대용량 분류에서 SVM과 신경망의 성능 비교

이 진 선[†] · 김 영 원[††] · 오 일 석[†††]

요 약

이 논문은 대용량 분류 문제를 위한 모듈러 신경망(modular feedforward MLP)과 SVM(Support Vector Machine)의 성능을 비교 분석하였다. 전반적으로 SVM이 상당한 성능 차이로 우수함을 확인하였다. 또한 부류 수가 많아짐에 따라 SVM이 신경망보다 완만하게 성능 저하가 있음도 확인하였다. 또한 기각에 따른 정인식률 추이를 분석하였고, 대용량 분류에 적합한 SVM 파라메터(kernel 함수와 관련 변수들)를 도출하였다.

Performance comparison of SVM and neural networks for large-set classification problems

Jin-Seon Lee[†] · Young-Won Kim[††] · Il-Seok Oh[†††]

ABSTRACT

In this paper, we analyzed and compared the performances of modular FFMLP(feedforward multilayer perceptron) and SVM(Support Vector Machine) for the large-set classification problems. Overall, SVM dominated modular FFMLP in the correct recognition rate and other aspects. Additionally, the recognition rate of SVM degraded more slowly than neural network as the number of classes increases. The trend of the recognition rates depending on the rejection rate has been analyzed. The parameter set of SVM(kernel functions and related variables) has been identified for the large-set classification problems.

키워드 : 대용량 분류(Large-set Classification), SVM, 모듈러 신경망(Modular Feedforward Neural Network), 성능 비교(Performance Comparison)

1. 서 론

웹의 발전, 멀티미디어 자료 처리에 대한 요구 증가 등과 같은 새로운 응용의 창출로 인해 패턴 인식 기술에 대한 요구가 갈수록 커지고 있다[1]. 현재 이러한 새로운 응용 분야를 중심으로 대용량 분류의 성능 향상에 대한 요구가 높다. 이러한 요구는 한글과 한자와 같이 부류 수가 많은 언어, 데이터 마이닝 분야와 같이 여러 응용에서 발생하고 있다.

이러한 대용량 분류 문제는 하나의 덩어리 구조를 갖는 분류기로는 만족스러운 성능을 얻을 수 없다. 이러한 문제에 대해서는 다단계 예비 분류(preclassification)[2], 자기 조직화 신경망 구조[3], 또는 모듈러 신경망 구조[4] 등이 성능 향상에 기여함이 이미 밝혀졌다. 하지만 이러한 분류기 구조만으로는 혼동이 발생하는(confusing) 부류들을 올바르게 분류할 수 없는 상황이 자주 발생하여 획기적인 성능 향상을 얻을 수 없다. 각각의 부류들을 독립적으로 고려하여 분류기를 설계하여 인식률을 향상시킬 수 있을 것이다.

이 논문에서는 부류마다 고유한 인식기를 갖는 OCOC (One-Class-One-Classifier) 분류 알고리즘을 채택하였다. 이 알고리즘에서는 부류마다 독립적인 이진 분류기를 가지며, 이 이진 분류기로 해당 부류와 그 이외의 부류를 분류하는 역할을 한다. 모듈러 신경망과 SVM(Support Vector Machine)으로 구현하고 성능을 비교하였다. SVM이 352 부류 문제에서 기존 모듈러 신경망의 성능을 약 10% 정도 향상시킴을 실험으로 확인하였다. 실험을 통해 모듈러 신경망과 SVM에 대해 부류 수에 따른 정인식률 추이와 기각률에 따른 정인식률 추이를 분석하였다. 또한 대용량 분류 문제에 적합한 SVM 파라메터를 실험적으로 찾아내어 제시한다.

