포섭구조기반 OVR SVM 결합을 통한 다중부류 암 분류

홍진혁O 조성배
연세대학교 컴퓨터과학과
생체인식연구센터
hjinh@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Multi-class Cancer Classification by Integrating OVR SVMs based on Subsumption Architecture

Jin-Hyuk HongO Sung-Bae Cho
Dept. of Computer Science, Yonsei University
Biometrics Engineering Research Center

요 약

지지 벡터 기계(Support Vector Machine; SVM)는 기본적으로 이진분류를 위해 고안되었지만, 최근 다양한 분류기 생성전략과 결합전략이 고안되어 다중부류 분류에도 적용되고 있다. 본 논문에서는 OVR(One-Vs-Rest) 전략으로 생성된 SVM을 NB(Naive Bayes) 분류기를 이용하여 동적으로 구성함으로써, OVR SVM을 이용한 다중부류 분류 시스템에서 자주 발생하는 동점을 효과적으로 해결하는 방법을 제안한다. 이 방법을 유전발현 데이터를 이용한 다중부류 암 분류에 적용하였는데, 고차원의 데이터로부터 NB 분류기 구축에 유용한 유전자를 선택하기 위해 Pearson 상관계수를 사용하였다. 14개의 암 유형과 16,063개의 유전발현 수준을 가지는 대표적인 다중부류 암 분류 데이터인 GCM 암 데이터에 적용하여 제안하는 방법의 유용성을 확인하였다.

1. 서론

최근 패턴인식의 중요한 문제인 다중부류 분류에 SVM을 많이 적용하고 있다. SVM은 다양한 이진분류 문제에서 우수한 성능을 보여 왔지만, 다중부류 분류를 위해서는 특별한 구성방식이 필요하다[1]. Vapnik이나 Crammer와 Singer는 SVM의 다중부류 버전을 제안하기도 하였지만[2], 매우 복잡한 연산이 필요하기 때문에 많은 연구자들은 주로 다수의 이진 SVM을 구성하고 효과적으로 결합하여 다중부류 분류를 수행하는 방법을 개발하였다[3]. OVR이나 쌍-방식(pair-wise), 완전코드(complete code) 등은 다수의 SVM을 학습하는 일반적인 전략들이며, 다수결 투표(majority vote), 승자독점(winner takes all), 오류정정코드(error correcting code) 등은 이들을 결합하는 대표적 방법이다.

Heu와 Lin은 OVR, 쌍-방식, DAGSVM(directed acyclic graph SVM) 등을 비교하였고[3], Bredensteiner와 Bennett은 선형 프로그래밍과 2차 프로그래밍을 결합하여 SVM을 다중부류 분류에 적용하였다[1]. Angulo 등은 K-SVCR(K classes-Support Vector Classification-Regression)을 제안하였고[4], Crammer와 Singer는 다중 분류기의 출력을 코딩기법을 이용하여 결합하였다[2]. 생물정보학에서도 SVM은 다중부류 암 분류에 많이 적용되었는데, Lee[5]나 Ramaswamy 등[6]은 OVR 전략을 이용하여 분류기를 생성하고 승자독점을 기반으로 분류를 수행하였고, Li 등은 SVM을 위한 다양한 전략을 비교하였다[7].

본 논문은 SVM과 NB을 효과적으로 결합한 다중부류 분류기법을 제안하며, 유전발현 데이터를 이용한 암 분류문제에 적용한다. 고차원의 데이터를 효과적으로 처리할 수 있는 SVM은 OVR 전략을 이용하여 원본 데이터로 학습을 시키고, NB는 Pearson 상관계수를 기반으로 추출된 유전자로 데이터를 구성하여 학습한다. 하나의 샘플이 입력되면 NB로부터 각 클래스의 확률이 계산되고, 높은 순서대로 대응하는 클래스의 OVR SVM을 평가하여 분류를 수행한다. 제안하는 방법을 대표적인 다중부류 암 분류 문제인 GCM 암 데이터에 적용하였으며, 기존의 방법보다 높은 분류성능을 보임을 확인하였다.

2. 배경

2.1 유전발현 데이터를 이용한 암 분류

최근에 개발된 마이크로어레이 기술은 대량의 유전발현 정보를 획득하여, 질병 등에 관련된 다양한 정보를 제공한다. 패턴인식 분야의 많은 연구자들은 획득된 유전발현 데이터를 이용한 암 분류 기술을 개발하고 있으며[5,6,7,8], 보통 유전자 선택 기술과 분류기술로 구분된다. 매우 많은 수의 유전자 정보로 구성되는 유전발현 데이터에서 특정 문제와 연관된 유전자의 수는 적기 때문에 유전자 선택이 분류 성능향상에 도움을 주며, 유전발현 데이터를 직접 분석하는 것이 매우 어렵기 때문에 신경망, 베이지안 기법, SVM, 결정트리 및 k 최근접 이웃 등의 많은 기계학습 기법이 활발히 적용되고 있다.

