論文94-31B-12-13

패턴분류기를 위한 최소오차율 학습알고리즘과 예측신경회로망모델에의 적용

(A Minimum-Error-Rate Training Algorithm for Pattern Classifiers and Its Application to the Predictive Neural Network Models)

羅 景 民*, 林 材 烈*, 安 秀 桔*

(KyungMin NA, JaeYeol RHEEM and SouGuil ANN)

要約

대부분의 패턴분류기들은 간단하면서도 비교적 성능이 우수한 ML (Maximum Likelihood) 개념의 학습알고리즘에 의해서 설계된다. ML 개념의 학습알고리즘은 분류기내의 각 계층(class)에 대한 모델들이 서로 독립이라는 가정하에 각 계층의 모델파라메터들을 개별적으로 추정하는 알고리즘이다. 그러나, 실제의 경우 그러한 독립가정이 보장되지 않는 경우가 많고 그에 따른 결정경계의 오차가 인식율을 저하시키는 주된 원인이 된다. 따라서, 본 논문에서는 어떤 입력에 대한 각 계층모델들의 출력을 정규화시켜서 그 입력에 대한 사후확률의 추정치(estimate of the a posteriori probability)로 보고, 그 사후확률의 추정치를 최대화시키는 MAP (Maximum a Posteriori) 개념의 최소오차율 학습알고리즘을 제안한다. Bayes decision theory에 의하면 이 알고리즘은 최소오차율분류를 만족시킨다. 제안된 알고리즘을 예측신경회로망모델의 일종인 NPM (Neural Prediction Model)에 적용시켜 새로운 학습알고리즘을 유도하여 한국어 숫자음에 대한 인식실험을 수행한 결과 기존의 학습알고리즘에서 발생한 오인식 갯수의 37.5 %가 감소되었다.

Abstract

Most pattern classifiers have been designed based on the ML (Maximum Likelihood) training algorithm which is simple and relatively powerful. The ML training is an efficient algorithm to individually estimate the model parameters of each class under the assumption that all class models in a classifier are statistically independent. That assumption, however, is not valid in many real situations, which degrades the performance of the classifier. In this paper, we propose a minimum-error-rate training algorithm based on the MAP (Maximum a Posteriori) approach. The algorithm regards the normalized outputs of the classifier as estimates of the a posteriori probability, and tries to maximize those estimates. According to Bayes decision theory, the proposed algorithm satisfies the condition of minimum-error-rate classification. We apply this algorithm to NPM (Neural Prediction Model) for speech recognition, and derive new discriminative training algorithms. Experimental results on ten Korean digits recognition have shown the reduction of 37.5 % of the number of recognition errors.

* 正會員, 서울大學校 大學院 電子工學科
(Dept. of Elec. Eng., Seoul Nat'l Univ.)

※ 본 연구는 통신개발연구원의 94′ 통신학술연구과제의 지원으로 이루어졌습니다.

接受日字 : 1994年 5月 13日

Ⅰ. 서 론

패턴분류기는 보통 파라메터집합과 결정규칙(decision rule)으로 구성되고 설계표본(design samples)을 이용하여 각 계층(class)에 대한 파라메터들을 추정한다. M개의 계층 C_m, m=1,2,…,M 에 대해서 k차원의 설계표본벡터 x_n이 어느 계층에 속하는지 알고있고, 설계표본벡터의 집합을 $\Omega = \{x_1, x_2, \cdots, x_N\}$라 하면, 최적분류기의 설계는 주어진 설계표본집합 Ω 에 기초해서 최적의 파라메터집합 $\wedge$와 그에 합당한 결정규칙을 찾는 것이 된다.

패턴분류에 있어서 잘 알려진 통계적 방법인 Bayes decision theory에 의하면 주어진 파라메터 집합 $\wedge$ 에 대해서 사후확률(a posteriori probability) $P_{\wedge}(C_m|x)$를 완전히 알고 있을때, 최소오차확률(minimum probability of error)을 얻을 수 있는 Bayes 결정규칙은 식 (1)과 같이 주어진다.[1]

