Pergamon

Neural Networks

Neural Networks 11 (1998) 209–213

Contributed Article

Multilayer neural networks and Bayes decision theory

Ken-ichi Funahashi*

Center for Mathematical Sciences, The University of Aizu, Aizu-Wakamatsu, Fukushima, 965 Japan

Received 9 November 1994; accepted 22 September 1997

Abstract

There are many applications of multilayer neural networks to pattern classification problems in the engineering field. Recently, it has been shown that Bayes a posteriori probability can be estimated by feedforward neural networks through computer simulation. In this paper, Bayes decision theory is combined with the approximation theory on three-layer neural networks, and the two-category *n*-dimensional Gaussian classification problem is studied. First, we prove theoretically that three-layer neural networks with at least 2*n* hidden units have the capability of approximating the a posteriori probability in the two-category classification problem with arbitrary accuracy. Second, we prove that the input–output function of neural networks with at least 2*n* hidden units tends to the a posteriori probability as Back-Propagation learning proceeds ideally. These results provide a theoretical basis for the study of pattern classification by computer simulation.

Keywords: Multilayer neural network; Bayes decision theory; Back-propagation algorithm; Bayes discriminant function; Gaussian probability density; A posteriori probability

1. Introduction

There are many studies of pattern classification problems using feedforward neural networks in the engineering field. Therefore, it is interesting to investigate the internal representation of the hidden layer of neural networks after learning by the Back-Propagation (B-P) algorithm (cf., Rumelhart et al., 1986). At present, there is a result by Ruck et al. (1990) that the statistical error between the input–output mapping of the neural network and the a posteriori probability densities of the classes decreases as the mean-squared error in the B-P algorithm decreases, when learning data are infinite (see also White, 1989). In connection with this result, Richard, and Lippmann (1991) show that a three-layer network can be used to estimate the a posteriori probability in the case of a two-category one-dimensional Gaussian classification problem by computer simulation. We study theoretically the following problems from a statistical viewpoint:

1. What is the input–output mapping of the feedforward neural network when it finished its learning in a pattern classification problem?
2. What relationship is there between the input–output mapping of the feedforward neural network and the corresponding a posteriori probabilities of the classes?

To solve these problems, we combine the approximation theory on three-layer neural networks introduced by Funahashi (1989) and the Bayes decision theory (cf., Duda and Hart, 1973) which is a statistical theory of pattern classification. In this paper we consider the case of the two-category pattern classification problem, in which the probability density functions of the classes are *n*-dimensional Gaussian, and we prove that the input–output function of the neural network with at least 2*n* hidden units can approximate the a posteriori probability in the statistical sense. Moreover, when we have infinite learning samples (i.e., in the theoretical setting), the input–output function tends to the a posteriori probability density when the B-P learning proceeds ideally.

2. Feedforward neural networks and a posteriori probabilities

In this section, we review the relationship between pattern classification using feedforward neural networks and a posteriori probabilities which is presented in Ruck et al. (1990). We denote R^n the *n*-dimensional Euclidean space.

* Requests for reprints should be sent to Dr K. Funahashi. Tel.: 00 81 0242 37 2719; Fax: 00 81 0242 37 2752; E-mail: funahashi@u-aizu.ac.jp.

0893–6080/98/$19.00
PII S0893-6080(97)00120-2

Let $x \in R^n$ be the input pattern and we define $F(x, w)$ to be the output of a feedforward neural network where w is the weight vector. In this paper, we consider the two-category pattern classification problem. Consider the input pattern x to be a random variable and let $p(x)$ be its probability density function. Let the probabilities of occurrence ω_1 and ω_2 be $p(\omega_1)$ and $p(\omega_2)$, respectively; these are the a priori probabilities. Let $p(x|\omega_i)$ $(i = 1,2)$ be the state-conditional probability density function for x. The conditional densities $p(\omega_i|x)$ are the a posteriori probabilities that a given pattern x belongs to class ω_i. We define X_1 and X_2 to be the finite patterns which are from ω_1 and ω_2, respectively. For simplicity, we only consider in this paper three-layer neural networks with one output unit. We assume that as teacher signals for the neural network, the value 1 is given when the input pattern x is from class ω_1 and the value 0 is given when the input pattern is from class ω_2. Under the above assumption, we define the mean-squared error function as follows:

$$E_s(\omega) = \sum_{x \in X_1} (F(x, w) - 1)^2 + \sum_{x \in X_2} F(x, w)^2.$$

If we assume that we have infinite learning samples, the mean-squared error is given by

$$E_a(w) = p(\omega_1) \int_{R^n} (F(x, w) - 1)^2 p(x|\omega_1)\, dx + p(\omega_2) \times \int_{R^n} F(x, w)^2 p(x|\omega_2)\, dx.$$

