2008년도 대한전기학회 하계학술대회 논문집 2008. 7. 16 - 18

입력패턴과 그 k 근방 원형상에서 최근접 결정법칙에 의한 패턴식별

김응규*
한밭대 정보통신공학전공*

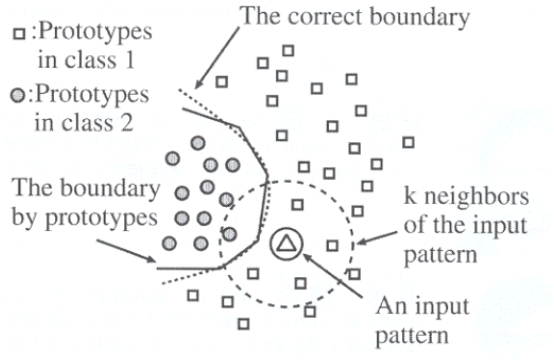
Pattern Classification using the Nearest Desion Method in Input Pattern and its k Neighbor Prototypes

Eung-Kyeu Kim*
Dept. of Information and Communiction, Hanbat National University

Abstract - 본 논문에서는 입력패턴과 그 k 근방 원형상에 잇어서 노름 평균에 기초한 최근접 결정법칙에 의한 패턴식별법을 제안한다. 이 방법은 식별경계 근방의 원형상에 있어서 분산의 차에 의한 가중치를 고려하기 때문에 패턴의 수가 적을 때 입력패턴을 정확하게 분류할 때 사용될 수 있다. 본 방법의 유효성을 평가하기 위해 인공적인 패턴과 실제패턴에 대해 k-NN 등 기존방법과 제안하는 방법을 적용하여 식별률에 의한 평가를 행한 결과, 특히 원형상의 분포가 희박한 경우 제안하는 방법이 기존방법에 비해 높은 식별률을 나타냈다.

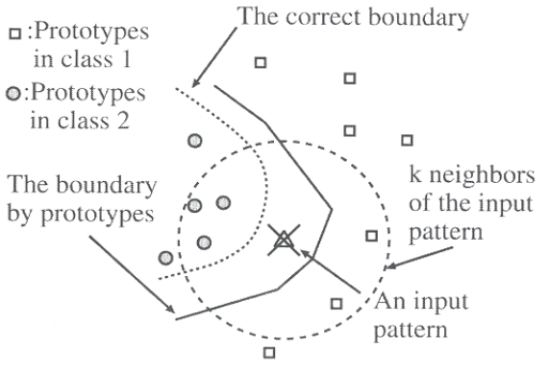
1. 서 론

선형분리가능한 패턴에 대한 패턴식별 방법중 k 근방 결정법(k-NN method)에 기초한 방법이 있다[1]-[4]. 이 방법은 우선, 분류된 클래스 패턴(class pattern)인 원형상(prototype)과 입력패턴과의 거리에 따라 그 거리가 작은 순서로 k개의 원형상을 선택한다. 다음으로, 이 k개의 원형상중에서 클래스별 다수결을 취해 최대가 된 클래스를 입력패턴이 속하는 클래스로 한다. 각 클래스의 원형상 분포가 집중될 최근접 결정법 k=1에 의해 고정밀도 식별이 가능하게 된다. 또한, 클래스별 최근접 원형상 근방의 원형상 조밀의 차가 각 클래스에서 작게 됨으로써 k 근방 결정법 k≥2에 의해 그림 1과 같이 고정밀도 식별이 가능하게 된다. 그러나, 각 클래스의 원형상 분포가 희박한 경우, 클래스 별 최근접 원형상에 의한 식별경계가 정확한 해의 식별경계로부터 크게 이탈하기 때문에 최근접 결정법 k=1에서는 식별 정밀도가 저하될 우려가 있다. 또한, 클래스별 최근접 원형상 근방의 원형상 조밀의 차가 각 클래스에서 크게됨으로써 k 근방 결정법 k≥2에서는 식별 정밀도가 그림 2에서와 같이 저하될 우려가 있다.



