SVM을 이용한 Microarray Gene Expression 자료의 Multiclass Classification

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
김 효 미

SVM을 이용한 Microarray Gene Expression 자료의 Multiclass Classification

지도 변 혜 란 교수

이 논문을 석사 학위논문으로 제출함

2002년 12월 일

연세대학교 대학원
의학전산통계협동과정
의학통계학전공
김 효 미

김효미의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2002년 6월 일

감사의 글

이제 막 본격적인 여름이 시작되는 지금, 작지만 저에게는 너무나 소중한 결실을 맺게 되었습니다. 이러한 결실을 맺게 해주신 하나님께 감사드립니다.

미숙한 저를 늘 지켜봐 주시고 지도해주신 김동기 교수님과 변혜란 교수님, SVM의 프로그램에 많은 도움을 주신 김동건 교수님께 감사드립니다.

언제나 마음 편하게 물심양면으로 도와주신 시부모님, 지금의 이러한 자리까지 올 수 있게 해주신 부모님, 멀리서나마 항상 저를 걱정해 주시는 할머니, 하늘에서도 저의 성공을 빌어주시는 할아버지께 감사의 말씀을 드립니다. 그리고 끊임없는 격려로 용기를 북돋아준 남편과 짓궂은 장난으로 항상 방해하던 나의 딸 현지와 나의 하나뿐인 동생 효남과 제부에게도 고맙다는 말을 전합니다.

학교생활에 있어 많은 도움을 주시고 챙겨주신 이효민 연구관님과 은경언니께 감사의 마음을 드립니다. 2년 반전 처음 의학통계학과에 들어와서 많은 도움을 주었던 시내언니와 기준오빠, 프로그램에 도움을 주었던 성민오빠, 힘들던 대학원 생활을 활력있게 만들어 주었던 경민, 윤주, 찬미, 민지에게도 감사의 말을 전합니다.

마지막으로 언제나 좋은 선배였던 일훈오빠와 영돈과 유의, 어릴 때부터 함께 해주던 나의 가장 좋은 친구 영순과 무영, 원열, 봉섭씨, 장섭씨, 희선씨등 연구실 식구들에게도 고마움을 전합니다.

2002년 7월

김효미 드림

차 례

그림 차례

표 차례

국 문 요 약

SVM을 이용한 Microarray Gene Expression 자료의 Multiclass Classification

본 논문에서는 Multiclass Microarray Gene Expression 자료를 분석하기 위하여 최근 생물공학(Bioinformatics)에서 대두되고 있는 Support Vector machine(SVM)을 소개하고, 또한 모의 실험을 통하여 얻은 Microarray gene expression 자료의 분류분석을 시행하였다. SVM을 사용하여 얻은 모형의 평가와 평가를 위한 microarray 자료의 생성은 Chen이 제시한 R-Package에 있는 SVM 라이브러리를 이용하였다.

모의 실험자료를 이용하여 평가항목(True Positive, True Negative, False Positive, False Negative, Correct Proportion, Miss Correct Proportion)에 대하여 평가한 결과 분류 기법(c-classification, ν-classification)에 따라 별 차이가 없었으며, 모든 모형에서 좋은 분류능력을 나타내었다. 또한 각 모형별 편차가 작을수록, 그룹별 차이가 클수록 더 좋은 분류능력을 나타내었다. 즉, 환자군에서의 Cy5의 강도가 높을수록, Cy5의 강도이 차이가 클수록 더 정확한 분류를 해냈고, 그리고 각 그룹에서의 microarray 실험오차(noise)가 적을수록 더 정확한 분류를 해냈다.

핵심되는 말 : Microarray Gene Expression, Support Vector Machine, 분류 기법

제 1 장 서 론

Human genome project를 포함한 많은 유기체의 genome project가 수행되어 수많은 유전정보가 쏟아져 나옴에 따라, 이를 어떻게 해석하고 서로 연관지을 수 있는가에 대해 많은 관심과 연구가 이루어 지고 있다. 이에 대한 분석의 한 종류로 유전자에 기능에 따라 어떠한 질병과 연관되어 있는가에 대해 많은 관심이 모아지고 있다. 현재의 DNA array 동시에 수천개의 유전자에 대한 발현수치를 측정할 수 있는 능력을 제공해 준다(Keller et al., 2000).

유전자의 발현의 비를 수치화 한 DNA microarray 기술이 나타남에 따라 생물학자들에서 하나의 실험에서 수천 개의 유전자에 대한 발현 수준을 측정할 수 있는 능력을 제공하여 주었다. 초기의 실험에서(Eisen et al., 1998) 유사한 기능을 가진 유전자들은 microarray 결합(hybridization) 실험에서 비슷한 발현 패턴을 나타낸다고 제안하였다. 이런 실험들에서 모아진 자료들은 유전자들의 생물학적인 중요성과 유전자의 기능의 판별을 위해 정확한 의미를 알아야할 필요가 있을 것이다(Brown et al., 1999).

Microarray를 사용한 한번의 실험으로 한 환경에만 발현하는 유전자를 찾을 수 있을뿐만 아니라 발현 정도까지도 알 수 있다. 모든 다른 종류의 세포는 서로 다른 유전자들을 발현하여 그들만의 특징을 나타낸다. 예를 들어 암 세포에만 특별히 발현되는 유전자는 이 암이 생성되는데 이 유전자가 어떠한 역할을 담당했다는 것을 의미하며 이들은 그 암의 진단을 할 때도 많은 도움을 줄 것이다. 이와 같은 암 연구 이외에도 각각 다른 장기로부터 얻은 세포들의 유전자 발현 정도를 알아냄으로서 생명의 신비를 좀더 분명하게 밝힐 수도 있을 것이다. 한마디로 요약해서 인간의 유전자 발현 청사진을 얻는 것이다. 이 청사진을 이용하면 이때까지 볼 수 없었던 유전자들

간의 복잡한 연결 고리들을 한결 쉽게 풀 수 있을 것이다.

DNA microarray 자료에 대한 분석의 문제는 class discovery 와 class prediction 의 두 가지로 나뉠 수 있다. class discovery는 유전자 자료에 대하여 class 가 결정되어지지 않을 경우 class를 정의하는 방법이고, class prediction 은 특정한 유전자가 들어왔을 때 이 유전자를 어떤 class로 할당하는 것에 대한 문제이다.

DNA microarray 자료를 분석하기 위해 사용되는 기계 학습(Machine Learning) 방법은 많은 양의 유전자 발현 자료를로부터 중요한 정보를 찾아내고 분석하기 위한 강력한 기술이다. DNA microarray 자료들을 분류분석하기 위한 여러 가지 방법들이 사용되고 있지만, 본 논문에서는 현재 생물공학에서 대두되고 있는 SVM을 이용하여 분석하고자 한다.

SVM은 최근에 집중적으로 연구되어 왔고 다른 여러 가지 방법들에 대하여 기준이 되어왔으며, 오늘날에는 가장 알려진 분류기법이다. 1995년 Vapnik에 의해 제안된 SVM 방법은 이원 패턴 인식문제를 해결하기 위해 제안된 학습방법으로 부정예제로부터 긍정예제를 분류해 낼 수 있는 결정면을 찾아내는 분류모형이다(Vapnik, 1999). 따라서 2개 이상의 그룹에 대한 분석을 하기 위해서는 이원 분류기를 여러개를 조합하여 다원분류기로 확장하게 된다.

본 논문에서 다루게 될 내용들을 소개하면 다음과 같다. 우선 최근 생물정보학에서 이슈가 되고 있는 SVM 기법에 대하여 소개한 후, 모의로 생성된 여러 가지 형태의 microarray 자료를 사용하여 SVM의 모수를 정하고 그 결과를 비교하였다.

제 2 장 Support Vector Machines

2. 1 Support Vector Machines 의 개요

패턴인식 문제는 주어진 데이터로부터 특정 정보를 자동적으로 찾아내는 문제(Detection)와 주어진 데이터를 2개 이상의 그룹으로 분리하여 특성을 파악하는 문제(Classification)로 크게 나뉘어 진다. 이러한 패턴인식 문제는 의료 이미지 정보의 자동해석(MRI, NMI, X-ray 등), 자동 생산시스템의 품질검사, 컴퓨터 음성인식, 물질 분류, 지질 변화 예측(지진 등), 지문, 홍채, 문자 인식 등의 다양한 응용분야를 가지고 있다. 그리고 이러한 패턴인식 문제를 해결하기 위한 방법으로 전통적으로 통계적 패턴인식과 구문론적 패턴인식 기법이 사용되고 있으며 점차 신경망의 활용이 확대되어 가고 있다.

신경망을 이용하여 패턴인식 문제를 해결하려 할 경우 일반적으로 역전파 알고리즘(Back Propagation Algorithm : BPA)이 가장 많이 쓰인다. 그러나 역전파 알고리즘을 쓸 경우 복잡한 입력 패턴의 분포를 추정하기가 어려우므로 학습단계에서 더 많은 양의 학습 데이터가 필요하게 된다. 지역적 최소값(Local Minimum)을 피하기 위한 초기화 작업이 거의 경험적으로 이루어지며, 수렴속도의 지연 그리고 근사화 및 수렴율에 영향을 미치는 커널 함수의 선택 등이 여전히 어려운 문제로 남아 있다. 또한 수렴성을 증명하는 문제도 풀어야 할 과제로 남아있었다.

이러한 문제들을 1995년 Vapnik에 의해 제안된 Universal Feed Forward 네트웍의 한 종류인 SVM 방법으로 인하여 새로운 연구의 전환기를 맞고 있다. SVM은 첫째, 명료한 이론적 근거에 기반하고 있다. 이는 입력으로부터 어떠한 학습 방법을 이용하는가에 대한 직관적인 해석을 제공해 준다. 즉, 간단하고 명료한 알고리즘을 통하여, 학습을 성공적으로 수행하는 데 영

향을 미치는 요소들을 규명할 수 있다. 둘째, 실제 응용 문제에서 높은 인식 성능을 나타낸다.

그러나 실제 응용에서는 신경망과 같이 보다 복잡한 구조의 패턴 분류를 요구한다. SVM 기법은 이런 Classifier를 이용하여 입력 공간의 비선형적인 높은 차수를 특징공간(feature space)에서 선형적으로 투영하여 해석할 수 있도록 하며, 각 특징공간 사이의 최적의 경계(최적분리면)를 제시한다. 이러한 특성으로 인해 SVM은 비선형 패턴 인식 문제, 함수 회귀문제, Human-Computer Interaction(HCI), 데이타마이닝, 웹 마이닝, 컴퓨터 비전, 인공지능, 예측, 의학진단 등의 분야에서 크게 활용될 것으로 보여, 최근 매우 활발하게 연구가 진행되고 있다.

현재까지 알려진 연구 결과로는 Polynomial Machine, Radial Basis Function Machine 그리고 Two-layer Network Machine 등의 세가지 형태의 Kernel 함수를 사용하고 있다. 그러나 패턴인식의 응용범위가 다양화됨에 따라, 이러한 제한된 방법으로는 많은 한계를 보이고 있다. 또한 SVM은 2차 최적화문제(Quadratic Optimization)를 풀어야 하는 과정을 포함하고 있는데, 패턴인식의 효율성과 정확성을 위하여 높은 차원의 특징 공간으로의 변환이 요구되고 있어 이에 따른 최적화 해법의 개발이 시급하다. 기존에 잘 알려진 비선형 최적화기법(예, Quasi-Newton 방법, Conjugate Gradient 방법)은 이러한 높은 차원의 특징 공간에서의 최적화문제를 해결하기에는 적합하지가 않다. 따라서, 새로운 커널을 이용한 SVM 모델 개발과 함께, 보다 안정적(Stable)인 고차원(High-Dimensional) 2차 최적화 해법의 연구가 필요하다.

2. 2 수용력(Capacity)과 VC 차원(Dimension)

수용력은 어떠한 훈련(training)집단을 오차 없이 학습(Learn) 할 수 있는 모형(Machine) 의 능력이라고 볼 수 있다. 모형이 많은 수용력을 갖고 있으면 과적합(Overfitting) 의 문제가 있는 반면, 낮은 수용력을 갖고 있으면 훈련집단의 에러를 일으킨다.

VC 차원이란 함수 $\{f(\alpha)\}$의 집합의 성질이며, 함수 f 의 여러 가지 클래스로 정의할 수 있다. 함수 $\{f(\alpha)\}$ 에서의 VC 차원은 $\{f(\alpha)\}$에 의해 분할(Shattering) 될 수 있는 최대한의 훈련자료의 수이다(Burges et al., 1998). 분할은 Indicator function들에 의해서 최대한의 학습 자료들을 가지도록 분리할 수 있는 함수의 집합이다. 이러한 VC 차원은 무한하며, 만약 N 개의 점들이 함수들의 집합에 의해서 분할될 수 있다면 N 은 얼마든지 커도 상관이 없다. 만일 N 개의 점들이 주어졌을 때 이 점들은 2^N 가지의 방법들로 분류될 수 있고, 이러한 분류를 올바르게 할당하는 $\{f(\alpha)\}$들의 집합을 찾을 수 있는데, 이러한 것을 점들의 집합이 함수들의 집합으로 분할되었다고 말한다.

$\{f(\alpha)\}$는 일반적으로 hypothesis 라고 불리며, $\{f(\alpha) : \alpha \in \Lambda\}$를 hypothesis space 라고 하며 S 라고 표기한다. 이러한 hypothesis space는 S에 의해서 분할될 수 있는 최대한의 학습 자료의 수라고 정의할 수 있다. VC 차원은 자료의 복잡도(complexity), 잘못된 경계범위(mistake bound), 신경망 에서의 수용력과 computational learning theory 등에서 자주 사용된다. R^n에서의 방향지어진 초평면(oriented hyperplane)의 집합의 VC 차원은 n+1 이다.

