통계적 패턴인식과 신경망을 이용한 필기 숫자 인식방법의 성능 비교

Comparative Analysis of Statistical and Neural Network Classifiers in Handwritten Digits Recognition

저자 (Authors)

정선화, 김수형, 조완현
Sun Wha Jeong, Soo Hyung Kim, Wan Hyun Cho

출처 (Source)

한국정보과학회 학술발표논문집 25(1B), 1998.4, 719-721(3 pages)

발행처 (Publisher)

한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE00627121

APA Style

정선화, 김수형, 조완현 (1998). 통계적 패턴인식과 신경망을 이용한 필기 숫자 인식방법의 성능 비교. 한국정보과학회 학술발표논문집, 25(1B), 719-721

통계적 패턴인식과 신경망을 이용한 필기 숫자 인식방법의 성능 비교

정선화* , 김수형, 조완현
전남대학교 전산학과

Comparative Analysis of Statistical and Neural Network Classifiers in Handwritten Digits Recognition

Sun Wha Jeong, Soo Hyung Kim, Wan Hyun Cho,
Department of Computer Science, Chonnam National University

요 약

본 논문에서는 필기 숫자 데이터베이스에 대한 인식률을 기준으로 통계적인 방법과 인공신경망 방법의 성능을 비교 분석하였다. 통계적인 방법으로서 선형판별함수(LDF), 이차형판별함수(QDF), 정칙화판별함수(RDF) 등의 모수적인 방법과, k-최근접이웃방법(k-NN) 등의 비모수적인 방법을 고찰하였고, 신경망 방법으로는 다층퍼셉트론(MLP)을 선택하여 이들 통계적인 방법과 비교하였다.

실험은 NIST 데이타베이스를 사용하여 이루어졌으며, 훈련용으로 SD3 중에서 64,000자를 그리고 테스트용으로 총 58,646자로 구성된 TEST1을 사용하였다. 실험결과 MLP가 가장 좋은 인식률을 보임을 알 수 있었지만, 1-NN 또한 이와 근접한 인식률을 보여주었다. 또한 RDF와 QDF가 그들과 비교하여 약간 낮은 인식 성능을 보였지만 이들은 각 부류에 대한 사후확률의 추정값을 제공한다는 장점을 가지고 있다.

Ⅰ. 서론

패턴분류에서 통계적인 기법들이 주로 사용되다가, 1980년 중반에 다층퍼셉트론과 오류 역전파 알고리즘이 개발되면서 신경망 기법이 패턴분류를 위한 보편적인 기법으로 사용되고 있다[5]. 한편 최근에는 Karhuen-Loève 변환방법을 도입한 통계적인 기법들이 신경망 기법에 비해서 우수한 성능을 보여 주는 사례가 나오고 있다[1,7]. 이들 통계적 기법 및 신경망을 이용한 기법들이 꾸준한 발전을 거듭하고 있음에도 불구하고, 특정 영역에서 이들 방법들을 서로 비교한 연구가 거의 이루어지지 않고 있다.

본 논문에서는 필기 숫자에 대한 인식률을 기준으로 통계적인 기법과 신경망 기법을 비교 분석하였다. 통계적인 방법은 크게 두 가지, 즉 모수적인 방법과 비모수적인 방법으로 구분할 수 있는데, 모수적인 방법으로는 부류에 대한 정규성과 등분산성을 가정하는 선형판별함수(LDF)와 부류에 대한 정규성만을 가정하는 이차형 판별함수(QDF), 그리고 이차형 판별함수에서 각 부류에 대한 공분산 행렬의 추정량으로 편의 추정량을 사용하는 정칙화판별함수(RDF)를 다루었다. 각 부류에 대하여 임의의 분포를 가정하지 않는 비모수적인 방법에서는 k-최근접 이웃방법(k-NN)을 채택하였다. 또한 신경망기법에서는 가장 많이 응용되고 있는 기법중의 하나인 다층퍼셉트론(MLP)을 선택하여 이들 통계적인 기법들과 비교하였다.

