응용통계연구 제 13권 2호, 2000년, pp. 415-425

향상된 PAIRWISE COUPLING 알고리즘에 의한 자료의 분류

최대우 [1] 윤중식 [2]

요 약

붓스트랩 표본추출과 pairwise coupling의 알고리즘을 결합한 새로운 분류 알고리즘을 제안하고, 이를 선형판별분석과 2차 판별분석에 적용하였다. 그리고 새로운 분류 알고리즘의 정확도를 비교하기위해 널리 사용되는 waveform 자료 등을 분석한 후, 그 결과를 기존 분류 방법과 비교하였다.

주요용어: 분류, 앙상블 기법, Pairwise coupling.

1. 서 론

결과에 대한 측정치를 y라하고, 그 원인에 해당되는 측정치들을 $\boldsymbol{t} = (t_1, t_2, \ldots, t_p)$ 라고 하자. 우리는 흔히 y를 종속변수, 혹은 반응변수라고 하고, $\boldsymbol{t}$는 입력변수, 설명변수 혹은 독립변수라 부른다. 그리고 y가 가격(price)이나 혈압(blood pressure)과 같이 연속형인 경우 회귀모형 문제(regression problem)로, y가 부도여부, 생사의 예측 등과 같이 순서 없는 유한한 상태일 경우 분류문제(classification problem) 라 칭한다.

분류예측을 위한 모형 C는 과거 자료 $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$를 이용하여 적합하는데 흔히 $\boldsymbol{x}$를 훈련자료(training data)라 하고, 각 $\boldsymbol{x}_i$는 설명변수 $\boldsymbol{t}^i$와 분류수준 y^i로 이루어져 있다. 즉, $\boldsymbol{x}_i = (\boldsymbol{t}^i, y^i)$이다. 훈련자료 $\boldsymbol{x}$에 의해 적합된 분류예측 모형은 $\boldsymbol{t}$라는 새로운 설명변수가 입력되면 분류결과 $C(\boldsymbol{x}, \boldsymbol{t})$를 출력하는 것이다.

분류와 관련된 모수적 통계분석 방법으로는 선형 판별 분석(linear discriminant analysis)과 로지스틱 회귀분석(logistic regression)등을 들 수 있다. 그리고 비모수적 방법으로는 나무 모형(tree model)과 신경망, projection pursuit regression등 비모수 회귀분석 (nonparametric regression)에 근거한 다양한 분류방법이 있다. 최근 들어서는 객체 지향적 프로그램밍(objected-oriented programming) 기법의 발달로 다수의 분류모형을 생성하고 그 결과를 결합하여 최종 분류를 결정하는 학습기 결합(combining learners)에 대한 연구가 활발히 진행되고 있다. 앙상블 기법(ensemble technique)이라고도 불리어지는 이 결합방법으로는 Breiman(1996)이 제시한 Bagging(Bootstrap Aggregation)을 시발점으로 Freund와 Shapire(1995)의 Boosting, Breiman(1996)의 Arcing(Adaptive Resampling and Combining)

1) (449-791) 경기도 용인시 모현면, 한국외국어대학교 정보통계학과, 조교수
E-mail: dachoi@popsmail.com
2) (449-791) 경기도 용인시 모현면, 한국외국어대학교 정보통계학과 대학원
E-mail: jsyun@stat.hufs.ac.kr

알고리즘 등이 개발되었다. 이 알고리즘들의 특징은 모형 혹은 학습기(learner) 도출에 사용되는 원 훈련자료(training data)에서 붓스트랩 표본추출(bootstrap sampling)에 의해 새로운 훈련자료를 생성한 후 분류 예측 모형을 구하는 과정을 반복하고, 반복을 통해 구해진 여러 예측모형에서 얻어진 분류결과를 적절히 결합하여 최종 분류를 결정한다는 점이다.

