Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 2, pp. 311 ~ 324

A Comparison Study of Multiclass SVM Methods in Microarray Data[1)]

Jinsoo Hwang[2)] · Jiyoung Lee[3)] · Jeeyun Kim[4)]

Abstract

The Support Vector Machine(SVM) is very functional and efficient classification method to any other classification analysis method. However, its optimal extension to more than two classes is not obvious. In this paper several multi-category SVM methods are introduced and compared using simulation and real data sets. Also comparison with traditional multi-category classification and SVM based methods is performed.

Keywords : Multi-category classification, SVM

1. 서론

마이크로어레이 데이터를 분석하기 위해 사용되는 기계 학습 방법은 많은 양의 유전자 발현 데이터로부터 중요한 정보를 찾아내고 분석하기 위한 강력한 기술이다. 이러한 데이터를 이용하여 분류분석을 하기 위해 여러 가지 방법들이 사용되고 있지만, 그 중 Support Vector Machine(SVM)은 최근에 집중적으로 연구되어 왔고 현재에는 가장 널리 알려진 분류 방법 중의 하나이다. SVM은 Vapnik (1998)에 의해 제안된 방법으로 명료한 이론적 근거에 기반을 두고 있어서 실제 응용문제에서 높은 성능을 나타내고 있다. SVM 방법은 복잡한 구조의 분류기를 사용하여 비선형적인 높은 차수의 문제를 선형적으로 투영하여 해석할 수 있도록 하며 각 집단 사이의 최적의 경계면을 구해준다.

SVM 방법은 원래 두 그룹간의 패턴 인식 문제를 해결하기 위해 제안된 학습 방법이다. 그렇기 때문에 실제 자주 마주하게 되는 3개 이상의 그룹을 가지고 있는 데이터의 경우에는 SVM 방법에 그대로 적용시킬 수 없다. 따라서 3개 이상의 그룹을 가

---

1) 이 논문은 인하대학교 교내연구비로 지원되었음.
2) 제 1 저자 : 인천광역시 남동구 용현동 253번지 인하대학교 통계학과 교수
3) 인천광역시 남동구 용현동 253번지 인하대학교 통계학과 석사과정
4) 교신저자 : 서울특별시 관악구 신림동 산56-1 서울대학교 복잡계통계연구센터 연구원
E-mail : jeeyun@inha.ac.kr

지고 있는 데이터를 분석할 경우에는 두 그룹에서 사용하던 SVM 이진 분류기를 여러 번 사용하여 다원분류기로 확장하여 사용하거나 모든 그룹을 동시에 고려하는 SVM 방법을 사용한다.

본 논문에서는 우선 제2절에서는 다중 그룹인 경우에 사용할 수 있는 SVM 방법들에 대해 정리를 해보았다. 제3절에서는 여러 방법들 간의 특징을 파악하고자 몇 가지 경우의 모의실험을 통하여 각 방법들 간의 오분류율을 계산한 뒤 비교해 보았으며, 제4절에서는 마이크로어레이 분석에서 사용되는 실제 데이터인 SRBCT(small round blue cell tumor) 데이터(Khan 등 (2001)), Leukemia 데이터(Golub 등 (1999)), GCM 데이터(Ramaswamy 등 (2001))를 가지고 여러 방법들 간의 오분류율을 비교하였다. 실제 데이터에서는 유전자의 수가 너무 많고 유전자 선택의 방법에 따라서 분류기의 성능도 달라질 수 있기 때문에 먼저 유전자 선택의 방법에 따른 결과를 제시한 후 2절에서 제시한 SVM을 이용한 다중분류법들을 적용하여 오분류율을 계산하여 비교해 보았다. 끝으로 SVM을 기반으로한 다중분류법들과 기존의 다중분류법에서 널리 쓰이는 k-Nearest Neighborhood(kNN)방법과 여러 분류기를 혼합하는 앙상블 방법인 BagBoosting(Dettling (2004)) 방법과도 오분류율을 계산하여 비교해 보았다. 마지막으로 제5절에서는 결론을 맺고 향후 발전 방향에 대해 제시하였다.

2. SVM을 이용한 다중분류법

SVM은 그룹이 2개인 경우의 문제를 해결하기 위해 제안된 학습 방법이다. 그룹이 2개인 경우의 SVM 이론은 생략하고, 이 절에서는 그룹이 3개 이상일 경우 분석할 수 있는 SVM 다중분류 방법에 대해 설명하겠다. SVM 다중분류 방법은 크게 두 가지 방법으로 나눌 수 있다. 하나는 SVM 이진 분류기를 여러 차례 반복적으로 사용하여 다중분류기처럼 사용하는 방법이다. 여러 그룹을 마치 두 그룹인 경우처럼 분류를 한 뒤, 그 결과들을 결합하는 방법으로 일대일 SVM(one-versus-one, OVO SVM), 일대다 SVM(one-versus-rest, OVR SVM), DAG SVM(directed acyclic graph SVM) 등이 이에 속한다. SVM 이진 분류기들 간의 차이는 자료의 형태에 약간씩 성능이 달라진다고 보고되고 있으나 커다란 차이는 없으므로 본 연구에서는 OVO SVM 방법을 사용하여 다른 분류기와 비교를 하고자 한다. 모든 그룹을 동시에 고려하는 방법으로는 Multi-category SVM(MSVM), Weston and Wakins SVM(WW SVM), Crammer and Singer SVM(CS SVM), 직접접근 SVM 등이 여기에 속한다.