2. OCOC 구조의 분류기

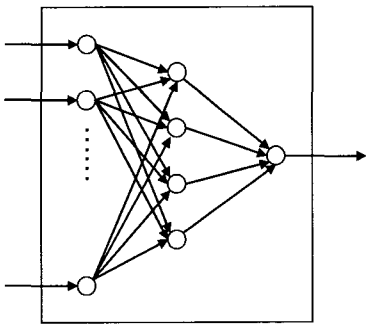
2.1 모듈러 신경망 분류 알고리즘

모듈러 신경망 구조는 K개의(K는 부류 수) 부 망

※ 이 연구는 한국과학재단 목적기초연구(R05-2002-000-00754-0)지원으로 수행되었음.
† 정 회 원 : 우석대학교 컴퓨터공학과 교수
†† 준 회 원 : 전북대학교 대학원 컴퓨터정보학과
††† 정 회 원 : 전북대학교 전자정보공학부 교수
논문접수 : 2004년 7월 30일, 심사완료 : 2005년 2월 7일

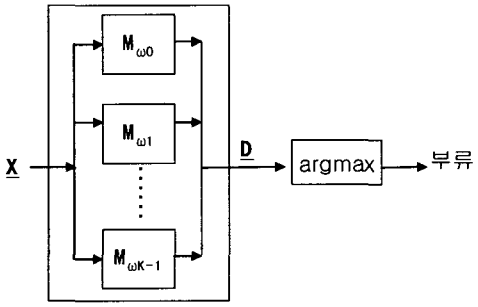
(subnetwork)으로 구성되고, 각 부 망은 K개의 부류 중 하나를 책임진다[4]. 부류 ω_i를 위한 부 망은 (그림 1)과 같으며 하나의 은닉 층, 하나의 출력 층을 갖는다. 이들 세층은 완전 연결되어 있다. 이 부 망의 기능은 두 개 부류군, $\Omega_1=\{\omega_i\}$과 $\Omega_2=\{\omega_k|1\le k\le K,\ k\neq i\}$를 분류하는 것이다. 입력층은 n-차원 특징 벡터를 받기 위해 n개의 노드를 갖는다. 출력층은 한 개의 노드를 갖는다.

K개의 부 망의 훈련은 각각 독립적으로 이루어지며, 오류 역전파 알고리즘으로 수행한다. 활성 함수(activation function)로는 이진 시그모이드(binary sigmoid)를 사용하였다. 부류 ω_i를 훈련하기 위해 훈련 집합을 두개의 군 $Z_{positive}$와 $Z_{negative}$로 나눈다. $Z_{positive}$는 Ω_1에 속한 샘플, $Z_{negative}$는 Ω_2에 속한 샘플을 갖는다. $Z_{positive}$에 속하는 샘플의 목표 출력(target output)은 +1.0, $Z_{negative}$에 속하는 샘플의 목표 출력은 0.0이다. K개 부류가 동일한 개수의 훈련 샘플을 갖는다고 가정했을 때 $Z_{negative}$는 $Z_{positive}$의 K-1배의 샘플을 갖는다. 각 부망 입장에서 살펴보면 1:K-1의 불균형이 있음에도 비 모듈러 신경망에 비해 우수한 인식 성능을 보인다.



(그림 1) 모듈러 MLP를 위한 부 망

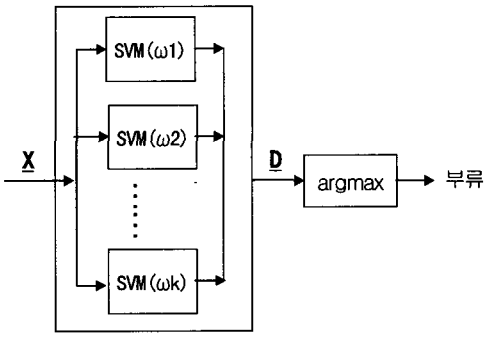
부류 ω_i의 부 망을 $M_{\omega i}$라 하면, 전체 인식기의 구조는 (그림 2)와 같다. 입력 패턴으로부터 특징 벡터를 추출하고, 이를 모든 부 망에 입력한다. 최종적으로 입력 패턴을 가장 큰 출력 값을 갖는 부류로 분류한다. X는 특징 벡터이고 D는 결정 벡터(decision vector) 이다. argmax는 결정 벡터를 보고 최고값을 갖는 부류 인덱스를 찾는 역할을 한다.



