A MULTICLASS CLASSIFICATION METHOD BY DISTANCE MAPPING LEARNING NETWORK

Kenji Suzuki and Shuji Hashimoto

Dept. of Applied Physics, Waseda University, Japan

ABSTRACT

In this paper, we propose a method of multiclass classification by utilizing a distance mapping learning network that is a distance-based multilayer perceptron [1]. The network can obtain the non-linear mapping between the input objects and the outputs by providing a pair of objects and the desired distance between them. It thus realizes multiclass classification based on pairwise classifications iteratively. We will show the validity of the model with two classification problems: Iris classification and facial expression classification.

1. INTRODUCTION

A number of techniques of multiclass classification have been proposed. In the conventional classification problems, one tries to distinguish between two (or more) classes of objects, and most methods try to estimate the probability density of the target set. On the other hand, a powerful and effective method of two-class classification is proposed such as Support Vector Machine [2]. One approach to classify multiple classes is to combine two-class classifiers. Another approach has been reported about multiclass classification by combining one-class classification [3] that tries to distinguish between a set of objects and all other objects. In these cases, although we can choose the classifiers so as to adapt each two-class classification of target set, the number of class is given or provisionally has to be set. However, it is often impossible or difficult to know the number of class in real-world problem.

In this paper, we propose a method of multiclass classification by utilizing *Distance Mapping Learning* network that is a distance-based multilayer perceptron [1]. The network can obtain the non-linear mapping between the input objects and the outputs by providing the desired distance between the objects, not the desired output. The desired distance represents the similarity between the input objects. The network thus realizes multiclass classification based on pairwise classifications.

Some learning methods utilizing a similarity-based distance have been reported, for instance, for image database organization [4], the classification method [5]. However, the proposed method differs from these approaches in the training of the network. It should be noted that the proposed model can deal with classification for an unknown number of class. Only by presenting a pair of objects and the desired distance, the network can map the objects onto the output space of arbitrary dimensions, which is regarded as data description space.

In this paper, we first introduce the mechanism of distance mapping learning. Secondly, we show its application to multiclass classification with some experimental results. Conclusions are then finally given.

2. DISTANCE MAPPING LEARNING

2.1. Problem specification

The framework of the acquirement of a non-linear mapping function between high-dimensional input feature space and a lower dimensional output description space is formulated as follows. In the m-dimensional input feature space $\boldsymbol{X}$, a object p is represented as the vector $\boldsymbol{x}^p = (x_1^p, x_2^p, \cdots, x_m^p)$. The target is to produce n-dimensional vector outputs $\boldsymbol{y}^p = (y_1^p, y_2^p, \cdots, y_n^p)$, $\boldsymbol{y}^q \in \mathbf{R}^n$ that preserve a desired distance $s_{pq} (\geq 0)$ with regard to given inputs $\boldsymbol{x}^p$ and $\boldsymbol{x}^q \in \mathbf{R}^m$.

The similarity s_{pq} approximates Euclidean distance d_{pq} between the non-linearly mapped objects $\boldsymbol{y}^p, \boldsymbol{y}^q$ $(=\Phi(\boldsymbol{x}^p),\ \Phi(\boldsymbol{x}^q))$ from the objects $\boldsymbol{x}^p, \boldsymbol{x}^q$ in the feature space. Here, $\boldsymbol{y}^p, \boldsymbol{y}^q$ are regarded as the description vector that describes $\boldsymbol{x}^p, \boldsymbol{x}^q$ in n dimensional space.

To solve the above problem, the following fitting value W is minimized under a provisionally determined dimension order n.

$$W \equiv \sum_p \sum_q (s_{pq} - d_{pq})^2 \tag{1}$$

* contact author: kenji@ieee.org
Address: 3-4-1, Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
http://www.phys.waseda.ac.jp/shalab/

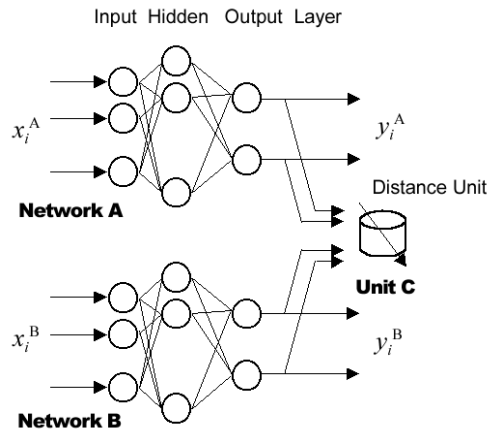

Figure 1: Structure of the DML network

$$d_{pq} = \|\boldsymbol{y}^p - \boldsymbol{y}^q\| \qquad (a)$$
$$d_{pq} = \|\Phi(\boldsymbol{x}^p) - \Phi(\boldsymbol{x}^q)\| \qquad (b)$$

