Redundancy Reduction of Features by Independent Component Analysis

Wataru Kusanagi, Manabu Kotani and Seiichi Ozawa

Faculty of Engineering, Kobe University, Japan

kusanagi@nn.cs.kobe-u.ac.jp, kotani@cs.kobe-u.ac.jp, ozawa@eedept.kobe-u.ac.jp

Abstract

We propose the method to reduce the redundancy between features extracted from images by the infomax algorithm. The method consisted of the addition of noise and the feature selection based on kurtosis. The addition of noise to images contributes to the reduction of redundancy and make it easy to set a criterion of feature selection. We adopt ETL-4 database sets of hand-written Japanese Hiragana characters, and performed the recognition experiment with the extracted features. Results of the recognition experiment showed that it was possible to reduce the redundancy of features by the proposed method.

1 Introduction

Principal component analysis (PCA) is a classical method for multivariate data analysis and it decorrelates signals based on the second order statistics. On the other hand, Independent component analysis (ICA) is useful generalization of PCA and can make signals be mutually independent based on higher order statistics.

Some ICA algorithms have been proposed in neural networks[1][2][3]. For example, Bell and Sejnowski proposed the infomax algorithm[1] which maximizes the information between inputs and outputs. Hyvärinen and Oja proposed the fastICA algorithm[2]. The fastICA involves PCA as preprocessing and can find independent components by maximization or minimization of the kurtosis.

Recently, there are some studies about applications of ICA to signal processing, such as image data. Olshausen and Field[4] showed the characteristics of basis functions extracted from natural scenes by sparse coding algorithm. Bell and Sejnowski[5] also showed similar results with the infomax algorithm. Bartlett

and Sejnowski[6] showed that the features extracted from face images by the infomax algorithm was better than features by PCA regarding to recognition accuracy. Ozawa et al.[7][8] applied the fastICA algorithm to Japanese hand-written Hiragana character, and performed recognition experiments with the extracted features. They showed the effectiveness of the addition of noise to the training data to obtain higher recognition performance, and discussed the possibility for redundancy reduction of features. The redundancy is defined as the similarity of basis functions.

However, it is considered that PCA as preprocessing of the fastICA algorithm badly affects on the result. The reason is that PCA decorrelates signals including noise. Therefore, Ikeda[9] proposed factor analysis instead of PCA as preprocessing.

The purpose of this paper is to examine redundancy reduction of features extracted by the infomax algorithm, which does not need PCA. We, here, propose the method of redundancy reduction consisted of the addition of noise and the feature selection based on kurtosis. Furthermore, we evaluate the effectiveness of the proposed method with experiment for Japanese hand-written Hiragana character.

2 Outline of Infomax Algorithm and Basis Functions

2.1 Infomax algorithm

The fundamental assumption made on the source vector is that it has individual components that are mutually independent,

$$X = AS, \quad (1)$$

where X is the observed signal (n -dimensional vector), S is the independent source (m -dimensional vector) and A is the mixing matrix.

From an information theoretic perspective, the infomax algorithm maximizes the information transferred through the units of network. The optimal weight is adjusted by stochastic gradient ascent on the entropy of the output with respect to weight.

The mutual information between the output and the input, $I(Y, X)$, is defined as

$$I(Y, X) = H(Y) - H(Y|X), \quad (2)$$

where $H(Y)$ is an entropy of the output, Y , and $H(Y|X)$ is a conditional entropy of Y given X . This equation is differentiated with respect to a weight, w ,

$$\frac{\partial}{\partial w} I(Y, X) = \frac{\partial}{\partial w} H(Y). \quad (3)$$

Since $H(Y|X)$ does not depend on w , $\frac{\partial}{\partial w} H(Y|X) = 0$. Thus the maximization of the mutual information between the output and the input leads to the maximization of the entropy of the output. Furthermore, the maximization of the entropy of the output minimizes the mutual information between the outputs. As a result, the outputs of the network are mutually statistical independent. The delta rule for the weight is defined as

$$\Delta W = [W^T]^{-1} + (1 - 2Y)X^T \quad (4)$$

$$W_{new} = W_{old} + \alpha \Delta W \quad (5)$$

$$Y = g(S), \quad S = WX \quad (6)$$

where α is the learning rate, $g(\cdot)$ is a sigmoid function, $g(u) = 1/(1 + e^{-u})$, and W is a weight matrix.