Friedman(1996)은 훈련자료 상의 여러 수준 중 두 수준만의 판별작업을 실시한 후 분류결과를 조합하는 소위 pairwise coupling이라고 부르는 간단한 학습기 결합방법을 제시하였다. 이 알고리즘은 분류의 수준이 3개 이상인 다수준 분류(polychotomous classification)문제에 있어 기존 방법보다 향상된 결과를 제공하는 것으로 알려져 있다. 그 후 Hastie와 Tibshirani(1996)는 Friedman이 제시한 방법의 한계를 지적하고 Bradley-Terry 계산법을 이용한 알고리즘을 소개하였다.

본 논문에서는 붓스트랩 표본추출에 기초한 학습기 결합방법과 pairwise coupling비교에 기초한 알고리즘을 결합하여 새로운 다수준 분류용 알고리즘을 제안하였다.

제 2절에서는 pairwise coupling 분류방법에 대하여 소개하였고, 3절에서는 새로운 알고리즘을 제안하였다. 제 4절에서는 분석대상인 자료들에 대하여 간단히 소개하고 해당 자료 분석결과를 다른 분류방법의 결과와 비교하였다. 제 5절에서는 본 연구결과의 의의 및 문제점, 그리고 앞으로의 연구방향에 대하여 언급하였다.

2. 기존의 PAIRWISE COUPLING 분류방법

Friedman(1996)이 제안한 pairwise coupling이란 짝을 이룰 수 있는 가능한 모든 수준 조합에 대하여 분류예측을 한 후, 가장 다수의 결과를 얻은 수준으로 최종 분류하는 방법이다. 예를 들어, 총 세 수준(수준 A_1, A_2, A_3)에 대한 분류예측 문제가 있다고 하자. Pairwise coupling이란, 수준 A_1과 A_2, 수준 A_2와 A_3, 수준 A_3과 A_1에 대한 두 수준(dichotomous) 분류예측 모형의 결과가 각각 수준 A_1, A_2, A_1을 예측했다면 가장 많은 예상을 한 수준 A_1으로 최종 결론을 내리는 방법인 것이다.

Hastie와 Tibshirani(1996)는 Friedman(1996)이 제안한 pairwise coupling 방법의 한계를 다음과 같이 지적하였다:

총 3개의 수준 A_1, A_2, A_3에 대한 분류문제를 생각해 보자. 확률 r_{ij}를

$$r_{ij} = P(A_i | \ A_i \ \text{or} \ A_j)$$

라고 하면, $r_{12} > 0.5$, $r_{23} > 0.5$, 그리고 $r_{31} > 0.5$인 경우 Friedman의 pairwise coupling은 분류예측을 할 수 없다. 아울러 Friedman의 pairwise coupling은 짝을 이룬 수준과의 비교를 통한 승리 회수(the number of winning)에 의존하므로 해당 수준에 속할 확률 값을 계산하지 못한다는 단점이 있다.

이상 언급한 단점을 보완하기 위해 Hastie와 Tibshirani(1996)는 Bradely-Terry 모형의 기본 아이디어를 이용하여 새로운 분류방법을 소개하였다. 그러나, Hastie와 Tibshirani(1996)의 알고리즘은 Friedman의 pairwise coupling에 의해 도출된 조건부 확률 r_{ij}의 추정치를 바탕으로 반복적인 계산에 의해 각 수준에 속할 확률값을 추정하고 있다. 이와 같이 Hastie와 Tibshirani가 제안한 방법은 각 수준에 속할 확률값을 분류모형이 추정하는 것

이 아니라 반복계산을 통하여 도출되기 때문에 자료에 대한 분류예측시 상당한 시간이 소요될 수 있다.

본 논문에서는 위와 같은 확률계산과 모형의 존재여부에 대한 분류예측 문제점을 보완하기 위한 새로운 알고리즘을 제안하였다.

