응용통계연구 제 13권 2호, 2000년, pp. 447-455

K-평균 군집방법을 이용한 가중커널분류기*

백장선[1] 심정욱[2]

요 약

본 논문에서는 커널분류기에 요구되는 다량의 계산량과 자료저장공간을 감소시키도록 고안된 최적군집방법을 적용한 K-평균 가중커널분류기법이 제안되었다. 이 방법은 원래의 훈련표본보다 작은 수의 참고벡터들과 그들의 가중값들을 찾아 원래 커널분류기준을 근사화하여 패턴을 인식하는 것이다. K-평균 가중커널분류기법은 가중파젠윈도우(WPW)분류기법을 개량한 것으로서 참고벡터들을 계산하기 위한 초기 부적절하게 군집된 관측값들을 최적으로 재군집화 함으로써 WPW기법의 단점을 극복하였다. 실제자료들에 제안된 방법을 적용한 결과 WPW 분류기법보다 참고벡터들의 대표성과 자료축소면에서 월등히 향상된 결과를 확인하였다.

주요용어: 커널패턴인식, K-평균 군집방법, 비모수 분류기.

1. 서론

통계적 패턴인식체계를 설계할 때 특징변수의 분포가 알려져 있지 않은 상황에서는 비모수적 분류기가 이용된다. 비모수적 방법은 특징변수의 밀도함수에 대한 특별한 형태를 가정하는 대신 훈련표본으로부터 직접 밀도함수를 추정한다. 커널추정량(kernel estimator)은 이러한 비모수적 방법의 대표적인 분류기라 할 수 있는데 이것의 이론적 배경과 성능은 이미 잘 알려져 있다. 그러나 불행하게도 대규모 훈련자료에 대하여 이 방법을 적용했을 때 분류성능은 크게 향상되지만 계산시간이 오래걸리고 대규모의 계산용량을 요구하는 단점을 가지고 있다.

위에 언급된 커널분류기의 단점을 극복하기위해 현재까지 미국의 학자들에 의해 여러가지 대안들이 제시되어 연구되고 있다. Fukunaga (1990)는 비모수적 자료감소기법을 적용하여 전체훈련자료를 축차적으로 분할하여 밀도함수를 추정할 수 있는 소위 Reduced Parzen(RP)분류기를 제안하였다. RP분류기에 사용되는 최종훈련자료는 초기분할에 일반적으로 의존한다. West (1993)는 훈련자료를 몇 개의 군집으로 분류한 다음 각 군집의 중심점들과 가중값들을 계산한 다음 그것들을 이용하여 커널 추정값을 계산하는 방법을 제안하였다. 또한 Fan and Marron (1994)은 등간격의 궤중심점을 사용한 궤커널추정량 (binned kernel estimator)을 제안하였다. 이 방법의 계산속도는 일반적으로 매우 빠르다.

* 이 논문은 1998년도 정보통신분야 우수학교 지원사업 연구비 지원에 의하여 연구되었음.

1) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 자연과학대학 통계학과, 정보통신연구소, 부교수
E-mail: jbaek@chonnam.chonnam.ac.kr

2) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 자연과학대학 통계학과, 정보통신연구소, 교수
E-mail: jwsim@chonnam.chonnam.ac.kr

최근 Babich and Camps (1996)는 Weighted Parzen Window(WPW)분류기를 제안하였는데, 이는 전체훈련표본을 적용했을 때의 커널 추정값을 근사시킬 수 있도록 훈련표본으로부터 참고벡터들(reference vectors)과 가중값들을 찾아서 그것들을 활용한 분류 방법이다. 이 방법은 West (1993)의 방법과 RP분류기와 유사한 점이 있으나 훈련단계에서 전체훈련표본에 대한 커널추정값을 이용하여 그것과 어느 한계내에서 근사할 때까지 참고벡터와 가중값을 찾아가는 점에서는 차이가 있다. 이 방법은 일단 참고벡터들과 가중값들이 결정되면 전체훈련표본을 대체하여 분류기를 설계할 수 있으므로 계산용량이 대폭 축소되어 계산속도 역시 빨라진다. WPW분류기의 참고벡터들을 찾는 방법은 일반적인 계보적 군집방법중의 하나인 중심연결법 (Centroid Linkage Method)이며 표본의 이상치(outlier)에 별로 영향을 받지 않은 장점은 있으나 일단 관측값이 부적절하게 군집되면 끝까지 그 군집에 속하게 되어 회복할 수 없는 단점이 있다.