2. 3 리스크의 최소화

N 개의 관측치가 있고 각각의 관측치 들은 벡터 $x_i \in R^n,\ i=1,\ \cdots, N$ 와 y_i 라는 클래스로 이루어 졌고, 이러한 자료들은 미지의 분포 $P(x,\ y)$ 에서 독립적으로 동등하게 뽑혔다고 가정하자. 여기서 x_i 를 y_i에 대응시키는 모형을 고려할 수 있는데. 이것은 x 를 조절 가능한 모수 α에 의해 분류되어지는 함수 $f(x,\ \alpha)$ 에 대응시키는 가능한 집합으로 정의할 수 있다.

학습집단에서의 test 오류의 기대값, 즉 리스크 함수(Risk function)는

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

이다. 이 함수의 목적은 함수들 $f(x,\ \alpha),\ \alpha \in A$에서 리스크를 최소화하는 $f(x,\ \alpha^*)$의 α^* 를 구하는 것이다. 하지만 $P(x,\ y)$의 분포를 모르기 때문에 모수 α 가 주어져도 리스크를 계산 할 수 없다. 그러므로 리스크를 최소화를 유도하는 원칙이 필요하다. 이러한 하나의 간단한 원칙은 경험적 리스크의 최소화 원칙이다.

경험적 리스크(Empirical risk) R_{emp} 는

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

이다.

경험적 리스크의 최소화(Empirical risk minimization : ERM)는 리스크 함수 $R(\alpha)$ 를 훈련집단 X에 의해 얻어지는 경험적 리스크 R_{emp} 로 구할 수 있다.

이 식의 N을 무한으로 보낸다는 가정이 주어진다면 이 학습 모형의 경험

적 리스크는 기대되는 리스크로 근사(converge)하게된다. 하지만 적은수의 자료에 대하여는 편차가 커지고 과적합의 현상이 발생한다. 이와 같은 과적합 딜레마를 피하기 위한 한가지의 방법으로 클래스의 함수의 복잡도를 제한하는 방법이 있다(Muller et al., 2001).

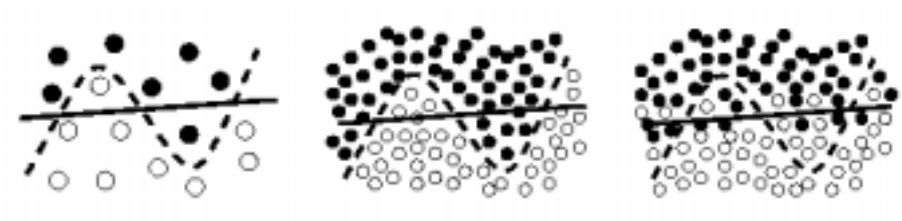

그림 2. 1 Overfitting의 딜레마(Muller et al., 2001)

기대되는 리스크를 줄이기 위해서는 경험적 리스크와 VC 차원과 자료의 수의 비가 작아야 한다. 적절한 VC 차원의 선택은, 특히 자료의 수가 적을 경우 좋은 성과를 얻기 위한 결정적 요인이 된다.

2. 4 구조적 리스크의 최소화

학습 모형(Learning machine)의 훈련자료의 오류가 적을 때 훈련자료에 없는 새로운 자료가 주어질 경우 과연 분류를 잘 할 수 있을 것인가? 만일 학습 모형이 수용력이 높다면 이 모형은 훈련집단은 100% 맞추겠지만, 새로운 자료에 대하여는 잘 맞추지 못할 것이다. 이와 비슷한 현상이 과적합 현상이다. 따라서 학습집단에서 얻어지는 정확도(Accuracy)와 모형의 수용력을 적절히 조절해야 한다. 경험적 리스크를 최소화하는 것으로 좋은 학습모형을 얻을 수 있는 것은 아니다. 또한 적절하게 VC 차원을 선택해야 한다. $\alpha \in \Lambda,\ l > h$ 에서 확률 $1-\eta$ 를 갖는 전형적인 uniform VC 경계는

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

$$\equiv R_{emp}(\alpha) + \sqrt{\frac{h(\eta)}{N}}$$

이다. parameter h 는 VC 차원이며, 오른쪽의 두 번째 더해 지는 식을 VC 신뢰도 라고도 한다. $\frac{1}{h}$ 이 작은 경우 적은 경험적 리스크는 실제의 적은 리시크를 보증하지 못한다. 이러한 경우 실제 리스크를 최소화하기 위해 위의 부등식 의 오른쪽의 부분을 최소화 해야한다. 즉 경험적 리스크와 VC 신뢰도를 동시에 최소화 해야 한다. 이러한 일반적인 방법은 구조적 리스크의 최소화(Structural Risk Minimization, SRM) 라고 한다.(Dibike, 2001) 구조적 리스크를 최소화하는 것의 원리는 경험적 리스크와 VC 차원을 최소화 시키는 것으로 전체 함수들의 클래스 $S=\{f(x, \alpha) : \alpha \in \Lambda\}$를 함수들의 부분집단

$$S_1 \subset S_2 \subset S_3 \ldots\ldots \subset S_n \ldots\ldots$$

으로 나눈다. 각각의 부분집합 S_i 에 대하여 h_i 와 h_j

$$h_i \leq h_2 \leq \ldots\ldots \leq h_n \leq \ldots\ldots \cdot$$

의 경계 범위를 계산한다

훈련집단이 주어졌을 때 SRM 은 리스크 함수의 경계 범위를 최소화 시키는 부분집합 S_k 를 선택하는 것이다. 적절한 부분집합을 선택하는 문제는

수용력의 조절과 관련되어 있다. 부분집합 S_k를 선택하는 SRM 은 경험적 리스크의 경계 범위를 최소화 함으로써 실제 리스크의 최상의 경계 범위를 만든다.

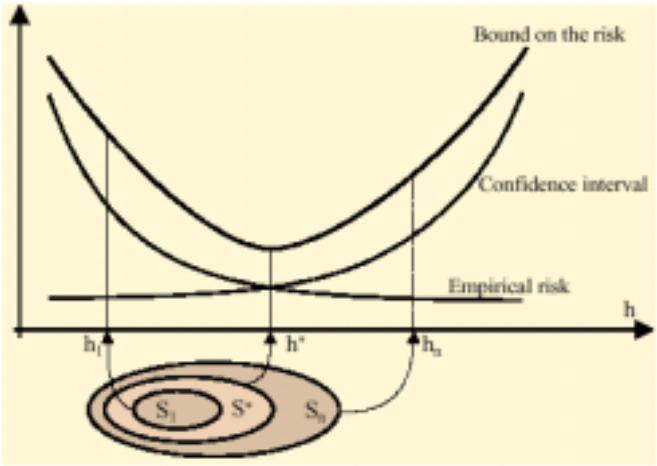

그림 2.2 실제 리스크와 경험적 리스크와 신뢰구간

SRM 은 VC 차원과 리스크의 경계 범위를 최소화 하는 경험적 리스크의 타협점을 마련한다. SRM 의 문제는 S_n 의 VC 차원을 계산하기가 어렵고, VC 차원을 계산 할 수 있는 모델(model)의 수가 적다. VC 차원의 값이 알려져 있다고 하여도 최소화 문제를 푸는 것이 쉽지 않다.

2. 5 라그랑제 이론

w^* 가 $f(w)$, $f \in C^1$ 의 최소값이 되기 위한 필요조건은

$$\frac{\partial}{\partial w} f(w*) = 0$$

이다. 라그랑지안 함수는 목적함수 $f(w)$ 와 등식 제약식 $h_i(w)$, $i=1, \ldots\ldots, m$ 의 선형결합의 합으로 정의된다. 다음의 식

$$L(w, \alpha) = f(w) + \sum_{i=1}^{m} \alpha_i h_i(w)$$

에서 α_i 는 라그상제 승수 이다.

목적함수 $f(w)$ 와 동등 제약식 $h_i(w)$, $i=1, \ldots\ldots, m$ 와 $f, h_i \in C^1$ 의 최적화 문제가 주어졌을 때 라그랑지안 함수의 w와 α 의 최적값은 편미분값

$$\frac{\partial}{\partial w} L(w^*, \alpha^*) = 0,$$

$$\frac{\partial}{\partial \alpha} L(w^*, \alpha^*) = 0$$

으로 구한다. 이 문제에 대하여 부등식제약을 포함하여 일반화 라그랑지안은 다음과 같다

$$\text{minimize} \quad f(w)$$

$$\text{subject to} \quad g_i(w) \le 0, \quad i=1, 2, , \ldots\ldots, k$$

$$h_j(w) = 0, \quad i=1, 2, , \ldots\ldots, m.$$

이 라그랑지안에대한 일반화된 라그랑지안 함수

$$L(w,\ \alpha,\beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{j=1}^{m} \beta_i h_j(w)$$
$$= f(w) + \alpha^t g(w) + \beta^t h(w)$$

이다. SVM 의 주요 요소인 2차(dual) 문제에 대하여 정의하면,

$$\text{minimize}\quad \theta(\alpha,\ \beta) = \inf L(w,\ \alpha,\beta)$$
$$\text{subject to}\quad \alpha > 0$$

이다.

Kuhn-Tuker(쿤-터커) 이론은, $f \in C^1$ 인 볼록함수이고 g_i, h_j가 주어졌을 때 1차 최적문제에서 w^* 가 최적이기 위한 필요충분 조건은 다음과 같은 α^*, β^* 가 존재하는 것이다

$$\frac{\partial}{\partial w} L(w^*,\ \alpha^*,\ \beta^*) = 0\ ;$$

$$\frac{\partial}{\partial \beta} L(w^*,\ \alpha^*,\ \beta^*) = 0\ ;$$

$$\alpha_i^* g_i(w^*) = 0,\ \ i = 1,\ 2,\ ,\ldots\ldots,k;$$

$$g_i(w^*) \le 0,\ \ i = 1,\ 2,\ ,\ldots\ldots,k\ ;$$

$$\alpha_i^* \ge 0,\ \ i = 1,\ 2,\ ,\ldots\ldots,k\ .$$

2. 6 선형 SVM

2. 6. 1 SVM 알고리듬

SVM 의 기원은 1960년대로 거슬러 올라간다. 하지만 1995년 Vladimir Vapnik에 의해 새롭게 대두되었다. 최근 몇 년간 SVM은 문자인식(hand-written character recognition), 영상 분류(image classification)와 biological sequence analysis 등의 실제 자료에서 훌륭하게 수행된다는 것이 증명되었다. SVM의 목적은 가능한 많은 데이터들을 가능한 멀리 두 개의 집단으로 분리시키는 최적의 초평면(hyperplane)을 찾는 것이다 (Markowetz, 2001).

자료들의 형태가 선형으로 판별이 가능한 경우, 분류 초평면 $(w,\ x)+b=0,\ w\in R^n,\ b\in R$ 에 대하여 대응하는 결정 함수

$$f(x)=sign((w,\ x)+b)$$

을 생각해 보자. 선형으로 2개의 클래스로 판별이 가능한 경우 두 클래스를 판별하는 방법은 여러 가지 이다. 이중 최적의 방법을 선택해야 하는 것이 중요한 문제이다. 분류 초평면과 훈련자료 사이들 간의 거리가 좁으면, 훈련자료들에 대해서는 훌륭하게 분류를 할 수 있지만, 새롭게 적용되는 자료에 대하여는 오분류의 가능성이 높아진다. 즉, 새롭게 적용되는 자료에 대하여도 우수한 판별력을 가지기 위해서는 최적의 분류 초평면(Optimal Separating Hyperplane)을 선택해야 한다.

2. 6. 2 OSH(Optimal Separating Hyperplane)의 구축

다음의 등식을 만족하는 단위벡터 w ($\|w\|=1$) 와 상수 b가 존재 할 때 훈련자료 $X=\{(x_1,\ y_1),\ \ldots\ldots,(x_N,\ y_N)\ :x_i\in R^n,\ y_i\in\{-1,\ 1\}\}$ 는 초평면 $(w,\ x)+b=0$ 에 의하여 분류 가능하다고 말하고, 이러한 초평면을 분류 초평면 H라고 한다(Markowetz, 2001).