3. 제안 알고리즘

N개의 관측치로 이루어진 훈련자료를 $\boldsymbol{x} = (\boldsymbol{x_1}, ..., \boldsymbol{x_N})$라 하자. 각 $\boldsymbol{x}_i = (\boldsymbol{t}^i, y)$는 설명변수 $\boldsymbol{t}^i = (t_1^i, \ldots, t_p^i)$와 반응변수 y로 구성되어 있다고 하자. Breiman(1996)의 Bagging과 pairwise coupling 방법을 결합한 알고리즘은 다음과 같다.

STEP 1

K개의 수준을 가진 훈련자료 $\boldsymbol{x}$에 대하여, 모든 가능한 두 개의 수준 조합으로 이루어진 훈련자료 $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_n)$로 분리, 확장한다. 여기서, $n = K(K-1)/2$이다.

STEP 2

각각의 훈련자료 $\boldsymbol{z}_i$에 대해 b회 붓스트랩 표본추출 하여

$$B_i^* = (B_{i1}^*, B_{i2}^*, ..., B_{ib}^*),\ i = 1, 2, ..., n$$

를 생성한다.

STEP 3

각각의 붓스트랩 자료들 B_{ij}^*, $(i = 1, \ldots, n; j = 1, ..., b)$에 대하여, 분류예측모형 $C(B_{ij}^*, \cdot)$를 적합한다. 입력(설명변수) $\boldsymbol{t}$에 대한 분류 결과로서, 분류 예측된 수준의 위치에 해당되는 좌표는 1, 나머지 $K-1$개는 0으로 이루어진 $K-$벡터 $C(B_{ij}^*, \boldsymbol{t})$를 생성한다.

STEP 4

입력 $\boldsymbol{t}$에 대한 최종 분류예측은 다음과 같이 결정한다.

$$C_{CB}(\boldsymbol{t}) = \frac{1}{nb} \sum_{i=1}^{n} \sum_{j=1}^{b} C(B_{ij}^*, \boldsymbol{t})$$

여기서, $C_{CB}(\boldsymbol{t})$는 관측값의 각 수준에 속할 가능성에 대한 추정치이며, 가장 높은 값에 해당되는 수준으로 분류한다.

4. 각종 자료 및 분석결과

4.1. WAVEFORM 자료 및 분석결과

세 개의 수준(Class 1, 2, 3)을 갖는 waveform 자료는 다음과 같은 21개의 설명변수로 이루어져 있다.

$$\begin{aligned} x_i &= uh_1(i) + (1-u)h_2(i) + \varepsilon_i : Class1 \\ x_i &= uh_1(i) + (1-u)h_3(i) + \varepsilon_i : Class2 \\ x_i &= uh_2(i) + (1-u)h_3(i) + \varepsilon_i : Class3 \end{aligned}$$

여기서, $i = 1, 2, ..., 21$이다. 그리고, u는 $[0,1]$ 사이의 균등분포를 따르는 변량이고, ε_i는 $N(0,1)$에서 추출된 변량이다. Waveform 자료 생성에 사용된 h_1, h_2, h_3은 다음과 같이 정의된다.

$$\begin{aligned} h_1(i) &= max(6 - |i-11|, 0) \\ h_2(i) &= h_1(i-4) \\ h_3(i) &= h_1(i+4) \end{aligned}$$

그림 4.1은 waveform 자료에서 표본 추출된 300개의 자료를 대상으로 주성분 분석을 실시한 후, 그 결과를 2차원 평면에 투영한 것이다. 평면상의 숫자는 자료의 해당 수준을 나타낸다. 그림 4.1 상의 원형 점선 안쪽 부분을 보면, 서로 다른 두 수준의 자료가 얽혀있는 것을 알 수 있다.