Ⅱ. 통계적인 분류방법

패턴분류의 목적은 p-차원의 벡터로 표현되는 미지의 패턴 $\mathbf{x} = \{x_1, \cdots, x_p\}$를 g개의 부류 중에 하나로 할당하는 것이다. 이때 주어진 패턴은 g개의 부류중 단 하나의 부류에만 속해야 하며 만약 소속 부류가 아닌 다른 부류에 할당되면 오류가 발생된다. 이러한 오류와 관련된 간단한 손실함수로

$$L(\pi_k, \hat{\pi}) = \begin{cases} 0, & \text{if } \pi_k = \hat{\pi} \text{ (옳은 분류일 경우)} \\ 1, & \text{if } \pi_k \neq \hat{\pi} \text{ (오분류일 경우)} \end{cases}$$

를 고려해 볼 수 있다. 여기서 π_k는 패턴 $\mathbf{x}$의 실제 소속 부류이며 $\hat{\pi}$는 분류기에 의해 할당된 부류이다. 위와 같은 손실함수를 갖는 베이즈 판별규칙은

$$\mathbf{x} \rightarrow \hat{\pi}, \quad \text{if } p(\hat{\pi} \mid \mathbf{x}) = \max_{1 \le j \le g} p(\hat{\pi} = \pi_j \mid \mathbf{x})$$

이다. 여기서 $p(\hat{\pi} \mid \mathbf{x})$는 부류 $\hat{\pi}$의 사후확률을 나타내며 이는 다음과 같은 베이즈 정리에 의해 구해질 수 있다.

$$p(\hat{\pi} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \hat{\pi}) P(\hat{\pi})}{\sum_{j=1}^{g} p(\mathbf{x} \mid \hat{\pi} = \pi_j) P(\hat{\pi} = \pi_j)}. \qquad (1)$$

식(1)의 분모항은 생략될 수 있으므로 베이즈 판별규칙은 부류의 조건부 확률밀도함수 $p(\mathbf{x}|\hat{\pi})$와 사전확률 $P(\hat{\pi})$의 추정방법에 의존하게 된다.

2.1 모수적인 방법

사전확률 $P(\pi_k)$는 일반적으로 훈련집합에서 각 부류의 크기 비율로 쉽게 추정될 수 있는 반면에 각 부류 분포에 대한 정보를 얻기는 매우 힘들다. 그래서 부류 π_k에 대하여 평균벡터 μ_k와 공분산 행렬 Σ_k를 갖는 다변량 정규분포

$$p_k(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma_k|^{-1/2} \times \exp[-1/2(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)]$$

를 가정하여 유도된 판별함수를 가장 많이 사용하고 있다. 위 식을 식(1)에 대입하면 다음과 같은 이차형 판별함수(Quadratic Discriminant Function: QDF)를 얻을 수 있다.

$$d_i(\mathbf{x}) = \min_{1\le k\le g}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k) + \ln|\Sigma_k| - 2\ln P(\pi_k) \quad (2)$$

만약 각 부류에 대하여 공분산 행렬을 Σ로 동일하다고 가정할 수 있다면 식(2)의 이차항은 생략 가능하여 다음과 같은 선형 판별함수(Linear Discriminant Functi-on: LDF)를 얻을 수 있다.

$$d_i(\mathbf{x}) = \min_{1\le k\le g} -2\ \mu_k^t\Sigma^{-1}\mathbf{x} + \mu_k^t\Sigma^{-1}\mu_k - 2\log P(\pi_k).$$

이때 모수 평균벡터 μ_i와 공분산 행렬 Σ에 대한 추정량으로 다음과 최우추정량(MLE)을 많이 사용한다.

$$\hat{\mu}_k = \bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_{j=1}^{n_k}\mathbf{x}_{k,j},$$

$$\hat{\Sigma} = \sum_{k=1}^{g}\frac{n_k}{n}\hat{\Sigma}_k = \frac{1}{n}\sum_{k=1}^{g}\sum_{j=1}^{n_k}(\mathbf{x}_{k,j}-\hat{\mu}_k)(\mathbf{x}_{k,j}-\hat{\mu}_k)^t.$$

여기서 $x_{k,j}$는 부류 π_k의 j번째 패턴을 가리키며, n_k는 부류 π_k의 크기, 그리고 n은 총 훈련집합의 크기를 나타낸다.

그런데 여기서 추정해야할 모수의 수는 LDF에서는 $gp+p(p+1)/2$인 반면, QDF에서는 $gp+gp(p+1)/2$로 매우 많이 증가하게 된다. 따라서 표본의 크기가 작을 때 QDF의 모수 추정량은 매우 불안정하고 큰 분산을 갖게 되므로 QDF가 LDF보다 우수한 성능을 보이기 어렵다[8].

QDF의 단점은 훈련집합의 크기가 작을 때 공분산 행렬 추정량의 정밀도가 매우 낮다는 것이다. 이를 해결하기 위한 정칙화방법은 모수의 추정량으로 불편 추정량을 사용하는 대신 실제로 매우 설득력 있는 편의 추정량을 사용함으로써 추정량의 정밀도를 향상시킨다.