(그림 2) 모듈러 신경망 인식기의 전체 구조

2.2 SVM 분류 알고리즘

SVM을 이용하여 OCOC 구조의 분류기를 (그림 3)과 같이 구성하였다[5]. X는 특징 벡터로서 K개의 SVM에 공통으로 입력된다. K개의 SVM은 각각 출력 값을 계산하며, argmax는 최대값을 갖는 부류 인덱스를 찾아 출력한다. SVM(ωi)의 훈련은 독립적으로 이루어진다. SVM(ωi)를 훈련하기 위해 훈련 집합을 두개의 군 $Z_{positive}$와 $Z_{negative}$로 나눈다. $Z_{positive}$는 ωi에 속한 샘플, $Z_{negative}$는 ωi를 제외한 모든 부류에 속한 샘플을 갖는다.



(그림 3) SVM을 이용한 대용량 분류기 구조

SVM 소프트웨어로는 웹에 공개되어 있는 SVM-light를 사용하였다[6]. Linux version을 다운받아서 사용하였다. 다음과 같은 옵션을 조절하여 성능을 측정하였다.

커널 옵션:

0 - 선형 (linear)
1 - 다항식 (polynomial (s a*b+c)^d)
d는 다항식 파라메터로서 정수형임
2 - radial basis function exp(-gamma || a-b ||^2)
g는 RBF의 gamma 파라메터로서 실수형임

학습 옵션:

c는 훈련 오류와 마진 사이의 trade-off를 조절하는 파라메터로서 실수형임
b [0,1] - (디폴트 1)
0 - 바이어스된 초평면
1 - 바이어스 안된 초평면
i [0,1] - (디폴트 0)
0 - 일관성없는 훈련 샘플 유지
1 - 일관성없는 훈련 샘플 제거

3. 실험 및 분석

3.1 실험 환경

OCOC 구조의 분류기를 모듈러 MLP와 SVM으로 구현하였다. 그리고 모듈러 MLP와 SVM을 여러 측면에서 비교 분석하였다. 분류기의 인식 성능은 아래 기준으로 측정하였다.

정인식률 (correct recognition rate) = C/(C+I) (1)

기각률 (rejection rate) = R/N (2)

N은 샘플 총수로서 C+I+R과 같고 C는 맞게 인식한 샘플 수, I는 틀리게 인식한 샘플 수, 그리고 R은 기각된 샘플 수이다.

실험용 특징 벡터 파일은 우리나라 우편 주소에 나타나는 352글자를 대상으로 하였다. 이 파일은 PE92 데이터베이스에서 추출하였으며, 샘플당 252-차원의 특징으로 구성되어 있다. 부류당 70개의 훈련 샘플(training sample)과 30개의 검사 샘플(test sample)로 구성된다. 따라서 24,640 샘플의 훈련 집합과 10,560 샘플의 검사 집합으로 구성된다. 이 파일은 전남대학교에서 구축하였다.

3.2 모듈러 신경망

(1) 부류 수에 따른 인식률

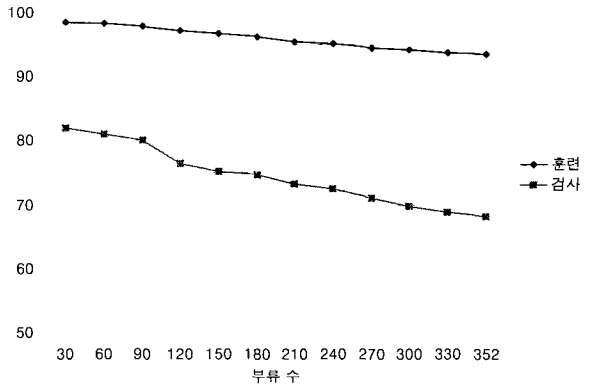
부류 수에 따른 인식률 추이를 알아보기 위해, 부류 수를 30개씩 증가시키며 실험하였다. 기각을 허용하지 않은 실험 결과이다. <표 1>에서 실험 결과를 보여주며 (그림 4)는 이를 그래프로 도시한 것이다.

〈표 1〉 신경망의 부류 수에 따른 인식률 추이
(기각 없음, 단위: %)

부류 수	훈련 집합에 대한 정인식률	검사 집합에 대한 정인식률
30	98.38	81.89
60	98.33	81.00
90	97.75	79.99
120	97.08	76.47
150	96.62	75.15
180	96.27	74.70
210	95.49	73.27
240	95.15	72.48
270	94.44	71.00
300	94.08	69.70
330	93.75	68.88
352	93.51	68.15

전체 352부류에 대해 훈련 집합과 검사 집합에 대해 각각 93.51%와 68.15%의 정인식률을 얻었다. 부류 수를 반으로 하여 180개일 때는 96.27%와 74.70%를 얻었다. 검사 집합에 대한 정인식률이 약6.55%가 상승하였다. 30부류인 경우 93.38%와 81.89%로서 문제 난이도가 낮아짐에 따라 정인식률이 상당히 상승하였다.