<그림 1> 각 클래스의 원형상 분포가 집중된 경우, 원형상에 의한 식별경계와 입력패턴 근방의 원형상

각 클래스에 있어서 원형상의 조밀을 고려한 식별방법으로서 매해러노비스의 거리(Mahalanobis)에 기초한 방법이 있다[5]-[8]. 이 방법은 각 클래스의 평균벡터와 입력패턴과의 거리를 각 클래스 전체의 공분산 행렬에 의해 가중치 부가를 행하기 때문에 원형상의 분포가 타원인 경우는 적당하지만, 타원이외에 임의 형상분포인 경우 식별정도가 저하된다고 생각된다. 또한, 선형분리가 불가능한 패턴에 효과가 있는 방법으로서 미타니(Mitani) 등은 각 클래스에 있어서 k 근방 패턴의 평균과 입력패턴과의 거리를 이용하는 CAP(Classification using Categorical Average Patterns)법을 제안했다[9]. 단, 이 방법은 각 클래스에 있어서 k 근방 패턴의 평균에 기초하여 입력패턴의 식별을 행하고 있기 때문에 패턴의 분포가 희박한 경우, 각 클래스에 있어서 k 근방 패턴의 평균이 희박한 패턴의 방향으로 크게 이동하기 때문에 식별률이 저하될 우려가 있다.



<그림 2> 각 클래스의 원형상 분포가 희박한 경우, 원형상에 의한 식별경계와 입력패턴 근방의 원형상

이에, 본 연구에서는 클래스별 원형상(prototype) 분포가 선형분리 불가능하고 동시에 분포가 서로 다르면서 빈약한 분포의 원형상에 있어서도 입력패턴의 고정밀도 식별을 행하는 것을 목적으로 한다. 여기에서 빈약한 원형상의 분포에 있어서 클래스별 최근방의 원형상에 의한 식별경계에 주목한다. 이 경우, 정확한 해의 식별경계에 대해서 클래스별 최근방 원형상에 의한 식별경계는 광범위하게 분포하는 클래스의 원형상 쪽으로 이동하는 경향이 있다. 따라서 입력패턴과 각 클래스에 있어서 원형상과의 거리에 기초한 식별을 행할 때, 어떤 클래스에 있어서 원형상의 분산이 클 수록 입력패턴과 그 클래스의 원형상 거리를 단축하는 듯한 가중치 부가를 행함으로써 고정밀한 식별가능성을 기대할 수 있다. 위에서 언급한 내용을 기초로, 본 연구에서는 입력패턴과 클래스별 최근접 원형상과의 거리를, 클래스별 최근접 원형상과 그 k 근방의 원형상에 있어서 평균벡터와 그 평균벡터의 산출에 이용한 각 원형상과의 노름(norm) 평균에 기초한 가중치 부가를 행하여 최근접 결정법clr에 의한 식별을 행한다.

이하에 각각 제안한 방법의 알고리듬 설명 및 인공적인 데이터와 실제 데이터에 대해서 기존의 방법인 k-NN법, 매해러노비스 거리[5]-[8], CAP[9], KCAP [10], SVM[11]의 각각에 기초한 방법과 제안하는 방법과의 비교를 통해 제안하는 본 방법의 유효성을 평가하고 결론을 맺는다.

2. 제안 방법의 알고리듬 및 유효성 평가

2.1 제안 방법 알고리듬

각 단계별 제안 알고리듬을 아래에 나타낸다.

[단계 1] 클래스별 최근접 원형상(prototype)의 결정.

[단계 2] 클래스별 최근접 원형상(prototype)과 그 k 근방 원형상 집합의 결정.

[단계 3] 클래스별 최근접 원형상(prototype)과 그 k 근방 원형상에 있어서 평균과 노름(norm) 평균의 산출.