표 4.1은 waveform 자료 중 300개를 훈련자료(training data)로, 500개를 검사자료(test data)로 생성하여 각 모형마다 5회 반복 실험한 결과이다. 사용된 자료에 포함된 각 수준의 비율은 거의 동일하다. 표의 오른쪽 부분은 오분류 비율의 평균과 괄호 안의 숫자는 오분류 비율평균의 표준오차를 나타낸 것이다. 상위 3개의 모형에 대한 결과는 Hastie와 Tibshirani (1996)에서 발췌한 것이며, 모의 실험은 동일한 조건하에서 이루어졌다. 모의 실험 결과의 비교에 사용된 모형 중 LDA(Linear Discriminant Analysis)는 선형 판별분석을,

표 4.1: Waveform 자료의 오분류 비율 (괄호 안의 숫자는 평균 오분류 비율의 표준오차)

모형	오분류 비율	
	Training	Test
LDA	0.148 (0.011)	0.214 (0.006)
Friedman	0.121 (0.010)	0.176 (0.006)
H&T	0.120 (0.010)	0.173 (0.005)
LDA/bagging	0.125 (0.008)	0.204 (0.015)
LDA/coupled with bagging	0.103 (0.010)	0.168 (0.006)

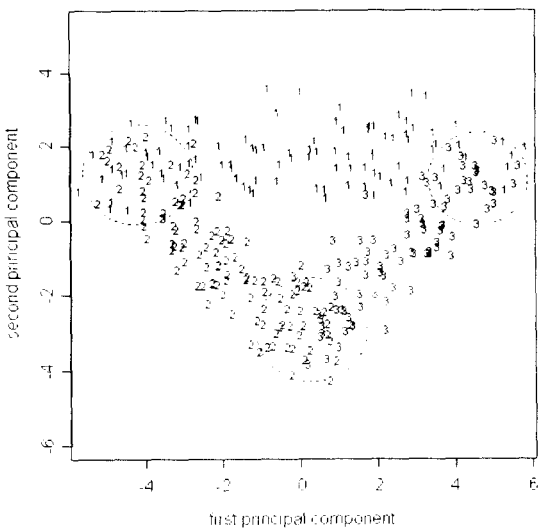

그림 4.1: waveform 자료

Friedman과 H&T는 각각 제 2절에서 소개한 Friedman(1996)과 Hastie와 Tibshirani(1996)의 방법을 의미한다. LDA/bagging은 Brieman(1996)의 Bagging 기법을 LDA에 적용한 것이고 LDA/coupled with bagging은 본 논문의 제 4절에서 제안하는 알고리즘을 의미한다. LDA/bagging과 LDA/coupled with bagging 모두 50회의 붓스트랩 표본추출을 실시하였다.

4.2. 3 CLASS 자료 및 분석결과

3 class 자료는 다음과 같은 방법으로 생성된다:
분포 *Uniform*$[-2,2]$에서 변량 2개를 추출하여 2차원 상의 점을 이룬다. 그리고, 평면상의 세 점 $(0,2),(-\sqrt{2},\sqrt{2}),(\sqrt{2},-\sqrt{2})$에 대하여 관측 값과의 거리 $d_j,(j=1,2,3)$를 구하고, 해당 점에 상응하는 수준은 $d_j^2-t_j$를 가장 작게하는 j로 결정한다. 즉, 수준 j는 다음과 같다.

$$j=\text{argmin}[d_j^2-t_j] \tag{4.1}$$

표 4.2는 3 class 자료 300개를 훈련자료(각 수준별로 100개씩)로, 각 수준별로 거의 동등한 수로 이루어진 300개의 검사자료를 생성하여 각 모형마다 5회 반복 실험한 결과이다. 상위 4개의 모형에 대한 결과는 Hastie와 Tibshirani (1996)에서 발췌한 것이며, 모의실험은 동일한 조건 하에서 이루어졌다.

표 4.2의 여러 가지 모형과 비교할 때, 본 연구에서 제안한 방법이 waveform 자료에서와는 달리 LDA와 QDA의 분류정확도를 모두 떨어뜨리는 결과를 가져왔다. 이와 같은 결과에 대한 이유를 밝히기 위하여, 몇가지 모형에 대한 분류 경계선을 그려 보았다.