본 논문에서 성능을 비교하고자 하는 SVM 기반 다중분류법은 우선 이진분류기를 반복적으로 이용하는 OVO SVM 방법과 다중분류방법인 MSVM, WW SVM, CS SVM이며 이진 SVM과 기존의 NON-SVM기반 다중분류기를 융합한 성격을 가지는 PairWise Coupling SVM(PWC SVM)이다.

2.1 일대일 SVM (OVO SVM)

OVO SVM은 SVM 이진 분류기를 이용한 가장 기본적인 방법으로 Clarkson이 제안한 분류 방법이다. 훈련 데이터 집합 $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$, $\boldsymbol{x} \in \mathbb{R}^p$, $y \in (1, \cdots, k)$에서 2개의 그룹으로 짝지어진 쌍에 대한 모든 이진 분류기가 요구된다.

$$(w_{ij})^T\Phi(\boldsymbol{x}) + \boldsymbol{b}_{ij} \geq \quad 1-\xi_{ij}, \quad \text{if } y_t = i$$
$$(w_{ij})^T\Phi(\boldsymbol{x}) + \boldsymbol{b}_{ij} \leq -1-\xi_{ij}, \quad \text{if } y_t = j$$

여기서 w_{ij}는 벡터의 방향을 b_{ij}는 기준벡터이다. ξ_{ij}는 구간변수(slack variable)를 나타내며 Φ는 커널함수(Kernel function)를 나타낸다. 이를 이용하면,

$$\min_{w_{ij},\, b_{ij},\xi_{ij}} \frac{1}{2} \mid w_{ij} \mid^2 + C\sum_t \xi_{ij}^t$$

의 새로운 형태의 함수식이 만들어진다. 여기서 C는 모델의 복잡성과 평활도에 대한 정도를 서로 보정해 주는 역할을 하며, 구한 w_{ij}를 support vector라고 한다. 결국 모든 각 그룹의 쌍 (i,j)에 대해 결정함수는

$$\phi_{ij}(\boldsymbol{x}) = sign((w_{ij})^T\Phi(\boldsymbol{x}) + b_{ij})$$

로 만약 $\boldsymbol{x}$가 i번째 그룹에 속한다면 i번째 그룹에 '1'을 더하고, 아닌 경우에는 j번째 그룹에 '1'을 더한다. 테스트 데이터로 계산할 경우 모두 $\frac{k(k-1)}{2}$개의 이진 분류기가 만들어지고 가장 많이 할당되어진 그룹으로 분류한다.

2.2 Multi-category SVM (MSVM)

이 방법은 다중 그룹에 대해 hinge 손실함수를 이용한 방법으로 Lee 와 Lee (2003)에 의해 제안되었다. 주어진 그룹이 k개인 경우를 생각해보자. 이 방법의 각 그룹 라벨(label)은 만약 i번째 표본이 그룹 j로 할당된다면, $y_{ij}=1$이고 나머지의 경우는 $-\frac{1}{k-1}$로 이루어진 k-차원의 $\boldsymbol{y}_i=(y_{i1},\cdots,y_{ik})$라고 표현되는 벡터 값을 갖는다. 주어진 $\boldsymbol{x}\in\mathbb{R}^d$에서 k짝(k-tuple)인 분리함수 $\boldsymbol{f}(\boldsymbol{x})=(f_1(\boldsymbol{x}),\cdots,f_k(\boldsymbol{x}))$에서 $f_j(\boldsymbol{x})=h_j(\boldsymbol{x})+b_j \;\; j=1,\ldots k$ 로 나타내며 $\sum_{j=1}^{k} f_j(\boldsymbol{x})=0$의 조건을 만족한다. 이때 $h_j\in H_k$(H_k는 reproducing kernel Hilbert space를 뜻함)에 속하는 함수이다. 모든 오분류의 비용이 같을 경우의 비용행렬을 Q로 하면 이 행렬은 대각에는 '0'이고 나머지는 '1'인 $k\times k$인 행렬이 된다. 만약 $\boldsymbol{y}_i$가 그룹 j를 나타낸다면 행렬 Q의 j번째 행을 클래스 $\boldsymbol{y}_i$라벨을 나타내는 손실함수를 $L(\boldsymbol{y}_i)$로 정의할 때, $L(\boldsymbol{y}_i)$는 j번째는 '0'이고 나머지는 '1'인 k-차원의 벡터가 되는 것이다. 그러면 최적화 문제는 다음과 같이 표현할 수 있다.