where d_{pq} denotes the distance between mapped vectors. The similarity value s_{pq} is provisionally given as the desired distance between the objects p and q. The geometric arrangement in Euclidean space is chosen because it is helpful for intuitive understanding of the data structure. In the conventional multidimensional scaling method [6], the target is to minimize equation (1) under the constraint of equation (a), not containing mapping function Φ with input vector $\boldsymbol{x}$. Therefore, the arranged vector $\boldsymbol{y}$ does not have any relationship with $\boldsymbol{x}$. On the other hand, the proposed method is to minimize equation (1) under the constraint of equation (b) so that the nonlinear mapping function Φ could be derived from the neural network learning.

2.2. Network structure

The structure of the Distance Mapping Learning (DML) network is illustrated in Figure 1. Network A and network B are identical three-layered perceptrons. The inputs are the feature parameters of input object A and B, respectively, and the outputs are the description parameters of object A and B. The nonlinear mapping between the input space and output space is therefore done by networks A and B. The outputs of A and B are connected in parallel to unit C which calculates the distance d_{AB} in the output space. The network outputs y_k^A and $y_k^B (k = 1, 2, \cdots, n)$ from each output layer, respectively, while the input vectors $x_i^A, x_i^B (i = 1, 2, \cdots, m)$ are given. Each output vectors of networks A, B can be of arbitrary dimensions. In the learning process, rather than providing networks A and B with absolute coordinates as a teacher signal, the desired distance between the inputs to network A and network B, s_{AB}, is given.

2.3. Formulation

2.3.1. Error backpropagation algorithm

Multi-layer perceptron (MLP) based on an error backpropagation algorithm for minimizing the mean square error is most popular [7]. The least-squares learning and regression are discussed in [8]. This states that among all the functions of $\boldsymbol{x}$, regression is the best predictor of $\boldsymbol{y}$ given $\boldsymbol{x}$, in the mean-squared-error sense.

A training set $(x^1, y^1), (x^2, y^2) ... (x^p, y^p) ...$ is a collection of observed $(\boldsymbol{x}, \boldsymbol{y})$ pairs. In other words, the pair $(\boldsymbol{x}, \boldsymbol{y})$ obeys some unknown joint probability distribution. To construct a non-linear function $\Phi(\boldsymbol{x})$ based on the training set is equivalent so that Φ satisfies the desired output $\boldsymbol{y}$. Φ is generally determined so as to minimize a given cost function. The learning process of a multi-layer perceptron is described as:

$$\varepsilon^2(\Phi) = \int \|y^p - \Phi(x^p)\|^2 p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} d\boldsymbol{x} \qquad (2)$$

where $p(\boldsymbol{x}, \boldsymbol{y})$ and ε denote the probability density function and mean-squared error of $\Phi(\boldsymbol{x})$ that represents a nonlinear functional of the neural network, in which the goal of learning process is to minimize ε^2. Equation (3) that is a simple quadratic formula regarding $\hat{\boldsymbol{y}}$ is delivered by probability distribution $\boldsymbol{y}$ conditioned by $\boldsymbol{x}$:

$$\hat{\boldsymbol{y}} = \Phi_{opt}(\boldsymbol{x}) = \int \boldsymbol{y} p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y} \qquad (3)$$

2.3.2. Distance mapping learning algorithm

On the other hand, the proposed neural network differs from the conventional MLP, because the desired distance between input objects is given as a teacher signal. The mean-squared error ε is described as follows with the probability density function $p(\boldsymbol{x}^p, \boldsymbol{x}^q, \boldsymbol{s}_{pq})$ where $\boldsymbol{s}_{pq}$ denotes the desired distance between $(\boldsymbol{x}^p, \boldsymbol{x}^q)$, which may not be a deterministic variable even though $\boldsymbol{x}^p$ and $\boldsymbol{x}^q$ are determined. It should be noted that the proposed method is independent of translation and rotation in the output space. Because the network that is trained by the relative distances, these transforms can be applied without missing the property.