$$C(x) = C_i, \; if \; P_{\wedge}(C_i|x) = \max_j P_{\wedge}(C_j|x) \qquad (1)$$

식 (1)에서 C(·)은 분류연산(classification operation)을 나타낸다. 그러나, 일반적으로 사후확률의 정확한 형태를 알기 어려우므로 Bayes 규칙을 이용해서 식 (1)의 $P_{\wedge}(C_i|x)$ 대신에 사전확률(a priori probability) $P(C_i)$과 계층조건부확률(class-conditional probability) $P_{\wedge}(x|C_i)$ 의 곱을 결정규칙으로 사용한다. 그러면, 사전확률들이 모두 같다는 가정하에서 결정문제는 결국 $P_{\wedge}(x|C_i)$ 에 의해서만 기술되고, 그에 따라서 사후확률 $P_{\wedge}(C_i|x)$ 대신 $P_{\wedge}(x|C_i)$를 추정하여 분류기를 설계하는 것을 ML (Maximum Likelihood) 학습방법이라 한다. 즉, 계층 i의 모델파라메터들을 λ_i 라 하고 각 λ_i가 서로 독립이라고 가정하면 전체파라메터집합은 $\wedge = \{ \lambda_1, \lambda_2, \cdots \lambda_M \}$이 되고 그에 따라서 각 계층의 설계표본들로부터 해당되는 모델파라메터들을 독립적으로 추정할 수 있다. 따라서, ML 개념의 학습시에는 각 계층의 모델파라메터집합들이 서로 독립이라는 가정이 필요하다. 그리고, Bayes decision theory 자체도 분류문제가 확률척도로 기술가능하고, 관련된 확률척도가 파라메터집합 $\wedge$의 함수로서 그 형태가 알려져있으며, 주어지는 설계표본집합에 대해서 그 모델파라메터들을 추정할 수 있는 효과적인 방법이 존재한다는 여러가지 가정들에 기초하고 있다. 그러나, 실제의 분류문제에 있어서 이러한 가정들은 대부분 유효하지 못하다.

그러므로, 확률척도에 대한 대안으로서 분별함수(discriminant functions)를 사용하여 분류기를 설계하는 경우가 많다. 일반적으로 적당한 분별함수의 집합 $g_m(x;\wedge)$, m=1,2,…,M이 주어지면, 결정규칙은 식 (2)와 같이 주어진다.

$$C(x) = C_i, \; if \; g_{\wedge}(x;\wedge) = \max_j g_{\wedge}(x;\wedge) \qquad (2)$$

여기서 최적분류기를 설계하는 문제는 sample risk를 최소화시키는 분별함수의 최적파라메터집합을 구하는 것이다. 그러나, 설계표본을 분류하는데서 발생하는 평균비용으로 정의되는 sample risk가 일반적으로 불연속적인 함수이기 때문에 경사법에 의한 최적화에 어려움이 있으므로 실제로는 다루기 쉬운 다른 스칼라 비용함수를 도입하여 최소화시키는 목적함수로 사용한다. 잘 알려진 스칼라 비용함수로는 perceptron criterion과 summed squared error criterion이 있다.[1] 이러한 스칼라 비용함수를 사용할 때의 문제점은 결정규칙이 함수의 형태로 스칼라 비용함수에 포함되어 있지 않고, 그러한 스칼라 비용함수와 최소오차율을 위한 비용함수가 일치되지 못하다는데 있다. 더우기 기존의 스칼라 비용함수를 이용한 학습은 ML 개념의 학습이어서 각 계층간의 결정경계(decision boundary)를 최적화시킬 수 없다는 문제점도 남아있다.

따라서 본 논문에서는 식 (2)의 결정규칙을 포함하는 새로운 결정규칙을 정의하여 식 (1)의 결정규칙을 근사하도록함으로서 최소오차율분류에 접근하고, 그러한 결정규칙이 함수의 형태로 포함되는 비용함수를 도입하는 MAP(Maximum a Posteriori) 개념의 최소오차율 학습알고리즘을 제안한다. 분류기의 각 계층에 대한 출력들을 정규화시켜 새로운 결정규칙으로 사용하면, 그 정규화된 양을 사후확률추정치로 볼 수 있으므로 식 (1)과 같은 결정규칙을 근사하도록 할 수 있다. 그러한 사후확률의 추정치들의 곱으로 전체 비용함수를 정의하면 식 (2)의 결정규칙이 비용함수에 함수의 형태로 포함됨을 알 수 있다. 이제 새로 정의된 비용함수를 최대로 하는 학습알고리즘을 유도하면 최적의 분류기를 얻을 수 있다. 경사법을 사용하기 위해서 모든 함수는 파라메터집합에 대해서 연속이 되도록 정의한다. 새로운 비용함수가 각 설계표본에서 발생하는 사후확률의 추정치들의 곱으로 정의되어 기존의 경사법을 적용하기 어려우므로 S. Amari의 확률적강하법(probabilistic descent method)을 적용한다.[2] 그러나, 전통적인 추정이론에서처럼 이 비용함수에 로그를 취하고 음의 부호를 붙이면 곱이 합으로 바뀌어 기존의 경사강하법을 사용할 수 있다. 그러한 두 비용함수에 대해서 각각 학

습알고리즘을 유도할 수 있으며, 수식적으로나 실험적으로 후자가 더 우수함을 알 수 있다. 자세한 것은 다음 장에서 다루어질 것이다. 제안하는 알고리즘은 분별함수에 의한 분류기를 확률적 해석에 기초한 최소오차율분류에 적합하도록 설계할 수 있게 하고 정적인 패턴분류기와 동적인 패턴분류기에 모두 적용이 가능하다는 점에서 그 적용범위가 넓다는 장점이 있다.