분류 초평면 H 와 훈련자료 x_i 사이의 거리를 마진(Margin) $\gamma_i(w,\ b)$ 이라고 하며, 벡터들의 집합 $S=\{x_1,\ x_2,\ \ldots\ldots\ ,x_N\}$에 대한 마진은 $\gamma_s(w,\ b)$ 이다. 마진이란 분류초평면과 자료사이의 거리의 최소값이며 그 식은 다음과 같다

$$\gamma_i(w,\ b)\ =\ y_i((w,\ x_i)+b)$$
$$\gamma_s(w,\ b)\ =\ \min_{x_i\in S}\ \gamma(w,\ b).$$

이에 대하여 훈련집단 X에 대하여 OSH 는 다음과 같이 정의되고,

$$(w^*,\ b^*)\ =\ \arg\max_{w,\ b}\ \gamma_x(w,\ b)$$

이러한 초평면은 유일하다. OSH는 마진을 최대화 하는 문제는 $\frac{1}{2}\|w\|^2$를 최소화 하는 문제와 같다. 따라서 OSH를 구축하는 최적화 문제는 다음과 같이 쓸 수 있다

$$\text{minimize} \quad \frac{1}{2}\|w\|^2$$

$$\text{subject to} \quad (w,\ x)+b \geq +1 \ \text{for} \ \ y_i = +1$$

$$(w,\ x)+b \leq -1 \ \text{for} \ \ y_i = -1.$$

위 식에서의 목적함수는 선형 제약을 갖는 볼록함수(Convex function)이다. 이 문제는 선형제약 2차 프로그래밍(Constrained Quadratic Programming) 에 의하여 구해 진다. 라그랑제 방법을 사용하여 최적화 문제를 풀면 라그랑지안 함수

$$L_P(w,\ \alpha,\ \beta) = \frac{1}{2}\|w\| + \sum_{i=1}^{N} \alpha_i [\, y_i((w,\ x_i)+b) - 1]$$

이다. L 에서의 w와 b을 최소화 시키는 안장점(saddle point)은 1차(primal)에 대하여 편미분 하여,

$$\frac{\partial}{\partial w} L_P(w,\ \alpha,\ \beta) = 0$$

$$\frac{\partial}{\partial b} L_P(w,\ \alpha,\ \beta) = 0$$

얻어낸 값들 $w = \sum_{i=1}^{N} \alpha_i y_i x_i$, $\sum_{i=1}^{N} \alpha_i y_i = 0$ 을 라그랑지안 함수에 대입하면 라그랑지안 함수

$$L_D(w,\ \alpha,\ \beta) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\alpha_i \alpha_j y_i y_j (x_{i,}\ x_j)$$

이다. 이 문제에 2차(dual)문제를 적용하면

$$\text{maximize} \quad L_D(w,\ \alpha,\ \beta) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\alpha_i\alpha_j y_i y_j (x_{i,}\, x_j)$$

$$\text{subject to} \quad \alpha_i \geq 0,\ \sum_{i=1}^{N} \alpha_i y_i = 0$$

와 같이 만들어 진다. OSH를 구축하기 위해 2차(Dual)문제에서 푼 계수 α_i^* 를 찾아야 한다. 이 값으로

$$w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i$$

을 구할 수 있다. 벡터 w^* 를 최적 초평면 이라고 하면 함수 $L_D(\alpha)$ 의 최대 값은

$$L_D(\alpha^*) = \frac{1}{2}\sum_{i=1}^{N} \alpha_i^*$$

이고, OSH 의 마진 $\gamma(w^*)$은

$$\gamma(w^*)^2 = (\sum_{i=1}^{N} \alpha_i^*)^{-1}$$

이다.

이렇게 만들어진 최적 초평면에서 마진 초평면에 놓이는 훈련자료의 x_i 와 대응되는 라그랑제 승수들은 0 보다 크지만, 다른 자료들에 대한 라그랑제 승수는 0이 된다. 이러한 자료들은 마진 초평면 H_1과 H_2에 놓이게 된다. 마진 초평면

$$H_1 : (w,\ x_i) = +1$$
$$H_2 : (w,\ x_i) = -1$$

위에 놓이는 벡터들을 지지벡터(Support Vector) 라 한다. 하지만 마진 초평면위에 놓여 있다고 해서 모두 지지벡터가 되는 것은 아니다. α_i 와 $y_i(w,\ x)+b-1$ 두 값이 0 이 되는 경우 이 점은 마진 초평면 위에 존재하지만 지지 벡터는 아니다(Markowetz, 2001). 지지벡터에 대한 라그랑제 승수의 값은 0보다 크게되므로 다음식을 얻을 수 있다

$$w^* = \sum_{i=1}^{\#\,sv} \alpha_i^* y_i x_i^{sv}.$$

선형으로 판별이 가능한 자료의 경우 위의 방법과 같이 적용하여 찾아낸 초평면을 하드 마진 초평면(Hard Margin Hyperplane)를 이용하여 판별할 수 있다.

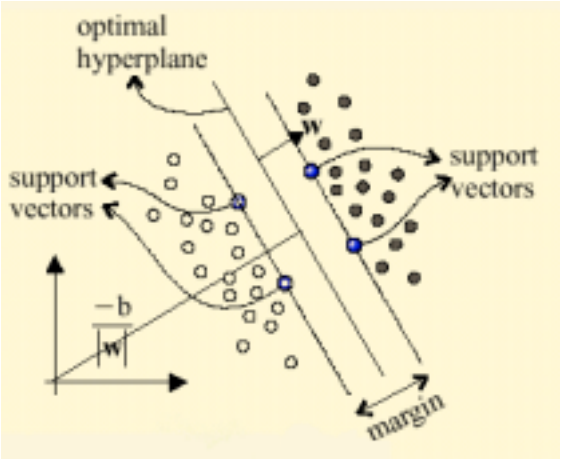

그림 2.3 하드 마진 분류기

선형으로 판별이 가능한 자료에서의 OSH 의 최대의 결점은 오분류를 허용하지 않는 것이다. 따라서 선형으로 판별이 불가능한 자료에 대해서 적용하게 되면 오분류에 대한 오류 패널티에 적용이 된다. 따라서 제약식의 완화가 필요하다. 제약식의 완화는 양수인 다음과 같은 형태의 완화변수 ξ_i

$$(w,\ x)+b \geq +1-\xi_i$$
$$(w,\ x)+b \leq -1+\xi_i$$
$$\xi \geq 0,\ \forall\ i$$
$$\xi = \max\{0,\ 1-y_i((w,\ x_i)+b)\}$$

를 사용한다. 완화 변수가 1보다 크게 되면 오분류로, 0과 1 사이의 값을 가지면 바르게 분류했지만 x_i 는 마진에 놓인 경우이다. 완화변수가 0의 값을 갖게 되면 바르게 분류했고, 마진의 경계 밖에 놓여 있는 것을 말한다.

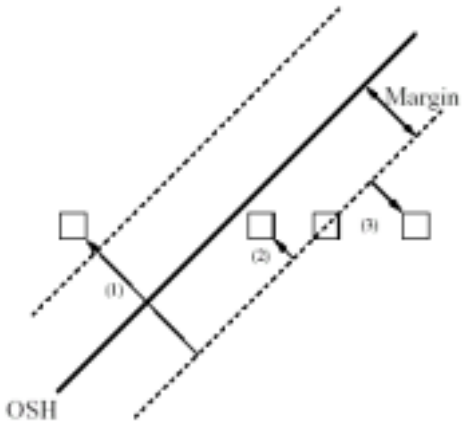

그림 2.4 완화변수의 해석. (1) 오분류
(2) 마진에 놓인 경우
(3) 마진의 경계밖에 놓인 경우

완화 변수를 넣은 라그랑지안 함수

$$L_P(w,\ b,\ \xi_i,\ \alpha,\ \beta)$$
$$= \frac{1}{2}\|w\| + \frac{1}{2}\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i[y_i((w,\ x_i)+b)-1+\xi_i]$$

이다. 이 식을 2차문제에 적용하여 풀면 최적화 문제는

$$\text{maximize}\quad L_D(\alpha^*)$$
$$= \frac{1}{2}\sum_{i=1}^{N}\alpha_i^* - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j((x_i,\ x_j) + \frac{1}{C}\delta_{ij}$$
$$\text{subject to}\quad \alpha_i \geq 0,\ \sum_{i=1}^{N}\alpha_i y_i = 0$$

이다. 이렇게 찾아낸 초평면을 소프트 마진 초평면(Soft Margin Hyperplane) 이라고 한다.

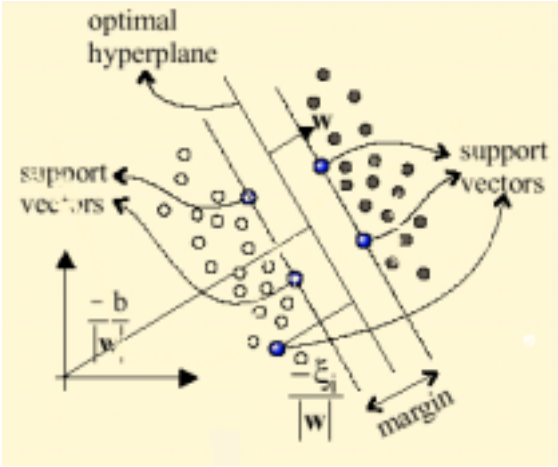

그림 2.5 소프트 마진 분류기

2. 7 비선형 SVM

선형 판별이 불가능한 경우, 자료들을 더 높은 차원의 공간인 특징공간에 사상(Mapping)함으로써 선형으로 판별이 가능하게 만든다. 사상이란 선형으로 분류가능하게 하기 위하여 현재의 입력공간을 더 차원이 높은 특징공간에 자료들을

$$\Phi: R^N \rightarrow F$$
$$x \mapsto \Phi(x).$$

와 같은 형태로 투영(Project) 시키는 것이다. 목적함수를 최대화 시키고 결

정함수를 평가하기 위해서 특징공간에서의 내적 $(\Phi(x_i), \Phi(x_j))$을 계산해야 한다. 이러한 복잡한 계산은 특정한 특징공간 F 와 대응하는 사상 Φ 는 커널함수

$$K(x_i,\ x_j) = \Phi(x_i)\,\Phi(x_j)$$

를 사용함으로서 특징공간에서의 스칼라곱(Scalar product)을 효과적으로 계산할 수 있다. 그러므로 결정함수의 형태는

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

이다. 일반적으로 많이 사용되는 커널함수로는 가우시안 RBF(Gaussian RBF(Radial basis function)) 커널함수 , 다항식(Polynomial) 커널함수, S 모형(Sigmoidal) 커널함수, 역 다이차(Inverse Multiquadric) 커널함수 등이 있다.

Gaussial RBF	$k(x,\ y) = \exp(\frac{-\|x-y\|^2}{c})$
Polinomial	$((x\cdot\ y)+\theta)^d$
Sigmoidal	$\tanh(k(x\cdot y)+\theta)$
inv. multiquadric	$\frac{1}{\sqrt{\|x-y\|^2+c^2}}$

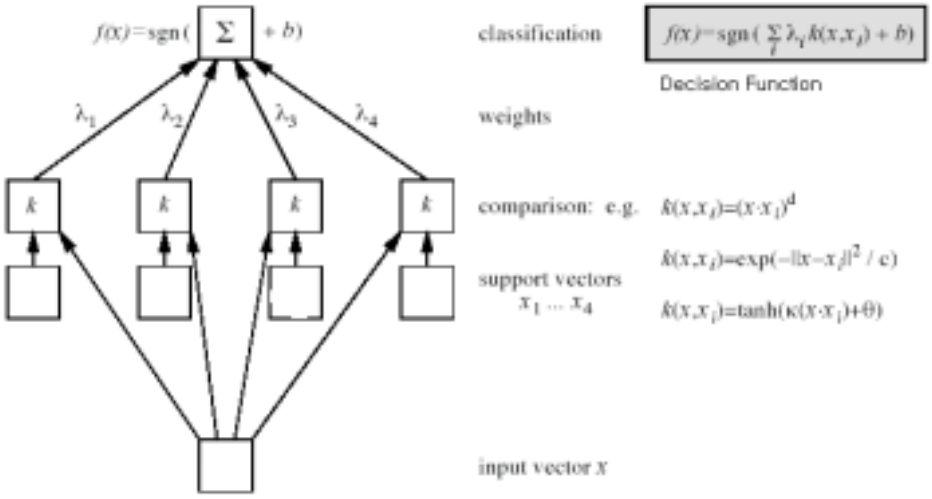
f(x)=sgn(Σ + b)
classification
f(x)=sgn(Σ λi k(x,xi) + b)
Decision Function
λ1
λ2
λ3
λ4
weights
k
k
k
k
comparison: e.g.
k(x,xi)=(x·xi)d
k(x,xi)=exp(-||x-xi||2 / c)
support vectors
x1 ... x4
k(x,xi)=tanh(κ(x·xi)+θ)
input vector x

그림 2.6 커널 함수

2. 8 Multiclass SVM

2. 8. 1 일 대 다(Rest) 검정

앞에서 설명한 SVM은 클래스가 2개일 경우로 제한하여 설명하였다. 하지만 실제 마주하게 되는 자료들은 2개 이상인 경우가 많다. 이러한 경우 자료들을 2개의 클래스를 가진 것 처럼 만들어서 분석하는 방법이 있다. 그 중 한가지 방법이 1 대 다 검정 법이다. 1 대 다 검정법은 n 개의 클래스를 가진 자료에 대하여 n 개의 2 클래스 결정 함수

$$f_k(x) = sign((w_k,\ x) + b_k),\ k = 1,\ 2,\ \ldots\ldots,\ n$$

$$f_k(x) = \begin{cases} +1 & \text{if } x \text{ belong to class } k \\ -1 & \text{otherwise} \end{cases}$$

를 만든다. 분류 규칙은 $f(x) = (f_1'(x),\ f_2'(x),\ \ldots\ldots,\ f_n'(x))$, $f_k'(x) = ((w_k,\ x) + b_k)$ 일 때, x 는 함수 $f_k'(x)$ 의 최대 값에 대응하는 클래스에 속한다고 결정하며 그 식은

$$k^* = \arg\max_k \ \mathrm{f_k}'(\mathrm{x})$$

이다.

2. 8. 2 일 대 일 검정

1 대 1 검정에서 클래스의 각 클래스의 쌍 $(k,\ l)$ $f_{kl}:\ R^n \rightarrow \{+1,\ -1\}$ 결정함수는

$$f_{kl}(x)=\begin{cases} +1 & \text{if } x \text{ belong to class } k \\ -1 & \text{if } x \text{ belong to class } l \end{cases}$$

이다. 이렇게 만들어진 결정함수는 대칭이므로 $f_{kl}=-f_{lk}$ 이고, $f_{kk}=0$ 이고, 만들어지는 결정함수의 개수는 $n(n-1)/2$ 이다. 분류규칙은 $f_k(x)$ 를 최대화 시키는 k 클래스에 x가 속한다고 결정하며 그 식은

$$f_k(x)=\sum_{i=1}^{n} f_{kl}(x)$$

이다.

