A THEORETICAL INTERPRETATION FOR LAYERED NEURAL NETWORK BASED IMAGE CLASSIFIER

Eihan SHIMIZU
Professor, Department of Civil Engineering
University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-8656
Email: shimizu@planner.t.u-tokyo.ac.jp
JAPAN

Commission V, Working Group 1

KEY WORDS: Image classifier, Layered Neural Network

ABSTRACT

Layered feed-forward neural networks (LNN) have been broadly applied to classification, prediction and other modeling problems. There have been so far, however, few studies that have provided a theoretical interpretation for the application of LNN. Most of the conventional studies have been empirical and the LNNs have been applied just like "black box" machines. This paper discusses the application of LNN to image or remotely sensed data classification. We provide a theoretical interpretation for the LNN classifier in comparison with the conventional classification or discriminant methods. The most distinguished part is the derivation of a generalized form of LNN classifier based on the maximum entropy principle. According to the generalized form, we discuss the relationship between the familiar type of LNN classifier employing the sigmoidal activation function and the other types of discriminant models such as the Multinomial Logit Model.

1. INTRODUCTION

Layered feed-forward neural networks (LNN) have been broadly applied into prediction, simulation, classification, pattern recognition and other modeling problems. Hill *et al.* (1994) gave a review of studies comparing LNNs with conventional statistical models. There have been so far, however, few studies that have provided a theoretical interpretation for the application of LNN except for comparisons with regression analysis. Most of the conventional studies have been empirical and LNNs have been applied just like "black box" estimation machines.

This paper discusses the applications of LNN to classification and pattern recognition problems which have been often attempted in the fields of remote sensing and digital image analysis. We provide a theoretical interpretation for the LNN based classifier mainly in comparison with Bayesian classifier and Multinomial Logit Model.

2. BASIC FORMULATION OF LNN CLASSIFIER

Let $\boldsymbol{x}$ represent a feature vector which is to be classified. Let the possible classes be denoted by $\omega_j (j = 1,2,..,J)$. Consider the discriminant function $d_j(\boldsymbol{x})$, then decision rule is

$$\boldsymbol{x} \in \omega_j, \quad \text{if} \quad d_j(\boldsymbol{x}) \geq d_{j'}(\boldsymbol{x}) \quad \text{for all} \quad j' \neq j. \quad (1)$$

A LNN is expected to be the I/O system corresponding to the discriminant function.

We show a typical LNN architecture which has been applied to a variety of classification problems. Let us consider the multi-layered neural network. A feature vector is input to the input layer, that is, the number of the neurons in the input layer is corresponding to the dimension of the feature vectors. The output layer has the number of neurons as same as the classes. The output signal from the jth neuron in the output layer is regarded as the discriminant value. Let the state of the jth output neuron be represented by

$$u_j = g(\boldsymbol{x}, \boldsymbol{w}) \tag{2}$$

where $\boldsymbol{w}$ is the parameter vector included in the designed LNN. These parameters are mainly constituted by the connection weights (synaptic weights) between neurons. We are not concerned here with the formulation of $g(\boldsymbol{x}, \boldsymbol{w})$. The output of LNN, $p_j(\boldsymbol{x}, \boldsymbol{w})$, under presentation of $\boldsymbol{x}$ is

$$p_j(\boldsymbol{x}, \boldsymbol{w}) = f(u_j), \tag{3}$$

where $f(u_j)$ is the activation function. The following sigmoid function, which is a bounded, monotonic and increasing, is frequently used,

$$f(u_j) = \frac{1}{1 + exp(-u_j)} \tag{4}$$

The feature vectors $\boldsymbol{x}_k (k = 1,2,..,K)$ for training the LNN are prepared. The classes which these feature vectors belong to are all known. Training data (target data) are given as follows;

$$d_i(\boldsymbol{x}_k) = \begin{cases} 1 & if \ \ \boldsymbol{x}_k \in \omega_i \\ 0 & otherwise \end{cases} \tag{5}$$