모듈러 신경망의 대용량 분류 문제에서의 일반화 (generalization) 능력 추이를 분석한다. (그림 4)의 그래프를 보면 부류 수가 많아질수록 훈련과 검사 집합의 곡선이 멀어짐을 볼 수 있다. 이는 부류 수가 많아짐에 따라 일반화 능력이 떨어짐을 의미한다. 훈련과 검사 집합의 정인식률 차이가 30부류의 경우 16.5%, 180부류에서는 21.6%, 그리고 352부류에서는 25.36%이다.



(그림 4) 신경망의 부류 수에 따른 인식률 추이(기각 없음)

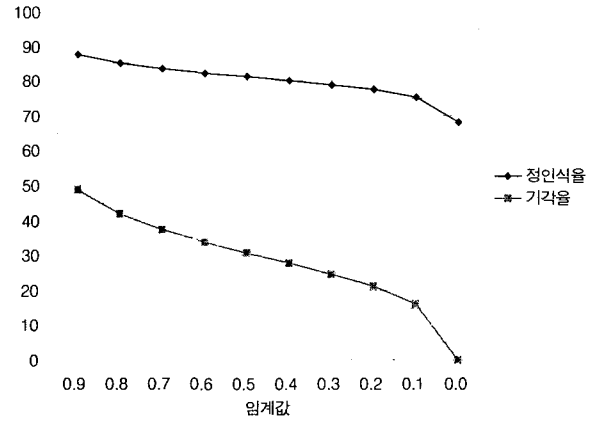
부류 수가 많아짐에 따른 전체적인 정인식률 변동 추세를 보면, 급격히 낮아지지 않고 완만한 선형을 그리며 낮아짐을 알 수 있다. 이러한 사실은 모듈러 신경망이 대용량 분류 문제에 적합함을 의미한다할 수 있다.

(2) 기각률에 따른 인식률

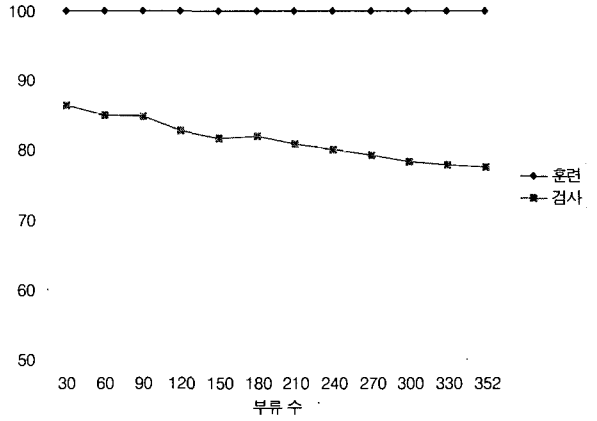
기각은 다음과 같은 방법으로 허용한다. K개 출력 중에서 가장 큰 값을 갖는 부류를 선택한 후, 이 부류의 출력값을 임계 값과 비교한다. 출력 값이 임계 값보다 크면 그 부류로 분류하고 그렇지 않으면 기각으로 결정한다. 출력값 범위가 0.0~1.0이므로 이 범위를 10개로 등분하여 이들을 임계 값으로 사용하였다. <표 2>는 실험 결과를 보여주며 (그림 5)는 이를 그래프로 도시한 것이다.