본 연구에서는 위에 언급된 WPW분류기의 단점을 극복할 수 있도록 중심연결법 대신 최적분리 군집방법을 적용한 새로운 커널분류기를 고안했으며 제 2장에 그 내용이 기술되어 있다. 제 3장에서는 본 연구에서 제안된 커널분류기를 실제자료에 적용하여 기존의 방법들과 성능을 비교한 후 우수성을 확인한 결과를 포함하고 있다. 마지막으로 제 4장에서는 결론을 기술하였다.

2. K-평균 군집방법을 이용한 커널패턴인식

2.1. 최적분리 군집방법과 가중커널분류기

패턴인식의 주된 기능은 입력된 패턴이 몇 개의 주어진 부류중 어느 곳에 속하는지 분류하는 것이다. 인식하고자 하는 새로운 패턴이 입력되면 훈련표본으로부터 구해진 판별함수들에 의해 가장 큰 값을 가진 판별함수에 대응되는 부류에 분류된다. 분류하고자하는 부류가 g개이며, 각 부류의 사전확률이 π_i이고, d-차원 특징변수 $\boldsymbol{x}$에 대한 각 부류의 확률밀도함수가 $f_i(\boldsymbol{x})$ 라하면 (i=1,2,$\cdots$,g) 각 부류에 대한 베이즈규칙에 의한 판별함수는 $\pi_i f_i(\boldsymbol{x})$이다. 따라서 비모수적 판별함수의 추정량은 각 부류의 사전확률과 밀도함수에 대한 추정량의 곱으로서 이루어져 있으며, 밀도함수가 커널방법에 의해 추정되었으면 우리는 그 추정량을 커널추정량이라 부른다. 만약 각부류의 사전확률이 모두 동일하면 밀도함수에 대한 커널추정량이 판별함수의 추정량, 즉 분류기가 된다.

n개의 d-차원 훈련표본 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$이 주어져 있다고 하면, 인식하고자 하는 패턴 $\boldsymbol{x}$에 대한 커널추정량 $\hat{f}_n(\boldsymbol{x})$(PW (Parzen Window) 분류기)는 다음과 같다 :

$$\hat{f}_n(\boldsymbol{x}) = \sum_{i=1}^{n} \frac{1}{n|H|^{1/2}} K\{H^{-1/2}(\boldsymbol{x} - \boldsymbol{x}_i)\} \tag{2.1}$$

이 때 $K(\cdot)$를 커널함수 또는 window라 하고, H를 $d \times d$평활모수(smoothing parameter, window width, 혹은 bandwidth) 행렬이라 한다. d-차원 커널함수는

$$\int K(\boldsymbol{x}) d\boldsymbol{x} \;=\; 1$$

을 만족해야하며, 종종 d변량 확률밀도함수를 사용한다. 평활모수행렬 H는 n의 함수, 즉,$H = H(n)$으로서 만약 $n^{-1}|H|^{-1/2}$ 과 H의 모든 요소들이 $n \to \infty$에 따라 0으로 접근하고, H의 고유값의 최소치와 최대치의 비율이 모든 n 에 대하여 유계(bounded)하고 몇가지 규칙을 만족하면 커널추정량 $\hat{f}_n(\boldsymbol{x})$는 진실된 밀도함수에 수렴한다. 따라서 훈련표본의 크기 n이 클수록 커널분류기 $\hat{f}_n(\boldsymbol{x})$는 보다 정확하게 되지만 반면에 계산량이 크게 증가하게 된다.