2.2 Basis functions

Fig.1 shows the concept of ICA. The observed signal is linear transformation by A from the unknown source signal. The basis function is defined as column of A . We can obtain the separating matrix, W , with the infomax algorithm. If the separated source signal is the unknown source signal, the mixing matrix would be the inverse of the separating matrix, that is, $A = W^{-1}$. Therefore, the basis function is column vector of W^{-1} .

Figure 1: The concept of ICA and the basis function.

3 Proposed Method and Recognition

3.1 Addition of noise

The redundancy between features may generally cause the deterioration of the recognition accuracy. Therefore, we attempt to reduce the redundancy of features by adding gaussian noise to training pattern. The gaussian noise is randomly generated from 0 to 15, and is added to the original images according to a probability, P . We call these images the “noisy images”. The noisy images were generated with various P . Examples of the noisy images are shown in Fig.2.

Figure 2: Examples of the noisy images for various P

3.2 Feature selection

We, here, describe the selection method of independent components (ICs), that is, features. We adopt the kurtosis of ICs as a criterion of feature selection. The kurtosis is widely used as a measure of a distribution, and it is zero for normal distribution. The kurtosis, K_i , is defined as

$$K_i = \frac{E\{s_i^4\}}{(E\{s_i^2\})^2} - 3, \quad (7)$$

where s_i denotes the estimated independent component and $E\{\cdot\}$ is the expectation.

3.3 Recognition method

To evaluate the effectiveness of the proposed method, we calculated the recognition accuracy with the k -Nearest Neighbor (k -NN) method ($k=3$). The k -NN is the majority decision algorithm which judges the class from k nearest candidates selected by a measure. Let a_i denotes the i -th template data, and let b_j denotes the j -th test data. We defined the similarity measure, $Q(i, j)$, between these data,

$$Q(i, j) = \frac{a_i \cdot b_j}{\|a_i\| \|b_j\|}. \quad (8)$$

The recognition accuracy is calculated based on the k -NN with the similarity measure.

4 Simulations and Discussion

4.1 Description of data

We adopt ETL-4 database sets of hand-written Japanese Hiragana characters as the inputs of the network. Since data set consists of 46 Hiragana character for 100 people, the number of image data is 4600 and these data are divided into halves, training data and test data. Each character image is composed of 76×72 pixel (PBM format).

These images were preprocessed through normalization of position, size and luminance. They were converted to 16×16 pixel of 16-bit gray-scale, and the rows of these images were concatenated to produce 256 dimensional vector.

4.2 Effectiveness of noise

We trained the fully connected single layer feedforward network by the infomax algorithm. We used two types of data, the original image and the noisy image. The matrix, W , was initialized to the identity, and updated according to equation (5) for 500 iterations with changing the learning rate each 100 iterations.

After the training, we got 256 basis functions for both types of data. Fig.3 (a) shows all the basis functions for the original image. There are some basis functions resembled each other and most of basis functions are the character bases. Here, "character basis" indicates that the basis function seems to correspond to a part of character or the whole of character.

(a) In the case of the original image of "あ"

(b) In the case of the noisy image

Figure 4: ICs from 16 basis functions shown in the first row of Fig.3 (b). Abscissa means basis function number. Ordinate means value of ICs.

Fig.3 (b) shows all the basis functions for the noisy image. The number of character bases dramatically decreased and the number of noise bases increased comparing to Fig.3 (a). "Noise basis" indicates that the basis function seems to correspond to a random pattern. The random pattern means the basis function like leftmost figure of top in Fig.3 (b).

Fig.4 shows ICs of 16 basis functions shown in the first row of Fig.3 (b). Since the 6-th basis function resembled the input of "あ", the output of the 6-th basis function was larger than any other basis functions. On the other hand, when the input was the noisy image, the output of the 6-th basis function was not large. Hence, the character bases like the 6-th basis function responded to not the noisy image, but the character image.