$$\arg\min_{h,b}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}L(\boldsymbol{y}_i)(f_j(\boldsymbol{x}_i)-y_{ij})_+ + \frac{1}{2}\lambda\sum_{j=1}^{k}\mid h_j\mid_{H_k}^{2} \qquad (1)$$
$$subject\ to \sum_{j=1}^{k} f_j(\boldsymbol{x}) = 0$$

여기에서 결정함수는 $\phi(x) = arg\min_j f_j(\boldsymbol{x})$이고, 이를

$$f_j(\boldsymbol{x}) = b_j + \sum c_{ij}K(\boldsymbol{x},\boldsymbol{x}_i),\ j=1,\cdots,k$$

의 형태에 적용시켜, 식 (1)을 만족하는 함수 $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}),\cdots,f_k(\boldsymbol{x}))$를 찾는다. 이 때 $K(\boldsymbol{x},\boldsymbol{x}_i)$는 각 커널함수의 내적을 나타내는 것이다.

2.3 Weston and Watkins SVM (WW SVM)

이 방법은 Weston 과 Watkins (1999)가 제안한 방법으로 $(k-1)n$개의 변수와 제약조건이 요구된다. SVM 이진 분류기를 직접적으로 확장한 모양과 비슷하며 모든 k개의 그룹에 대해 다음과 같은 조건에서의 적절한 w, b를 구한다.

$$\min_{w,\xi}\frac{1}{2}\sum_{p=1}^{k}w_p{}^Tw_p + C\sum_{i=1}^{n}\sum_{p\neq y_i}\xi_i^p$$
$$subject\ to\ \ w_{y_i}^T\Phi(\boldsymbol{x}_i) + b_{y_i} \le w_p^T\Phi(\boldsymbol{x}_i) + b_p + 2 - \xi_i^p,\ \ \xi_i^p \ge 0$$
$$\text{for}\ \ i=1,\cdots,n,\quad p\in(1,\cdots,k)$$

최적화 문제를 풀어 값들이 주어지면, 새로운 $\boldsymbol{x}$에 대한 결정함수는

$$f(\boldsymbol{x}) = arg\max_{p=1,\cdots,k}(w_p{}^T\Phi(\boldsymbol{x}) + b_p)$$

가 된다. WW SVM의 경우 $\sum_{m=1}^{k} b_m^2$항이 더해져 얻어진 수정된 알고리즘을 사용하게 되는데, 이를 bounded formulation이라 부른다. 이는 2차 최적화 문제를 훨씬 간단하게 풀어줄 수 있다. (Hsu 와 Lin (2002))

2.4 Crammer and Singer SVM (CS SVM)

이 방법은 Crammer 와 Singer (2000)가 제안한 방법으로 앞의 WW SVM과 비슷하다.

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}\sum_{p=1}^{k} {w_p}^T w_p + C\sum_{i=1}^{n}\xi_i \\ &subject\ to\ \ {w_{y_i}}^T\Phi(\boldsymbol{x}_i) - {w_p}^T\Phi(\boldsymbol{x}_i) \le e_i^p - \xi_i,\ \ \xi_i \ge 0 \\ &\text{for}\ \ i=1,\cdots,n, \quad p \in (1,\cdots,k) \end{aligned} \tag{2}$$

여기서 $e_i^p = \begin{cases} 0 & \text{if } y_i = p \\ 1 & \text{if } y_i \neq p \end{cases}$ 이다.

식 (2)로 최적화 문제를 풀면 결정함수는 다음과 같이 된다.

$$f(\boldsymbol{x}) = arg\max_{p=1,\cdots,k} {w_p}^T\Phi(\boldsymbol{x})$$

CS SVM은 WW SVM보다 컴퓨팅 시간에 중점을 두어 수정된 알고리즘 bounded formulation을 사용하여 2차 최적화 문제를 푼다.

2.5 SVM을 이용한 PairWise Coupling (PWC SVM)

만약 각 이진 분류기의 결과가 사후확률로 설명되어질 수 있다면 PairWise Coupling 분류기를 사용할 수 있다. 사후확률인 $p_i = P(y_i|\,\boldsymbol{x}),\ i=1,\cdots,k$를 얻기 위해서 OVO SVM 분류의 확률적인 결과를 결합하며, 분류규칙은 가장 큰 p_i가 있는 그룹으로 할당하는 것이다.