위의 커널분류기가 전체 훈련표본 모두를 사용하게되어 나타나는 계산용량의 증대를 감소시키기 위해 WPW분류기는 보다 작은 표본으로서 커널추정량을 근사시킬 수 있는 기법이다. 즉, 전체 훈련 표본을 대표할 수 있도록 군집방법이라는 자료감소기법을 통하여 보다 작은 표본을 생성하여 그것들을 이용하여 새로운 커널 추정량을 구축하였다. 군집방법을 통하여 d-차원 공간상에 흩어져있는 훈련표본들을 가까운 것끼리 군집화하여 각 군집의 중심점을 구하고 그것을 참고벡터라 불렀다. 또한 각 참고벡터의 가중값을 구하여 이 두가지를 이용하여 새로운 커널 추정량을 정의하였다. 이렇게 정해진 참고벡터집합을 $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_m\}$, $m \leq n$이라 하고 각 i번째 참고벡터$\boldsymbol{r}_i$의 가중값을 w_i라 하면 WPW분류기는

$$\hat{f}_m(\boldsymbol{x}) = \sum_{i=1}^{m} \frac{w_i}{n|H|^{1/2}} K\{H^{-1/2}(\boldsymbol{x} - \boldsymbol{r}_i)\} \tag{2.2}$$

로 정의된다. 이 때 가중값 w_i는 원래 훈련 표본 중 i번째 군집에 속하게 되는 개체들의 수를 나타낸다. 이때 $\int |\hat{f}_n - \hat{f}_m| d\boldsymbol{x}$의 L_1거리가 일정한 임계치를 넘지 않을 때까지 군집화를 계속하여 m을 결정하게 된다.

WPW의 군집화과정에서 사용된 방법은 중심연결법으로서 일단 원래 훈련표본의 어느 개체가 특정 군집에 속하게 되면 이후의 군집화과정에서 보다 적절한 다른 군집으로 재배치가 되지 못하여 원래 전체 훈련표본의 자료감소가 불합리하게 되고 따라서 참고벡터들의 대표성 역시 낮아질 가능성이 있다. 따라서 본 연구에서는 보다 대표성이 있는 참고벡터들을 생성하기 위해 군집내 산포행렬의 대각합(trace)을 최소화하는 최적분리 군집방법을 적용하고자 한다.

2.2. K-평균 가중커널분류기 설계

본 연구에서 채택한 최적분리방법인 군집내 산포행렬의 대각합을 최소화하는 것은 $\boldsymbol{a} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m)$에 관한 다음 식(2.3)을 최소화하는 m개의 점 $\boldsymbol{a} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m)$을 찾아낸 다음, 객체 $\boldsymbol{x}_i$, $i = 1, \cdots, n$ 들을 그 m개의 점들 중 가장 가까운 거리에 있는 점끼리 묶는 K-평균 군집방법과 동일한 것이다.

$$\psi(\boldsymbol{a}) \;=\; (1/n) \sum_{j=1}^{n} min\{(\boldsymbol{x}_j - \boldsymbol{a}_i)'(\boldsymbol{x}_j - \boldsymbol{a}_i); 1 \leq i \leq n\} \tag{2.3}$$

이 때 K-평균이라 함은 앞에서 얻은 m개의 점들 $\boldsymbol{a} = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m)$ 이 해당하는 군집에 속한 객체들의 평균이 되기 때문이다. (여기서는 m개의 군집평균들이 생성되기 때문에 m-평

균이라 기술해야 하지만 일반적으로 이 방법을 K-평균 군집방법으로 부르기 때문에 K-평균이라는 용어를 사용하였다.)