〈표 2〉 기각률에 따른 정인식률 추이(352 부류, 단위: %)

임계값	정인식율	기각율
0.9	87.56	48.86
0.8	85.20	41.70
0.7	83.56	37.26
0.6	82.43	33.90
0.5	81.37	30.72
0.4	80.13	27.69
0.3	78.95	24.75
0.2	77.43	21.20
0.1	75.43	16.11
0.0	68.15	0.00



(그림 5) 기각률에 따른 정인식률 추이



(그림 6) SVM의 부류 수에 따른 인식률 추이(기각 없음)

임계 값을 0.1로 한 경우, 기각률이 16.1%이다. 이를 달리 말하면, 83.9%의 샘플은 출력 최대값이 0.1을 넘은 상태로 인식됨을 의미한다. 기각률 16.1% 희생에 7.3%의 정인식률 향상을 얻었다. 임계 값을 0.5로 하는 경우 기각률 30.7%에 정인식률 81.37%로서, 13.2%의 정인식률 향상을 얻었다. 임계 값을 0.9로 하여 인식한 결과 기각률이 48.9%로서 약 반절의 샘플이 기각되었다. 대신 19.4%의 정인식률 향상을 얻을 수 있었다.

3.3 SVM

(1) 부류 수에 따른 인식률

부류 수를 30개씩 증가시키며 부류 수에 따른 인식률 추이를 알아보았다. 기각을 허용하지 않았으며, <표 3>에서 실험 결과를 보여주며 (그림 6)은 이를 그래프로 도시한 것이다. 아래에 있는 파라메터 집합을 사용했으며, 이 값들은 실험에 의해 좋은 인식률을 보이도록 설정한 것이며, 파라메터에 따른 인식률 변화는 항목 (3)에서 제시할 것이다.

<표 3> SVM의 부류 수에 따른 인식률 추이(기각 없음, 단위: %)

(커널 옵션=다항식, d=170, c=6, b=0, i=1)

부류 수	훈련집합에 대한 정인식률	검사집합에 대한 정인식률
30	100	86.44
60	100	84.94
90	100	84.88
120	100	82.80
150	100	81.66
180	100	81.98
210	100	80.79
240	100	80.07
270	100	79.31
300	100	78.36
330	100	77.91
352	100	77.48

훈련 집합에 대해서는 모든 경우 100% 정인식률을 얻었다. 전체 352부류에 대해 검사 집합에 대해 77.48%의 정인식률을 얻었다. 부류 수를 반으로 하여 180개일 때는 81.98%를 얻어 약4.5% 상승하였다. 30부류인 경우 86.44%로서 약 9% 상승하였다.

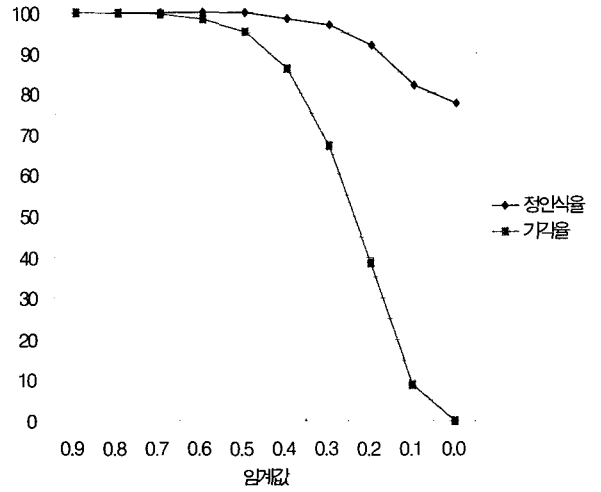
부류 수가 많아짐에 따른 전체적인 정인식률 변동 추세를 보면, 급격히 낮아지지 않고 완만한 선형을 그리며 낮아짐을 알 수 있다. 이러한 사실은 SVM이 대용량 분류 문제에 매우 적합함을 의미한다.

(2) 기각률에 따른 인식률

기각은 다음과 같은 방법으로 허용한다. K개 출력 중에서 가장 큰 값을 갖는 부류를 선택한 후, 이 부류의 출력 값을 임계 값과 비교한다. 출력 값이 임계 값보다 크면 그 부류로 분류하고 그렇지 않으면 기각으로 결정한다. 출력 값 범위를 사전에 파악한 후, 이 범위를 10개로 등분하여 이들을 임계 값으로 사용하였다. <표 4>는 실험 결과를 보여주며 (그림 7)은 이를 그래프로 도시한 것이다.