[단계 4] 입력패턴과 클래스별 최근접 원형상과 노름 평균에 의

한 가중치를 부가한 거리의 산출.
[단계 5] 산출한 거리에 기초한 입력패턴의 식별.

2.2 제안한 방법의 유효성 평가

2.2.1 인공 패턴에 대한 실험

우선, 인공패턴에 대한 각 방법의 특성평가를 행한다. 여기에서는 두 클래스 패턴의 식별을 행한다. 두 클래스의 선형분리 불가능한 인공패턴을 생성하기 위해, 우선, 다음과 같은 초반구(super-hemisphere)를 사고한다.

$$x_n = \sqrt{r_c^2 - x_1^2 - x_2^2, \cdots, x_{n-1}^2} \qquad (1)$$

각 클래스에 있어서 패턴의 벡터는 다음의 확률분포를 만족하하는 d 를 각각 부가하여 생성한다.

$$p(d) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{d^2}{2\sigma^2}\} \qquad (2)$$

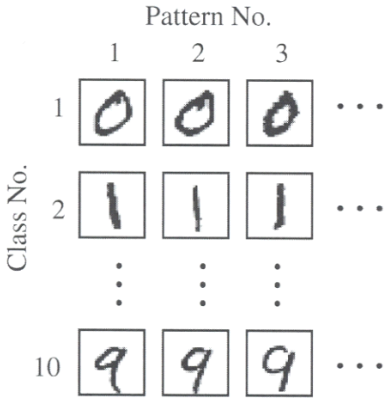
또한, 식별률 C_r 은 다음 식으로 표시한다.

$$C_r[\%] = \frac{1}{C_{Num}} \sum_{i=1}^{C_{Num}} [\frac{1}{P_{Num}} \{\sum_{j=1}^{P_{Num}} D(p_j^{(i)})\} \times 100] \qquad (3)$$

먼저, 각 클래스의 원형상(prototype) 수에 대한 k-NN에 의한 식별률을 나타내고, 다음으로 차원수에 대한 각 방법의 식별률을 나타낸다. 차원수에 대한 각 방법에 대한 식별률 실험결과, 각 클래스에서 원형상의 수는 20개, 입력패턴은 각 클래스별 100개로 한 경우, 제안하는 방법, SVM에서는 차원수가 증가하더라도 식별률은 고정밀도로 유지되고 있음을 알게 되었다.

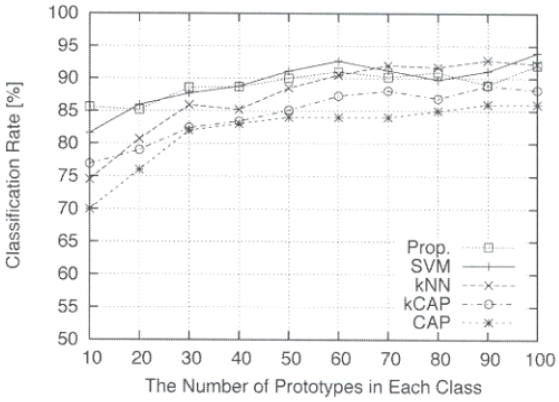
2.2.2 실제 패턴에 대한 실험

실제패턴에 대한 각 방법의 특성평가를 행하였다. 실험에 사용한 데이터는 문자식별용 데이터베이스인 미니스티(MINIST)를 이용했다[8]. 이것은 문자가 0-9까지 10개의 클래스, 각각 28×28 픽셀 사이즈 (pixel size), 8[bit/pixel] 계조영상에 대해 라스터 스캔(raster scan)에 의한 벡터화를 취하여 실험을 행하였다(그림 3). 여기에서 패턴의 차원수는 영상의 화소수 784로 하여 원형상의 수를 변화시킨 경우의 식별률에 의해 각 방법에 대한 평가를 행하였다. 더구나 SVM은 기본적으로 두 클래스의 식별문제를 취급하였기 때문에 각 클래스 1,000개 패턴의 평균에 대해 최단거리가 된 다른 클래스와의 식별률을 산출하여 그 식별률을 전체 클래스로 평균한 것을 여기에서의 각 방법의 식별률로 하였다.