가중커널분류기에 사용되는 참고벡터들과 가중값을 계산하기 위하여 K평균 군집방법을 적용한 훈련과정이 그림 2.1에 나와있다. 훈련과정은 참고벡터의 크기, 즉 군집의 수가 표본크기 n부터 시작하여 한 개씩 감소하면서 원래의 표본을 대표할 수 있도록 K-평균 군집방법을 적용하여 참고벡터들을 생성한다. 원래표본을 이용한 커널추정값 $\hat{f}_n$과 참고벡터들을 이용하여 계산된 가중커널추정값 $\hat{f}_m$의 차이가 일정한 임계치를 넘지 않을 때까지 계속된다. 이러한 훈련과정은 매단계마다 새롭게 구성된 최적의 군집으로부터 참고벡터들이 생성되므로 어느 한 표본객체가 전단계의 참고벡터를 계산하기 위한 군집에 속해 있어야 하는 WPW 훈련과정과는 달리 다음단계의 새로운 군집으로 이동하여 참고벡터를 생성하는데 사용될 수 있다. 이러한 과정을 우리는 K-평균가중파젠윈도우(K-means Weighted Parzen Window : KWPW) 훈련과정이라 부르며, 최종선택된 참고벡터들과 가중값을 이용한 식 (2.2)의 커널분류기를 K-평균 가중커널분류기라고 명명하였다.

1. H의 선택과 e_{max}값 결정 ($e_{max} \geq 0$)
2. 초기화 : $m \leftarrow n$, $\boldsymbol{R} \leftarrow \boldsymbol{X}$, $\boldsymbol{w} = [w_1, w_2, \cdots, w_m]^T = \mathbf{1}$.
3. (식 2.1)을 이용하여 $\hat{f}_n(\boldsymbol{x}_i)$, $i = 1, \cdots, n$ 계산
4. $\boldsymbol{x}_i^* = H^{-1/2}\boldsymbol{x}_i$ 들에 대하여 K-평균 군집방법 적용
   $\boldsymbol{R} = \{r_i : r_i \leftarrow i$번째군집의 $K-$평균점$, i = 1, \cdots, m\}$갱신
   $\boldsymbol{w} = \{w_i : w_i \leftarrow i$번째군집의 크기$, i = 1, \cdots, m\}$갱신
5. (식 2.2)를 이용하여 $\hat{f}_m(\boldsymbol{x}_i)$계산
6. $e = \frac{1}{n}\sum_{i=1}^{n} \frac{|\hat{f}_n(\boldsymbol{x}_i) - \hat{f}_m(\boldsymbol{x}_i)|}{\hat{f}_n(\boldsymbol{x}_i)}$ 계산
7. $e \leq e_{max}$이면 , $m \leftarrow m - 1$,단계 4.로 감
   $e > e_{max}$이면 이전단계의 $\boldsymbol{R}$과 $\boldsymbol{w}$추출

그림 2.1 : 단일집단의 KWPW 훈련 알고리즘

3. 실제자료 적용

KWPW 훈련 알고리즘에 의하여 새롭게 구해진 참고벡터들과 가중값들을 이용하여 구축한 커널 분류기의 성능을 확인하기 위하여 실제자료들을 활용하여 오분류율을 계산하고 다른 분류기들과 비교한다.

성능실험에 사용될 실제자료는 붓꽃자료(Iris data)와 입학자료(Admission data : Johnson and Wichern (1992) p 567)이다. 붓꽃자료는 세가지 품종의 붓꽃으로부터 각