Using the Bayes formula

$$p(x|\omega_1) = \frac{p(\omega_1|x)p(x)}{p(\omega_1)},$$

where $p(x) = p(x|\omega_1)p(\omega_1) + p(x|\omega_2)p(\omega_2)$, Ruck et al. (1990) obtained the following formula

$$E_a(w) = e^2(w) + \int_{R^n} p(\omega_1|x)(1 - p(\omega_1|x))p(x)\, dx,$$

where

$$e^2(w) = \int_{R^n} [F(x, w) - p(\omega_1|x)]^2 p(x)\, dx.$$

The B-P learning algorithm modifies the weight vector w so that the mean-squared error $E_a(w)$ decreases. As the second term of $E_a(w)$ is independent of w, $e^2(w)$ decreases as the learning proceeds. That is, the statistical error between the network output $F(x,w)$ and the a posteriori probability $p(\omega_1|x)$ decreases. However, we do not know whether the error tends to zero or not.

3. Bayes decision theory

In this section, we briefly review Bayes decision theory (cf., Duda and Hart, 1973). The Bayes decision is a method that assigns discriminant functions $g_i(x)$ $(i = 1,2,...,r)$ to classes ω_1 $(i = 1,...r)$, respectively. For a pattern x, we decide x belongs to ω_1 if $g_i(x) > g_j(x)$ (for all $j \neq i$). As the discriminant functions $g_i(x)$ $(i = 1,...,r)$, we usually use $g_i(x) = p(\omega_1|x)$: a posteriori probability density for x. We can use $f(p\omega_i|x)) + h(x)$ as $g_i(x)$ where f is any monotone increasing function, which leaves the resulting classification unchanged. For example, from the Bayes formula, we obtain

$$\log p(\omega_i|x) = \log p(x|\omega_1) + \log p(\omega_i) - \log p(x).$$

Hence, we can use $g_i(x) = \log p(x|\omega_i) + \log p(\omega_i)$ $(i = 1,...,r)$ as discriminant functions. Especially in the two-category case, we may use $g(x) = g_1(x) - g_2(x)$ as a discriminant function. Then we decide x belongs to ω_1 if $g(x) > 0$ and x belongs to ω_2 if $g(x) < 0$. To apply the Bayes decision method, one needs to know

1. a priori probabilities $p(\omega_i)$ $(i = 1,...,r)$ and
2. a posteriori probability densities $p(x|\omega_i)$ $(i = 1,...,r)$

If we assume $p(x|\omega_i)$ is an n-dimensional Gaussian probability density, then $p(x|\omega_i)$ is given by

$$p(x|\omega_i) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_i|}} \exp\{ -\frac{1}{2}(x - m_i)^t \Sigma_i^{-1} (x - m_i)\},$$

where m_i is the mean vector of patterns which are from class ω_i and Σ_i is the covariance matrix of class ω_i, then the Bayes discriminant function $g_i(x)$ is given by the following:

$$g_i(x) = -\frac{1}{2}(x - m_i)^t \Sigma_i^{-1} (x - m_i) - \frac{n}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| + \log p(\omega_i).$$

In the special case of the two-category classification problem, a Bayes discriminant function $g(x)$ is given by

$$g(x) = {}^t x w_1 x - {}^t x w_2 x + {}^t u_1 x - {}^t u_2 x + w_{10} - w_{20}, \quad (1)$$

where

$$w_i = -\frac{1}{2}\Sigma_i^{-1},$$

$$u_i = \Sigma_i^{-1} m_i,$$

and

$$w_{i0} = -\frac{1}{2} m_i^t \Sigma_i^{-1} m_i - \frac{1}{2}\log|\Sigma_i| + \log p(\omega_i).$$

Because $g(x)$ is a quadratic polynomial of n variables, the decision surface $g(x) = 0$ is generally a quadratic hypersurface in R^n.

4. Approximation theory of the Bayes discriminant function in the case of a two-category Gaussian classification problem

4.1. Relationship between the input–output function of three-layer network and Bayes discriminant function

In the rest of this paper, the following notation will be used. Let $\phi(x)$ be a sigmoid function, that is, a bounded,

nonconstant, and strictly increasing continuous function such that $\phi(R) = (0,1)$. We define $\phi_\lambda(x) = \phi(\lambda x)$. The sigmoid function which is usually used is denoted by $\sigma(x)$, i.e.,

$$\sigma(x) = \frac{1}{1+\exp(-x)}.$$

The following theorem is easily derived from the approximation theorem on three-layer neural networks by Funahashi (1989).