Ⅱ. 최소오차율 학습알고리즘

1. 최소오차율분류(Minimum-error-rate Classification)

어떤 입력 x의 실제 계층이 C_j 인 경우에 그 입력을 C_i 로 분류하는 어떤 행위를 α_i 라고 정의하고, α_i 라는 행위를 함으로써 발생하는 손실(loss)을 $\lambda(\alpha_i|C_j)$ 라 정의하자. 즉, 어떤 x를 관찰하고 α_i 라는 행위를 했는데 실제로는 x가 C_j 에 포함되면, $\lambda(\alpha_i|C_j)$의 손실이 일어났다고 할 수 있다. 그러면, conditional risk라고 알려진 손실의 기대치 $R(\alpha_i|x)$는 다음과 같이 정의된다.

$$R(\alpha_i|x)=\sum_{j=1}^{M}\lambda(\alpha_i|C_j)P_\wedge(C_j|x) \tag{3}$$

최소오차율분류를 위해서는 특별히 대칭손실함수 혹은 one-zero 손실함수라고 부르는 다음과 같은 손실함수가 사용된다.

$$\lambda(\alpha_i|C_j)=\begin{cases}0 & i=j\\ & i,j=1,2,\cdots,M\\ 1 & i\neq j\end{cases} \tag{4}$$

식 (4)의 손실함수는 정확한 결정에 대해서는 아무런 손실도 주지 않고, 틀린 결정에 대해서는 1의 손실을 준다. 이 손실함수에 대한 risk는 다음과 같이 평균오차확률(average probability of error)이 된다.

$$\begin{aligned}R(\alpha_i|x)&=\sum_{j=1}^{M}\lambda(\alpha_i|C_j)P_\wedge(C_j|x)\\ &=1-P_\wedge(C_i|x)\end{aligned} \tag{5}$$

Bayes 결정규칙은 식 (5)의 conditional risk를 최소화시키는 것이다. 따라서, 최소오차율분류를 위해서는 사후확률 $P_\wedge(C_i|x)$ 를 최대화시키는 계층 C_i 를 택해야 한다.[1]

2. 최소오차율 학습알고리즘의 유도

분별함수를 기반으로하는 분류기의 출력은 확률값이 아니므로 위와 같은 최소오차율분류가 불가능하다. 그러나, 분류기의 출력들을 정규화시켜서 식 (1)의 사후확률의 추정치로 사용하면 근사적으로 최소오차율분류를 위한 학습알고리즘을 얻을 수 있다. 패턴분류기를 사후확률추정기로 보고 어떤 입력패턴에 대해 그에 대응하는 계층의 사후확률을 최대로하도록 학습시키는 것이다.

그러므로, 먼저 분류기의 출력을 효과적으로 정규화시키는 과정이 필요하다. 이 과정은 여러가지 방법으로 수식화가 가능한데 예를 들어서 출력이 항상 영보다 크거나 같다면 다음과 같은 수식화로 가능하다. scaling factor를 α 라 하면,

$$f_m(x;\wedge)=\frac{(g_m(x;\wedge))^\alpha}{\sum_{i=1}^{M}(g_i(x;\wedge))^\alpha} \tag{6}$$

분류기의 출력의 부호에 상관 없이 사용하려면 다음과 같은 정규화된 지수형태인 "softmax" 함수가 적절하다.[3]

$$f_m(x;\wedge)=\frac{\exp(\alpha\, g_m(x;\wedge))}{\sum_{i=1}^{M}\exp(\alpha\, g_i(x;\wedge))} \tag{7}$$

각각의 설계표본들이 서로 독립이라고 가정하면, 정규화 과정을 거쳐서 다음과 같은 전체 비용함수를 정의할 수 있다.

$$L(\wedge)=\prod_{i=1}^{M}\prod_{x\in C_i}f_i(x;\wedge) \tag{8}$$

식 (8)의 비용함수를 최대화시키는 것이 설계의 목표이다. 그러나, 비용함수가 식 (7)의 곱으로 표현되어 있으므로 일반적인 경사법을 적용하기 어렵다. 따라서, 확률적강하법을 적용하여 분류기를 설계한다. 이 방법에 의하면 개별적인 사후확률추정치인 $f_m(x;\wedge)$들에 대해서만 강하법을 적용한다.[12]

$$\wedge_{t+1}=\wedge_t-\eta_t U\nabla(-f_i(x;\wedge)) \tag{9}$$

학습계수 η_t 는 $\sum_{t=1}^{\infty}\eta_t\to\infty$ 와 $\sum_{t=1}^{\infty}\eta_t^2<\infty$ 를 만족하도록 시간에 따라 줄어나간다. U는 positive-definite 행렬로서 본 논문에서는 항등행렬을 사용했다. ∇는 gradient 연산을 나타낸다. 또한, 강하법은 비용함수를 최소화시키는 방법이므로 $f_i(x;\wedge)$앞에 (-) 부호를 붙여서 수식화했는데 그것은 경사상승법과 같은 효과를 내서 식 (8)을 최대화시키는 목적에 합당하다.