그림 4.2에서 점선은 식 (4.1)에 의해 생성된 각 수준을 구분하는 경계선이고, 실선은 각 모형에 의해 생성된 분류 경계선이다. 좌측 두 개의 그림과 우측 두 개의 그림 각각에 대해

표 4.2: 3 class 자료의 오분류 비율 (괄호 안의 숫자는 평균 오분류 비율의 표준오차)

모형	오분류 비율	
	Training	Test
LDA	0.065 (0.004)	0.063 (0.004)
Friedman	0.069 (0.004)	0.068 (0.005)
H&T	0.069 (0.004)	0.068 (0.005)
QDA	0.032 (0.004)	0.029 (0.003)
LDA/bagging	0.060 (0.006)	0.066 (0.008)
LDA/coupled with bagging	0.067 (0.008)	0.069 (0.006)
QDA/bagging	0.021 (0.004)	0.043 (0.010)
QDA/coupled with bagging	0.024 (0.004)	0.045 (0.010)

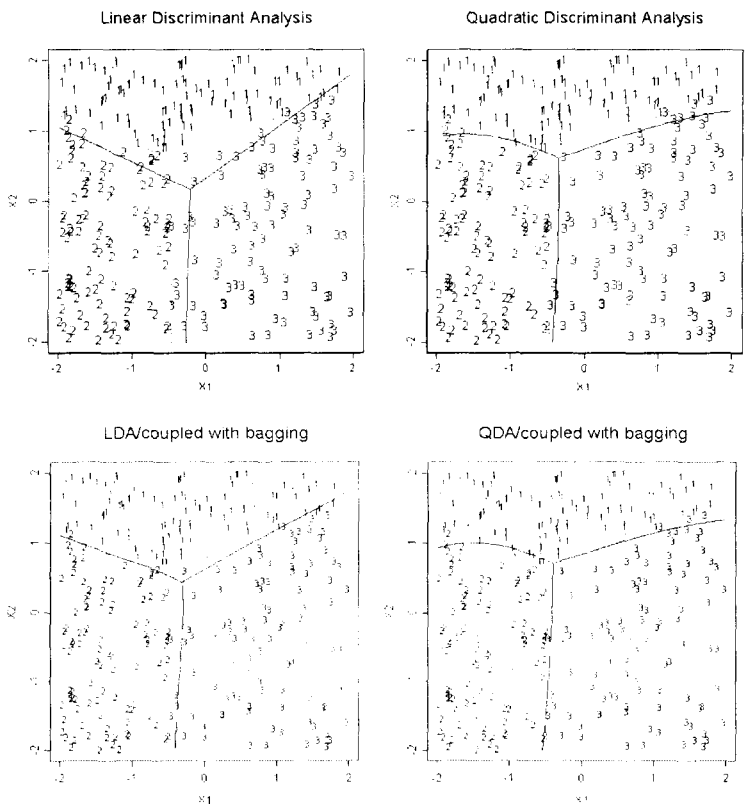

그림 4.2: 3 class 자료에 대한 모형별 분류 예측

서 점선과 실선 사이의 면적을 비교해 보면, 그 면적이 제안 알고리즘에 의해 작아진 것을 알 수 있다. 전체 면적에 대한 점선과 실선 사이의 면적의 비율을 각 모형의 오분류 비율이라고 할 수 있으므로 앞의 실험 결과와는 달리 본 논문에서 제안한 방법이 기존의 LDA와 QDA의 분류 정확도를 향상 시켰음을 알 수 있다. 표 4.2에서 제안 알고리즘의 오분류 비율이 높았던 것은 다음의 이유에서 비롯되었다고 할 수 있다:

(i) 검사자료 300개의 수준별 비율이 거의 동일한데 반해 각 수준에 해당되는 면적은 다르다.