<표 4> SVM의 기각률에 따른 정인식률 추이 (단위: %)

임계값	정인식율	기각율
0.9	100.00	99.97
0.8	100.00	99.92
0.7	100.00	99.64
0.6	100.00	98.52
0.5	100.00	95.24
0.4	98.36	86.11
0.3	96.74	67.19
0.2	91.67	38.49
0.1	81.92	8.98
0.0	77.48	0.00



(그림 7) SVM의 기각률에 따른 정인식률 추이

임계 값을 0.1로 한 경우, 기각률이 8.98%이다. 기각률 8.98% 희생에 4.44%의 정인식률 향상을 얻었다. 임계 값이 0.2인 경우 38.49% 기각률에 14.19%의 정인식률 상승을 얻었다. 임계 값이 증가함에 따라 기각률이 매우 빠르게 상승하였다.

(3) SVM 파라메터 설정에 따른 인식률 변화

SVM 파라메터를 변화시키며 정인식률 변화 추이를 관찰하였다. <표 5>는 다항식 커널을 사용할 때 파라메터 d에 따른 정인식률 변화를 보여준다.

〈표 5〉 다항식 커널에서 d 변화에 따른 정인식률

(c=6, 나머지 변수는 디폴트 값, 즉 b=1, i=0)

d	정인식률	d	정인식률	d	정인식률
2	67.24	15	69.38	130	75.46
3	67.71	20	70.21	150	75.65
4	67.87	25	70.70	160	75.74
5	67.97	30	71.20	165	75.80
6	68.25	35	71.73	167	75.81
7	68.35	45	72.50	170	75.84
8	68.42	70	73.91	180	75.83
9	68.65	90	74.69	190	75.78
10	68.70	110	75.14		

d가 증가함에 따라 완만하게 정인식률이 향상되었다. d=170 근방에서 최고값을 보였으며, 더 커짐에 따라 정인식률이 떨어지는 현상을 보였다. 따라서 d=170으로 정하고, b와 i 파라메터를 변화시키며 정인식률 변화를 관찰하였다. <표 6>은 그 결과를 보여준다. b=0, i=0과 b=0, i=1에서 가장 좋은 정인식률을 보였다.

〈표 6〉 다항식 커널에서 b와 i의 변화에 따른 정인식률

(c=6, d=170)

b	i	정인식률
0	0	77.48
1	0	75.84
0	1	77.48
1	1	77.09

<표 7>은 RBF 커널을 사용하는 경우의 정인식률을 보여준다. g 값에 따른 추이를 측정하였는데, g가 커짐에 따라 정인식률이 상승하였으며 g=6.0 근방에서 최고값을 보였다. RBF 커널은 다항식 커널에 비해 약 8% 정도 차이로 열등하였다.

〈표 7〉 RBF 커널에서 g의 변화에 따른 정인식률

(c=6, 나머지 변수는 디폴트 값)

g	정인식률
1.0	67.51
2.0	67.89
3.0	68.41
4.0	68.71
5.0	69.08
6.0	69.32
7.0	69.32

<표 8>은 RBF 커널에서 g=6.0으로 고정하고 c 값을 변화시키며 정인식률 추이를 측정한 것인데, c에 따라 큰 차이를 보이지 않음을 알 수 있다.

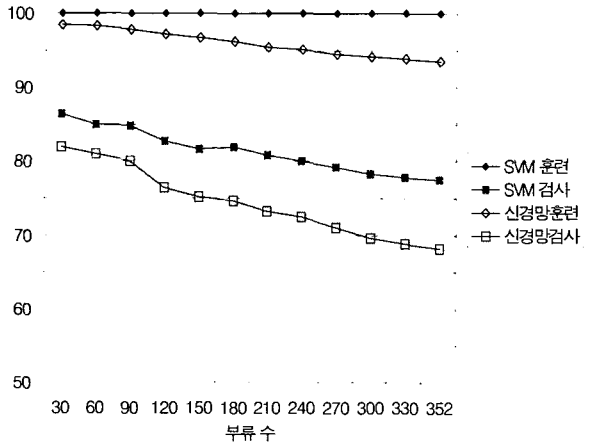
〈표 8〉 RBF 커널에서 c의 변화에 따른 정인식률(g=6.0)

c	g	정인식률
1	6.0	68.95
2	6.0	69.05
3	6.0	69.06
4	6.0	69.08
5	6.0	69.08
6	6.0	69.08

3.4 결과 및 고찰

(그림 8)은 SVM과 모듈러 신경망의 인식 성능을 한 눈에 비교하기 위해 (그림 4)와 (그림 6)을 하나로 그린 것이다. 검사 집합에 대한 비교를 해 보면, 모든 부류 수에서 SVM이 월등히 우수함을 알 수 있다. SVM은 30부류에서 4.55%, 180부류에서 7.28%, 그리고 352부류에서 9.33%의