2.2 SVM을 이용한 다중부류 분류

SVM은 통계적 학습 이론을 기반으로 최근 패턴분류와 선형 모델에 많이 사용되고 있다. 입력된 샘플을 비선형 함수를 이용하여 고차원의 특징공간으로 맵핑하고, 학습 데이터에 대해 인식오류가 최소화되는 초평면을 찾는다[4].

$$X : x = (x_1,\ldots,x_n) \rightarrow F : \Phi(x) = (\Phi_1(x),\ldots,\Phi_n(x)) \quad (1)$$

학습 샘플의 수가 n일 때, 클래스 레이블 $c_i \in \{1,-1\}$를 갖는 샘플 x_i에 대해 SVM은 다음과 같이 마진을 계산한다.

$$f(x) = \sum_{i=1}^{n} \alpha_i c_i K(x,x_i) + b, \quad K(x,x_i) = \Phi(x)\cdot\Phi(x_i) \quad (2)$$

수식 (2)에서 상관계수 a_i는 x_i가 초평면을 구성하는 지지 벡터인 경우에는 0이 아니며, 지지 벡터일 경우 0값을 갖는다. 커널 함수 $K(x, x_i)$는 비선형 맵핑 함수의 내적에 의해 간단히 계산되면, 대표적으로 선형, 다항식, 가우시안, 시그모이드 함수 등이 있다.

기본적으로 SVM은 이진 분류기이기 때문에, 다중부류 분류를 위해서는 OVR, 쌍-방식, 완전코드 등의 분류기 생성전략이 필요하다[2]. 대표적으로 OVR 전략이 많이 사용되는데, 이는 M개의 클래스로 구성되는 데이터에 대해 M개의 SVM을 학습한다. 각 SVM은 대응하는 클래스와 나머지 클래스로 구분하는 이진분류를 수행한다. j번째 SVM의 결정함수 $f_j(x)$는 수식 (2)의 c_i를 다음과 같이 t_i로 치환하여 얻는다.

$$t_i = \begin{cases} +1 & if \quad c_i = j \\ -1 & if \quad c_i \neq j \end{cases} \quad (3)$$

OVR SVM을 모두 학습한 후에는 이들 출력을 하나로 결합하는 기법이 필요하다. 대표적으로 다수결 투표, 승자독점, 오류정정코드, 행위지식공간(behavior knowledge space), 결정템플릿(decision template) 등이 있다.

1) 다수결 투표: 분류기들로부터 가장 많이 선택된 클래스가 입력 샘플의 부류로 선택된다. Condorcet의 이론을 통해 검증되었으며, Kuncheva는 다수결 투표를 이론적으로 정리하였다. 단순한 방식으로 우수한 성능을 종종 보이지만, 동일 표를 얻은 클래스가 발생할 수 있는 단점이 있다.

2) 승자독점: 다수결 투표에서의 동점 문제를 해결하기 위해서 가장 높은 값을 출력한 분류기가 선택한 클래스로 분류한다. M 클래스 문제에서 L개의 분류기가 있을 경우, 다음과 같이 클래스 레이블을 선택한다. $\mathrm{ind}_{i,j}(x)$는 j번째 SVM에 대응하는 클래스에 대해서는 1값을, 그렇지 않은 클래스들에 대해서는 0값을 갖는다.

$$c = \arg\max_{i=1,\ldots,M} \sum_{j=1}^{L} ind_{i,j}(x) d_j(x) \quad (4)$$

3. 다중부류 암 분류를 위한 결합분류기

기존의 정적인 분류모델과는 달리, 제안하는 방법은 OVR SVM의 모호성을 해결하기 위해 그림 1에서와 같이 NB로부터 얻어진 각 클래스에 대한 확률을 고려한 분류기이다. 다중부류 분류에서 OVR SVM을 사용할 때 발생하는 동점은 분류 성능을 저하시킨다. 이를 해결하기 위해서 본 논문에서는 OVR SVM을 각 클래스의 가능성을 고려한 포섭구조로 구성하여 분류를 수행하기 때문에 복수의 OVR SVM이 비슷한 값을 출력하더라도 이를 효과적으로 처리한다. 포섭구조는 다수의 행동이 활성화될 경우, 적절한 행동을 선택하는 기술로 Brooks가 제안하였다[9].