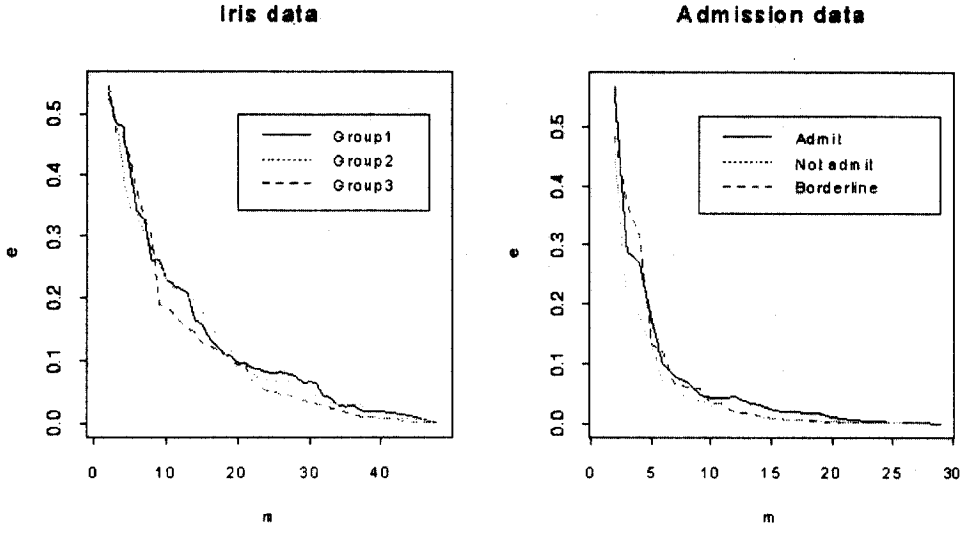

그림 3.1 : 두가지 실제자료들의 그룹별 참고벡터의 수(m)에 대한 오차(e).

각 50개씩 총 150개 개체들을 추출해서 꽃받침 조각의 길이, 꽃받침 조각의 폭, 꽃잎의 길이, 꽃잎의 폭의 4개 특징변수를 mm단위로 측정한 것이다. 이 자료는 지금까지 패턴인식에서 널리 사용되고 가장 많이 인용된 대표적인 것이다. 입학자료는 어느 경영대학원 응시생들 중에서 31명의 합격생(Admit), 28명의 불합격생 (Not admit), 26명의 입학보류생(Borderline)들의 GPA와 GMAT의 두가지 성적특징변수로 구성되어 있다.

본 실험에서 사용된 커널함수는 다변량 표준정규분포 밀도함수이다. 평활모수행렬 H는 S를 표본공분산행렬이라 할 때 $h > 0$에 대하여 $H = h^2S$로 선택하였다. 다양한 커널함수와 평활모수행렬의 선택과 그들의 특징들은 Wand and Jones (1995)에 정리되어 있다. h는 교차타당성(leave-one-out cross validation) 방법에 의하여 (2.1) 커널분류기의 최소오분류율을 달성하는 것($\hat{h}_{opt}$)으로 선택하였으며, 붓꽃자료의 경우 $\hat{h}_{opt}$= 1.30이며, 입학자료의 경우 $\hat{h}_{opt} = 0.70$이었다. 입학자료의 경우에는 각 집단의 훈련표본의 크기가 다르므로 판별함수 $\pi_i f_i(\boldsymbol{x})$의 사전확률 추정값으로서 전체훈련표본 크기에 대한 각 집단의 상대적 표본크기의 비율을 사용하였다.

KWPW 훈련 알고리즘에는 참고벡터를 이용한 가중커널추정값과 원래의 전체훈련표본을 이용한 커널추정값과의 상대적 평균차이에 대한 최대허용한계인 e_{max}를 정해야 한다. 두 추정값의 상대적 평균차이는 군집화를 더욱 진행시켜 갈수록(m이 작아질수록) 크게 될 것이며 만약 어느 단계까지는 상대적 평균차이가 조금씩 증가하다가 그 단계를 지나서 급격히 증가한다면 참고벡터의 대표성이 크게 손상되는 것을 의미하므로 그 단계의 상대적 평균차이가 최대허용한계인 e_{max}로 적합할 것이다. 본 실험자료의 각 부류의 표본크기로부터 차례로 하나씩 줄여가면서 K-평균군집방법에 의해 참고벡터를 구해서 저장한 후 가중커널 추정값을 구하여 상대적 평균차이 e를 계산한다.