The LNN is trained by minimizing a mean squared error;

$$min. \sum_{k=1}^{K} \sum_{j=1}^{J} \left\{ p_j(\boldsymbol{x}_k, \boldsymbol{w}) - d_j(\boldsymbol{x}_k) \right\}^2. \tag{6}$$

Training of the LNN is performed through the adjustment of connection weights. The most common method is so-called 'back propagation' which is gradient descent in essence. After the completion of training, the LNN plays a role of the discriminant function. Let the output of trained LNN be denoted $p_j(\boldsymbol{x}, \hat{\boldsymbol{w}})$.

3. INTERPRETATION FOR LNN CLASSIFIER

3.1 Relationship between LNN and Bayesian classifier

The Bayesian optimal decision rule, in the sense of minimizing the probability of classification error, is to choose the class which maximizes the posterior probability;

$$p(\omega_j \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \omega_j) \cdot p(\omega_j)}{p(\boldsymbol{x})}. \tag{7}$$

If the prior probabilities $p(\omega_j)$ are equal, then the conditional probability density function $p(\boldsymbol{x} \mid \omega_j)$ corresponds to the optimal discriminant function. Maximum likelihood classifier, in which a multivariate normal distribution is assumed, is frequently applied.

How the LNN classifier is related to the Bayesian optimal classifier? This question has been already discussed by Wan (1990) and Ruck *et al.* (1990). The conclusion is that the output of the LNN, $p_j(\boldsymbol{x}, \hat{\boldsymbol{w}})$, approximates the Bayesian posterior probability. According to Wan (1990), we show a short proof. Consider the training data given in the form of (5). Suppose that the training data are random variables and samples from the probability density function $p(\boldsymbol{x}, d_j(\boldsymbol{x}))$, where

$$d_j(\boldsymbol{x}) = \begin{cases} 0 & if \ \ \boldsymbol{x} \in \omega_j \\ 1 & otherwise \end{cases} \tag{8}$$

Since $p_j(\boldsymbol{x}, \hat{\boldsymbol{w}})$ is the least squares estimate of $d_j(\boldsymbol{x})$, then $p_j(\boldsymbol{x}, \hat{\boldsymbol{w}})$ is the conditional expectation of $d_j(\boldsymbol{x})$ given $\boldsymbol{x}$ Therefore,

$$\begin{aligned} p_j(\boldsymbol{x}, \hat{\boldsymbol{w}}) &= E\left[d_j(\boldsymbol{x}) \mid \boldsymbol{x} \right] \\ &= \sum_{d_j(\boldsymbol{x}) \in \{0,1\}} d_j(\boldsymbol{x}) \cdot p(d_j(\boldsymbol{x}) \mid \boldsymbol{x}) \\ &= p(d_j(\boldsymbol{x}) = 1 \mid \boldsymbol{x}) \\ &= p(\omega_j \mid \boldsymbol{x}) \end{aligned} \tag{9}$$

This means that $p_j(\boldsymbol{x}, \hat{\boldsymbol{w}})$, in the sense of minimizing a mean squared error, approximates the posterior probability $p(\omega_j \mid \boldsymbol{x})$. This provides a theoretical interpretation for the LNN based classifier. It is proved that a three-layered neural network, when the appropriate number of neurons are set in the hidden layer and sigmoidal activation functions are used in the hidden layer, can approximate any continuous mapping (e.g. Gallant et al., 1988; Funahashi, 1989; Cybenko, 1989; Hornik *et al.*, 1989). It is expected that LNN approximates accurately the posterior probability.

Up to this point, however, the derivations have been for an arbitrary mapping trained by $d_j(\boldsymbol{x}_k) \in \{0,1\}$. The result is well-known in the field of statistics (Wan, 1990). The above proof provides a theoretical justification for any non-parametric discriminant function trained by the least squares criteria. The following section will discuss the interpretation for the activation functions used in LNN.