PairWise Coupling은 다음과 같다. 먼저 $r_{ij} = P(y_i|\,y_i\ or\ y_j)$의 확률이 주어져 있다면 PairWise Coupling의 목적은 이 주어진 확률 r_{ij}로부터 확률 $p_i = P(y_i)$의 집합을 구하는 것이다. 이 문제를 풀기 위해서는 보조변수 $\mu_{ij} = \dfrac{p_i}{p_i + p_j}$가 필요하며, 결국 r_{ij}에 근사하는 $\hat{\mu}_{ij}$를 찾아 이것과 대응하는 $\hat{p}_i$을 찾는 것이다. 이때, r_{ij}와 $\hat{\mu}_{ij}$사이의 적당한 근사법은 Kullback-Leibler distance로 그 식은 다음과 같다.

$$D^{KL} = \sum_{k<l} r_{kl}\log\frac{r_{kl}}{\hat{\mu}_{kl}} + (1-r_{kl})\log\frac{1-r_{kl}}{1-\hat{\mu}_{kl}} \tag{3}$$

연관된 점수 식은 다음과 같다.

$$\sum_{l\neq k}\hat{\mu}_{kl} = \sum_{l\neq k} r_{kl}, \quad k=1,\cdots,M, \quad subject\ to \sum_k p_k = 1$$

그다음 $\hat{p}_i$과 $\hat{\mu}_{ij}$의 초기값을 정하고, 다음 과정의 반복을 통해 식 (3)의 값을 최소화 시키는 p_i을 찾는다.

1. $\hat{p}_k = \hat{p}_k \cdot (\sum_{l \neq k} r_{kl}) / (\sum_{l \neq k} \hat{\mu}_{kl})$
2. 다시 $\hat{p_i}$와 $\hat{\mu}_{ij}$을 설정한다.

여기서 초기값 $\hat{\mu}_{ij} = r_{ij}$가 되고 OVO SVM 방법을 통해 얻어진 $r_{ij} = P(y_i | \, y_i \, or \, y_j)$가 사용되어 진다. 하지만 SVM의 결과 값은 확률 값이 아니기 때문에 OVO SVM을 통해 얻어진 값을 그대로 PairWise Coupling에 적용할 수 없다. 그래서 Platt (1999)이 제안한 SVM 결과를 시그모이드 함수(sigmoid function)에 적용시켜 확률로 사용되어 질 수 있도록 바꾸어주는 방법을 사용한다. 즉,

$$P(w_k \mid \boldsymbol{x}) = \frac{1}{1 + e^{Af + B}}.$$

여기서 f는 OVO SVM의 결과 값이 되고 A와 B는 Platt (1999)에서 주어진 알고리즘을 사용하여 결정하였다. 결국 이 방법은 OVO SVM 방법으로 얻은 결과 값을 Platt (1999)이 제시한 방법을 이용하여 확률로 바꾸어 PairWise Coupling에 적용한 방법이라고 할 수 있다.

3. 모의실험 결과 및 분석

데이터간의 겹침 정도에 대해 2절에서 설명한 SVM 방법들을 비교하기 위해 3가지 모의실험을 시행하였다. 3가지 모의실험은 모두 각 집단에서 훈련 데이터와 테스트 데이터를 따로 뽑아 적용시켰으며, 이러한 과정을 50번 반복하여 분류기에 적용시킨 후 오분류율을 산출하여 그 평균을 계산하였다.

본 논문에서 SVM 계산을 위해 필요한 커널함수는 RBF 커널함수를 사용했으며, SVM 모델의 복잡성과 평활도에 대한 정도를 서로 보정해주는 C값과 커널함수를 사용할 때 필요한 모수의 값은 훈련 데이터마다 정해진 범위 안에서 3-fold Cross Validation을 통해 가장 최적의 값으로 정하였다. (Hsu 와 Lin (2002))

MSVM은 R package를 이용하였으며, OVO SVM, WW SVM, CS SVM은 GEMS 프로그램(Alexander 등 (2005))을, PWC SVM의 경우에는 SVMTorch 프로그램(Platt (1999))을 이용하여 계산하였다.

먼저 퍼짐의 정도가 모두 다르고 집단간 겹침 정도의 차이가 작은 2개의 집단과 전혀 겹치지 않는 1개 집단으로 3개의 그룹을 가진 600개의 훈련 데이터와 300개의 테스트 데이터를 생성하였다.

$$\underset{\text{그룹 1 } (n_1=100)}{N_2\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right]} + \underset{\text{그룹 2 } (n_2=200)}{N_2\left[\begin{pmatrix}2\\2\end{pmatrix}, \begin{pmatrix}4 & 0\\0 & 4\end{pmatrix}\right]} + \underset{\text{그룹 3 } (n_3=300)}{N_2\left[\begin{pmatrix}-4\\-4\end{pmatrix}, \begin{pmatrix}9 & 0\\0 & 9\end{pmatrix}\right]}$$

분류방법	오분류율(%)	표준오차(%)
OVO SVM	11.92	0.265
MSVM	11.98	0.263
PWC SVM	10.82	0.315
WW SVM	12.16	0.222
CS SVM	11.51	0.250

표 1 . Case1 (k=3, training data=600, test data=300)

표[1]을 보면, PWC SVM 방법이 10.82%로 5가지 SVM을 이용한 방법 중 가장 좋은 오분류율을 나타내고 그 다음으로 CS SVM, OVO SVM 방법의 순으로 오분류율이 좋게 나타났다. WW SVM 방법이 가장 나쁜 오분류율을 나타냈다. 하지만 5가지 방법 모두 거의 비슷한 오분류율을 나타내며 통계적으로 유의한 차이를 보이는 방법은 PWC SVM과 WW SVM 방법뿐이다.