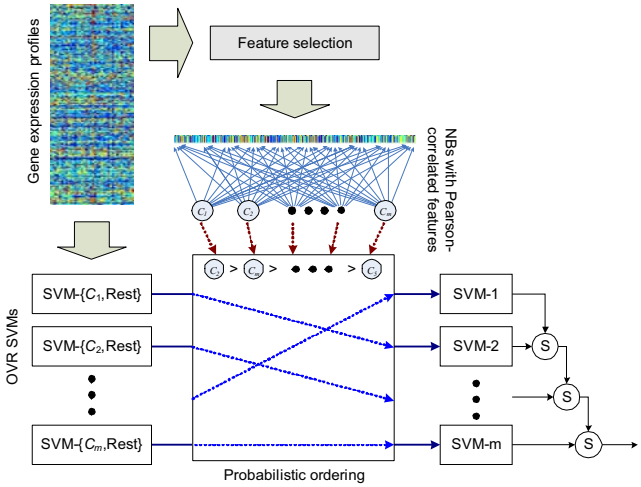

그림 1. 제안하는 방법

그림 1에서와 같이 NB는 Pearson 상관계수로 선택된 유전자를 이용하여 클래스의 사후 확률 $prob = \{p_1, p_2, \ldots, p_m\}$를 측정하고, SVM은 학습 데이터 전체를 사용하여 2.2절에서 설명한 OVR 전략으로 학습한다. OVR SVM을 이용한 분류를 통해 입력샘플의 SVM에 대한 마진 $o\text{-}svm = \{ma_1, ma_2, \ldots, ma_m\}$을 계산한다. 그림 2와 같이 클래스 사후 확률을 따라 OVR SVM을 정렬하고, 마진값을 분석하여 임계치를 넘으면 해당 클래스로 분류한다.

```
prob[m] = {p1, p2, ..., pm}          // prob[] : NB 계산
order[m] = {0, 1, 2, ..., m-1}
o-svm[m] = {ma1, ma2, ..., mam} // o-svm[] : OVR SVM 계산

// 확률기반 OVR SVM 순서 결정
for(i=0; i<m i++)
  for(j=i+1; j<m j++)
    if(prob[i] < prob[j])
    {
      int iTemp = prob[i]; prob[i] = prob[j];  prob[j] = iTemp;
      iTemp = order[i];    order[i] = order[j]; order[j] = iTemp;
    }

// 포섭구조기반 OVR SVM 분류
if(prob[order[0]] < r1)          // r1 : 거부 임계치
  return reject;
for(i=0; i<m i++)
  {
  if(o-svm[order[i]] >= a)       // a : 임계치
  {
    if(o-svm[order[i]] < r2)     // r2 : 거부 임계치
      return reject;
    return order[i];
}}
return order[0];
```

그림 2. 분류 과정

DNA 마이크로어레이 데이터는 수천 내지 수만의 유전자로 구성되기 때문에, 모든 정보를 이용하여 NB를 설계하는 것은 매우 비효율적이다. 본 논문에서는 [8]에서와 같은 방식으로 해당 암 분류에 관련이 있는 유전자를 선택하여 데이터를 축약한 후 NB를 학습한다.

4. 실험 및 결과

4.1 평가 데이터

제안하는 방법을 평가하기 위해서 대표적 다중부류 유전발현 데이터인 GCM 암 데이터를 사용하였다. GCM 암 데이터는 144개의 학습 샘플과 54개의 테스트 샘플로 구성되며, 16,063개의 유전발현 수준을 가진다[6]. 데이터는 총 14가지 암(breast adenocarcinoma, prostate, lung adenocarcinoma, colorectal adenocarcinoma, lymphoma, bladder, melanoma, uterine adenocarcinoma, leukemia, renal cell carcinoma, pancreatic adenocarcinoma, ovarian adenocarcinoma, pleural mesothelioma, central nervous system)으로 구분되어 있다. 샘플은 적은데 반해 특징 차원은 매우 커서 높은 성능을 얻기 어려우며, 이를 위해 많은 기계학습 연구자들이 활발히 연구하고 있다. Ramaswamy 등은 승자독점 방식의 SVM을 이용하여 78%의 분류 정확율을 획득하였고[6], Li 등은 다양한 전략을 이용한 SVM 모델로 63.3%의 분류 정확율을 얻었다[7]. 본 논문에서는 NB의 학습을 위해 각 클래스를 가장 잘 구분하는 특징을 Pearson 상관계수를 기반으로 총 140개 선택하였고, SVM의 기본 커널함수로 선형함수를 사용하였다. 데이터는 모두 0에서 1사이로 정규화하여 실험을 수행하였다.