또한, 기존의 강하법을 사용하기 위해서 식 (8)의 자연로그를 취하고 음수부호를 붙여서 새로운 비용함

수를 정의하면 다음과 같다.

$$L'(\wedge) = -\ln L(\wedge) = -\sum_{i=1}^{M}\sum_{x\in C_i} \ln f_i(x;\wedge) \qquad (10)$$

식 (10)에 경사강하법을 적용하면 다음과 같은 학습식을 얻을 수 있다.

$$\nabla L'(\wedge) = \sum_{i=1}^{M}\sum_{x\in C_i} \nabla(-\ln f_i(x;\wedge)) \qquad (11)$$

$$\wedge_{t+1} = \wedge_t - \eta\,\nabla(-\ln f_i(x;\wedge)) \qquad (12)$$

제안하는 알고리즘의 특성을 더 자세히 파악하기 위해서 식 (7)을 사용하여 구체적인 학습알고리즘을 유도하겠다. 다른 정규화 방법에 대해서도 쉽게 유도할 수 있으므로 분류기의 특성에 따라서 설계자가 사용할 식을 결정하면 된다. 먼저 식 (9)에 의한 학습알고리즘은 다음과 같다.

$$(\lambda_m)_{t+1} = (\lambda_m)_t + \eta_t \alpha\, f_m(1.0 - f_m)\frac{\partial\, g_m(x;\wedge)}{\partial \lambda_m} \text{ for } x \in C_m, \qquad (13.a)$$

$$(\lambda_l)_{t+1} = (\lambda_l)_t + \eta_t \alpha\, f_m f_l \frac{\partial\, g_l(x;\wedge)}{\partial \lambda_l} \text{ for } all\ l \neq m. \qquad (13.b)$$

또한, 식 (12)에 의한 학습알고리즘도 다음과 같이 쉽게 유도할 수 있다.

$$(\lambda_m)_{t+1} = (\lambda_m)_t + \eta\,\alpha\,(1.0 - f_m)\frac{\partial\, g_m(x;\wedge)}{\partial \lambda_m} \text{ for } x \in C_m, \qquad (14.a)$$

$$(\lambda_l)_{t+1} = (\lambda_l)_t - \eta\alpha\, f_l \frac{\partial\, g_l(x;\wedge)}{\partial \lambda_l} \text{ for } all\ l \neq m. \qquad (14.b)$$

편의상 $f_i = f_i(x;\wedge)$이라고 표현했다. 식 (13)과 (14)의 기본적인 차이는 학습에 참여하는 비용에 있다. 식 (13)에서는 학습에 참여하는 비용이 $f_m(1.0-f_m)$과 $f_m f_l$ 인데 비하여 식 (14)에서는 각각 $(1.0-f_m)$과 f_l 이다. 학습에 의해서 f_m이 1에 가까워지면 질수록 f_l 은 0에 가까워진다. 학습 초기에는 같은 f_m과 f_l 에 대해서 식 (13)의 비용보다 식 (14)의 비용이 크므로 비용함수가 빨리 줄어들고, 학습이 어느정도 진행된 후에는 $(1.0-f_m)$과 f_l 이 모두 0에 가까워짐으로 적어도 국부적으로 최적인 해를 얻을 수 있다.[12-15] 따라서, 식 (14)의 학습알고리즘이 식 (13)보다 빨리 수렴하고 두 알고리즘이 모두 최소오차율 분류에 근사되는 학습알고리즘임을 알 수 있다.

Ⅲ. 예측신경회로망모델에의 적용

1. 예측신경회로망모델

최근에 다양한 음성인식문제에 있어서 우수한 성능을 보이는 예측신경회로망모델(predictive neural network models)과 그 학습알고리즘들이 제안되었다.[3,4,6] 이 모델은 다층퍼셉트론(multilayer perceptron: MLP)을 연속되는 음성특징벡터간의 비선형예측기로 사용하고, 여러 개의 MLP 예측기의 열로 하나의 단어단위(word unit)를 구성하는 동적인 인식모델이다. 한 단어단위 내의 MLP 간의 천이는 예측오차행렬에 대한 동적프로그래밍 기법(dynamic programming technique)에 의해서 결정되고, 결정된 천이경로를 따라서 각 MLP 예측기들은 잘 알려진 오차역전파 학습알고리즘(error backpropagation algorithm)으로 학습된다. 인식시에는 입력을 각 단어단위의 모델에 가하여 그에 대응되는 예측오차행렬들을 구하고 각 행렬에 동적프로그래밍 기법을 사용하여 대응되는 최소누적예측오차들을 구한후 그 중에서 가장 작은 값을 출력하는 단어단위를 인식결과로 한다.