In this framework, we realize the algorithm of Sammon's nonlinear mapping [9] by a modified multilayer perceptron.

$$\varepsilon^2(\Phi) = \int \left\| s_{pq} - d_{pq} \right\|^2 p(\boldsymbol{x}^p, \boldsymbol{x}^q, \boldsymbol{s}_{pq}) d\boldsymbol{x}^p d\boldsymbol{x}^q d\boldsymbol{s}_{pq} \quad (4)$$

$$
\begin{aligned}
d_{pq}(\Phi(\boldsymbol{x}^p), \Phi(\boldsymbol{x}^q)) &= \|\Phi(\boldsymbol{x}^p) - \Phi(\boldsymbol{x}^q)\| \\
&= \|(A\Phi(\boldsymbol{x}^p) + B) \\
&\quad -(A\Phi(\boldsymbol{x}^q) + B)\|
\end{aligned}
$$

(A : rotation matrix, B : translation matrix)

2.4. Learning rule

The connection weights in network A and B are tuned to make the distance d_{AB} close to teacher signal s_{AB}. The sigmoid output function is used for each cell, and a modified back propagation method is applied for the learning. Networks A and B start from the same initial connection weights and are trained in the same manner to give the same mapping.

The learning rule is described below. In the following equations, subscript i, j and k correspond to the cell number in input, hidden and output layers, respectively. w_{ij} represents the connection weight from unit i in input layer to unit j in hidden layer . The cost functional of DML E_d is defined as equation (5).

$$
\begin{aligned}
E_d &= \sum^{n}(s_{AB} - d_{AB})^2/2 \qquad (5) \\
d_{AB} &= \sqrt{\sum_{k=1}^{n}(y_k^A - y_k^B)^2}
\end{aligned}
$$

Updating the connection weights in each iteration is performed as follows:

$$
\Delta w_{ij}(t+1) = \alpha\frac{\partial E_d}{\partial w_{ij}} + \eta\Delta w_{ij}(t) \qquad (6)
$$

The partial derivative $\partial E_d/\partial w_{ij}$ with respect to y_k^A and y_k^B can be described as:

$$
\begin{aligned}
\frac{\partial E_d}{\partial w_{ij}} &= \sum_{k=1}^{n}\{\frac{\partial E_d}{\partial y_k^A}\frac{\partial y_k^A}{\partial w_{ij}} + \frac{\partial E_d}{\partial y_k^B}\frac{\partial y_k^B}{\partial w_{ij}}\} \qquad (7) \\
&= (s_{AB} - d_{AB})\sum_{k=1}^{n}(y_k^A - y_k^B)\{\frac{\partial y_k^A}{\partial w_{ij}} - \frac{\partial y_k^B}{\partial w_{ij}}\}
\end{aligned}
$$

The learning parameter α and the momentum η are empirically chosen for the stable convergence.

The network does not need absolute coordinates as the desired output, only the desired distance between the input objects. It should be noted that the structure and initial condition of network A and network B, and the definition of distance are identical, while the number of cells in the output layers of network A and network B can be set arbitrarily, depending upon the structure of the required output space. In the present work, we adopt a two-dimensional output space and Euclidean distance for the distance metric in Unit C.

3. EXPERIMENTS

3.1. Iris Classification problem

In this experiment, the proposed model is applied to Iris Classification problem [10]. An Iris flower that has 4 attributes (length and width of the flower's sepal and petal) is classified into one of three classes (Iris-Setosa, Iris-Versicolor, Iris-Virginica). One class is linearly separable from the other two classes; the latter are not linearly separable from each other.

In the learning phase, a pair from 75 instances (contains evenly three classes) and the teacher signal s_{AB} as described below are given for training the network.

$$
s_{AB} = \begin{cases} 0 & \\ \rho & (\rho > 0) \end{cases} \qquad (8)
$$

where ρ represents the scaling parameter that is the scale of the distance between two classes, A and B. If a training pair is chosen from the same class, the teacher signal is set to 0.0. While, if they are chosen from different classes, the value is set to ρ. In this experiment, ρ is set to 0.6. The number of cells in the input, hidden and output layers are 4, 6 and 2, respectively. The learning rate is set to 0.3, and the momentum is set to 0.2. After training the network, the model holds a non-linear mapping from 4 attributes of Iris flower to 2 dimensional output space.

Figure 2. shows the mapped objects in the 2 dimensional output space after the 10,000 iterations. The objects are clearly classified into three classes, and the center point of each class (called *spot*) forms a regular triangle.

In order to verify the generalization ability, we experimented with test data that are not used for the training. 75 test data are mapped onto the output space as illustrated in Figure 3. We displayed the Voronoi diagram with regard to the spot of each class for giving an indication of the class. The Voronoi diagram has the property that for each spot, every point in the region around that spot is closer to that spot than to any of the other spots. Although the network does not give the identification of the class, the aspect of classification can be displayed with the aid of the diagram.