표 3.1: 실제자료들에 대한 분류결과

자료	QDA 오분류율	PW 오분류율	e_{max}	WPW 오분류율	WPW $Ave\left(\frac{M}{N-1}\right)$	KWPW 오분류율	KWPW $Ave\left(\frac{M}{N-1}\right)$
Iris			0.10	0.0133	0.402	0.0133	0.286
($\hat{h}_{opt}$=1.30)	0.0267	0.0133	0.20	0.0200	0.290	0.0133	0.128
			0.30	0.0200	0.214	0.0133	0.100
Admission			0.07	0.0706	0.389	0.0353	0.271
($\hat{h}_{opt}$=0.70)	0.0471	0.0353	0.10	0.0706	0.323	0.0471	0.228
			0.20	0.0706	0.261	0.0588	0.166

이렇게 계산된 e를 참고벡터의 수 m의 함수로 나타내면 그림 3.1과 같다. 그림 3.1에서 볼 수 있듯이 붓꽃자료의 경우 $e < 0.1$에서 대체적으로 선분들이 안정적이며 $e \geq 0.1$에서 증가하기 시작하므로 e_{max}의 적절한 값은 0.1이라 할 수 있겠다. 입학자료에서는 $e_{max} = 0.07$이 선택되었다. 일단 e_{max}가 결정되면 이전에 저장된 참고벡터의 수 m과 그때의 참고벡터를 이용하여 KWPW 분류기를 시행할 수 있다.

그림 3.1로부터 몇가지 e_{max}를 선택하여 실험자료들에 대한 각 분류기들의 오분류율을 계산한 결과가 표 3.1에 정리되어 있다. 각 분류기들의 오분류율은 교차타당성 방법에 의해 계산된 것이다. 우선 두 자료 모두 전체자료를 이용한 커널분류기(PW분류기)의 성능이 모수적 분류방법인 이차판별방법(QDA)보다 좋음을 확인할 수 있다. (참고로 선형판별방법 (LDA)의 결과는 이차 판별방법보다 좋지 않으므로 비교에서 제외하였다.) N이 각 집단 훈련표본들을 합한 전체 훈련표본의 크기를 나타내고 M이 각 집단의 참고벡터들의 총수를 나타낼 때, 표 3.1의 $Ave\left(\frac{M}{N-1}\right)$은 원래 표본에 비해 축소된 필요 참고벡터 자료량의 평균비율을 나타낸다. 표 3.1로부터 먼저 우리는 KWPW 분류기의 참고벡터들이 WPW 분류기의 그것들에 비하여 우수한 대표성을 가지고 있음을 알 수 있다. 즉 붓꽃자료에서는 WPW 분류기의 경우 $e_{max} = 0.20$ 에서 이미 오분류율이 0.0133에서 0.0200으로 증가하였으나 KWPW 분류기의 경우 $e_{max} = 0.30$ 까지 오분류율 0.0133을 그대로 유지하고 있다. 이는 WPW 훈련과정에서 계보적 군집방법을 사용한 결과 초기에 부적절하게 군집된 관측값들이 끝까지 그 군집에 속하게 되는 반면 KWPW 훈련과정에서는 최적분리 군집방법에 의해 군집의 대표성이 높아졌음을 확인시키는 결과이다. 두번째로 우리는 제안된 KWPW 분류기의 뛰어난 자료감소 성능을 확인할 수 있다. WPW 분류기의 경우 원래 훈련표본의 약 40 % 자료를 가지고 원래표본 전체를 이용한 PW 분류기의 오분류율을 달성하고 있는 반면, KWPW 분류기는 약 10% 자료를 사용하여 동일한 오분류율을 달성한다. 즉, KWPW 분류기는 PW 분류기에 비해 1/10 , WPW 분류기에 비해 1/4의 자료저장 공간만 있으면 동일한 인식기능을 수행할 수 있는 것이다.