예측신경회로망모델에 관한 연구는 크게 K. Iso 등이 제안한 NPM (Neural Prediction Model)[4], J. Tebelskis 등이 제안한 LPNN (Linked Predictive Neural Network)[5], E. Levin의 HCNN (Hidden Control Neural Network)[6] 등으로 나눌 수 있다. NPM과 LPNN이 음성의 시변성을 MLP 예측기 자체의 천이로 모델링하는데 반하여 HCNN은 입력에 추가된 제어신호의 변화로 모델링한다. 그러나, 기본적으로 모든 모델들의 학습은 오차역전파학습과 동적프로그래밍 기법의 결합으로 이루어진다.

DTW (Dynamic Time Warping), HMM (Hidden Markov Models), TDNN (Time-Delay Neural Networks) 등과 같은 기존의 음성인식모델들에 비해서 예측신경회로망모델은 다음과 같은 점에서 우수하다. (1) 연속되는 음성특징벡터간의 일시적 상관관계가 인식에 이용된다. (2) 발성시간의 차이에 따른 음성의 왜곡이 효과적으로 흡수된다. (3) 학습에 필요한 데이타양이 상대적으로 적다. (4) 연속음성인식으로의 확장이 용이하다. (5) 새로운 계층의 추가시 전체 모델을 재학습시킬 필요가 없다.

그러나, 기존의 학습알고리즘에 의해 학습된 모델은 음운학적으로 유사한 음성구간에서 변별력이 떨어지는 문제점이 있다. 그것은 기존의 학습알고리즘이 각 계층의 모델들을 학습시킬때 대응되는 계층의 학습데이타들로만 학습시키는 ML (Maximum Likelihood) 방법이기 때문이다. J. Tebelskis 등은 이러한 문제점을 지적하고 변별력있는 학습알고리즘

개발의 필요성을 언급했다.[5] E. Levin도 기본적인 ML 방법 대신에 Bayesian approach의 필요성을 제기했다.[6] 또한, 최근에 이러한 문제점을 개선하기 위한 몇가지 변별력있는 학습알고리즘들이 제안되었다.[7-12] H. Jun과 H. Leich는 barrier function을 도입하여 어떤 입력에 대해서 그 입력이 해당되는 계층의 모델은 누적예측오차를 최소화시키도록 학습시키고 동시에 다른 계층의 모델들은 누적예측오차를 최대화시키도록 학습시키는 알고리즘을 제안했다.[7] Y. Liu 등은 확률척도를 도입하여 단층 퍼셉트론에 기초한 예측신경회로망모델의 변별력있는 학습알고리즘을 제안했다.[8] A. Mellouk과 P. Gallinari는 비용(costs)과 정정규칙을 도입한 학습알고리즘을 제안했다.[9] K. Iso는 주어진 제어명령열에 대한 음운학적 관찰확률을 Gauss 분포로 모델링하고 그에 기초한 MAP(Maximum a Posteriori) 학습알고리즘을 제안했다.[10] B. Petek과 A. Ferligoj는 실험에 의해서 예측오차가 백색잡음이 아니라는 것을 보이고, 예측오차의 중요한 성분을 예측하는 HCNN을 추가하여 변별력을 높였다.[11] 나 경민 등은 최소분류오차 수식화와 GPD (Generalized Probabilistic Descent) 알고리즘에 기초한 변별력있는 학습알고리즘을 유도했다.[12]

본 논문에서는 예측신경회로망의 일종인 NPM에 대해서 변별력을 높이기 위한 최소오차율 학습알고리즘을 유도하고 한국어숫자음인식에 대한 실험을 실시하였다. 그러나, 유도된 알고리즘이 변형된 오차역전파 학습알고리즘으로 해석될 수 있으므로 다른 예측신경회로망모델들에도 쉽게 적용할 수 있다.

2. NPM과 기존의 학습알고리즘

NPM은 MLP를 음성의 비선형예측기로 사용하고, 그러한 MLP의 순서열(ordered sequence)로 하나의 단어단위를 모델링하는 동적인 인식모델이다. 시간 t 에서의 음성특징벡터 S_t에 대하여 τ 시간 이전까지의 특징벡터들인 $S_{t-\tau}, \cdots, S_{t-1}$ 들의 비선형결합으로 V를 계산하고, 예측오차 $\|\hat{S}_t - S_t\|^2$들로 부터 예측오차행렬을 얻는다. 그리고, 예측오차행렬에 동적프로그래밍 기법과 backtracking 기법을 적용시켜서 최적분할경로(optimal segmentation path)를 찾고, 그 경로를 따라서 해당되는 MLP 예측기들을 학습시킨다. 수식의 정리를 위해서 M개의 계층 C^m, m=1,2,…,M, N개의 3층-MLP 예측기 $1 \le n(t) \le N$, 최적분할경로 (t,n(t))가 주어지고, 계층 C^m을 위한 모델에 있어서 최적분할경로상의 n(t)번째 MLP 예측기의 입력-은닉층간의 가중치를 $V = (V^m_{ij,n(t)})$, 은닉출력층간의 가중치를 $W = (W^m_{jk,n(t)})$, 은닉층의 출력을 $H^m_{j,t}$, $S_{t-\tau}, S_{t-\tau+1}, \cdots, S_{t-1}$ 로 구성된 입력벡터의 i번째 성분을 $\bar{s}_{i,t}$, S_t 의 k번째 성분을 $s_{k,t}$, $\hat{S}_t$ 의 k번째 성분을 $\hat{s}^m_{k,t}$라 하면 누적예측오차 D(m), 각 층의 출력 그리고 가중치의 학습량은 각각 (15), (16), (17)과 같다.