Theorem 1. *Let $g(x)$ be a bounded Bayes discriminant function such that $0 < g(x) < 1$, and $p(x)$ be the associated probability function. Then for any $\epsilon > 0$, there exists a three-layer neural network with a sigmoid output function $\phi(x)$ such that*

$$\|g - \tilde{g}\|^2 = \int_{R^n} [g(x) - \tilde{g}(x)]^2 p(x)\, dx < \epsilon,$$

where $\tilde{g}$ is the input–output function of the network.

The above theorem is purely an existence theorem, so the number of hidden units necessary for the approximation is not mentioned.

In the following, we will consider the two-category classification problem with n-dimensional Gaussian probability densities. In this case, we consider the Bayes discriminant function $g(x)$ given by (1) in Section 3 and the approximation to it. For this purpose, we propose the following:

Proposition 1. *Any quadratic polynomial function $f(x_1,...,x_n)$ of n variables can be approximated by the input–output function of three-layer neural networks with at least $2n$ hidden units whose output function is C^2-sigmoid and one linear output unit, with any precision on any compact subset K or R^n. In other words, for any $\epsilon > 0$, and any compact subset K of R^n, there exist real constants c_i, θ_i $(i = 1,...,2n)$, w_{ij} $(i = 1,...,2n, j = 1,...,n)$, and τ such that*

$$\max_{x\in K} |f(x_1,\ldots,x_n) - \sum_{i=1}^{2n} c_i \phi(\sum_{j=1}^{n} w_{ij}x_j + \theta_i) - \tau| < \epsilon.$$

Moreover, the above uniform approximation also applies to the first order differential.

Because the Bayes discriminant function $g(x)$ in our problem is a quadratic polynomial of n variables, $g(x)$ can be approximated by three-layer networks with at least $2n$ hidden units and a linear output unit, by the above proposition. For $g(x)$ given by (1) in Section 3, $\phi_\lambda(g(x))$ is also a Bayes decision function, we obtain the following theorem, using the above proposition.

Theorem 2. *Consider a two-category classification problem with n-dimensional Gaussian probability densities. For any $\lambda > 0$, a Bayes discriminant function $f(x) = \phi_\lambda(g(x))$ can be approximated in the statistical sense by the input–output functions of three-layer neural networks with at least $2n$ hidden units whose output function is $\phi(x)$, where the approximation of statistical sense means the L^2-approximation with the density function $p(x)$.*

For the sigmoid function ($\sigma(x) = 1/(1 + \exp(-x))$), we consider the Bayes discriminant function $\sigma(g(x))$ where $g(x)$ is given in Section 3. If $\sigma(g(x)) > 1/2$ for x, we assign ω_1, and if $\sigma(g(x)) < 1/2$ for x, we assign ω_2. As $\sigma(g(x)) = 1/(1 + \exp(-g(x)))$, where

$$g(x) = g_1(x) - g_2(x) = \log \frac{p(x|\omega_1)p(\omega_1)}{p(x|\omega_2)p(\omega_2)},$$

we obtain

$$\sigma(g(x)) = \frac{p(x|\omega_1)p(\omega_1)}{p(x|\omega_1)p(\omega_1) + p(x|\omega_2)p(\omega_2)} = p(\omega_1|x).$$

That is, $\sigma(g(x))$ is equal to the a posteriori probability density function for the category ω_1. Therefore we obtain the following theorem, which follows from the above discussion and Theorem 2.

Theorem 3. *Consider a two-category pattern classification problem with n-dimensional Gaussian probability densities by the use of three-layer neural networks whose output function is $\sigma(x) = 1/(1 + \exp(-x))$. Then, the a posteriori probability density of category ω_1: $p(\omega_1|x)$ $(=1 - p(\omega_2|x)$, where $p(\omega_2|x)$ is a posteriori probability density of $\omega_2)$ can be approximated in the statistical sense by a three-layer network with at least $2n$ hidden units.*

As stated in Section 2, if there are infinite learning samples, then the mean-squared error function $E_a(w)$ is given by

$$E_a(w) = e^2(w) + \int_{R^n} p(\omega_1|x)(1 - p(\omega_1|x))p(x)\, dx,$$

where $e^2(w)$ depends only on w and is given by

$$e^2(w) = \int_{R^n} [F(x,w) - p(\omega_1|x)]^2 p(x)\, dx.$$

Therefore, $E_a(w) > E_a(w')$ implies $e^2(w) > e^2(w')$. On the other hand, Theorem 3 implies $\inf_w e^2(w) = 0$. Hence we obtain the following theorem which is the main theorem of this paper.