2. 8. 4 직접접근법

훈련집단 $X=\{(x_i,\ y_i):\ i=1,\ 2,\ldots,\ l\},\ x_i\in R^n,\ y_i\in\{1,\ldots,n\}$ 에서의 선형 분류가능하지 않은 자료의 경우 최대 마진 분류기의 최적화 문제에 대한 일반화의 형식은

$$\text{minimize}\ \ \frac{1}{2}\sum_{k=1}^{n}\|w_k\|^2+C\sum_{k=1}^{n}\sum_{i=1}^{l_k}\xi_i^k$$

$$\text{subject to}\ \ (w_k,\ x_i)+b_k-(w_m,\ x_i)-b_m\geq 1-\xi_i^k$$

$$\text{for} \quad y_i = k$$

$$\text{where} \quad \xi_i^k \geq 1, \quad m \neq k, \quad i = \{1, \ldots, l_k\}$$

이다. 이에 대한 결정함수는

$$f(x) = \arg\max_k f_k(x), \quad f_k(x) = ((w_k, x) + b_k)$$

이다. 여기에서 클래스가 2개 일 때와 다른점은 모든 n 개의 클래스에 대해서 더해준다는 것이다. 1차 라그랑지안 함수에 대하여 안장점을 찾고 2차 (Dual) 에 대한 라그랑지안 함수는

$$\text{maximize} \quad L_D(\alpha)$$

$$= \sum_{k=1}^{n} \sum_{m \neq k} [\sum_{i=1}^{l_k} \alpha_i^{k,m} - \frac{1}{2} \sum_{m^* \neq k} (\sum_{i,j=1}^{l_k} \alpha_i^{k,m^*} \alpha_j^{k,m} (x_i^k, x_j^k)$$

$$+ \sum_{i=1}^{l_m} \sum_{j=1}^{l_m^*} \alpha_i^{m,k} \alpha_j^{m^*,k} (x_i^m, x_j^{m^*}) - 2 \sum_{i=1}^{l_k} \sum_{j=1}^{l_m^{\alpha_i^{k,m}}} \alpha_j^{k,m^*} (x_i^k, x_j^m))]$$

$$\text{subject to}$$

$$0 \leq \sum_{m \neq k} \alpha_i^{k,m} \leq C$$

$$\sum_{m \neq k} \sum_{i=1}^{l_k} \alpha_i^{k,m} = \sum_{m \neq k} [\sum_{j=1}^{l_m} \alpha_j^{m,k}, \quad k = 1, \ldots, n$$

이며, 최종적으로 구해지는 함수는

$$f_k(x) = \sum_{m \neq k} \sum_{i=1}^{l_k} \alpha_i^{k,m}(x_i^k, x) - \sum_{m \neq k, j=1}^{l_m} \alpha_j^{m,k}(x_j^m, x) + b_k$$

이다.

제 3 장 Microarray

3. 1 Microarray와 유전자 발현 (Gene Expression)

1953년 Watson 과 Crick에 의해 밝혀진 DNA의 구조는 분자생물학, 생화학 등과 같은 생명과학 분야에서 매우 중요한 위치를 차지해왔다. 최근들어 인체가 갖고 있는 약 10만 개로 추정되는 모든 유전자 정보를 규명하고자 하는 노력과 함께 유전적 질병을 진단하고 치료, 예방하는데 있어서 막대한 양의 유전자 정보를 신속히 제공할 수 있는 방법에 대한 요구가 크게 증가하고 있다.

그러나 현재 사용되고 있는 분석 방법은 과정이 번거롭고 많은 시간과 노력, 비용 그리고 고도의 숙련도를 필요로 하기 때문에 이러한 단점을 극복할 수 있는 새로운 분석 시스템이 모색되어 왔다. 이러한 필요에 의하여 지난 수년간 미국에서 DNA 칩의 제작 및 이용 기술과 관련된 괄목할 만한 진보가 있었다. DNA 칩은 작은 면적의 고체표면에 염기서열이 알려진 oligonucleotide 탐침(probe)을 정해진 위치에 미세집적(micro-array)시킨 것을 통칭한다.

이러한 DNA 칩에 분석하고자 하는 시료 DNA 단편을 결합시키면 DNA 칩에 부착되어 있는 probe와 시료 DNA 단편상의 염기서열의 상보적 정도에 따라 각기 다른 결합(hybridization) 상태를 이루게 되는데 이를 광학적인 방법 혹은 방사능 화학적 방법 등을 통해 관찰 해석함으로써 시료 DNA의 염기 서열을 측정할 수 있다. 이러한 DNA 칩을 이용한 방법은 PNA 분석 시스템의 소형화를 이루어 극미량의 시료만으로도 진단이 가능하며 시료 DNA 상의 여러 군데의 염기서열을 동시에 규명할 수 있게 함으로써 간편하고도 저렴하고 신속하게 유전 정보를 제공할 수 있다.

인간 유전자의 정확한 개수는 더 많은 실험을 통해 결정돼야 하겠지만, 현재 다국적팀과 셀레라는 대략 3만-3만5천개 정도가 게놈에 존재한다고 예상하고 있다. 이 새로운 추정치는 초파리에 비해 단지 2배 정도에 불과한 것이다. 게놈에서 실제 발현된 염기서열조각(EST, expressed sequence tags)을 연구한 결과에 따르면, 인간의 경우 하나의 유전자에서 하나의 단백질을 생성하는 것이 아니라 평균 3개의 서로 다른 단백질을 만든다고 예측되고 있다. 게놈연구에서 추정된 인간 유전자 수는 이미 잘 알려진 유전자의 구조를 바탕으로 새로운 유전자의 존재 가능성을 분석한 것이다. 따라서 유전자 발현정도가 매우 낮아 현재의 유전자 선별기법으로는 검증되지 못하는 유전자들은 제외될 수밖에 없었기 때문에 유전자 수만으로 인간의 생물학적 복잡성을 설명하기에는 다소 논란의 소지가 있다. 그러므로 인간의 생물학적 복잡성은 단순히 유전자 개수로 설명할 수 있는 사안이 아니고, 유전정보를 담고 있는 유전자의 암호를 풀어 단백질을 만들어내는 과정인 유전자 발현에서 설명되는 것이 더 설득력이 있다.

유전자 발현 분석은 생물학적 과학분야에서 매우 중요하다. 유전자 발현 분석은 유전자 발현의 형태(pattern)의 변화에 수반되는 장기(organism)의 생리적인 변화에 의한 세포의 많은 함수적인 관계를 밝혀준다(Chan 2000). 유전자 발현의 정도의 차이를 찾아내고 다르게 표현되는 유전자를 확인하기 위한 여러 가지 방법들이 개발되었다. 이러한 방법들로는 RDA(representational difference analysis), SAGE(suppression substractive hybridization serial analysis of gene expression), DD-PCR(differential display PCR) 그리고 cDNA-microarray 가 있다.

DNA microarray 자료란 서로 다른 두 실험환경 하에서 여러 유전자들의 발현정도가 어떻게 달라지는지에 대한 비를 수치적으로 표현한 것을 말한다. 즉 수천 개의 유전자에 대한 DNA 의 시퀀스를 두 개의 클라스에 깔아

놓고, 특정 실험환경에서 각각 다른 시각에 채집된 mRNA를 역전사 하여 만든 cDNA를 hybridization 하면 특정 유전자들이 이 cDNA와 특별히 많이 결합되어 발현 수치가 놓아진다. 즉, 수천 개의 유전자에 대해 서로 다른 조건(일반적으로 한 조건은 백그라운드(background)조건으로 하고 다른 한 조건은 heat shock 과 같은 특정한 조건으로 한다)의 cDNA가 얼마나 발현 수준비를 보이는가가 DNA microarray 자료인 것이다.

특정세포의 형질이나 질병의 특성은 한가지 유전자의 이상에 의해서 나타나기보다는 여러 가지 유전자 발현의 변화가 축적되어 결정된다고 할 수 있으며, 유전자 발현의 변화를 총체적으로 이해하는 것이 중요하다. 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, 이 중 가장 보편적인 것이 SAGE (serial analysis of gene expression) 나 cDNA microarray 이다.

3. 2 DNA microarray의 제작 및 분석

Watson and Crick에 의하여 DNA의 이중 나선 구조가 밝혀진 이래 제한 효소의 발견, 결합(hybridization) 기법, PCR등의 발전은 생명 현상의 분자 수준에서 이해에 크게 기여하였다. 그러나 복잡한 조절 기능을 갖는 생명 현상의 단편적 이해는 인간 게놈 프로젝트(HGP:Human Genomic Project)와 같은 전체적 이해를 할 수 있는 실험의 필요성이 대두되었다. 2003년 혹은 그 이전에 모두 밝혀질 염기서열은 그 기능을 이해하는 것이 필수적이고 이러한 과정에서 DNA 칩이 개발되었다. HGP와 DNA 칩의 결과를 효율적으로 활용하기 위하여 생물공학과 Functional Genomics의 연구도 활발하게 진행되고있다.

Microarray는 DNA 칩은 혹은 바이오 칩이라고도 불리우며 유리판, nitrocellulose membrane 혹은 실리콘 위에 target DNA (cDNA 또는 Oligonucleotide)를 붙인 것이다. 형광물질 혹은 방사선 동위 원소로 표식된 탐침(probe)과 결합 시켜 유전자의 발현 정도, 돌연 변이의 확인, single nucleotide polymorphism (SNP), 질병의 진단, high-throughput screening (HTS)등에 사용할 수 있다.

Microarray 자료는 대부분이 방대한 양의 자료들을 가지고 있다. 예를 들어 5개의 샘플과 2번의 반복이 있는 작은 스케일의 실험에서 약 100,000 개의 자료가 생성된다. 따라서 한가지 선호되고 있는 컴퓨터를 이용하는 것이 이러한 방대한 양의 자료를 해석하고 관리하는 것이 필요하다. 자료의 해석을 포함한 3가지 단계는 자료의 표준화(Data normalization), 자료의 필터링(Data filtering) 그리고 패턴인식(Pattern recognition) 이다. 효과적으로 발현수준을 비교하기 위하여 반드시 표준화(normalized) 되어야 한다. 다음으로 data는 적정한 크기로 조절이 되어야 한다.

3. 2. 1 Microarray의 제작

Chip을 제작하는 방식은 크게 기판 위에 oligonuceotide를 직접 합성하는 방식과 합성 또는 증폭된 target DNA를 기판 위에 심는 방식으로 나뉜다. 전자는 반도체 chip을 제작하는 방식에서 유래된 photolithographic 방법을 근간으로 개발된 것으로 고 밀도 집적이 가능한 반면 target DNA의 길이가 20 nucleotides 내외로 제한되는 단점이 있다. 질병의 진단 혹은 SNP의 연구 등에 적합하다. 반면 cDNA microarray는 특이현 유전자의 발현의 연구에 많이 응용되며 poly L-lysine, amine, 혹은 aldehyde로 coating된 슬라이드 위에 target DNA를 심는다. DNA를 plotting하는 방법은 piezoelectric 방

법을 이용한 micropipetting 법 (Gesim, Nano-plotter) 등이 있으며 이때 Spot 의 직경 크기는 100 ㎛ 내외가 되며 약 1000 spots/㎠ 정도가 심어진다. 이때 슬라이드 이외에 plotting 용액도 고려하여야 한다. 일반적으로 SSC용액을 사용하나 Sodium bicarbonate buffer 혹은 MicroSpotting Solution (manufactured by Telechem)등을 시도하여 최적의 실험 조건을 확립하여야 한다

3. 2. 2 탐침의 제작

탐침은 DNAChip의 염기 서열과 서로 상보적인 서열을 갖고 있어 이들의 결합정도가 유전자의 발현량을 결정한다. 따라서 증폭하고자 하는 염기서열은 다른 유전자와 유사성이 적고 해당 유전자에 특이적인 서열을 선택하여야 한다. 또한 mRNA 혹은 cDNA로부터 PCR 또는 RT-PCR을 통하여 탐침 제조 시 Scanner의 LASER 종류에 따라서 표식 하고자하는 형광 물질의 종류를 결정하여야 한다. 즉 Excitation wavelengths, Emission wavelenths 및 Emission filters를 고려하여야 한다. PCR 산물은 정제 하여 (PCR purification Kit Telechem) 사용하는 것이 낮은 background를 나타낸다.

3. 2. 3 Hybridization

Cover glass 밑에서 소량의 probe를 사용하여 이루어지므로 Hybridization cassette (Telechem)를 사용하여 시험 도중 probe의 건조를 방지하여야 한다. Target DNA의 양이 probe보다 약 10배 이상 유지되어야 발현되는 양의 linearity를 확인할 수 있다. 또한 monovalent cation은

Heteroduplex의 형성에 도움을 주므로 최적의 상태를 선택하여야 한다. hybridization 온도는 염기 서열에 따라 다르므로 경험적으로 결정하여야 하나 일반적으로 oligomicroarray (25 - 42 ℃)가 cDNA microarry (55 - 70 ℃)보다 낮은 온도에서 진행된다.

3. 2. 4 Scanning

Microarray를 읽는 장치는 CCD camera system, non-confocal laser scanner, confocal laser scanner 로 크게 나뉜다. CCD camera system은 빠르지만 numerical apeture가 낮은 단점이 있고, (Courtesy from GSI Lumonics) onconfocals은 작동이 간단하나 background artifact가 높다. 현재로는 artifact의 문제나 background측면에서 볼 때 confocal system이 가장 바람직하다. Scanner에서 정밀도가 높은 상을 얻기위해작은 pixcell resolution이 필요하며 현재 5 ㎛까지 가능하다 .(ScanArray 4000, 5000 GSI Lumonics).