$$D(m) = \min_{n(t)} \frac{1}{2} \sum_{t=1}^{T} \left\| \hat{S}_t - S_t \right\|^2 \tag{15}$$

$$H^m_{j,t} = f\left(\sum_i \bar{s}_{i,t} V^m_{ij,n(t)} \right) \tag{16.a}$$

$$\hat{s}^m_{k,t} = \sum_j H^m_{j,t} W^m_{jk,n(t)} \tag{16.b}$$

$$\delta W^m_{jk,n(t)} = \eta \left(s_{k,t} - \hat{s}^m_{k,t} \right) \tag{17.a}$$

$$\delta V^m_{ij,n(t)} = \eta \delta^m_{j,t} H^m_{j,t} \left(1.0 - H^m_{j,t} \right) \bar{s}_{i,t} \tag{17.b}$$

$f(\cdot)$은 sigmoid 함수이고, $\delta^m_{j,t}$는 $\delta^m_{j,t} = \sum_k \left(s_{k,t} - \hat{s}^m_{k,t} \right) W^m_{jk,n(t)}$ 이다.

3. NPM의 최소오차율 학습알고리즘의 유도

먼저 NPM을 이용한 분류기의 분별함수는 다음과 같이 정의된다.

$$g_i(x;\wedge) = -\min_{n(t)} \sum_{t=1}^{T} D^{\theta_i}(t,n(t)), \tag{18}$$

$$D^{\theta_i}(t,n(t)) = \frac{1}{2} \sum_{k=1}^{K} \left(s_{k,t} - \hat{s}'_{k,t} \right)^2. \tag{19}$$

식 (7)과 식 (8)로 부터 다음의 학습식을 유도할 수 있다.

$$\delta W^m_{jk,n(t)} = \eta_i \alpha f_m \left(1.0 - f_m \right) \left(s_{k,t} - \hat{s}^m_{k,t} \right) H^m_{j,t} \quad \text{for } S \in C_m, \tag{20.a}$$

$$\delta V^m_{ij,n(t)} = \eta_i \alpha f_m \left(1.0 - f_m \right) \delta^m_{j,t} H^m_{j,t} \left(1.0 - H^m_{j,t} \right) \bar{s}_{i,t} \quad \text{for } S \in C_m, \tag{20.b}$$

$$\delta W^l_{jk,n(t)} = -\eta_i \alpha f_m f_l \left(s_{k,t} - \hat{s}^l_{k,t} \right) H^l_{j,t} \quad \text{for all } l \neq m, \tag{20.c}$$

$$\delta V^l_{ij,n(t)} = -\eta_i \alpha f_m f_l \delta^l_{j,t} H^l_{j,t} \left(1.0 - H^l_{j,t} \right) \bar{s}_{i,t} \quad \text{for all } l \neq m. \tag{20.d}$$

또, 식 (7)과 식 (9)로 부터 또 하나의 학습식을 유도할 수 있다.

$$\delta W^m_{jk,n(t)} = \eta \alpha \left(1.0 - f_m \right) \left(s_{k,t} - \hat{s}^m_{k,t} \right) H^m_{j,t} \quad \text{for } S \in C_m, \tag{21.a}$$

$$\delta V^m_{ij,n(t)} = \eta \alpha \left(1.0 - f_m \right) \delta^m_{j,t} H^m_{j,t} \left(1.0 - H^m_{j,t} \right) \bar{s}_{i,t} \quad \text{for } S \in C_m, \tag{21.b}$$

$$\delta W^l_{jk,n(t)} = -\eta_i \alpha f_l \left(s_{k,t} - \hat{s}^l_{k,t} \right) H^l_{j,t} \quad \text{for all } l \neq m. \tag{21.c}$$

$$\delta V^{l}_{ij,n+(t)} = -\eta\alpha\, f_i \delta^{\,l}_{j,t} H^{l}_{j,t} \left(1.0 - H^{l}_{j,t}\right) \bar{s}_{i,t} \quad \text{for all } l \neq m. \quad (21.d)$$

위의 식 (20)과 식 (21)은 결과적으로 바른 계층의 모델파라메터들에 대해서는 경사강하법을 적용시키고, 동시에 그외의 계층 파라메터들에 대해서는 경사상승법을 적용시킴을 알 수 있다.