Theorem 4. (Main Theorem) *Consider a two-category pattern classification problem with n-dimensional Gaussian probability densities by the use of three-layer neural networks with one output unit. As teacher signals, we assign 1 when the input data is from ω_1 and 0 when the input data is from ω_2. Suppose that we use the usual sigmoid function*

$\sigma(x) = 1/(1 + \exp(-x))$ for both output and hidden layer, and the hidden layer has at least 2n units. Then, if the learning proceeds ideally (i.e., the mean-squared error decreases to its infimum), the input–output function of the network tends to the a posteriori probability density $p(\omega_1|x)$ for class ω_1 in the statistical sense, that is, the L^2-distance with weight $p(x)$ between the input–output function and $p(\omega_1|x)$ tends to zero.

To complete the proofs of theorems 2, 3 and 4, it remains to prove Proposition 1. In the next subsection, we shall give the proof.

4.2. Proof of the proposition

A general quadratic polynomial of n variables is given by

$$f(x_1, \ldots, x_n) = \sum_{i,j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n} b_i x_i + c, \tag{2}$$

where $a_{ij} = a_{ji}$ $(i,j = 1,...,n)$.

Because any quadratic form can be transformed into a canonical form by an orthogonal transform $x = Py$ where $P = (p_{ij})$ is an orthogonal matrix, $x = {}^t(x_1,...,x_n)$ and $y = {}^t(y_1,...,y_n)$, $f(x_1,...,x_n)$ can be transformed to the following:

$$F(y_1, \ldots, y_n) = \sum_{i=1}^{n} \lambda_i y_i^2 + \sum_{i=1}^{n} \mu_i y_i + \tau, \tag{3}$$

where $\{\lambda_i\}$ are eigenvalues of the symmetric matrix $A = (a_{ij})$.

For $\lambda_i \neq 0$, $\lambda_i y_i^2 + \mu_i y_i = \pm(\sqrt{|\lambda_i|}y_i + \mu_i/2\sqrt{|\lambda_i|})^2 + \nu_i$ where $\nu_i = -\frac{\mu_i^2}{4\lambda_i}$. Therefore, for the approximation of the quadratic polynomial function, we must consider the approximations of both the linear function $h(x) = x$ and the second-power function $f(x) = x^2$. On the approximations to these functions, we have Lemmas 1 and 2 as follows.

Lemma 1. *The linear function $h(x) = x$ can be approximated uniformly with any precision on any compact subset K of R, by the use of a three-layer network with one linear output unit and one hidden unit whose output function is a C^1-sigmoid function.*

For proof of this Lemma, see Funahashi (1990), Proposition 1.

Moreover, in Toda et al. (1991), it has been proven that $f(x,y) = xy$ can be approximated uniformly on any compact subset K or R^2 by three-layer neural networks with two hidden units whose output function is a C^2-sigmoid function. By modifying this proof, we obtain Lemma 2 given below. Although a similar result has been proven by Kreinovich (1991), we include the proof of Lemma 2 for completeness.

Lemma 2. *The second power function $f(x) = x^2$ can be approximated uniformly on any compact subset K of R by the use of three-layer neural networks with one linear output unit (i.e., the output function of the output unit is a linear function) and two hidden units whose output function $\phi(x)$ is a C^2-sigmoid function. That is, for any $\epsilon > 0$, and any compact subset K of R, there exist constants c_i, w_i, θ_i $(i = 1,2)$ and τ such that*

$$\max_{x \in K} |x^2 - \sum_{i=1}^{2} c_i \phi(w_i x + \theta_i) - \tau| < \epsilon.$$

Moreover the approximation is uniform including to the first differential.