3.2. Facial expression classification

The proposed model is applied to facial expression classification. Figure 4 shows a line drawing model [11] of the facial expression and five facial images of typical emotional categories such as *happiness, anger, fear, surprise* and *sadness*. The nine parameters indicated by P_i are used to generate an image of facial expression.

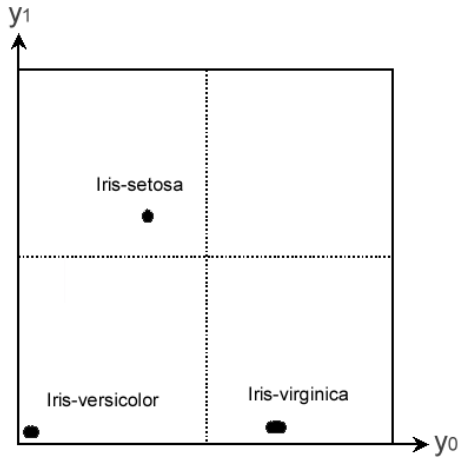

Figure 2: Performance of Iris Classification

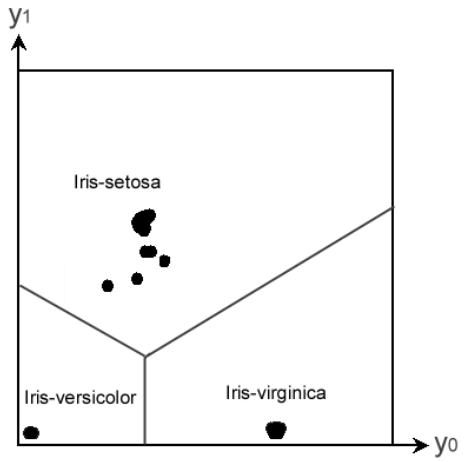

Figure 3: Generalization ability of the model in Iris Classification

The movement of each parameter which is allocated to the eyebrows, eyes and mouth changes the intensity of facial expression. Each face has line symmetry with respect to the vertical central line on which the nose is located and fixed. Each feature point is consistently connected to the others, introducing the spline interpolation. Using this line drawing model, five facial images of typical emotional categories are then acquired from images drawn by 36 subjects in the previous work [11]. Subjects moved the nine feature points in order to create the desired emotional face. Thus, an object has 9 attributes and is classified into one of five classes.

In the learning phase, a pair from 50 instances (contains evenly five classes) is given for the training. The scaling parameter ρ is set to 0.6. The number of cells in the input, hidden and output layers are 9, 12 and 2,

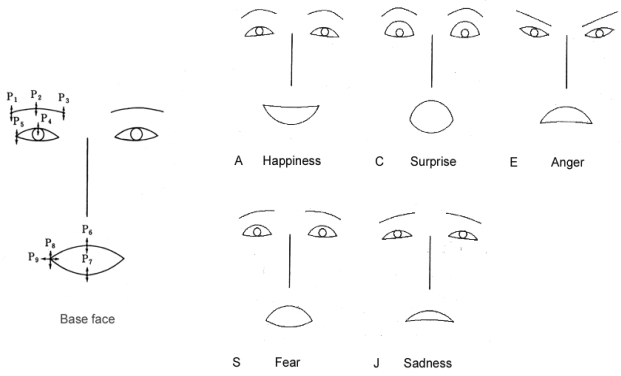

Figure 4: A line drawing model of facial expression and five facial images of typical emotional categories

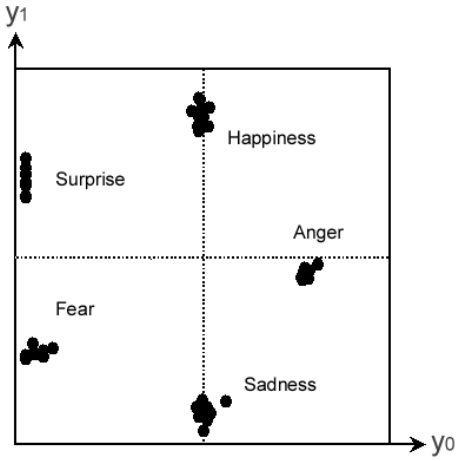

Figure 5: Performance of facial expression classification

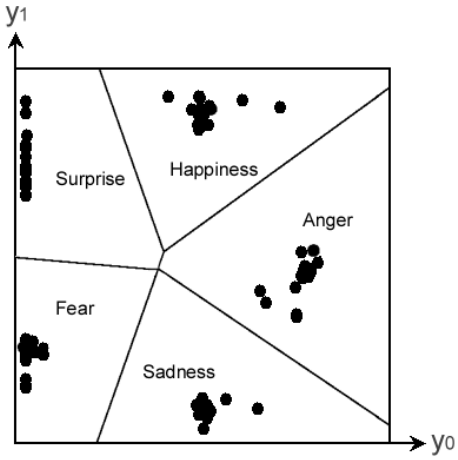

Figure 6: Generalization ability of the model in facial expression classification