Friedman이 제안한 각 부류의 공분산 행렬 Σ_k에 대한 정칙화는

$$\Sigma_k(\lambda,\gamma) = (1-\gamma)\Sigma_k(\lambda) + \frac{\gamma}{p}tr[\Sigma_k(\lambda)]I \quad (3)$$

이고, 여기서

$$\Sigma_k(\lambda) = \frac{(1-\lambda)n_k\hat{\Sigma}_k + \lambda n\hat{\Sigma}}{(1-\lambda)n_k + \lambda n}.$$

이때 정칙화 모수 λ는 0과 1사이의 값을 취하고, 이는 각 부류의 공분산 행렬에 대하여 합동 공분산 행렬의 기여도를 나타낸다. 그리고 가법 정칙화 모수 γ를 0과 1사이 값으로 적절히 취함으로써 추정량의 편의를 감소시킬 수 있다. 대부분의 경우 λ와 γ값으로 작은 값을 취할 경우 좋은 성능을 얻는다.

QDF에서 각 부류에 대한 공분산 행렬 추정량으로 식(3)을 사용하는 판별함수를 정칙화 판별함수(Regularized Discriminant Function : RDF)라고 한다[4].

2.2 비모수적인 방법

지금까지는 각 부류의 분포를 가정하는 베이즈 판별규칙에 대해 설명하였다. 그러나 실제 문제에서는 각 부류의 분포를 가정할 수 없는 경우가 많다.

Fix & Hodges가 제안한 최근접 이웃 방법[3]은 베이즈 결정규칙에서 사후확률 $p(\pi_i|\mathbf{x})$를 바로 추정하는 방법이다.

$$\mathbf{x}\rightarrow\pi_j, \quad \text{if } k_j = \max_{1\le i\le g}k_i.$$

여기서 k_i는 패턴 $\mathbf{x}$와 근접한 k개의 이웃중 부류 π_i에 소속된 훈련패턴의 개수이다. 이때 우리가 무한 훈련집합을 가정하면 k_i/k는 부류 π_i의 사후확률 $p(\pi_i|\mathbf{x})$의 불편추정량이 된다[6].

$k=1$인 경우 위 규칙을 최근접이웃 결정규칙이라 한다. 최근접 이웃규칙은

$$\mathbf{x}\rightarrow\pi_i, \quad \text{if } k_i = 1,\ k_j = 0,\ j\neq i, j = 1,\cdots,k$$

으로 나타낼 수 있다. 이 방법은 직관적으로 이해하기 쉽고 k-최근접 이웃 결정규칙보다 계산이 훨씬 간단하기 때문에 많이 사용되고 있는 방법이다.

Ⅲ. 신경망을 이용한 분류방법

본 연구에서는 입력층과 출력층 그리고 그 사이에 한 개의 은닉층을 가지는 다층 퍼셉트론(Muti-Layer Perceptron: MLP)을 선택하였다. 입력 유니트는 두 번째 층에 있는 은닉층에 입력값을 분배하는 역할을 한다. 각 유니트들은 그들의 입력값을 합하고 그 값에 편기항(bias)을 더한 다음 그 결과에 은닉층의 활성화 함수 f_h를 취한다. 출력 유니트 역시 동일한 형태지만 활성화 함수로 f_o를 취한다는 점이 다르다. 이를 식으로 표현하면 다음과 같다.

표1. 훈련집합의 크기를 증가시키면서 계산된 각 방법의 인식률

인식함수 \ 표본의 크기	1000	2000	4000	8000	16000	32000	64000
LDF	77.68	82.26	83.99	85.31	87.77	87.96	87.95
QDF	46.37	68.98	83.61	88.93	92.02	92.49	92.84
RDF	73.67	83.00	88.21	90.73	93.11	93.21	93.37
1-NN	80.94	86.00	88.54	89.39	92.30	93.40	94.52
MLP	81.42	85.36	88.17	90.05	93.38	93.65	94.67

$$y_i = f_o(\alpha_i + \sum_h w_{hi} f_h(\alpha_h + \sum_j w_{jh} x_j)). \qquad (4)$$

일반적으로 은닉층과 출력층의 활성함수로 시그모이드 함수가 많이 사용된다.

$$f(x) = \frac{\exp(x)}{1+\exp(x)}$$

식(4)로 표현되는 네트워크를 훈련하기 위해서 우리는 출력값 y_i와 목표값의 차이에 대한 평균제곱합(Mean Square Error: MSE)을 최소로 하도록 가중치를 조정한다. 이때 MSE에 대한 최소값을 찾기 위하여 오류 역전파 학습 알고리즘[2]을 많이 사용한다.