다음으로는 동일한 퍼짐 정도를 갖는 2개의 집단과 그 두 집단을 포함시키는 1개의 집단으로 3개의 그룹을 가진 300개의 훈련 데이터와 150개의 테스트 데이터를 생성하였다.

$$N_2\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}36 & 0\\ 0 & 36\end{pmatrix}\right] + N_2\left[\begin{pmatrix}5\\0\end{pmatrix},\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right] + N_2\left[\begin{pmatrix}-5\\0\end{pmatrix},\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right]$$

그룹 1 (n_1=190) 그룹 2 (n_2=55) 그룹 3 (n_3=55)

분류방법	오분류율(%)	표준오차(%)
OVO SVM	11.53	0.347
MSVM	11.20	0.417
PWC SVM	10.94	0.404
WW SVM	11.61	0.348
CS SVM	11.55	0.457

표 2 . Case2 (k=3, training data=300, test data=150)

표[2]에서 보면, PWC SVM 방법이 10.94%로 5가지 SVM을 이용한 방법 중 가장 좋은 오분류율을 나타내고 그 다음으로 MSVM, OVO SVM 방법의 순으로 좋은 오분류율이 나타났다. WW SVM은 가장 나쁜 오분류율을 나타냈다. 그러나 여기에서도 5가지 방법의 통계적인 차이는 발견할 수 없었다.

마지막으로는 모두 동일한 퍼짐 정도를 갖고 서로 모두 겹침이 있는 3개의 집단으로 3개의 그룹을 가진 300개의 훈련 데이터와 150개의 테스트 데이터를 생성하였다.

$$N_2\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right] + N_2\left[\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right] + N_2\left[\begin{pmatrix}1\\0.5\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right]$$

그룹 1 (n_1=100) 그룹 2 (n_2=100) 그룹 3 (n_3=100)

분류방법	오분류율(%)	표준오차(%)
OVO SVM	44.35	0.477
MSVM	46.52	0.528
PWC SVM	43.26	0.551
WW SVM	43.87	0.442
CS SVM	43.76	0.486

표 3 . Case3 (k=3, training data=300, test data=150)

표[3]에서 보면, PWC SVM 방법이 43.26%로 5가지 SVM을 이용한 방법 중 가장 좋은 오분류율을 나타내고 그 다음으로 CS SVM, WW SVM 방법의 순으로 좋은 오분류율이 나타났다. MSVM은 가장 나쁜 오분류율을 나타냈다. 전반적으로 자료들의 겹침이 많은 경우이기 때문에 분류의 결과가 좋지 않으며 MSVM은 다른 방법과 통계적으로 유의한 차이를 나타내고 있다.

4. 실제 데이터 적용

2절에서 설명한 방법들을 가지고 실제 데이터에 적용시켰다. 실제 데이터로는 SRBCT 데이터, Leukemia 데이터, GCM 데이터를 이용하였으며, SVM 다중분류법을 적용시키기에 앞서 각 데이터에 대해 유전자 선택 방법을 사용하였다. 유전자는 150개를 뽑았으며 뽑힌 유전자만을 가진 데이터로 3-fold Cross Validation을 통해 훈련 데이터와 테스트 데이터를 랜덤하게 적용시켰다. 이러한 과정을 20번 반복하여 분류기에 적용시킨 후 오분류율을 산출하여 그 평균을 계산하였다. 각 분류기에 적용시키는 데이터는 다음과 같다.

데이터	샘플(n)	유전자(p)	그룹(k)
Leukemia	72	5327	3
SRBCT	83	2308	4
GCM	198	16063	14

표 4 . data set

Golub (1999)의 Leukemia 데이터는 총 72개의 샘플 데이터이며, 각 샘플은 5327개의 유전자 발현 정보를 갖고 있다. 백혈병에 대한 데이터로 AML(acute myelogenous leukemia), ALLB(acute lymphoblastic leukemia B-cell), ALLT(ALL T-cell)의 3개의

그룹으로 구성되어 있다. Khan (2001)의 SRBCT 데이터는 5개의 SRBCT가 아닌 데이터를 제외하고 총 83개의 샘플 데이터이며, 각 샘플은 2308개의 유전자 발현 정보를 갖고 있다. 이 자료는 20세 미만이 걸리는 악성 종양에 관한 자료로서 NB(neuroblastoma), RMS(rhabdomyosarcoma), NHL(non-Hodgkin lymphoma), EWS(Ewing family of tumors)의 4개의 그룹으로 구성되어 있다. Ramaswamy (2001)의 GCM 데이터는 총 198개의 샘플 데이터이며, 각 샘플은 16063개의 유전자 발현 정보를 갖고 있다. GCM 데이터는 흔히 볼 수 있는 암에 대한 데이터로 prostate, bladder, melanoma, uterine, leukemia, breast, colorectal, renal, ovarian, pancreatic, lung, lymphoma, central nercous system, pleural mesothelioma의 14개의 그룹으로 구성되어 있다.