Fig.5 shows the kurtosis of ICs for the original image and the noisy image. The kurtosis of ICs for the original image gradually decreased. On the contrary, the kurtosis of ICs for the noisy image sharply decreased up to about 40 components. This result means that it is relatively easy to set the criterion of feature selection.

Table 1 shows the relation between the number of noisy images (training data) and the number of character bases. We adopted the variance of basis functions to discriminate character basis and noise basis in this table. The variance of character bases were

(a) For the original image

(b) For the noisy image

Figure 3: Basis functions for the image. Black denotes negative value and white denotes positive value. Shading denotes the magnitude of value.

Figure 5: Kurtosis of ICs. Ordinate denotes the magnitude of kurtosis and abscissa denotes the component number which are ordered by the magnitude of kurtosis. The magnification of ordinate is not identical to both figures.

usually larger than those of noise bases. When the number of noisy images were large, the number of character bases was small. However, the number of noisy images exceeded 1000, the number of character bases hardly changed. From this result, it was important for the reduction of redundancy to train the network with many noisy images.

Table 1: The relation between the number of noisy images and the number of character bases.

The number of noisy images	The number of character bases
100	132
200	83
500	69
1000	56
2300	54
10000	53

4.3 Results of feature selection

Table 2 shows the recognition results with ICs of basis functions selected by kurtosis. As you can see in Table 2, we got the best performance with 34 basis functions.

These basis functions are shown in Fig.6. All the basis functions in Fig.6 were character bases. This

Table 2: Recognition results for selected basis functions[%]. “No noise” means 256 bases trained for the original images.

The number of selected bases	Recognition accuracy
10	34.0
20	66.7
34	83.7
50	79.8
100	78.0
256	72.8
No noise	55.3

Figure 6: Basis functions selected by kurtosis. These figures were ordered by the magnitude of kurtosis.

means that kurtosis of ICs were large for the character bases. It may depend on the number of noisy images. The number of original images was much smaller than the number of noisy images. The outputs of character bases were almost small for many noisy images, but the outputs were large for a few original images. Therefore, the kurtosis of character bases became large.

5 Conclusion

We discussed the redundancy reduction of features of the image extracted by the infomax algorithm. The results were obtained as follows:

- The addition of noise to image data was effective in discrimination of character bases and noise bases.
- The feature was suitably selected by the kurtosis.

From these results, the proposed method was useful to reduce redundancy of features.

References

- [1] A.J.Bell and T.J.Sejnowski: An information maximization approach to blind separation and blind deconvolution, *Neural computation*, **7**, 1129/1159 (1995)
- [2] A.Hyvärinen and E.Oja: A fast fixed-point algorithm for independent component analysis, *Neural computation*, **9**, 1483/1492 (1997)
- [3] S.Amari, A.Cichocki and H.Yang: A new learning algorithm for blind signal separation, *Neural information processing*, **8**, 757/763 M.I.T Press.(1995)
- [4] B.Olshausen and D.Field: Wavelet-like receptive fields emerge from a network that learns sparse codes for natural images, *Nature*, **381**, 607/609 (1996)
- [5] A.J.Bell and T.J.Sejnowski: The 'Independent components' of natural scenes are edge filters, *Vision research*, **37**, 3327/3338 (1997)
- [6] M.S.Bartlett and T.J.Sejnowski: Independent component representations for face recognition, *Proc. of SPIE*, 528/539 (1998)
- [7] S.Ozawa, T.Tsujimoto, M.Kotani and N.Baba: Application of independent component analysis to hand-written Japanese character recognition, *Proc. of IJCNN*, CD-ROM (1999)
- [8] S.Ozawa and M.Kotani: A study of feature extraction and selection using independent component analysis, *Proc. of ICONIP*, (2000) in press
- [9] S.Ikeda: ICA on noisy data: a factor analysis approach, *Advances in independent component analysis*, 201/215 (2000)