입학자료의 경우에도 KWPW분류기의 우수성을 비슷하게 확인할 수 있다. 즉 KWPW의 경우 전체훈련표본 중 약 27%만 가지고서도 PW분류기의 오분류율 0.0353을 달성할 수 있는

반면, WPW분류기는 비슷한 자료량(약 26%)으로 PW분류기 오분류율보다 높은 0.0706의 오분류율을 나타낸다. 또한 약 17% 자료를 이용한 KWPW분류기의 오분류율이 0.0588로서 약 26% 자료를 사용한 WPW분류기의 오분류율 0.0706보다 낮으므로 KWPW의 자료감소 능력이 WPW의 그것에 비해 더 우수함을 다시한번 확인할 수 있다.

4. 결론

본 논문에서는 K-평균가중커널(KWPW) 분류기법을 소개하였다. 이 방법은 최적분리 군집방법의 하나인 K-평균 군집방법을 이용하여 원래의 훈련표본보다 작은 수의 참고벡터들과 그들의 가중값들을 찾아 커널분류기준을 근사화하여 패턴을 인식하는 것이다. KWPW 분류기법은 커널분류기에 요구되는 다량의 계산량과 자료저장 공간을 감소시키도록 고안된 가중파젠윈도우(WPW) 분류기법을 개량한 것으로서 참고벡터들을 계산하기 위한 초기 부적절하게 군집된 관측값들을 최적으로 재군집화함으로써 WPW 기법의 단점을 극복하였다. KWPW 훈련과정은 원래 훈련표본을 이용한 커널 추정값과 참고벡터들을 이용한 가중커널추정값과의 상대적 평균차이가 최대허용한계인 e_{max}보다 크게되면 종료하게 되며 그때의 참고벡터들과 가중값들을 이용하여 K-평균가중커널분류기를 설계한다. e_{max}는 상대적 평균차이를 참고벡터들의 개수 m의 함수로 나타낼 때 m이 감소함에 따라 급격히 증가하는 상대적 평균차이값으로 선택된다. 실제자료들에 KWPW 분류기법을 적용한 결과 WPW분류기법보다 참고벡터들의 대표성과 자료축소면에서 월등히 향상된 결과를 확인하였다.

감사의 글

본 논문에 대하여 조언을 해주신 두분의 심사위원님께 감사를 드립니다.

참고문헌

[1] Babich, G.A. and Camps, O.I. (1996). Weighted Parzen Windows for Pattern Classification, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, 567-570.

[2] Fan, J. and Marron, J.S. (1994). Fast Implementation of Nonparametric Curve Estimators, *J. Computational and Graphical Statistics*, vol. 3, 35-56.

[3] Fukunaga, K. (1990). *Statistical Pattern Recognition* , 2nd edition, Academic Press Inc, San Diego, Calif.

[4] Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*, 3rd edition, Prentice Hall, New Jersey.

[5] Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*, Chapman & Hall, London.

[6] West, M. (1993). Approximating Posterior Distributions by Mixtures, *J. Royal Statistical Soc. B*, vol. 55, 409-422.

[1999년 7월 접수, 2000년 6월 채택]

Kernel Pattern Recognition using K-means Clustering Method *

Jangsun Baek[1] Jungwook Sim[2]

ABSTRACT

We propose a weighted kernel pattern recognition method using the K-means clustering algorithm to reduce computation and storage required for the full kernel classifier. This technique finds a set of reference vectors and weights which are used to approximate the kernel classifier. Since the hierarchical clustering method implemented in the Weighted Parzen Window (WPW) classifier is not able to rearrange the proper clusters, we adopt the K-means algorithm to find reference vectors and weights from the more properly rearranged clusters. We find that the proposed method outperforms the WPW method for the representativeness of the reference vectors and the data reduction.

Keywords: Kernel pattern recognition; K-means clustering method; Nonparametric classifier.

* This research was supported by the Ministry of Information and Communication Fund in 1998.

1) Associate Professor, Department of Statistics, Chonnam National Univ.
E-mail: jbaek@chonnam.chonnam.ac.kr

2) Professor, Department of Statistics, Chonnam National Univ.
E-mail: jwsim@chonnam.chonnam.ac.kr