또한 높은 sensitivity를 얻기 위하여 다음의 사항을 고려하여 선택하여야 한다.즉 excitation power, the numerical aperture, the transmission filters, the PMT voltage level, number of scans of the image that are summed together. 이와 같은 점을 고려하여 현재 가장 높은 sensitivity는 0.1 molecule fluor/㎛2로 되어있다. 또한 다양한 LASER source는 probe의 표식과 관련이 있으므로 충분히 고려하여야 한다. Microarray 의 분석에는 QuanArray (GSI Lumonics)와 Imagene (Biodiscovery)이 사용되고 있다. Biodiscovery web site를 방문하면 Demonstration 목적의 제한된 용도 프로그램을 down load 받아 사용할 수 있다.

3. 2. 5 cDNA microarray chip 생산 및 검색

cDNA microarray chip은 두가지 다른 환경에서 발현되는 독특한 유전자들을 분석하는데 엄청난 도움이 된다. 수천개 이상의 유전자 발현변이를 단 한번의 실험으로 검색 할 수 있는 것이다. 실험과정을 살펴보면 먼저 두 개의 다른 환경에서 얻은 세포들로부터 mRNA를 추출한다 (Duggan DJ et al., 1999). 이들 mRNA를 역전사 (reverse transcription) 시킬 때 각각 다른 색깔의 형광 물질을 띤 염기를 집어넣어 빨간 색 (Cy5)이나 녹색 (Cy3)을 띤 cDNA를 합성한다. 이와같이 합성된 두 개의 cDNA를 똑 같은 양으로 섞어서 하나의 cDNA microarray chip에 결합시킨다. 결합이 안된 유전자들을 씻어낸 chip은 laser fluorecence scanner에 의하여 읽혀진다. 각각 유전자의 형광 정도는 그 유전자의 발현정도를 알려주는 것으로 이들 정보는 컴퓨터에 의하여 분석되어 진다.

이 방법으로 1:50,000의 빈도로 발현하는 유전자까지 검색할 수 있다. 이렇게 cDNA microarray chip을 사용한 한번의 실험으로 한 환경에만 발현하는 유전자를 찾을 수 있을뿐만 아니라 발현 정도까지도 알 수 있다. 이와같은 방법은 인간의 새로운 암 유발 유전자를 찾을 때나 진단에도 널리 사용할 수 있다. 미국에서 진행되고 있는 CGAP (cancer genome anatomy project)에서도 이 cDNA microarray chip 기술을 사용하여 암관련유전자들의 발현정보를 모으고 있다. 모든 다른 종류의 세포는 서로 다른 유전자들을 발현하여 그들만의 특징을 나타낸다. 이와 같이 암세포에만 특별히 발현되는 유전자는 이 암이 생성되는데 이 유전자가 어떠한 역할을 담당했다는 것을 의미하며 이들은 그 암의 진단을 할 때도 많은 도움을 줄 것이다. 이와 같은 암 연구 이외에도 각각 다른 장기로부터 얻은 세포들의 유전자 발현 정도를 알아냄으로서 생명의 신비를 좀더 분명하게 밝힐 수도 있을 것이

다. 한마디로 요약해서 인간의 유전자 발현 청사진을 얻는 것이다. 이 청사진을 이용하면 이때까지 볼 수 없었던 유전자들간의 복잡한 연결 고리들을 한결 쉽게 풀 수 있을 것이다.

3. 2. 6 Microarray 자료의 표준화

Microarray 데이터의 표준화 작업에는 크게 두가지로 구분할 수 있다. 첫 번째는 데이터의 표준규정을 제안하는 것이고, 두 번째는 컴퓨터데이터의 포맷을 개발하는 것이다. 데이터의 표준규정의 제안은 Minimal Information About a Microarray Experiment (MIAME)으로 불리는 모임에서 microarray 데이터 분석과 응용을 위해서 microarray 실험에서 수록되어야 할 데이터의 종류와 그 데이터를 공유하도록 해줄 수 있는 표준 포맷, 즉 데이터를 수집할 때의 시료와 실험 조건들에 대한 규정의 초안을 제안하였고, 이 모임을 주도한 유럽 생물정보학 연구소(European Bioinformatics Institute)의 Alvis Brazma을 대표로 하여 Nature Genetics 저널에 "Minimum information about a microarray expriment (MIAME)-toward standards for microarray data"라는 제목으로 출판하였다.

두 번째는 microarray 데이터를 전송하고 보관할 수 있는 컴퓨터 포맷을 마련하는 일이다. 이 문제를 해결하기 위한 제안은 MGED(Microarray Gene Expression Database) 협회에 참여한 OMG 그룹에서 초기에 제안되었으며, 현재 미국 버클리에 있는 캘리포니아 주립대학의 Paul Spellman이 이끄는 그룹에 의해 MIAME 데이터를 실험실간에 전달하기 위한 표준 데이터 교환 포맷인 MAGE-ML(Microarray Gene Expression Markup Language)가 제안되어 있다. 이 MAGE-ML는 XML(eXtensible Markup Language)에 기반을 둔 microarray 데이터용 데이터 포맷이며, 대부분의 생물학적 정보들은

웹형식으로 제공하고 있어서 모든 브라우저에 호환성을 지닌 XML형식으로 저장을 하여 다른 응용 프로그램을 이용한 가공이 용이하도록 한 것이다.

그 외에도 microarray 기술을 개발하는 두 개의 생명공학 회사인 Rosetta Inpharmatics Inc.와 NetGenics Inc.에서도 독립적으로 제안을 발표했다.

3. 2 .7 Microarray의 활용분야

20세기 후반에 가장 큰 발전을 한 유전자 조작 기술의 발달로 암이나 질병에 유전자가 관여 한다는 것은 밝혀진 지 오래되었고 병원성 세균의 감염 여부 및 항생제 내성 검사, 신약개발, 유전자의 기능 연구, 동식물 검역, 범죄자 확인 등 많은 부분에 DNA와 유전자 감식 기술이 이용되고 있다. 지금까지는 이러한 사실을 밝히기 위하여 DNA sequencing, RFLP, Southern blot, orthern blot등의 기술을 이용하였다. 그러나 이러한 기술을 사용하여 한번에 여러 개의 유전자 발현변이나 돌연변이를 확인한다는 것은 쉽지 않았다. 하지만 DNA chip은 동시에 이러한 문제점들을 해결해 줄 수 있다. 먼저 DNA chip을 이용한 게놈 차원에서의 유전자 발현 확인은 유전자 발현 청사진을 제공함으로서 과학 기술적인 측면뿐만 아니라 인류의 건강과 생명의 신비를 해석하는데 결정적인 기여를 할 것이다. 또한 인간의 질병이나 암과 관련된 하나이상의 돌연변이들을 동시에 검색할 수 있을 것이다. 이와 같은 일이 현실로 빨리 다가오기 위해서는 생물 공학의 발달이 가장 시급하다. 엄청나게 쏟아지는 유전 정보들의 효율적인 관리가 필요한 것이다(이석기 등, 2002).

제 4 장 모의실험자료를 이용한 Multiclass classification

4. 1 모의 자료의 생성

Microarray Gene Expression 자료를 SVM Multiclass 분류에 적용을 위한 실험을 위하여 여러 가지 형태의 모의 자료를 R.1.5.0 프로그램을 이용하여 생성하였다. 모의 생성 자료는 대조군과 2개의 환자군으로 생성되었고, 각 군별 50명씩, 유전자의 수는 200개로 생성하였다. 모의 자료에 대한 분포는 다변량 정규분포를 가정하였다. 대부분의 Microarray 자료들은 Cy3와 Cy5의 로그비를 사용하므로, 모의 자료에서도 로그비의 값을 구하였다. 2개 그룹의 환자군에 대하여 20%의 유전자를 유의한 유전자로 두었다. 즉 20%에 대해 유전자에 대한 Cy5의 비를 강하게 두었다. 대조군의 경우 로그비를 0으로 두어 유의하지 않은 군으로 설정하였다. 생성된 여러 가지의 자료의 형태는 각 군별 유의한 유전자에 대해 로그비의 차이를 동등하게 준 경우, 군별 로그비의 차이를 다르게 줄 경우, 로그비의 차이를 크게 줄 경우와 작게 줄 경우 등의 형태의 자료를 생성하였다.

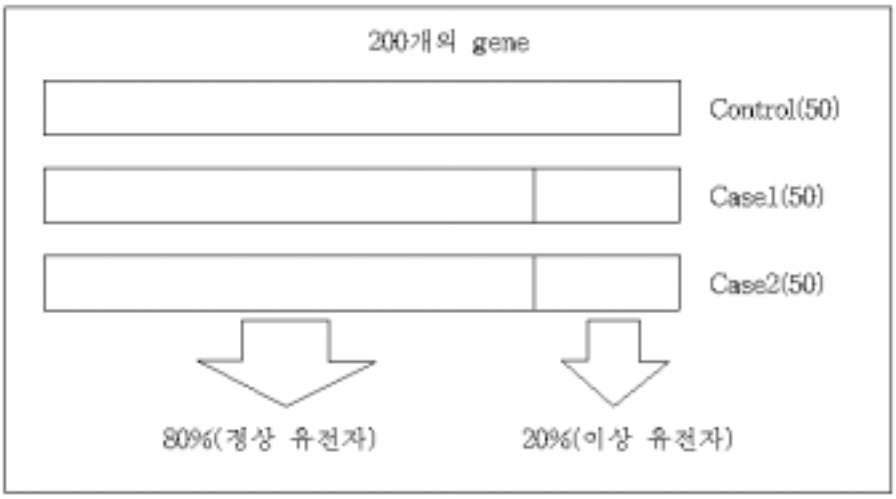

그림 4.1 모의 자료의 생성

유의한 유전자의 경우 자료들의 분포는 다음과 같다.

$$y_i = \log\left(\frac{R}{G}\right) + \varepsilon\ ,\ \ \varepsilon \sim MVN\left(\hat{\mu}_i,\ \hat{\Sigma}\right)\ \ i=2,\ 3$$

유의하지 않은 자료들에 대해서는 다음의 분포에서 생성되었다고 할 수 있다.

$$y_i = \varepsilon\ ,\ \ \varepsilon \sim MVN\left(\hat{\mu}_i,\ \hat{\Sigma}\right)\ \ i=1,\ 2,\ 3$$

생성된 자료는 훈련자료와 검정자료로 나누어 훈련자료를 이용하여 얻은 모형에 검정자료를 이용하여 모형을 평가 하였다. 훈련자료로 사용된 자료의 수의 100개이고, 검정자료로 사용된 자료는 50개이다.

4. 2 모의 실험자료의 평가 방법

SVM은 분류기법에 따라 오류가중치 C 값을 이용하는 c-classification과 모형의 복잡도를 결정하는 ν 값을 이용하는 ν-classification 두 가지의 종류의 분류기법이 있는데, 이 두 모형에 대하여 모수를 추정한 뒤 모형을 평가 하였다. 모수의 추정은 교차타당도(cross validation) 방법의 한가지인 LOO(Leave-one-out)를 사용하여 지지벡터의 수와 정확도를 이용하여 평가하였다.

분류기법에 따라 적용되는 C 값과 ν 값은 그 값이 클수록 분류가 잘 된다. C 값은 오류 페널티 값으로 무한대의 값을 가질 수 있고, ν 값은 지지백터의 수와 오류를 조절하는 값으로 훈련집단에서의 오류의 상한값이며 지지벡터의 하한값이다.(Chang, 2002) 그 값이 0과 1 사이의 값을 가진다. C 값이 무한대까지의 값을 가지고 있기 때문에 ν 보다는 값을 조절하기가 어렵다. 따라서 ν-classification 방법을 이용하여 모수값을 조절하였다. ν 값은 값이 커질수록 모형이 복잡해진다. 커널함수의 scale을 조절해주는 값인 또다른 모수인 gamma 값은 그 값이 작아질수록 지지벡터의 수가 줄어들게 된다.

본 논문에서 모의로 생성된 자료의 경우 linear 커널함수를 사용하여도 분류가 잘 되어 지므로 linear 커널함수를 사용하여 모형을 평가하였다. 그러므로 추정되는 모수는 ν 와 C 값이다. 모형의 평가하기 위하여 구해진 3×3 table을 각 클래스(case)에 대하여 마진합을 구하여 2×2 table로 나타내었다. 평가항목들은 각 클래스에 대하여 True negative(TN), True Positive(TP), False Negative(FN), False Positive(FP), 정분류율(Correct Proportion:CP), 오분류율(Miss Correct Proportion:MP) 값을 구하였다. TN은 평가하고자 하는 하나의 환자군(case1 또는 case 2)에 대하여 그 환자군

이 아닌 다른군에 있는 사람을 평가하고자 하는 환자군이 아닌 군으로 분류하는 것이며, TP는 평가하고자 하는 환자군에 있는 사람을 평가하고자 하는 환자군으로, FN은 평가하고자 하는 환자군을 다른 군으로, FP 는 평가하고자 하는 환자군이 아닌 군에 속한 사람들을 평가하고자 하는 환자군으로 분류한 경우이다.