큰 차이로 우수성을 보여주고 있다. 부류 수가 많아짐에 따라 SVM이 보다 완만하게 정인식률이 떨어진다는 사실은 매우 특기할 만하다. 이는 SVM이 대용량 분류 문제에서 신경망보다 우수한 성질을 가지고 있음을 의미한다.



(그림 8) SVM과 모듈러 신경망의 인식률 비교(기각 없음)

이러한 분석을 통하여 다음과 같은 두 가지 결론을 도출하였다. 첫째, 대용량 분류 문제에서 SVM은 신경망 분류기보다 우수하며, 특히 부류 수가 많아짐에 따라 성능 격차는 커진다. 둘째, 대용량 분류에서 SVM은 커널 함수로 매우 고차원의 다항식을 사용해야 한다.

4. 결 론

OCOC 구조의 분류기로 대용량 분류 문제를 해결하고자 노력하였다. 모듈러 MLP와 SVM을 여러 측면에서 인식 성능을 비교 분석하였다. 결과적으로 SVM이 대용량 분류 문제에서 모듈러 신경망보다 훨씬 우수함을 확인하였고 특히 부류 수가 많아짐에 따라 보다 완만한 속도로 인식률이 떨어짐을 확인하였다.

향후 연구로서, 1부류와 나머지 K-1개 부류를 분류하는 이진 분류기 K를 사용하는 OCOC 구조 대신 k부류와 K-k 부류를 분류하는 이진 부류기를 사용하는 방법을 고안한다. k=2인 경우 K/2개의 이진 분류기와 두 개 부류만을 구분하는 K/2개의 이진 분류기가 필요하다.

참 고 문 헌

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd Ed., Wiley-Interscience, 2001.

[2] Hahn-Ming Lee, Chin-Chou Lin, and Jyh-Ming Chen, "A preclassification method for handwritten Chinese character recognition via fuzzy rules and SEART neural net," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol.12, No.6, pp.743-761, 1998.

[3] Hee-Heon Song and Seong-Whan Lee, "A self-organizing neural tree for large-set pattern classification," *IEEE Transactions on Neural Networks*, Vol.9, No.3, pp.369-380, 1998.

[4] Il-Seok Oh and Ching Y. Suen, "A class-modular feedforward neural network for handwriting recognition," *Pattern Recognition*, Vol.35, No.1, pp.229-244, 2002.

[5] C. J. C. Burges, A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, Vol.2, pp.121-167, 1998.

[6] http://svmlight.joachims.org

이 진 선

e-mail : jslee@core.woosuk.ac.kr

1985년 전북대학교 전산통계학과(학사)

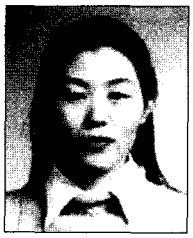
1988년 전북대학교 대학원 전산통계학과 (이학석사)

1995년 전북대학교 대학원 컴퓨터공학과 (공학박사)

1988년~1992년 한국전자통신연구원 연구원

1995년~현재 우석대학교 컴퓨터공학과 교수

관심분야 : 패턴인식, 영상처리, 멀티미디어



김 영 원

email : ywkim@dahong.chonbuk.ac.kr

2001년 전북대학교 컴퓨터과학과(학사)

2003년 전북대학교 대학원 컴퓨터정보학과 (이학석사)

2003년~현재 전북대학교 대학원 컴퓨터정보학과 박사과정

관심분야 : 워터마킹, 문자인식, 컴퓨터비젼

오 일 석

email : isoh@moak.chonbuk.ac.kr

1984년 서울대학교 컴퓨터공학과(학사)

1992년 KAIST 전산학과(공학석사, 박사)

1992년~현재 전북대학교 전자정보공학부 교수

관심분야 : 문서영상처리, 패턴인식, 유전알고리즘의 패턴인식 응용