4.1 유전자 선택

마이크로어레이 데이터에서는 보통 유전자의 수가 굉장히 많다. 하지만 분류분석을 하는데 있어 모든 유전자가 다 사용되어지는 것은 아니며, 다 사용한다면 오히려 중요하지 않은 유전자에 의해 분류가 잘못되는 경우도 있다. 그래서 유전자 선택의 과정은 마이크로어레이 데이터에서 중요한 과정 중의 하나이다. 특정 그룹 안에서는 상관관계가 높고 다른 그룹과의 상관관계가 낮은 유전자를 선택하는 것이 유전자를 선택하는 가장 보편적인 방법이다.

유전자 수는 유전자 선택에서 첫 번째로 결정해야 하는 사항이다. 적절한 유전자의 수를 결정하는 것은 실험이나 경험에 의존하게 되므로 매우 어렵다. 본 논문에서는 이 실험은 직접 하지 않고 Tao 와 Zhang (2004)의 실험을 통해 결정된 150개로 유전자를 선택하였다.

유전자 선택 방법으로는 Rankgene 프로그램에 나와 있는 총 8가지의 방법을 사용하였다.

1. information gain : $\sum_{i=1}^{k}(\frac{l_i}{n}\log\frac{l_i}{n_l}+\frac{r_i}{n}\log\frac{r_i}{n_r})-\sum_{i=1}^{k}(\frac{l_i+r_i}{n})\log(\frac{l_i+r_i}{n})$
2. towing rule : $\frac{n_l n_r}{n^2}(\sum_{i=1}^{k}\mid\frac{l_i}{n_l}-\frac{r_i}{n_r}\mid)^2$
3. sum minority : $\sum_{i=1}^{k}l_i-\max_i l_i+\sum_{i=1}^{k}r_i-\max_i r_i$
4. max minority : $\max(\sum_{i=1}^{k}l_i-\max_i l_i,\sum_{i=1}^{k}r_i-\max_i r_i)$
5. Gini index : $\frac{n_l}{n}(1-\sum_{i=1}^{k}(\frac{l_i}{n_l})^2)+\frac{n_r}{n}(1-\sum_{i=1}^{k}(\frac{r_i}{n_r})^2)$
6. sum of variance : $\sum_{i=1}^{n_l}c_i^2-\frac{1}{n_l}(\sum_{j=1}^{n_l}c_j)^2+\sum_{i=1}^{n_r}c_i^2-\frac{1}{n_r}(\sum_{j=1}^{n_r}c_j)^2$
7. t-statistic : t-통계량을 이용하여 절대값이 가장 작은 순서
8. one-dimensional SVM : 각각의 유전자에 대해 SVM을 적용

여기서 k는 그룹의 수, n은 샘플의 수(두 집단 n_l과 n_r로 구성)이고 c_i는 샘플의 i번째 그룹, l_i는 그룹 i에 속하는 값의 수를 나타낸다.

4.2 유전자 선택 방법 결과

유전자 선택 방법을 결정하기 위해서 이번 절에서는 SVM 다중분류법 중 OVO SVM, MSVM, PWC SVM을 이용하여 그 결과를 비교하였으며, 비교하기 위한 데이터는 SRBCT 데이터와 GCM 데이터를 사용하였다. Rankgene 프로그램을 이용하여 각각의 유전자 선택 방법으로 선택된 150개의 유전자에 대해 3가지 분류 방법으로 그 결과를 비교해 보았다. 다음 표는 그 결과이며, 1-8의 숫자는 앞 절에서 설명한 유전자 선택 방법을 나타낸 것이다.

분류방법	1	2	3	4	5	6	7	8
OVO SVM	2.0	2.4	2.6	2.1	2.1	2.6	2.1	2.3
MSVM	2.2	2.4	2.4	2.4	2.1	2.1	2.0	2.3
PWC SVM	1.8	2.4	2.6	2.4	2.2	2.7	2.1	2.3

표 5 . SRBCT data set의 유전자 선택 비교

표 [5]에서 8가지 유전자 선택 방법을 살펴보면, OVO SVM 방법에서는 information gain 방법이 제일 좋게 나왔고 그 다음으로는 max minority, Gini index, t-statistic 방법들이 좋다고 나타났다. MSVM 방법에서는 t-statistic 방법이 가장 좋게 나왔고 그 다음으로 Gini index, sum of variance 방법들이 좋게 나타났다. PWC SVM 방법에서는 information gain 방법이 가장 좋게 나왔으며 그 다음으로 t-statistic, Gini index 방법 순으로 좋은 오분류율을 나타내었다. SRBCT 데이터의 경우는 information gain과 t-statistic의 방법이 가장 좋은 오분류율을 나타내는 것으로 볼 수 있다.