3.2 Interpretation for activation functions

Let the activation function, $f(u_j)$, be a monotonic increasing function. Then, the state of the output neuron, u_j, and the posterior probability, $p(\omega_j \mid \boldsymbol{x})$, have a one-to-one mapping, and $u_j = g(\boldsymbol{x}, \boldsymbol{w})$ becomes also an optimal discriminant function.

The activation function should be a probability distribution given a certain level of state. This is analogous to the probability distribution of a particle being in a certain state given the energy level of each state in the statistical mechanics. In the statistical mechanics different probability distributions are derived from so-called maximum entropy principle. We derive the activation forms from maximum entropy principle.

Consider the maximization of Kapur's generalized measure of entropy under the expected discriminant value (Kapur, 1986).

$$max. H(\boldsymbol{p}) = -\sum_{j=1}^{J} p_j \cdot ln\, p_j + \frac{1}{a}\sum_{j=1}^{J}(1+ap_j)\cdot ln(1+ap_j), \quad a \geq -1 \tag{10}$$

$$s.t. \quad \sum_{j}^{J} p_j u_j = U \tag{11}$$

where $H(\boldsymbol{p})$ is the Kapur's generalized entropy in which the constant term is omitted, $p_j (j = 1,2,..,J)$ is a probability distribution corresponding to $p_j(\boldsymbol{x}, \boldsymbol{w})$, a is a parameter prescribing the type of entropy, that is, the type of probability distribution, and U is an expected discriminat value. Here, we do not explicitly give the constraint;

$$\sum_{j}^{J} p_j = 1 \tag{12}$$

to the maximization problem, because p_j approximates the posterior probability.

From (10) and (11), we get

$$p_j = \frac{1}{-a + exp(-\beta u_j)} \tag{13}$$

where β is a Lagrange multiplier associated with (11). The parameter β is the so-called temperature parameter. When a is fixed and the LNN with the activation function (13) is trained, β is estimated being included in the connection weights in a training process, since u_j is generally defined by the linear function of the connection weights between the output neuron concerned and the hidden neuron.

Now, assume that β is constant given, the probability distribution (13) is equivalent to an optimal solution of the following maximization (Brotchie, 1979);

$$max. \quad U = \sum_{j=1}^{J} p_j u_j + \frac{1}{\beta} H(\boldsymbol{p}). \tag{14}$$

Therefore, the activation function form (13) is interpreted as the representation of the above expected discriminant value maximization taking into account the uncertainty shown as the Kapur's entropy.

Now, let us return to the activation form (13) and discuss the meaning of the parameter a. For a =-1, (13) gives;

$$p_j = \frac{1}{1 + exp(-\beta u_j)}. \tag{15}$$

This is just the sigmoid function (i.e., (4)) most frequently used in the applications of LNNs. In addition, if a =-1, it is well-known that (10) subject to (11) and (12) gives Fermi-Dirac (F-D) distribution. Note that $p_j(\boldsymbol{x}, \hat{\boldsymbol{w}})$ approximates the posterior probability; thus the familiar sigmoid function is interpreted as the representation of the expected discriminant value maximization under the F-D type entropy.

Similarly, for a =1, (13) is

$$p_j = \frac{1}{-1 + exp(-\beta u_j)}. \tag{16}$$

It is known that, for a =1, (10) subject to (11) and (12) gives Bose-Einstein (B-E) distribution. Thus (16) approximates the B-E Distribution.

Next, consider the case of a =0, that is,

$$p_j = \frac{1}{exp(-\beta u_j)}. \tag{17}$$

As a tends to zero, (10) approaches Shannon's measure of entropy. It is well-known that the maximization of the Shannon's entropy subject to (11) and (12) gives Maxwell-Boltzmann (M-B) probability distribution;

$$p_j = \frac{exp(\beta u_j)}{\sum_{j'} exp(\beta u_{j'})}. \tag{18}$$

Accordingly, (17) approximates the M-B distribution. In addition, (18) gives the structural similarity with so-called Multinomial Logit Model which is familiar in the field of the

discrete choice behavioral modeling (Anas, 1983). Hence the LNN classifier with the activation function (17) is interpretated as the approximate of the Multinomial Logit Model.