Proof. For the sigmoid function $\phi(x)$, here we define $\phi_\theta(x) = \phi(x + \theta)$ unlike the definition given at the beginning of Section 4. We can choose θ so that $\phi_\theta''(0)$ is not zero, because ϕ is bounded and not constant. A Taylor expansion to second order in the neighborhood of $x = 0$ gives:

$$\phi_\theta(x) = \phi_\theta'(0) + \phi_\theta(0)x + \frac{1}{2}\phi_\theta''(0)x^2 + o(x^2), \tag{4}$$

where $o(x^2)$ is the term which tends to zero faster than x^2 as x tends to zero. We replace x by wx and $-wx$ in Eq. (4) and obtain

$$\phi_\theta(\pm wx) = \phi_\theta(0) \pm \phi_\theta'(0)wx + \frac{1}{2}\phi_\theta''(0)w^2x^2 + o(w^2x^2).$$

If x is included in the compact subset K of R, we obtain

$$\phi_\theta(\pm wx) = \phi_\theta(0) \pm \phi_\theta'(0)wx + \frac{1}{2}\phi_\theta''(0)w^2 + o(w^2),$$

where $o(w^2)$ is the term which tends to zero uniformly on K faster than w^2 as w tends to zero. For the sum of $\phi_\theta(wx)$ and $\phi_\theta(-wx)$ we obtain

$$\phi_\theta(wx) + \phi_\theta(-wx) = 2\phi_\theta(0) + \phi_\theta''(0)w^2x^2 + o(w^2).$$

Hence,

$$x^2 + \frac{o(w^2)}{\phi_\theta''(0)w^2} = \frac{1}{\phi_\theta''(0)w^2}\{\phi_\theta(wx) + \phi_\theta(-wx)\} - \frac{2\phi_\theta(0)}{\phi_\theta''(0)w^2}. \tag{5}$$

This implies that $f(x) = x^2$ can be approximated uniformly with any precision on any compact subset K of R by three-layer neural networks with two hidden units.

Let the right hand side of Eq. (5) be $g(x)$. Then $g'(x)$ can be expressed as

$$g'(x) = \frac{1}{\phi_\theta''(0)w}\{\phi_\theta'(wx) - \phi_\theta'(-wx)\}$$

By applying Taylor's formula to $\phi_\theta'(\pm wx)$, we obtain

$$g'(x) = 2x + \frac{o(w)}{\phi_\theta''(0)w},$$

where $o(w)$ is the term which tends to zero uniformly on K faster than w as w tends to zero. This shows that the approximation of the first differential is also uniform. *q.e.d.*

From Lemma 1 and Lemma 2 we see that the quadratic polynomial $F(y_1,...,y_n)$ can be approximated by input–

output functions of three-layer neural networks with at least $2n$ hidden units on any compact subset K of R^n and hence the quadratic polynomial function $f(x_1,...,x_n)$ can be approximated by input–output functions of neural networks with hidden units of the same number, because the transformation from the input layer to the hidden layer of the three-layer networks is an affine transformation.

5. Summary

In this paper, we considered the two-category pattern classification problem with n-dimensional Gaussian probability densities and proved the following.

1. The a posteriori density probability can be approximated in the statistical sense by three-layer neural networks with at least $2n$ hidden units and one output unit whose output function is the usual sigmoid function.
2. If we have infinite learning samples and the learning proceeds ideally (i.e., the mean-squared error decreases to the infimum), then the input–output function of the neural network with at least $2n$ hidden units tends to the a posteriori probability density in the statistical sense.

These results provide a theoretical basis for the study of pattern classification by computer simulation. In the future, we would like to study the extension of our theory to the case of multi-category pattern classification problem.

Acknowledgements

We thank T. Yokoyama for discussions on this research and testing of our theory by computer simulation. We also thank Prof. P. Möller for critical reading of the manuscript.

References

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: John Wiley and Sons.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks, 2*, 183–192.

Funahashi, K. (1990). Approximate realization of identity mappings by three-layer neural networks. *Electronics and Communication in Japan,* **73** (3), 61–68. Translated from *Transactions of the Institute of Electronics, Information and Communication Engineers of Japan,* **73-A,** 139–145 (1990).

Kreinovich, V.Y. (1991). Arbitrary nonlinearity is sufficient to represent all functions by neural networks: A theorem. *Neural Networks, 4*, 381–383.

Richard, M.D., & Lippmann, R.P. (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation, 3*, 461–483.

Ruck, D. W., Rogers, S., Kabrisky, M., Oxley, H., & Suter, B. (1990). The multilayer Perceptron as an approximator to a Bayes optimal discriminant function. *IEEEE Transactions on Neural Networks, 1*, 296–298.

Rumelhart, D.E., McClelland, J.L., & the PDP research group (1986). *Parallel distributed processing,* Vol. 1, Cambridge: MIT Press.

Toda, N., Funahashi, K., and Usui, S. (1991). Polynomial functions can be realized by finite size multilayer feedforward neural networks. In *Proceedings of International Joint Conference on Neural Networks,* Singapore (pp. 343–348).

White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural Computation, 1*, 425–464