(ii) 실제 분류 경계선 부근에서의 자료 빈도가 적다.

위와 같은 문제점을 해결하기 위하여 $[-2,2]\times[-2,2]$에서 균등하게 분포하는 검사자료를 사용하였고, 그 개수를 500, 1000, 3000개로 늘려가며 각 모형마다 5회 반복 실험을 하였다. 표 4.3의 결과로부터 전체 면적에 대하여 각 모형에 대한 점선과 실선에 의해 형성된 면적의 비율을 근사해 볼 수 있다.

표 4.3: 자료의 수에 따른 오분류 비율 (괄호 안의 숫자는 평균 오분류 비율의 표준오차)

모형	오분류 비율		
	n =500	n =1000	n =3000
LDA	0.113 (0.003)	0.112 (0.007)	0.107 (0.003)
QDA	0.054 (0.005)	0.053 (0.005)	0.051 (0.004)
LDA/coupled with bagging	0.071 (0.008)	0.066 (0.007)	0.061 (0.005)
QDA/coupled with bagging	0.036 (0.003)	0.035 (0.002)	0.034 (0.001)

표4.3에 의하면 제안 알고리즘을 LDA와 QDA에 적용하는 경우 각각의 분류 정확도가 4.9%와 1.8% 증가하였다.

4.3. Vowel 자료 및 분석결과

Vowel 자료는 11개의 모음을 수준으로하는 자료로써, 10개의 설명변수를 갖는다. 11개의 모음과 그에 대응하는 단어는 다음과 같다.

수준	1	2	3	4	5	6	7	8	9	10	11
모음	i	O	I	C:	E	U	A	u:	a:	3:	Y
단어	heed	hod	hid	hoard	head	hood	had	who'd	hard	heard	hud

각각의 단어에 대해서 6번씩, 8명의 연설자의 발음을 디지털화하여 10개의 설명변수를 갖는 훈련자료가 생성되었고, 같은 방법을 통해 7명의 연설자의 발음으로 검사자료가 생성되었다. 그림 4.3은 vowel 자료 중 528개의 훈련자료를 대상으로 주성분 분석을 실시한 후, 그 결과를 2차원 평면에 투영한 것이다.

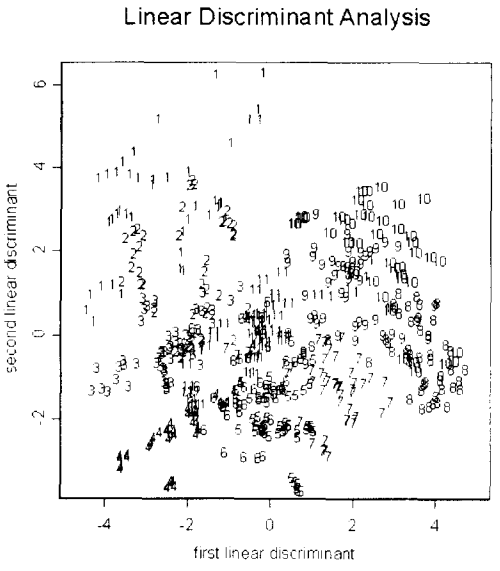

그림 4.3: Vowel 자료

표 4.4는 주어진 vowel 자료 중 528개를 훈련자료로, 462개를 검사자료로 사용하여 각 모형마다 5회 반복 실험한 결과이다. 상위 4개의 모형에 대한 결과는 Hastie와 Tibshirani (1996)에서 발췌한 것이며, 제안 알고리즘에 대한 모의 실험은 동일한 조건 하에서 이루어졌다. LDA/coupled with bagging, QDA/coupled with bagging 모두 50회의 붓스트랩 표본 추출을 실시하였다.

표 4.4: Vowel 자료의 오분류 비율 (괄호 안의 숫자는 평균 오분류 비율의 표준오차이다.)