As mentioned above, the choice of a = -1, 0 and 1 leads to Fermi-Dirac (F-D), Maxwell-Boltzmann (M-B), and Bose-Einstein (B-E) probability distributions respectively in statistical mechanics. Let us compare the characteristics of the above representative distributions in statistical mechanics. These three distributions are all derived from Jaynes's maximum entropy principle (Kapur, 1992). One distribution differs from another due to the constraints to Shannon's measure of entropy. In the M-B distribution, the expected energy of a particle in the system is only prescribed. The F-D and B-E distributions are derived by the constraints with respects to the expected energy of the system and the expected number of the particles in the system. In the F-D distribution the maximum number of the particles allowed in a certain state is assumed to be one, while in the B-E distribution the maximum number is assumed to be infinite.

Thus, the parameter a is associated with the constraints to the maximization of the Shannon's entropy. This gives us an implication that, for a lying between -1 and 1, we can get the various types of probability distributions, though it may be difficult to provide the significant interpretation for the distributions in the framework of the statistical mechanics. We have a choice of infinite types of models corresponding to different values of a. A possible method is to choose the parameter a to get the best fit to the training data. Regardless as the selected parameter, we can provide the interpretation to the activation function as the representation of the expected discriminant value maximization under the Kapur's generalized entropy.

4. CONCLUSION

This paper has provided an interpretation for the LNN classifier. The output of the LNN under the completion of training approximates the Bayesian posterior probability. Therefore, if we assume the activation function of the output neuron to be monotonic increasing, the state of the output neuron is also Bayesian optimal discriminant function. From the maximum entropy principle, we can provide the interpretation for the activation function. The familiar sigmoid function is approximate to the Fermi-Dirac distribution. The LNN classifier using the activation function of the Maxwell-Boltzmann distribution approximates the Multinomial Logit Model. The maximization of Kapur's generalized measure of entropy gives the generalized form of the probability distributions including the Maxwell-Boltzmann, Fermi-Dirac, and Bose-Einstein distributions. In the practical sense, it is proposed to apply the Kapur's generalized distribution into the generalized activation function and to fix the function form in the process of training. Regardless as the resulting selected function form, we can provide the interpretation for that as the representation of the maximization of the expected discriminant value under the Kapur's generalized entropy.

REFERENCES

Anas, A., 1983. Discrete choice theory, information theory and the multinomial logit and gravity model. Transportation Research, 17B(1), pp.13-23.

Brotchie, J. F. and Lesse, P. F., 1979. A unified approach to urban modeling. Management Science, Vol.25, 1, pp.112-113.

Funahashi, K., 1989. On the approximate realization of continuous mappings by neural networks. Neural Networks, 2, pp.183-192.

Gallant, A.R. and White, H., 1988. There exists a neural network that does not make avoidable mistakes. Proc. Int. Conf. Neural Networks 1 (July, 1988), pp.657-666.

Hill, T., Marquez, L., O'Connor, M. and Remus, W., 1994. Artificial neural network models for forecasting and decision making. Int. Jour. Forecasting, Vol.10, pp.5-15.

Hornik, K.M., Stinchcombe, M. and White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), pp.359-366.

Kapur, J. N., 1986. Four families of measures of entropy. Ind. Jour. Pure and Applied Mathematics, 17, pp.429-449.

Kapur, J. N. and Kesavan, H. K., 1992. Entropy optimization principles with applications. Academic Press, Inc., pp.77-97.

Ruck, D. W., Rogers, et al., 1990. The multilayer perceptron as an approximation to a Bayes optimal discriminant function. IEEE Transactions on Neural Networks, 1(4), pp.296-298.

Wan, Eric A., 1990. Neural network classification: a Bayesian interpretation. IEEE Transactions on Neural Networks, 1(4), pp.303-305.