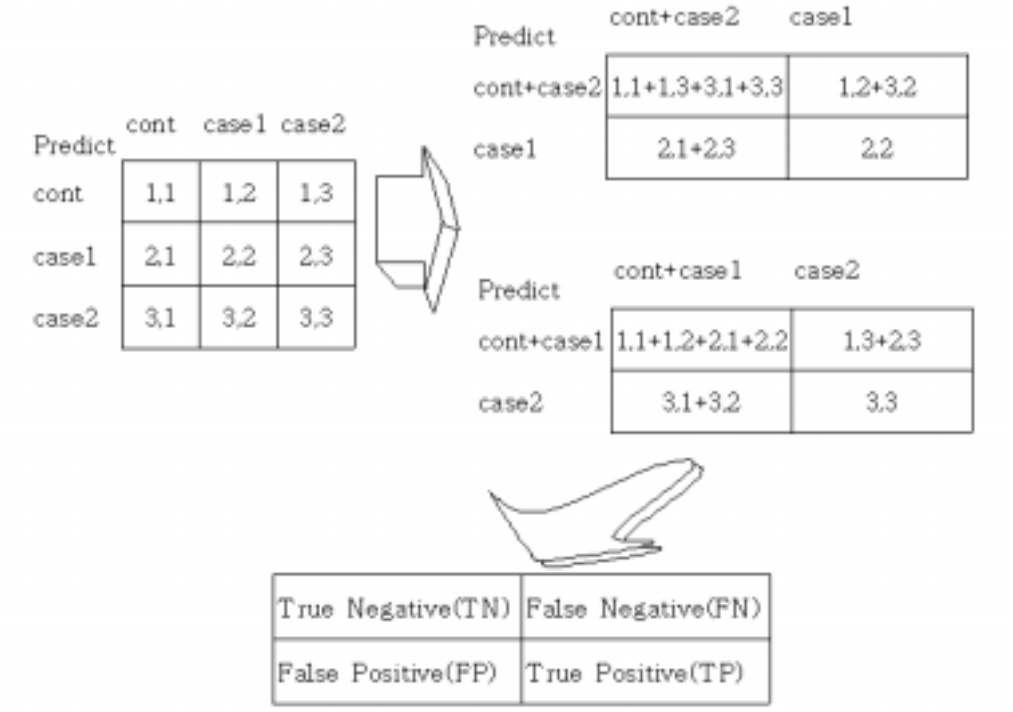

그림 4.2 모형 평가 방법

4. 3 모의 실험자료의 사전 실험

여러 가지 형태의 자료에 대하여 ν-classification 방법을 적용할 경우 추정된 모수값은 다음과 같이 정해졌다.