Ⅳ. 실험 및 결과

NIST 데이터베이스 상의 필기 숫자에 대한 인식 실험을 통해 지금까지 설명한 통계적 분류방법 및 신경망 기법의 성능을 비교하였다. 훈련용으로 총 223,125자로 구성된 SD3중 64,000자를, 테스트용으로는 58,646자로 구성된 TEST1을 사용하였다. TEST1은 고등학교 학생들로부터 수집된 것으로, 인간조차 1.5%의 오분류율을 가지는 것으로 알려진 데이터베이스이다.

분류를 위해 128×128 이진 영상에 비선형 방향 기여도 특징추출법을 적용하여 얻은 120차원의 특징벡터를 사용하였으며, 1,000자부터 64,000자까지 훈련집합의 크기를 두배로 증가시켜 가면서 각각의 방법에 대한 인식률을 얻었다. 이 결과는 표1에 제시되어 있다.

표1에서 훈련집합의 크기가 증가하면서 인식률이 향상되고 있음을 알 수 있다. 또한 QDF는 훈련집합의 크기가 작을 경우 다른 방법들과 비교하여 매우 낮은 인식률을 보여주고 있다. 이는 작은 크기의 훈련집합의 크기으로 QDF를 구성하고 있는 각 부류의 추정된 공분산 행렬의 정밀도가 낮기 때문이다. LDF는 전반적으로 비교적 낮은 인식률을 보여주고 있는데 이는 자료가 등분산성이라는 가정을 만족하지 않기 때문이다. QDF와 LDF가 갖는 이러한 단점을 보완하고 있는 RDF의 경우 QDF와 LDF보다 항상 더 높은 인식률을 보여주고 있다. 그러나 각 부류에 대하여 정밀도가 높은 공분산 행렬을 추정할 수 있을 만큼 큰 크기의 훈련집합을 가지고 있을 경우는 QDF가 RDF보다 더 높은 인식률을 보여주게 된다. 이는 정칙화 추정량의 단점인 편의성 때문이다.

표1에서 볼 수 있듯이 MLP가 가장 좋은 인식률을 보여주고 있으며, 1-NN 또한 이와 매우 비슷한 인식률을 보여주고 있다. 이들은 다른 방법들과는 달리 자료에 대한 가정을 거의 가지고 있지 않기 때문이다. 또한 QDF와 RDF는 MLP와 1-NN에 비교하여 약간 낮은 인식률을 가지고 있지만 이들은 각 부류에 대한 사후확률을 추정해주는 장점을 가지고 있다.

Ⅴ. 결론

본 논문에서는 NIST 필기 숫자 데이터 베이스를 사용하여 여러 가지 통계적 방법과 신경망 방법의 성능을 비교하고 그들 각각의 특징을 고찰하였다. 실험결과 MLP가 가장 좋은 인식률을 보여 주었으며, 1-NN 또한 이와 매우 비슷한 인식률을 보여주고 있다. QDF와 RDF가 그들과 비교하여 약간 낮은 인식률을 가지고 있지만 이들은 각 부류에 대한 사후확률을 추정해주는 장점을 가지고 있다.

참고문헌

[1] C. Smith, et al., Handwritten character classification using nearest neighbor in large database, *IEEE T-PAMI*, vol. 16, no. 9, pp. 915-919, 1994.

[2] D. Rumelhart, G. Hitton and R. Williams, Learning representations by backpropagation error, *Nature*, vol. 332, pp. 533-536, 1986.

[3] E. Fix and J. Hodges, Discriminant analysis nonparametric discrimination: consistency properties, *Report no. 4, U.S. Air Force School of Aviation Medicine*, Random Field, Texas, 1951.

[4] H. Friedman, Regularized discriminant analysis, *JASA*, vol. 84, no. 405, pp.165-175. 1989.

[5] J. Blue, G. Candela, et al., Evaluation of pattern classifiers for fingerprint and OCR applications, *Pattern Classification*, vol. 27, no. 4, pp. 485-501, 1994.

[6] J. Kittler, Statistical pattern recognition in image analysis, *Advances in Applied Statistics*, vol 21, pp. 61-75, 1994.

[7] P. Grother, G. Candela, and J. Blue, Fast implementations of nearest neighbor classifiers, *Pattern Recognition*, vol. 30, no. 3, pp. 459-465, 1997.

[8] S. Marks and J. Dunn, Discriminant functions when covariance matrices are unequal. *JASA*, vol. 69, 555-559. 1974.