Ⅳ. 실험결과

III 장에서 유도된 식 (20)과 식 (21)들을 적용해서 "공"을 포함하는 한국어 숫자음 10 개에 대한 인식 실험을 수행하였다. 음성신호의 전처리과정은 다음과 같다. 총 20 명의 남성화자의 1 회 발성 음성시료를 10 kHz로 표본화하고 Rabiner와 Sambur의 끝점 검출 알고리즘으로 끝점을 검출한다. 그리고 128 표본씩 중첩시키면서 256 표본씩 pre-emphasis, Hamming 창, 그리고 12 차의 LPC 분석을 거쳐서 0 차 켑스트럼을 제외한 12 차의 가중켑스트럼 계수를 얻어서 입력으로 사용했다. 8 명의 데이타로 학습을 시키고 나머지 12 명의 데이타로 화자독립 인식실험을 수행했다. NPM은 2차의 예측기를 사용했으며, 단어당 MLP 예측기의 수는 Iso의 방법대로 단어들의 평균 길이의 절반 정도인 12 개로 정했다.

$$L(W,V) = \sum_{t=0}^{9} \sum_{x \in C} f_m(x,W,V) \qquad (22)$$

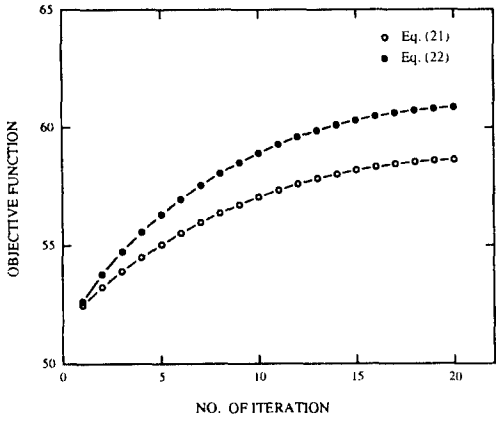

그림 1. 목적함수의 최대화
Fig. 1. Maximization of objective function.

먼저 기존의 알고리즘에 의해서 학습계수 0.01, 학습횟수 1000 회의 학습을 시키고, 그 후 α = 1.0, 학습계수 0.0002로 식 (20)과 식 (21)을 이용하여 각각 20 회의 학습을 시켰다. 기존의 알고리즘에 대해서는 약 93.3 %(112/120)의 인식율을, 식 (20)과 (21)에 의한 학습결과로는 약 95.8 %(115/120)의 인식율을 보여서 제안한 알고리즘이 기존의 알고리즘을 사용했을때 발생한 오인식의 갯수를 약 37.5 % 정도 감소시켰다. 식 (20)과 식 (21)의 최종적인 인식결과는 같았으나, 식 (21)에 의한 학습이 식 (20)에 의한 학습보다 빨리 전체 비용함수를 최대화시키고 최적의 해에 도달함을 알 수 있었다. 그림 1과 그림 2에 그에 대한 결과가 나타나 있다. 이해하기 쉽게 식 (8)과 (10) 대신에 다음의 식을 그림 1의 목적함수로 사용했다.

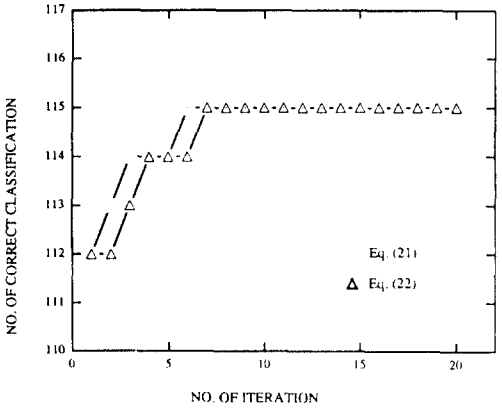

그림 2. 오인식 갯수의 감소
Fig. 2. Reduction of the number of recognition errors.

Ⅴ. 결 론

본 논문에서는 일반적인 패턴 분류기의 설계에 적용될 수 있는 최소오차율 분류에 근사하는 MAP 개념의 학습알고리즘을 유도하고 음성인식을 위한 예측신경회로망모델에 적용하여 실험적으로 제안한 알고리즘의 타당성을 보였다. 특히, 같은 이론적 배경에서 유도된 식 (20)과 식 (21)의 차이점을 분석한 결과 식 (21)이 더욱 빨리 국부적인 최적해로 수렴할 것이라는 것을 알 수 있었고 실험결과도 그러한 분석을 뒷받침해준다.