분류방법	1	2	3	4	5	6	7	8
OVO SVM	22.2	22.2	38.9	20.4	24.1	44.4	17.6	24.1
MSVM	22.2	24.1	37.0	24.1	24.1	37.0	22.2	22.2
PWC SVM	37.0	24.1	44.4	20.4	24.1	42.5	20.4	22.2

표 6 . GCM data set의 유전자 선택 비교

표 [6]의 결과를 살펴보면, OVO SVM 방법에서는 max minority 방법이 제일 좋게 나왔고 그 다음으로는 information gain과 towing rule 방법이 좋다고 나타났다. MSVM 방법에서는 information gain, t-statistic, one-dimensional SVM 방법이 가장 좋게 나왔다. PWC SVM 방법에서는 max minority와 t-statistic 방법이 가장 좋게 나왔으며 그 다음으로 one-dimensional SVM 방법이 좋은 오분류율을 나타내었다.

GCM 데이터의 경우는 information gain, max minority, t-statistic, one-dimensional SVM 방법이 좋은 오분류율을 나타내는 것으로 볼 수 있다. 결국 GCM 데이터의 경우에는 오분류 결과가 가장 나쁜 sum minority와 sum of variance 방법을 제외한 나머지 방법들에 많은 차이점이 있지 않다는 것을 알 수 있다.

두가지 데이터에 대해 실험을 한 결과, 각 SVM의 방법에 대해서는 유전자 선택이 큰 영향을 미친다고 볼 수는 없다. 하지만 유전자 선택의 방법에서만 본다면 8가지의 유전자 선택 방법 중 information gain과 t-statistic 방법이 가장 좋았음을 볼 수 있다.

4.3 실험 결과

이번 절에서는 앞에서 설명한 SVM을 이용한 다중분류법을 3가지 데이터를 이용하여 비교하였다. 또한 SVM이 아닌 기존의 분류법인 kNN과 BagBoosting도 함께 적용시켜 비교하였다. 모든 데이터는 4.2절에서 가장 좋았던 t-statistic 방법을 통해 150개의 유전자를 선택하여 분류기에 적용시켰다.

분류방법	오분류율(%)	표준오차(%)
OVO SVM	1.95	0.305
MSVM	2.18	0.282
PWC SVM	2.21	0.493
WW SVM	3.13	0.458
CS SVM	2.50	0.373
kNN	3.81	0.442
BagBoosting	10.21	0.379

표 7 . Leukemia data set (n=72, k=3)

표 [7]에서 보면, OVO SVM, MSVM, PWC SVM의 순으로 좋은 오분류율을 보임을 알 수 있다. 일반적으로 SVM 기반의 방법들은 통계적으로 차이가 없으나 SVM을 사용하지 않은 방법에 비하여는 좋은 결과를 보여주고 있다. 가장 나쁜 결과를 보여준 방법은 BagBoosting이며 다른 방법에 비해 굉장히 좋지 않은 오분류율을 보인다.

분류방법	오분류율(%)	표준오차(%)
OVO SVM	2.30	0.379
MSVM	2.10	0.379
PWC SVM	2.10	0.377
WW SVM	1.92	0.393
CS SVM	2.00	0.402
kNN	2.01	0.321
BagBoosting	1.97	0.450

표 8 . SRBCT data set (n=83, k=4)

표[8]에서 보면, WW SVM, BagBoosting, CS SVM의 순으로 좋은 오분류율을 보인다. SRBCT 데이터의 경우에는 SVM을 이용한 방법보다 kNN과 BagBoosting을 이용한 방법이 더 좋은 오분류율을 나타낸다. 하지만 각 분류 방법간의 통계적인 차이는 보이지 않는다.

분류방법	오분류율(%)	표준오차(%)
OVO SVM	17.64	0.414
MSVM	22.22	0.374
PWC SVM	20.43	0.571
WW SVM	24.21	0.379
CS SVM	23.51	0.391
kNN	42.70	0.301
BagBoosting	37.00	2.146

표 9 . GCM data set (n=198, k=14)

표 [9]에서 보면, OVO SVM, PWC SVM, MSVM의 순으로 좋은 오분류율을 보인다. GCM 데이터의 경우에는 SVM을 이용한 방법이 그렇지 않은 두 방법보다 훨씬 좋은 오분류율을 나타낸다. GCM 데이터는 앞의 두 데이터에 비해 굉장히 많은 그룹을 가지고 있다. 결국 그룹이 많아지면 SVM을 이용한 방법이 그렇지 않은 방법보다 좋은 결과를 보여줌을 알 수 있다.

Leukemia 데이터와 GCM 데이터의 경우에는 SVM을 이용한 방법이 그렇지 않은 방법에 비해 좋은 오분류율을 나타내지만, SRBCT 데이터의 경우 각 분류 방법 사이에 큰 차이가 보이지 않는다. 이는 SRBCT 데이터의 경우에는 SVM을 이용한 방법인지 아닌지에 대해 큰 영향을 받지 않는다는 것이다. 사실 SRBCT 데이터의 경우, 본 논문에서처럼 필요한 모수의 값을 정해놓지 않고 각 훈련 데이터, 각 SVM 방법마다의 최적의 모수값을 사용한다면, 모든 분류 방법에서 오분류율은 '0'에 가까운 값이 나오게 된다.