모형	오분류 비율	
	Training	Test
Linear Discriminant Analysis (LDA)	0.296 (0.020)	0.500 (0.018)
Friedman	0.132 (0.013)	0.479 (0.019)
H&T	0.128 (0.013)	0.480 (0.017)
Quadratic Discriminant Analysis (QDA)	0.023 (0.002)	0.490 (0.014)
LDA/coupled with bagging	0.369 (0.009)	0.360 (0.011)
QDA/coupled with bagging	0.058 (0.004)	0.072 (0.004)

위의 여러 가지 모형과 비교할 때, 본 연구에서 제안한 방법이 기존의 LDA와 QDA의 분류 정확도를 크게 향상시켰으며, 특히 QDA/coupled with bagging의 경우 상당히 높은 분류정확도(검사자료에 대한 오분류 비율 0.072)의 향상을 가져왔다.

표 5.1: Vowel 자료에 대한 모의 실험 소요시간

모형	Modeling	Scoring/observation
LDA/coupled with bagging	42분	26.9초
QDA/coupled with bagging	53분	83.6초

5. 결론 및 향후 연구과제

본 논문에서는 LDA와 QDA에만 제안 알고리즘을 적용하였으나 tree model 등 다양한 방법에 적용하여야 할 것이다. 그 이외에 현재 제안한 알고리즘이 지니고 있는 다음과 같은 문제점에 대하여도 깊은 연구가 필요하다.

(i) Pairwise coupling이 갖고 있는 공통된 문제로써, 수준이 많으면 많은 계산을 하여야 한다. 예를 들어, vowel 자료의 경우 수준 $K = 11$이므로, 기본적으로 $K(K-1)/2 = 55$의 비교가 필요하다. 게다가 각 비교마다 50회의 bagging을 실시하면 총 $55 \times 50 = 2750$개의 모형계산이 필요하다. 표 5.1는 vowel 자료에 대한 모의 실험에서 소요된 시간을 나타낸다. 즉, 표5.1의 결과는 각 모형별로 528개의 훈련자료를 사용하여 모형을 생성하고 한개의 입력자료에 대한 분류 예측하는데 소요되는 시간을 Pentium III 500MHz Dual processor와 RAM 512Mb의 환경 하에서 측정한 것이다.

(ii) 총 $K(K-1)/2$회의 짝을 이룬 수준 비교 중, $K-1$개를 제외한 나머지는 해당 수준과 관련 없는 모형이다. 이와 같이 실제 수준과 관련 없는 모형으로부터 산출되는 결과를 제거하거나 비중을 줄여주는 방법을 고려하여야 한다.

참고문헌

[1] Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and regression trees*, Chapman and Hall.

[2] Breiman, L. (1996). Bagging predictors, *Machine Learning*, 26, 123-140.

[3] Breiman, L. (1998). Arcing classifiers,(discussion paper) *Annals of Statistics*, 26, 801-824.

[4] Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119-139.

[5] Friedman, J.H. (1996). Another approach to polychotomous classification, Technical report, Stanford University.

[6] Hastie, T. and Tibshirani, R. (1996). Classification by pairwise coupling, Technical report, University of Toronto.

[2000년 2월 접수, 2000년 5월 채택]

On the Classfication by an Improved Pairwise Coupling Algorithm

Daewoo Choi[1] Joongsik Yoon[2]

ABSTRACT

We proposed a new classification algorithm based on bootstrap sampling and pairwise coupling method. Also, for comparing the accuracy of a proposed algorithm with those of old methods, we conducted classification with waveform data and others.

Keywords: Classification; Ensemble technique; Pairwise coupling.

1) Assistant Professor, Department of Statistics, Hankuk University of Foreign Studies. E-mail: dachoi@popsmail.com

2) Graduate Student, Department of Statistics, Hankuk University of Foreign Studies. E-mail: jsyun@stat.hufs.ac.kr