4.2 결과분석

표 1은 제안하는 방법의 기존 기법들에 대한 성능 비교를 보여준다. 승자독점 기반의 OVR SVM의 경우 75.9%의 분류 정확률을 얻었고, Pearson 상관계수를 이용한 NB는 SVM보다 약간 낮은 72.2%의 분류 정확율을 획득하였다. 기계학습의 대표적 분류기법인 신경망은 지역해에 빠지는 경우가 많았으며, 특징과 클래스 수는 많은데 반해 학습 샘플이 너무 적어 모델 학습이 어려웠으며 분류 정확율도 64.8%로 매우 낮았다. SVM과 NB의 출력을 내적한 경우에는 66.7%로 어느 한쪽에 잡음이 있을 때 성능을 저하시켜 단일 분류기의 성능보다 떨어졌다. 제안하는 방법은 79.6%의 분류 정확율로 두 분류기의 성능을 극대화하여 기존 방법보다 높은 성능을 획득하였다.

표 1. 분류 정확율 비교

기법 (특징 수)	MLP (140)	SVM (16,063)	NB (140)	곱 결합 (SVM+NB)	제안하는 방법
분류율	64.8%	75.9%	72.2%	66.7%	79.6%

표 2는 SVM과 NB에 의해 분류된 샘플을 분석한 것으로, 두 분류기가 모두 분류에 실패한 샘플은 2개였다. 제안하는 방법은 SVM만 가지고 분류했을 때 실패한 7개의 샘플에서 5개를 정확히 분류하였고, NB만 가지고 분류했을 때 실패한 7개의 샘플 중 3개를 제대로 분류하였다. 표 3은 제안하는 방법의 테스트 데이터에 대한 혼동행렬이다. 결과에서 보이듯이 lung, colorectal, uterus, mesothelioma와 CNS는 모두 정확히 분류되었고, prostate, lymphoma와 leukemia는 83%의 분류 정확률이 확인되었다.

표 2. 분류결과 분석

(정확/부정확) : 제안하는 기법		SVM	
		O	X
NB	O	38 (38/0)	7 (5/2)
	X	7 (3/4)	2 (0/2)

표 3. 테스트 데이터에 대한 혼동행렬

%	0	1	2	3	4	5	6	7	8	9	10	11	12	13	n
0	75										25				4
1		83										17			6
2			100												4
3				100											4
4			17		83										6
5			33			67									3
6							50				50				2
7								100							2
8		17							83						6
9								33		67					3
10				33		33					34				3
11				25			25					50			4
12													100		3
13														100	4
n	3	6	6	6	5	3	2	3	5	2	3	3	3	4	54

5. 결론

다중부류 분류는 패턴인식에서 매우 도전적인 과제로 이진분류기인 SVM을 이용한 다양한 기법이 연구되고 있다. 특히 이진분류기를 이용한 다중부류 분류를 위해서 복수의 분류기를 생성하고 이들의 출력을 결합하는 다양한 기술이 개발되고 있다. 본 논문에서는 OVR 전략을 바탕으로 SVM과 NB를 학습하고 이들을 효과적으로 결합하여 다중부류 분류를 수행하는 분류기술을 제안하였다. 제안하는 방법을 생물정보학의 대표적인 다중부류 암 분류 데이터인 GCM 암 데이터에 적용하여 기존 방법보다 높은 분류 정확율을 획득하였다. 향후에는 보다 다양한 다중부류 데이터에 적용할 것이다.

감사의 글

본 연구는 생체인식연구센터(BERC)를 통해 한국과학재단(KOSEF)에서 지원받았음.

참고문헌

[1] E. Bredensteiner and K. Bennett, "Multicategory classification by support vector machines," *Computational Optimization and Applications*, vol. 12, no. 1, pp. 53-79, 1999.

[2] K. Crammer and Y. Singer, "On the learnability and design of output codes for multiclass problems," *Machine Learning*, vol. 47, no. 2-3, pp. 201-233, 2002.

[3] C. Hsu and C. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415, 425, 2002.

[4] C. Angulo, et al., "K-SVCR, A support vector machine for multi-class classification," *Neurocomputing*, vol. 55, no. 1-2, pp. 57-77, 2003.

[5] Y. Lee and C. Lee, "Classification of multiple cancer types by multicategory support vector machines using gene expression data," *Bioinformatics*, vol. 19, no. 9, pp. 1132-1139, 2003.

[6] S. Ramaswamy, et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. National Academy of Science*, vol. 98, no. 26, pp. 15149-15154, 2001.

[7] T. Li, et al., "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429-2437, 2004.

[8] S.-B. Cho and J. Ryu, "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features," *Proc. of the IEEE*, vol. 90, no. 11, pp. 1744-1753, 2002.

[9] R. Brooks, "A robust layered control system for a mobile robot," *IEEE J. of Robotics and Automation*, vol. 2, no. 1, pp. 14-23, 1986.