5. 결론

본 논문은 3개 이상의 그룹을 가진 데이터에서 SVM을 이용한 다중분류법을 적용하는 방법에 대해 알아보았으며, 유전자를 선택하는 여러 가지 방법에 대해서도 알아보았다. 또한 3가지의 시뮬레이션 데이터와 3가지 실제 데이터를 통해 각 방법을 비교해보았다.

모의실험에서는 퍼짐 정도가 모두 다른 3개의 그룹에서 2개의 집단은 작은 겹침이 있고 다른 1개의 집단은 겹침이 없는 경우, 동일한 퍼짐 정도를 갖는 2개의 집단과 그 두 집단을 포함하는 1개의 집단인 경우, 그리고 모두 동일한 퍼짐 정도를 갖고 서로 모두 겹침이 많은 경우를 고려하였다. 각 실험의 결과에서 보다시피 대부분의 방법들 간의 통계적인 차이가 크게 보이지 않았다. Case1에서는 PWC SVM과 WW SVM만이 통계적으로 차이를 보였고 Case3에서는 PWC SVM과 MSVM만이 차이를 보였다. 자료의 형태에 따라서 분류기 성능이 약간의 차이가 있었으나 제한된 모의실험에서의 결과로 인하여 이를 일반적으로 확장하기에는 무리가 있다고 생각된다.

실제 자료에 대한 결과에 따르면, 주어진 8가지 유전자 선택 방법 중에서 information gain과 t-statistic 방법이 가장 좋은 방법임을 알 수 있었다. 그러나 이 유전자 선택 방법은 각각의 분류 방법에는 크게 영향을 미치지 않음을 보였다. 하지만 컴퓨팅 시간을 줄이거나 원활한 프로그램의 이용을 위해서 유전자 선택이 필요함을 알 수 있다. 실제 자료에서의 분류결과는 MSVM, OVO SVM PWC SVM등이 비교적 좋은 결과를 보여주고 있으며 자료의 분류가 비교적 명확한 경우에는 기존의 분류방법인 kNN, Bagboosting도 좋은 결과를 보이고 있다. 그러나 GCM자료 같은 경우를 보면 NON SVM기반 방법은 아주 좋지 않은 결과를 보여주고 있다.

본 논문에서 비교 연구한 분류 방법은 크게 SVM을 이용한 방법과 이용하지 않은 방법으로 나눌 수 있다. 3가지 실제 데이터에 대한 결과를 보면, 그룹의 수가 많아지면 SVM을 이용한 방법이 그렇지 않은 방법보다 훨씬 좋은 성능을 나타낸다. 또한 각 SVM을 이용한 방법들 사이에서는 PWC SVM이 비교적 좋은 결과를 나타냈다. 그러나 각 방법들 간의 우열을 나누는 것이 현재까지의 실험에서 명확하게 드러나지는 않았으므로 앞으로 다양한 상황의 모의실험을 통하여 여러 다른 SVM 방법들 간의 차이점에 대한 정확한 분석이 앞으로 해결해야 할 과제라고 생각한다.

참고문헌

1. Alexander,S., Constantin,F.A., Lannis,T., Douglas,H. and Shawn,L. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, *Bioinformatics*, 21, 631-643.
2. Crammer,K. and Y.Singer. (2000). On the learnability and design of output codes for multiclass problems, *Proceedings of the Computational Learning Theory*, 35-46.
3. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M.,

Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999). Molecular classification of cancer : class discovery and class prediction by gene expression monitoring, *Science*, 286, 531-537.

4. Hsu, Chih-Wei and Chih-Jen Lin. (2002). A comparison of methods for multi-class support vector machines, *IEEE Transactions in Neural Networks*, 13, 415-425.
5. Khan,J., Wei,J.S., Ringner,M., Saal,L.H., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C. and Meltzer,P.S. (2001). Classification and diagnostic prediction of cancer using expression profiling and artificial neural networks, *Nat.Med.*, 7, 673-679.
6. Lee,Y. and Lee,C.-K. (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data, *Bioinformatics*, 19, 1132-1139.
7. M.Dettling. (2004). BagBoosting for tumor classification with gene expression data, *Bioinformatics*, 20, 3583-3593.
8. Platt,J. (1999). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods, *Advances in Large Margin Classifiers*, 61-74.
9. Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,C.H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J.P. (2001). Multiclass cancer diagnosis using tumor gene expression signatures, *Proc.Natl Acad. Sci.*, 98, 15149-15154.
10. Tao,L., C.Zhang and Mitsunori,O. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, *Bioinformatics*, 20, 2429-2437.
11. Vapnik,V. (1998). Statistical learning theory, *Wiley-Interscience*.
12. Weston,J. and Watkins,C. (1999). Support vector machines for multi-class pattern recognition, *Technical Report*.

[2006년 2월 접수, 2006년 4월 채택]