제안된 알고리즘은 ML 개념으로 설계된 분류기의 결정경계를 설계표본 내에서 최적화시키는 알고리즘으로 볼 수 있다. 대부분의 경우 각 계층들의 결정경

계가 겹쳐져 있는데 ML 개념으로 학습하면 최적인 경계를 얻을 수 없고 그에 따라서 오인식되는 결과가 증가하는 것이다. 제안하는 알고리즘도 본래부터 겹쳐져 있는 경계내에서 발생하는 고유의 오인식까지 줄일 수는 없으나 최적의 결정경계로 경계를 이동시킴으로써 결정경계의 오차로 인한 오인식을 최소한으로 줄이는 것이다.

결과적으로 제안하는 학습알고리즘은 바른 계층의 모델파라메터들에 대해서는 그에 합당한 비용이 가중된 경사강하법을 적용시키고, 동시에 그외의 계층파라메터들에 대해서는 각각에 합당한 비용이 가중된 경사상승법을 적용시킴을 알 수 있다. 그에 따라서 각 모델간의 경계가 최적화되고 변별력이 향상됨을 알 수 있다.

이런 종류의 다른 알고리즘들에서처럼 제안된 알고리즘도 학습에 소요되는 시간이 기존 방법보다 길다는 문제점이 있다. 실제로 분류하고자하는 계층의 수가 N개라면 학습시간도 N배 정도 늘어난다. 또한, scaling factor α도 분류기 출력의 동적인 범위를 결정하여 학습에 영향을 미치므로 분류기의 목적에 맞도록 잘 선택해야 한다.

제안된 알고리즘은 정적인 패턴과 동적인 패턴의 분류기에 모두 적용할 수 있다. 앞으로 DTW나 HMM 같은 다른 음성인식모델이나 정적인 패턴을 분류하는 응용분야에 제안된 알고리즘을 적용하는 연구가 계속될 것이다. 또한, 분류기의 출력을 각 분류기의 특성에 맞게 정규화시키는 연구도 진행될 것이다.

參 考 文 獻

[1] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. New York: Wiley. 1973.

[2] S. Amari. "A theory of adaptive pattern classifiers," *IEEE Trans. Electronic Computers*. vol. EC-16. no. 3. 1967.

[3] H. Bourlard and C. J. Wellekens. "Links between Markov models and multilayer perceptrons," *IEEE Trans. Pattern Analysis and Machine Intelligence*. vol. 12. no. 12. pp. 1167-1178. 1990.

[4] K. Iso and T. Watanabe. "Large vocabulary speech recognition using neural prediction model," *Proc. ICASSP-91*. pp. 57-60. 1991.

[5] J. Tebelskis, A. Waibel, B. Petek and O. Schmidbauer. "Continuous speech recognition using linked predictive neural network," *Proc. ICASSP-91*. pp. 61-64. 1991.

[6] E. Levin. "Hidden control neural architecture modeling of nonlinear time varying systems and its applications," *IEEE Trans. Neural Networks*. vol. 4. no. 1. pp. 109-116. 1993.

[7] H. Jun and H. Leich. "A discriminative training algorithm for speech recognizer based on the neural prediction model," *Proc. EUSIPCO-92*. pp. 423-426. 1992.

[8] Y. D. Liu. Y. C. Lee. H. H. Chen and G. Z. Sun. "Discriminative training algorithm for predictive neural network models," *Proc. IJCNN-92*. pp. 685-690. 1992.

[9] A. Mellouk and P. Gallinari. "A discriminative neural prediction system for speech recognition," *Proc. ICASSP-93*. pp. 533-536. 1993.

[10] K. Iso. "Speech recognition using dynamical model of speech production," *Proc. ICASSP-93*. pp. 283-286. 1993.

[11] B. Petek and A. Ferligoj. "Exploiting prediction error in a predictive-based connectionist speech recognition system," *Proc. ICASSP-93*. pp. 267-269. 1993.

[12] K. M. Na. J. Y. Rheem and S. G. Ann. "A discriminative training algorithm for predictive neural network models," will appear in *Proc. ISCAS-94*. May 1994. London. UK.

[13] P. C. Chang and B. H. Juang. "Discriminative training of dynamic programming based speech recognizer," *IEEE Trans. Speech and Audio Processing*. vol. 1. no. 2. pp. 135-143. 1993.

[14] B. H. Juang and S. Katagiri. "Discriminative learning for minimum error

classification," *IEEE Trans. Signal Processing*, pp. 3043-3054, 1992.

[15] P. C. Chang, S. H. Chen and B. H. Juang, "Discriminative analysis of distortion sequences in speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 3, pp. 326-333, 1993.

著者紹介

羅 景 民(正會員)

1968年 3月 5日生. 1990年 2月 서울대학교 전자공학과 졸업. 1994年 2月 서울대학교 대학원 전자공학과 공학석사. 1994年 3月 서울대학교 대학원 전자공학과 박사과정. 주관심 분야는 음성인식, 패턴분류, 신경회로망, 유전자 알고리듬 등임.

林 材 烈(正會員) 第 31卷 第 2號 參照

•

安 秀 桔(正會員) 第 31卷 第 2號 參照