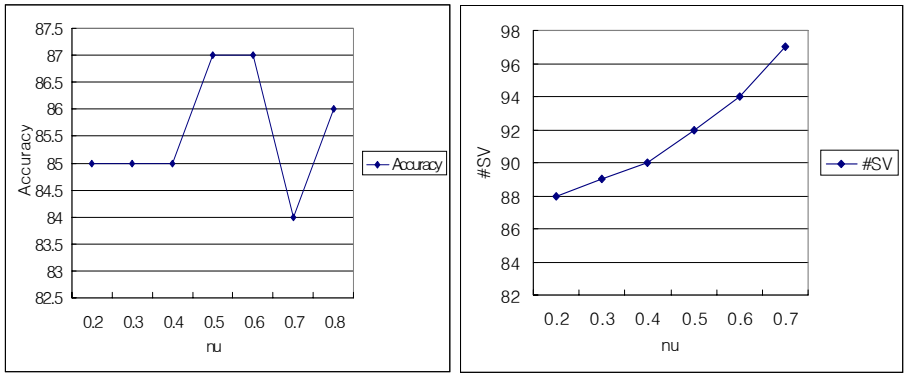

그림 4.3 군별 유전자의 log비의 차이가 log2.0, σ=1일 때의 모수추정

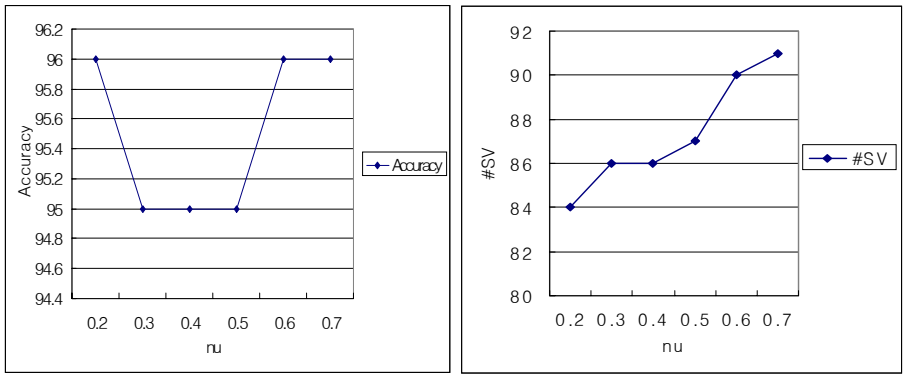

그림 4.4 군별 유전자의 log비의 차이가 log3.0, σ=1일 때의 모수추정

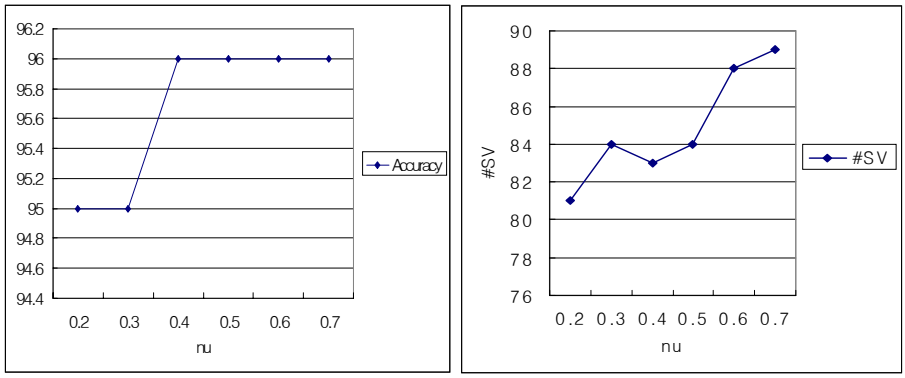

그림 4.5 군별 유전자의 log비의 차이가 log4.0, σ=1일 때의 모수추정

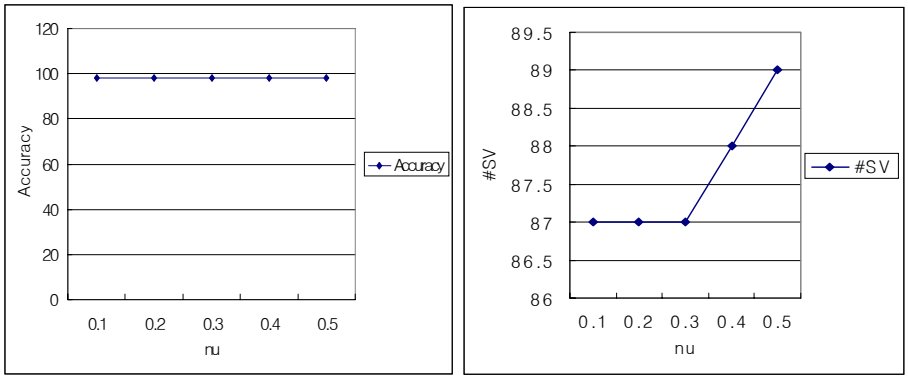

그림 4.6 군별 유전자의 log비의 차이가 log2.0, σ=0.5일 때의 모수추정

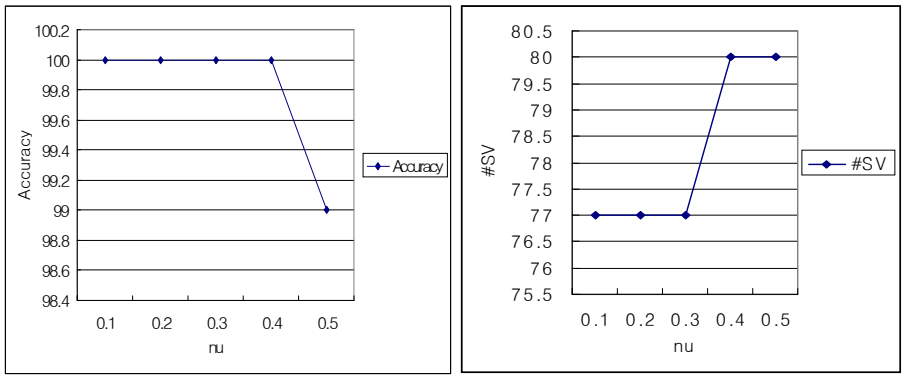

그림 4.7 군별 유전자의 log비의 차이가 log3.0, σ=0.5일 때의 모수추정

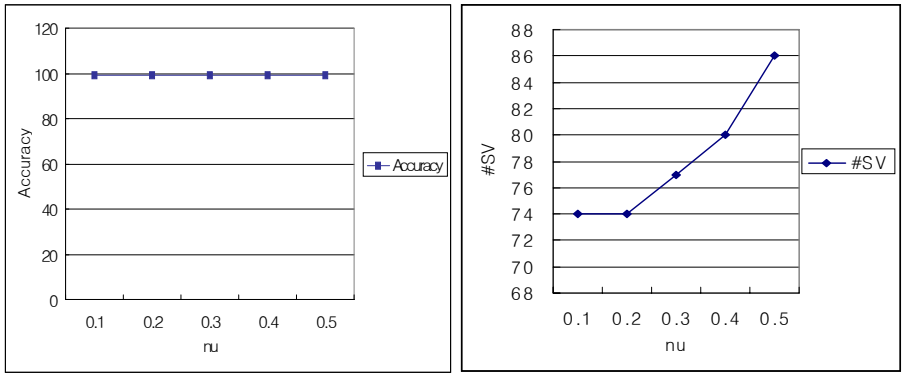

그림 4.8 군별 유전자의 log비의 차이가 log4.0, σ=0.5일 때의 모수추정

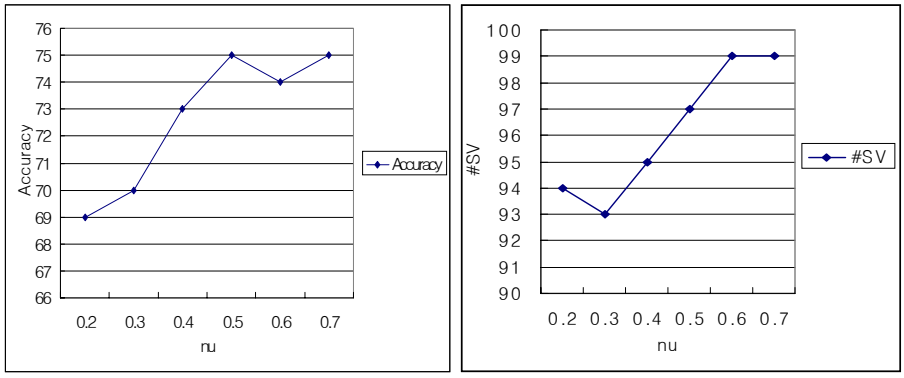

그림 4.9 군별 유전자의 log비의 차이가 log2.0, σ=2일 때의 모수추정

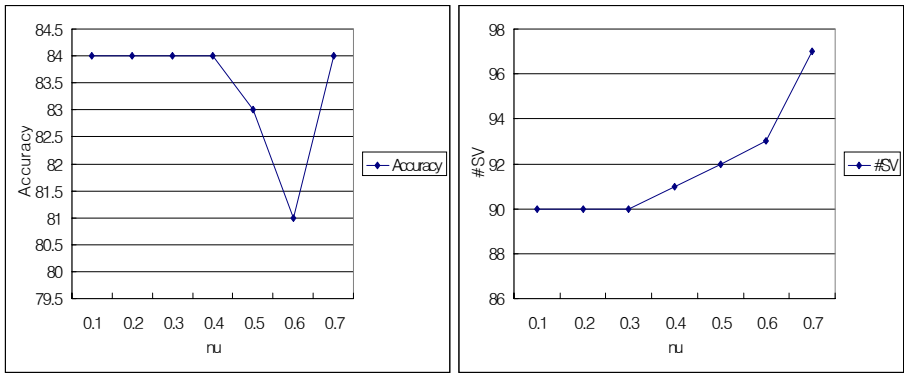

그림 4.10 군별 유전자의 log비의 차이가 log3.0, σ=2일 때의 모수추정

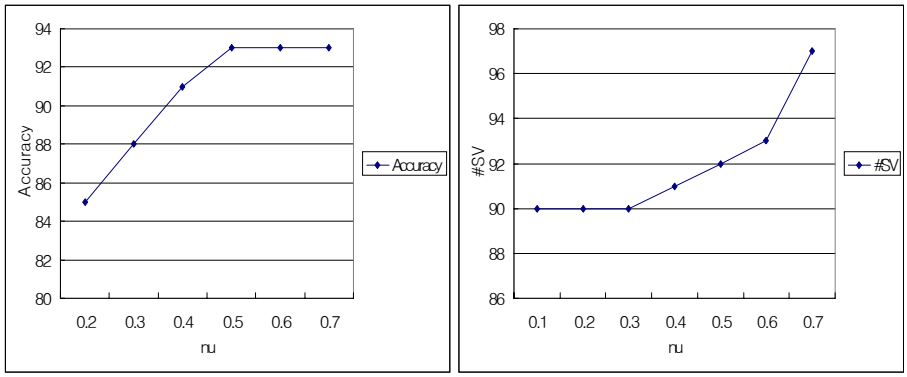

그림 4.11 군별 유전자의 log비의 차이가 log4.0, σ=2일 때의 모수추정

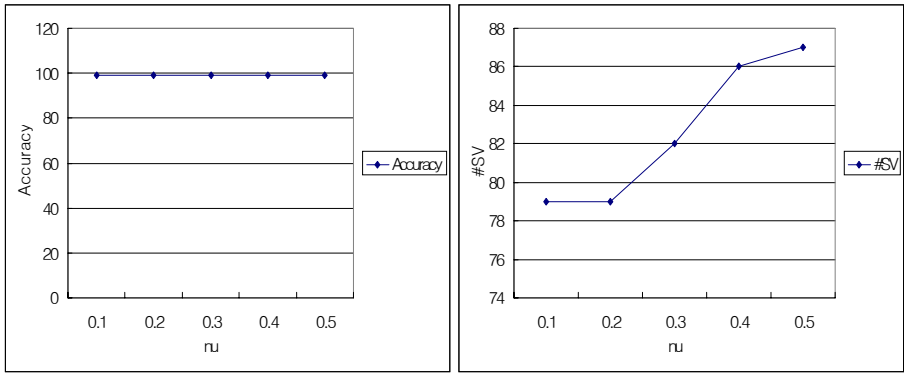

그림 4.12 유전자의 log비의 차이가 log2.0, log6.0, σ=0.5일 때의 모수추정

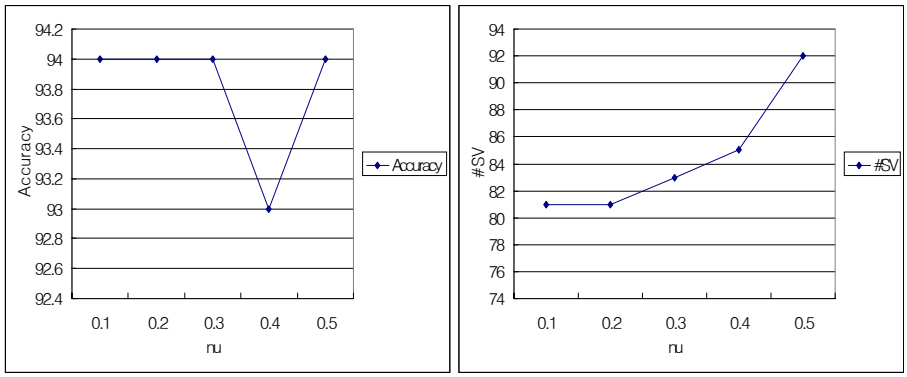

그림 4.13 유전자의 log비의 차이가 log4.0, log6.0, σ=0.5일 때의 모수추정

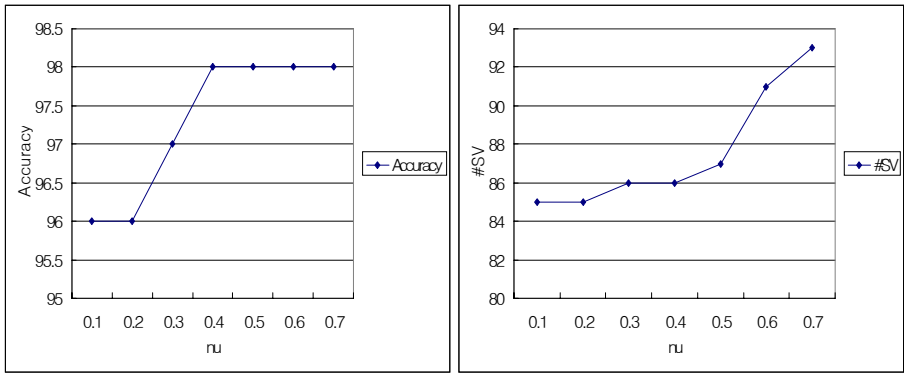

그림 4.14 유전자의 log비의 차이가 log2.0, log6.0, σ=1일 때의 모수추정

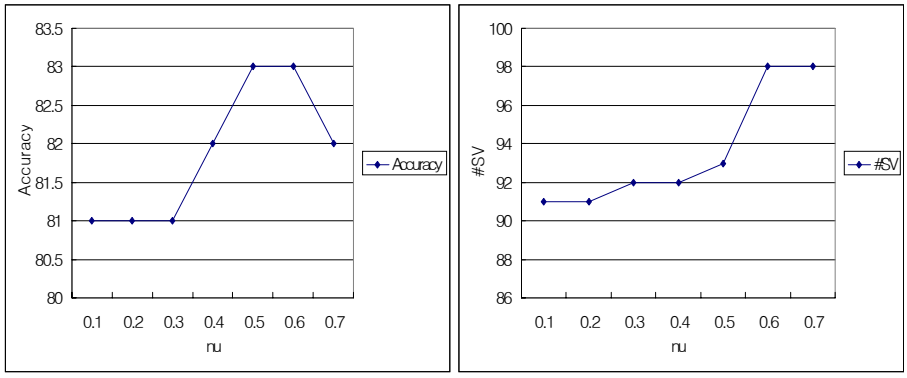

그림 4.15 유전자의 log비의 차이가 log4.0, log6.0, σ=1일 때의 모수추정

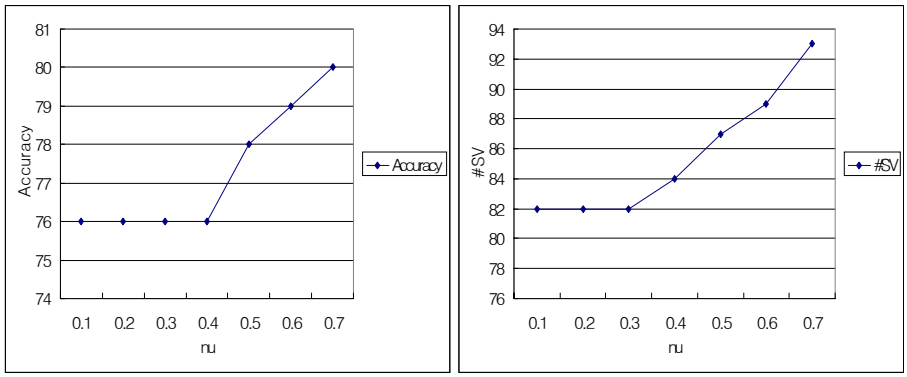

그림 4.16 유전자의 log비의 차이가 log2.0, log6.0, σ=2일 때의 모수추정

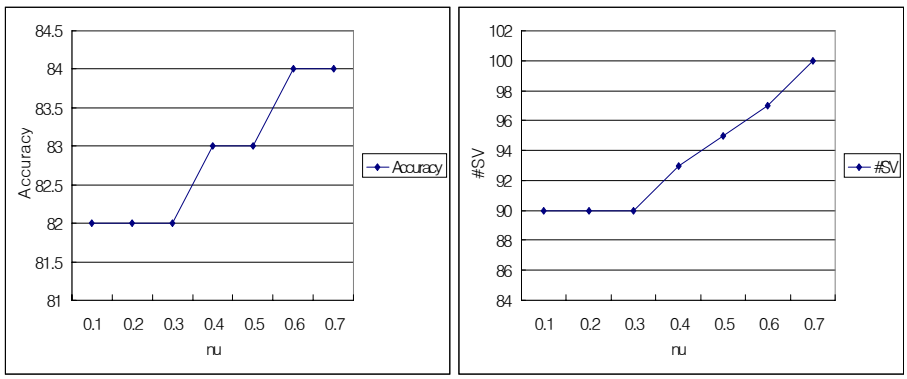

그림 4.17 유전자의 log비의 차이가 log4.0, log6.0, σ=2일 때의 모수추정

표 4.1 모의 자료를 이용하여 추정된 모수(ν-classification)

log ratio		σ	ν
case1	case2		
0.2	0.4	0.5	0.1
		1.0	0.5
		2.0	0.5
0.3	0.6	0.5	0.1
		1.0	0.5
		2.0	0.1
0.4	0.8	0.5	0.1
		1.0	0.4
		2.0	0.4
0.2	0.6	0.5	0.1
		1.0	0.4
		2.0	0.5
0.4	0.6	0.5	0.1
		1.0	0.5
		2.0	0.4

C-classification 방법에서의 추정된 cost 값은 LOO 방법을 사용할 경우 cost 값의 변화에 따라 적중률과 지지벡터의 수의 변화는 없었다. 또한 분류함수에서 사용되는 지지벡터의 놈(norm)의 값들의 변화에 따라서 가장 적절한 cost 값을 정하였다. 이렇게 정해진 cost 값은 1로 정해졌다.

4. 4 모의 실험자료의 모형평가와 결과

모형의 평가 항목 중 정분류율(CP)과 오분류율(MP)의 정확도를 높이기

위하여 100번 반복 실험한 결과를 같이 계산하였다. 모형에 있어서 TN과 TP 값을 클수록, FN 과 FP 값은 작을수록 좋은 모형이다. [표 4.2] 는 유의한 유전자 즉, Cy5의 log비의 차이를 동등하게 두고, 편차를 1로 두었을 때의 최적의 모수에 대한 모형의 평가 결과를 나타내었다.

표 4.2 군별 log 비의 차이를 동등하게 두었을 경우의 모형의 평가

Classification Parameter	Class	Parameter	Gap of log ratio		
			log2.0	log3.0	log4.0
nu	Case1	TP	20	19	18
		TF	27	41	32
		FP	3	0	0
		FN	0	0	0
		CP	0.94	1	1
			0.960±0.02*	0.999±0.004*	1.0±0.0*
		MP	0.06	0	0
			0.040±0.02*	0.001±0.004*	0.0±0.0*
	Case2	TP	11	15	16
		TF	36	33	33
		FP	3	0	1
		FN	0	2	0
		CP	0.94	0.96	0.98
			0.962±0.023*	0.962±0.023*	0.958±0.025*
		MP	0.06	0.04	0.02
			0.037±0.023*	0.037±0.023*	0.041±0.025*
C	Case1	TP	14	24	18
		TF	32	26	32
		FP	1	0	0
		FN	3	0	0
		CP	0.92	1	1
			0.9592±0.0275*	0.9986±0.0058*	0.9998±0.002*
		MP	0.08	0	0
			0.0408±0.0275*	0.0014±0.0058*	0.0002±0.002*
	Case2	TP	17	11	18
		TF	30	37	31
		FP	2	0	1
		FN	1	2	0
		CP	0.94	0.96	0.98
			0.9526±0.0269*	0.9486±0.0354*	0.9562±0.0289*
		MP	0.06	0.04	0.02
			0.0474±0.0269*	0.0514±0.0354*	0.0438±0.0289*

* : number of iteration = 100, (mean±s.d.)

위의 표에서 나타난 결과 유의한 유전자에 대한 log비의 차이가 클수록 더 좋은 결과값을 나타내었다. 즉 Cy5의 강도가 높게 나타날수록 더 정확하게 분류해 냈다. classification의 종류인 C-classification 과 ν-classification 두가지 종류의 분류방법에 대하여 그 차이는 크게 나타나지 않음을 알 수 있다. 즉 분류방법에 따라 최적의 모수에서는 분류능력에는 별 차이가 없었다.

다음의 [표4.3] 부터 [표 4.5] 는 분류방법에 따른 최적의 모수에서의 각 Cy5의 로그비의 차이에서의 여러 가지 편차에 따른 모형의 평가 결과를 나타내었다.

표 4.3 군별 log비의 차이가 log2.0 일 경우 편차에 따른 모형의 평가

Classification Parameter	Class Parameter		Standard Error (σ)		
			0.5	1.0	2.0
nu	Case1	TP	15	20	15
		TF	34	27	29
		FP	0	3	1
		FN	0	0	5
		CP	0.98	0.94	0.88
			0.994±0.009*	0.960±0.02*	0.8654±0.0516*
		MP	0.02	0.06	0.12
			0.005±0.009*	0.040±0.02*	0.1346±0.0516*
	Case2	TP	21	11	10
		TF	29	36	32
		FP	0	3	1
		FN	0	0	7
		CP	1	0.94	0.84
			0.995±0.009*	0.962±0.023*	0.8752±0.0520*
		MP	0	0.06	0.16
			0.004±0.009*	0.037±0.023*	0.1248±0.0520*
C	Case1	TP	17	14	17
		TF	33	32	27
		FP	0	1	4
		FN	0	3	2
		CP	1	0.92	0.88
			0.995±0.0095*	0.9592±0.0275*	0.8728±0.0535*
		MP	0	0.08	0.12
			0.005±0.0095*	0.0408±0.0275*	0.1272±0.0535*
	Case2	TP	16	17	7
		TF	33	30	36
		FP	0	2	3
		FN	1	1	4
		CP	0.98	0.94	0.86
			0.9936±0.0093*	0.9526±0.0269*	0.8762±0.0446*
		MP	0.02	0.06	0.14
			0.0064±0.0093*	0.0474±0.0269*	0.1238±0.0446*

* : number of iteration = 100, (mean±s.d.)

표 4.4 군별 log비의 차이가 log3.0 일 경우 편차에 따른 모형의 평가

Classification Parameter	Class	Parameter	Standard Error (σ)		
			0.5	1.0	2.0
nu	Case1	TP	20	19	15
		TF	30	41	33
		FP	0	0	2
		FN	0	0	0
		CP	1	1	0.96
		MP	1.0 ±0.0*	0.999±0.004*	0.9758±0.0222*
			0	0	0.04
			0.0 ±0.0*	0.001±0.004*	0.0242±0.0222*
	Case2	TP	19	15	18
		TF	31	33	27
		FP	0	0	4
		FN	0	2	1
		CP	1	0.96	0.9
		MP	0.993±0.0134*	0.962±0.023*	0.8696±0.0445*
			0	0.04	0.1
			0.007±0.0134*	0.037±0.023*	0.1304±0.0445*
C	Case1	TP	14	24	20
		TF	36	26	27
		FP	0	0	1
		FN	0	0	2
		CP	1	1	0.94
		MP	1.0±0.0*	0.9986±0.0058*	0.8728±0.0535*
			0	0	0.06
			0.0±0.0*	0.0014±0.0058*	0.1272±0.0535*
	Case2	TP	20	11	9
		TF	30	37	35
		FP	0	0	3
		FN	0	2	3
		CP	1	0.96	0.88
		MP	0.9954±0.0093*	0.9486±0.0354*	0.8762±0.0446*
			0	0.04	0.12
			0.0046±0.0093*	0.0514±0.0354*	0.1238±0.0446*

* : number of iteration = 100, (mean±s.d.)

