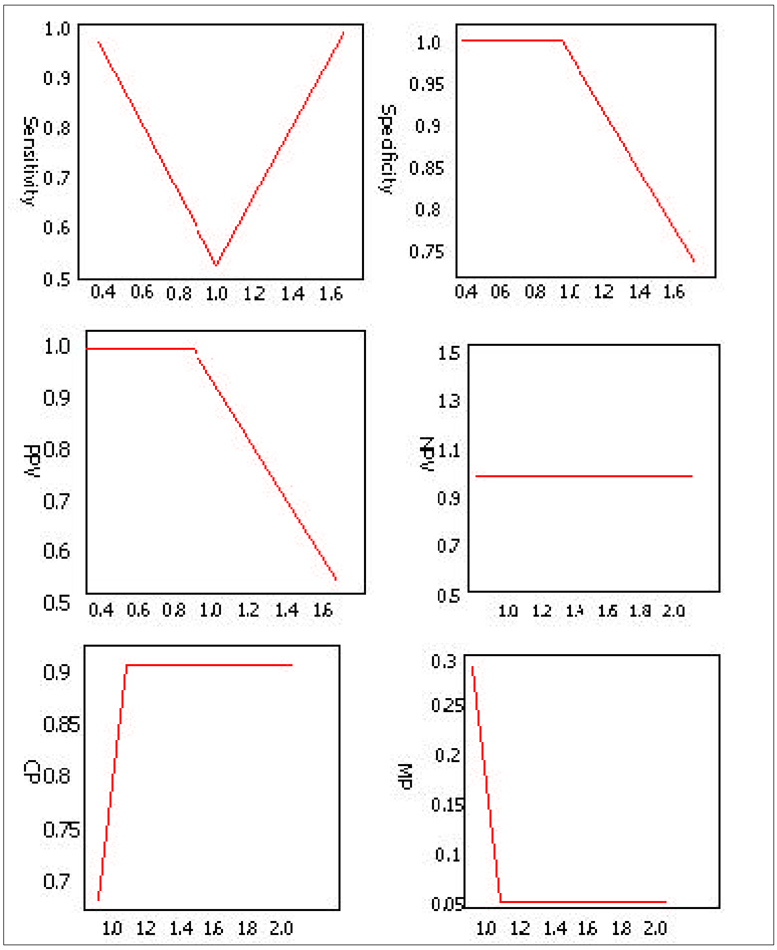
Yee Hwa Yang, Normalization for cDNA Microarray Data. In *SPIE BioE*, 2001



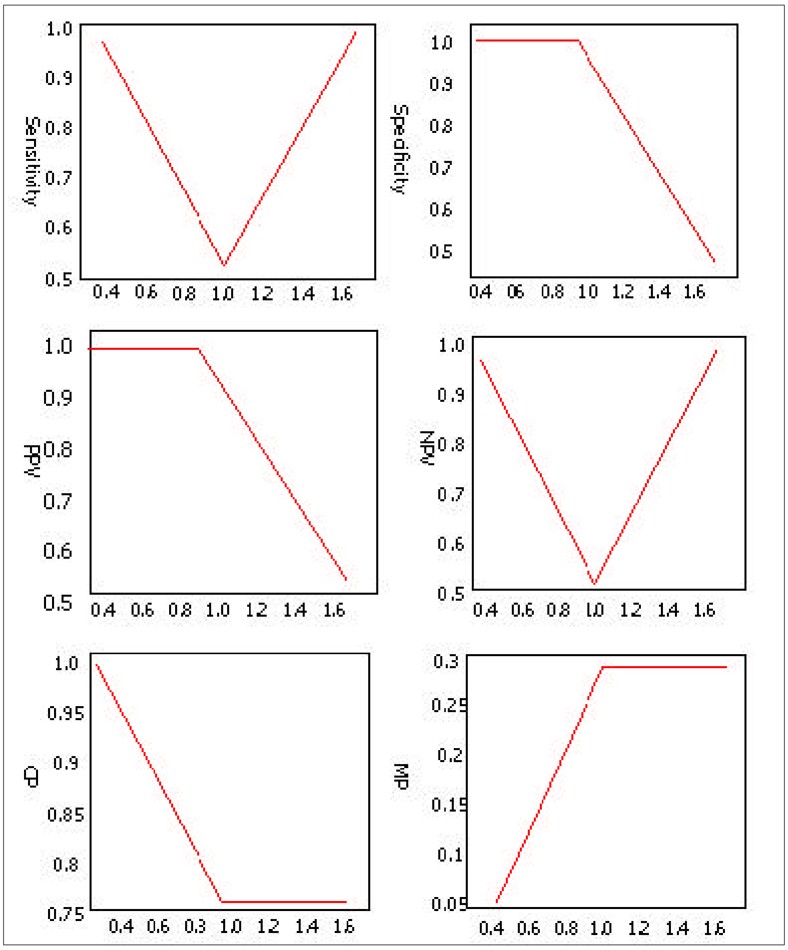
부록 1. 표준편차에 따른 항목들의 변화 radial (cost=1.4)



부록 2. **표준편차에 따른 항목들의 변화** polynomial (cost=1.4)



부록 3. **표준편차에 따른 항목들의 변화** radial (ν =0.3)



부록 4. 표준편차에 따른 항목들의 변화 polynomial (ν =0.3)

ABSTRACT

Microarray Gene Expression Data Classification Using Support Vector Machine

Ku, Kyung Min
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

In the thesis, we introduce the Support Vector Machine(SVM) classification from microarray data and use simulation of microarry data for kernel-function in order to evaluate SVM. The point of this thesis is to evaluate by SVM classifier using microarray data which is generated by adopting S-PLUS and R-Package.

In conclusion, the simulation result has the following result.

First, the increase of log ratio(Cy5/Cy3), the value of each evaluation item (sensitivity, specificity, Positive Predicted Value, Negative Predicted Value, Correct Proportion, Miss Correct Proportion) was improved. The intensity of Cy5 appeared high in microarray experiment.

Second, classification was more accurate but, there was no significant difference between the kernel-function and classification method. With the increase of standard deviation, the value of each evaluation item was decreased. And the classification became poorer as the noise in microarray experiment increased.

Key Word : Support Vector Machine, Microarry, Kernel-function, SVM classifier