respectively. The learning rate is set to 0.3, and the momentum is set to 0.2. Figure 4. shows the mapped objects in the 2 dimensional output space after the 30,000 iterations. It can be seen that the objects are classified into five classes, and the spot of each class forms a deformed pentagon. In case that the number of spot becomes more than four, every spot that have a constant distance to other spots cannot be arranged on 2 dimensional space. Therefore, the training of the network is terminated when the number of iteration steps becomes 30,000 in this experiment.

Figure 6. illustrates 50 test data that are mapped onto the output space. The Voronoi diagram is also displayed as well as the previous experiment. The result proves that the trained network holds a non-linear mapping from 9 attributes of facial expression model to 2 dimensional output description space.

4. DISCUSSION AND CONCLUSIONS

We described a method of multiclass classification by utilizing Distance Mapping Learning network. We have introduced a method to obtain a nonlinear mapping between input data and the description parameter by using a neural network model in which the desired distances between the input pair are given as a teacher signal set. The novel aspect of the proposed method is that the number of class is not needed for the classification.

With regard to the structure of the proposed neural network, the distance-based learning has some distinguishing properties compared with conventional multilayer perceptrons. The convergence is relatively stable, but the converged arrangement often depends upon the initial conditions of weights due to the characteristics of the network.

We have been considering that the proposed method can be applied to the categorical perception of facial expressions [12]. The analytical evaluation of the network is one of the future issues. Although the Euclidean distance is given as a teacher signal in the present work, the algorithm allows any other distance metric such as Hamming distance, city-block distance and Minkowsky distance. The influence of non-Euclidean distance functions will be also considered.

This approach utilizing non-linear mapping delivered good results in terms of fitting rate compared with the conventional statistical analysis. In addition, new objects which are not used in the training in the network can also be evaluated by the generalization ability of the network. We consider that the proposed method can extend the scheme of the multidimensional scaling method.

Acknowledgements: The presented work was partly supported by the JSPS Research for the Future Program in the Area of KANSEI (Intuitive & Affective) Human Interfaces, and by a grant-in-aid from the Japanese Ministry of Education, Science, Sports and Culture (A-11305021).

5. REFERENCES

[1] Suzuki, K., Yamada, H. and Hashimoto, S., Interrelating physical feature of facial expression and its impression, *in Proc. of IEEE/INNS International Conference on Neural Networks 2001*, Washington D.C., USA, pp. 1864-1869 (2001)

[2] Vapnik, V.N., *The nature of statistical learning theory*, Springer Verlag, New York (2000)

[3] Tax, D.M.J. and Duin, R.P.W., Combining one-class classifiers, *in Proc. of the Second International Workshop Multiple Classifier systems*, Vol. 2096, Springer Verlag, Berlin, pp. 299-308 (2001)

[4] Squire, D., Learning a similarity-based distance measure for image database organization from human partitionings of an image set, *in Proc. of Fourth IEEE Workshop on Applications of Computer Vision*, Princeton, USA, pp. 88-93 (1998)

[5] Duch, W., Neural minimal distance methods, *in Proc. of Third Conference on Neural Networks and Their Applications*, Kule, pp. 183-188 (1997)

[6] Kruskal, J.B. and Wish, M., *Multidimensional Scaling*, Sage Publications, Calif., USA (1978)

[7] Rumelhart, D.E., Learning internal representation by error propagation, *Parallel Distributed Processing*, Vol. 1, pp. 318-362, MIT Press (1986)

[8] Geman, S., Bienenstock, E. and Doursat, R., Neural networks and the bias/variance dilemma, *Neural Computation*, 4, pp. 1-58 (1992)

[9] Sammon, J.W., Nonlinear mapping algorithm for data structure analysis, *IEEE Trans. Computer*, vol. C-18, pp. 401-409 (1969)

[10] Fisher, R.A, The use of multiple measurements in taxonomic problems, *Annual Eugenics*, 7, Part II, 179-188 (1936)

[11] Yamada, H., Visual information for categorizing facial expression of emotion, *Japan Psychology Rev.*, 35, pp.172-181 (1993)

[12] Etcoff, N. and Magee, J., Categorical perception of facial expressions, *Cognition*, 44, 227-240 (1992)