표 4.5 군별 log 비의 차이가 log4.0일 경우 편차에 따른 모형의 평가

Classification Parameter	Class	Parameter	Standard Error (σ)		
			0.5	1.0	2.0
nu	Case1	TP	20	18	13
		TF	30	32	35
		FP	0	0	2
		FN	0	0	0
		CP	1	1	0.96
			1.0±0.0*	1.0±0.0*	0.995±0.0095*
		MP	0	0	0.04
			0.0±0.0*	0.0±0.0*	0.005±0.0095*
	Case2	TP	13	16	16
		TF	37	33	32
		FP	0	1	0
		FN	0	0	2
		CP	1	0.98	0.96
			0.9924±0.0119*	0.958±0.025*	0.8792±0.0473*
		MP	0	0.02	0.04
			0.0076±0.0119*	0.041±0.025*	0.1208±0.0473*
C	Case1		16	18	13
		TP	34	32	34
		TF	0	0	2
		FP	0	0	1
		FN	1	1	0.94
		CP	1.0±0.0*	0.9998±0.002*	0.8728±0.0535*
		MP	0	0	0.06
			0.0±0.0*	0.0002±0.002*	0.1272±0.0535*
	Case2		17	18	13
		TP	33	31	28
		TF	0	1	7
		FP	0	0	2
		FN	1	0.98	0.82
		CP	0.9948±0.0096*	0.9562±0.0289*	0.8762±0.0446*
		MP	0	0.02	0.18
			0.0052±0.0096*	0.0438±0.0289*	0.1238±0.0446*

* : number of iteration = 100, (mean±s.d.)

위의 결과에서 보듯이 편차가 커질수록 분류율이 낮아지는 것을 알 수 있다. 그러나 군별 log비의 차이가 커질수록 군별 로그비의 차이보다 적은 분산에 대하여는 분류율에 큰 차이가 나타나지 않았다. log비의 차이가 log3.0 과 log4.0 인 경우에서의 편차가 0.5 와 1.0 인 경우에 대하여 분류율에 대한 값이 비슷하게 구해졌고, 편차가 2.0인 경우에는 분류율이 떨어지는 것을 알 수 있다. 즉, microarray 실험에서 구해진 Cy5의 값의 오차(noise)가 많을수록 분류율이 떨어진다는 것이다. 이번 결과에서도 앞에서와 마찬가지로 분류방법에 대하여 분류율의 값은 별 차이가 없었다.

다음의 [표 4.6] 과 [표 4.7] 은 군별 유의한 유전자의 로그비의 차이가 동등하지 않을 경우에 대한 결과이다.

표 4.6 군별 log비의 차이가 다를 경우(0, log2.0, log6.0) 편차에 따른 모형의 평가

Classification Metod	Class	Parameter	Standard Error (σ)		
			0.5	1.0	2.0
nu	Case1	TP	16	12	8
		TF	34	36	36
		FP	0	1	5
		FN	0	1	1
		CP	1	0.96	0.88
			0.9972±0.0069*	0.9542±0.0298*	0.8824±0.0533*
		MP	0	0.04	0.12
			0.0028±0.0069*	0.0458±0.0298*	0.1176±0.0533*
	Case2	TP	16	20	22
		TF	34	30	26
		FP	0	0	0
		FN	0	0	2
		CP	1	1	0.96
			1.0±0.0*	0.9988±0.0055*	0.9758±0.0209*
		MP	0	0	0.04
			0.0±0.0*	0.0012±0.0055*	0.0242±0.0209*
C	Case1	TP	15	17	18
		TF	34	32	28
		FP	0	0	4
		FN	1	1	0
		CP	0.98	0.98	0.92
			0.995±0.0091*	0.995±0.0091*	0.995±0.0091*
		MP	0.02	0.02	0.08
			0.005±0.0091*	0.005±0.0091*	0.005±0.0091*
	Case2	TP	16	15	17
		TF	34	35	32
		FP	0	0	1
		FN	0	0	0
		CP	1	1	0.98
			1.0±0.0*	0.0±0.0*	1.0±0.0*
		MP	0	0	0.02
			0.0±0.0*	0.0±0.0*	0.0±0.0*

* : number of iteration = 100, (mean±s.d.)

표 4.7 군별 log비의 차이가 다른 경우(0, log4.0, log6.0) 편차에 따른 모형의 평가

Classification Metod	Class	Parameter	Standard Error (σ)		
			0.5	1.0	2.0
nu	Case1	TP	19	16	13
		TF	31	34	37
		FP	0	0	0
		FN	0	0	0
		CP	1	1	1
			1.0±0.0*	0.9998± 0.002*	0.9968±0.0093*
		MP	0	0	0
			0.0±0.0*	0.0002± 0.002*	0.0032±0.0093*
	Case2	TP	17	13	10
		TF	34	28	30
		FP	0	6	2
		FN	3	3	8
		CP	0.94	0.82	0.8
			0.908±0.0414*	0.827±0.0515*	0.749±0.0599*
		MP	0.06	0.18	0.2
			0.092±0.0414*	0.173±0.0515*	0.251±0.0599*
C	Case1	TP	17	13	16
		TF	33	37	34
		FP	0	0	0
		FN	0	0	0
		CP	1	1	1
			1±0*	1±0*	0.9958±0.0091*
		MP	0	0	0
			0±0*	0±0*	0.0042±0.0091*
	Case2	TP	12	12	12
		TF	31	28	25
		FP	4	7	7
		FN	3	3	6
		CP	0.86	0.8	0.74
			0.9148±0.0352*	0.8158±0.0601*	0.7514±0.0701*
		MP	0.14	0.2	0.26
			0.0852±0.0352*	0.1842±0.0601*	0.2486±0.0701*

* : number of iteration = 100, (mean±s.d.)

위의 표에서 보듯이 유의한 유전자의 log비의 차이가 다를 경우에서도 편차에 대해서는 편차가 클수록 분류율이 떨어지는 것을 알 수 있다. 유의한 유전자의 log비를 다르게 두어 생성한 자료중 첫 번째의 자료에 대하여 첫 번째 환자군 보다 두 번째 환자군에서의 값들이 더 좋은 값들을 나타내었고, 두 번째 자료에서는 첫 번째 자료에서와 반대되는 첫 번째 환자군에서 더 좋은 값들을 나타내었다. 즉 분류하고자 하는 그룹이 다른 그룹들에 비하여 멀리 떨어진 경우에 대하여 좋은 분류능력이 더 좋다는 것이다.

제 5 장 결론 및 고찰

특정세포의 형질이나 질병의 특성은 한가지 유전자의 이상에 의해서 나타나기보다는 여러 가지 유전자 발현의 변화가 축적되어 결정된다고 할 수 있으며, 유전자 발현의 변화를 총체적으로 이해하는 것이 중요하다. cDNA microarray 는 유전자 발현 검색 시스템의 하나로 이 실험의 결과로 여러 개의 유전자 발현변이나 돌연변이를 확인할 수 있다. 이 실험으로 생성된 자료를 분석하기 위한 하나의 방법이 SVM을 이용한 분류분석이다.

SVM은 최근에 집중적으로 연구되어 왔고 다른 여러 가지 방법들에 대하여 기준이 되어왔으며, 오늘날에는 가장 알려진 분류기법이다.

microarray 자료에 대하여 SVM 기법을 적용하기 위하여 여러 가지 형태의 모의 실험 자료를 R 1.5.0 Package를 이용하여 생성하여 R 1.5.0 Package 의 SVM 라이브러리의 Multiclass classification 기법에 적용하였다.

모의 실험자료를 이용하여 평가한 결과 분류 기법에 따라 별 차이가 없었으며, 모든 모형에서 좋은 분류능력을 나타내었으며, 각 모형별 편차가 작을수록, 그룹별 차이가 클수록 더 높은 분류능력을 나타내었다. 즉, 환자군에서의 Cy5의 강도가 높을수록 더 정확한 분류를 해낸다는 것과, 각 그룹에서의 실험오차(noise)가 적을수록 더 정확한 분류를 해낸다는 것이다.

모의 실험자료를 이용한 실험은 유전자의 수를 200개로 정하였다. 하지만 실제 자료들의 경우 더욱 많은 유전자에 대해 다루고 있다. 모의실험에서 사용된 유전자들 중 결과에 영향을 주는 유전자는 고작 40개에 지나지 않는다. 즉 나머지 160개의 유전자는 결과에 영향을 주지 않는다. 이러한 변수들을 분석을 통해서 제거된다면 더 적은 수의 지지벡터를 이용하여 같은 결과를 도출할 것이다. 즉 이러한 변수는 제거되는 것이 바람직하다고 생각된다.

수많은 유전자들 중에서 판별 유전자를 찾아내는 것은 기초적이고 실재적인 관심사이다. 생물학과 의학 등의 연구에서는 수많은 유전자들 중에서 어떠한 것이 클래스를 잘 구별해 줄 수 있는 가장 우선 순위의 유전자를 조사하면 이득이 될 것이다.

참고문헌

이석기, 박정애, 김규원, DNA microarray 실험 데이터의 표준화 작업. *Biowave*(*http://bric.postech.ac.kr/webjine*), 2002, 4(4)

Andrew D. Keller, Michel Schummer, Bayesian Classification of DNA array Expression Data. *Technical Report* UW-CSE-2000-08-01, 2000

Bernhard Scholkopf, Alexander J. Smola, Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, 2001

Bernhard Scolkopf, Statistical Learning and Kernel Methods. *Technical Report* MSR-TR-2000-23, 2000

Brazma A., Volo J., Gene expression data analysis. *Federation of European Biocemical Societies : Letters*. 2000, 480(1): 17-24

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

Chin-Chung Chang, Chih-Jen Lin, LIBSVM: a Library for Support Vector Machines., 2002

Colin Campbell, Nello Cristianini, Simple Learning Algorithms for Training Support Vector Machines. *Machine Learning*, 1998

Eisen, M., Spellman, P., Brown, P., and Botstein, D. Cluster analysis and displa of genome-wide expression pattern. *pnas*, 1998, 95:14863-14868

Florian Markowetz, Support Vector Machines in Bioinformatics. Mathematische Fakultat der Ruprecht-Karls Universitat Hidelberg, 2001

Giorgio Valentini, Classification of Human Lymphoma Using Gene Expression Data. *DISI*(Dipartimento Di Informatica e Scienze dell'Informazione), 2001

I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning* January, 2002, 46(1/3):389-422

J. Weaton, C. Watkins, Multi-class Support Vector Machine. *Technical Report* CSD-TR-98-4, 1998

Klause-Robert Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, Bernhard Scholkopf, An Introduction to Kernel-Based Learning Algorithms. *IEEE Transactions on Neural Networks*, vol. 12, no. 2, 2001

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

M. O Stitson, J. A. E. Weston, A. Gammerman, V. Vovk, V. Vapnik, Theory of Support Vector Machines. *Technical Report* CSD-TR-96-17, 1996

Nello Cristianini, John Shawe-Talor, An Introduction to Support Vector Machine. Cambridge University Press, 2000

O. L. Mangasarian, Generalized Support Vector Machines. Advances in Large Margin Classifiers. MIT Press, 1999

Simin M. Lin, Method of Microarray Data Analysis. Kluwer Academic Publishers, 2002

Sridhar Ramaswamy et al., Multiclass cancer diagnosis using tumor gene expression signatures. *pnas*, 2001, 98: 15149-15154

Thorsten Joachims, Macking Large-Scale SVM Learning Practical. Advances in Kernel Methods-Support Vector Learning. MIT Press. 1998

T. T Golub, D. K. Slonim, P. Tamao, C. Huard. M. Gaasenbeek, J. P. Mesirov, 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. *Science,* 1999, 15; 286: 531-537(in Reports)

Viann Chan, Nick Hontzeas, Vincent Park, Gene Expression, 2002.

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

Yann Guermeur, A new Multiclass SVM Based on a Uniform

Convergence Result. IJCNN(International Joint Conference on Neural Network), 2000

ABSTRACT

Classification of Multiclass Microarray Gene Expression Data Using SVM

Kim, Hyo-Mi

Dept. of Biostatisticts and Computing

The Graduate School

Yonsei University

In the thesis, we introduce the Support Vector Machine(SVM) that lately being raised in Bioinformatics and carrying out classification analysis using simulation of multiclass microarray gene expression data for classification method in order to evaluate the model obtained by SVM. For the microarray simulation data and evaluation of SVM model using SVM-library in R-Package.

In conclusion, the value of each evaluation item(True Positive, True Negative, False Positive, False Negative, Correct Proportion, Miss Correct Proportion) was no significant difference between the classification method(c-classification, ν-classification). With the decreasing of standard deviation and with the increasing of log ratio, the value of evaluation item was improved. In other words, with the increasing of intensity of Cy5, the value of evaluation item was improved. And the value of

evaluation item became improved as the microarray experiment error(noise) decreased.

Key Ward : Support Vector Machine, Microarray Gene Exprssion Data, Classification method