<그림 3> 실험에 이용한 실제패턴의 예

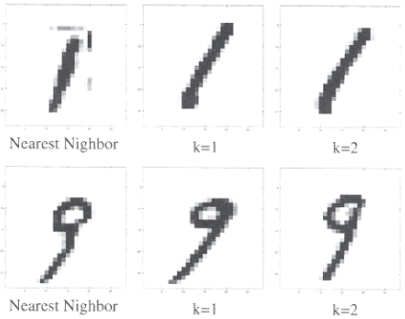
그림 4에 원형상(prototype)의 수 10으로부터 100개까지 매해러노비스(Mahalanobis)외의 방법에 있어서 식별률을 나타낸다. 실험결과로부터 제안하는 방법에 의한 식별률은 원형상의 수가 적은 경우에 특히 유효하다는 것을 알 수 있다. 또한 원형상의 수 785로부터 1,200개까지 각 방법에 있어서 식별률을 나타낸다.



<그림 4> 원형상 수 10에서 100개까지 각 방법에 있어서 식별률

실험결과로부터, 제안하는 방법(Prop.)에서 약간 높은 식별률을 나타냈지만, 매해러노비스(Mahalanobis)외의 방법은 거의 동등한 정도의 식별률을 나타냄을 알 수 있다. 이상의 실험으로 부터 제안하는 방법은 원형상의 수가 적고 희박한 경우에 특히 유효함을 알 수 있다.

다음으로, 각 클래스의 원형상 수를 10개로 한 경우 각 클래스의 최근접 원형상의 수와 그 k 근방 k=2의 패턴을 그림 5에 나타냈다. 이 그림 5로 부터 각 클래스의 원형상 수가 많은 경우 각 클래스의 최근접 원형상과 그 k 근방 원형상의 유사성이 높다는 것을 알 수 있다.



<그림 5> 원형상의 수 10개에 있어서 각 클래스의 최근접 원형상과 그 k 근방의 패턴

3. 결 론

클래스(class)별 원형상(prototype)의 분포가 선형분리 불가능하고 동시에 분산이 서로 다른 희박한 분포의 원형상에서도 입력패턴에 대한 고정밀도의 식별을 행하기위해 클래스별 최근접 원형상과 그 k 근방 원형상에서 노름(norm) 평균에 기초한 최근접 결정법칙에 의한 패턴식별법을 제안했다. 제안하는 방법의 유효성을 평가하기위해 인공적인 패턴과 실제 패턴에 대해 일반적인 k-NN법, CAP 등 각각에 기초한 방법과 제안하는 방법을 적용하여 식별률에 의한 평가를 행하였다. 그 결과, 인공적인 패턴에 대한 실험에 있어서 제안하는 방법은 타 방법들과 비교하여 원형상의 수를 변화시킨 경우와 차원수를 변화시킨 경우 양자 모두 식별률이 향상되었고, 원형상의 분포가 희박한 경우에 있어서 특히 유효 했다. 금후의 과제로서 노름 평균에 대한 가중치 부여 등 보다더 식별률을 향상시킬 수 있는 방안 모색이 남아있다.

[참 고 문 헌]

[1] S. M. Weiss, "Small sample error rate estimation for k-NN Classifiers", IEEE Trans. on Patt. Analy. and Machin. Intel. Vol.PAMI-13, No.3, pp.285-289, March 1991.

[7] F. Sun, S. Omachi, N. Kato, H. Aso, S. Kono and T. Tagagi, "Two-Srage Computational Cost Reduction Algorithm based on Mahalanobis Distance Approximations", Proc. of ICPR, Vol.2, pp.700-703, 2000.

[8] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition", Proc. of IEEE, Vol.86, No.11, pp.2278-2324, 2005.