Feature Extraction Using ICA

Nojun Kwak, Chong-Ho Choi, and Jin Young Choi

School of Electrical Eng., ASRI, Seoul National University
San 56-1, Shinlim-dong, Kwanak-ku, Seoul 151-742, Korea
{triplea,chchoi}@csl.snu.ac.kr, jychoi@neuro.snu.ac.kr

Abstract. In manipulating data such as in supervised learning, we often extract new features from original features for the purpose of reducing the dimensions of feature space and achieving better performances. In this paper, we propose a new feature extraction algorithm using independent component analysis (ICA) for classification problems. By using ICA in solving supervised classification problems, we can get new features which are made as independent from each other as possible and also convey the output information faithfully. Using the new features along with the conventional feature selection algorithms, we can greatly reduce the dimension of feature space without degrading the performance of classifying systems.

1 Introduction

In supervised learning, one is given an array of attributes to predict the target value or output class. These attributes are called features, and there may exist irrelevant or redundant features to complicate the learning process, thus leading to wrong prediction. Even in the case when the features presented contain enough information about the output class, they may not predict the output correctly because the dimension of feature space is so large that it requires numerous instances to figure out the relationship. This problem is commonly referred to as the 'curse of dimensionality' [1], and can be avoided by selecting only the relevant features or extracting new features containing the maximal information about the output from the original ones. The former methodology is called feature selection or subset selection, while the latter is named feature transformation whose variants are feature extraction and feature construction [2].

This paper considers the feature extraction problem since it often results in improved performance, especially when small dimensions are required, thus can be directly applied to make oblique decision borders as optimal classifiers like support vector machines do.

Recently, in neural networks and signal processing societies, independent component analysis (ICA), which was devised for blind source separation problems, has received a great deal of attention because of its potential applications in various areas. Bell and Sejnowski have developed an unsupervised learning algorithm performing ICA based on entropy maximization in a single-layer feedforward neural network [3]. Though this ICA method can be directly applied

G. Dorffner, H. Bischof, and K. Hornik (Eds.): ICANN 2001, LNCS 2130, pp. 568–573, 2001.

to transform the original feature set into a new set of features as in [4]-[5], the performance cannot be guaranteed, since it does not utilize output class information like the PCA. These ICA-based feature extraction methods just focus on the visual representation of the data [6].

In this paper, we propose a new ICA-based feature extraction algorithm which utilizes the output class information in addition to having the advantages of the original ICA method. This method is well-suited for classification problems in the aspect of constructing new features that are strongly related to output class. By combining the proposed algorithm with existing feature selection methods, we can greatly reduce the dimension of feature space while improving classification performance.

This paper is organized as follows. In Section 2, we propose a new feature extraction algorithm. In Section 3, we give some simulation results showing the advantages of the proposed algorithm. Conclusions follow in Section 4.

2 Feature Extraction Based on ICA

The main idea of the proposed feature selection algorithm is very simple. In applying ICA to feature extraction, we include output class information in addition to input features.

ICA is classified as unsupervised learning because it outputs a set of maximal independent component vectors. This unsupervised method by nature is related to the input distribution, but cannot guarantee good performance in classification problems though the resulting independent vectors may find some application in such areas as visualization [4] and source separation [3]. Instead of using only the input vectors, we treat the output class information as one of the input features and append it to the inputs for ICA. To illustrate the advantage of this method, consider the following simple problem.

Suppose we have two input features x_1 and x_2 uniformly distributed on [-1,1] for a binary classification, and the output class y is determined as follows:

$$y = \begin{cases} 0 & \text{if } x_1 + x_2 < 0 \\ 1 & \text{if } x_1 + x_2 \geq 0 \end{cases}$$

Plotting this problem on a three dimensional space of (x_1, x_2, y) leads to Fig. 1 where the class information as well as the input features correspond to each axis respectively. The data points are located in the shaded areas in this problem. As can be seen in the figure, this problem is linearly separable and we can easily pick out $x_1 + x_2$ as an important feature.

Performing ICA on a set of data provides us with N-dimensional vectors indicating the directions of independent components on feature space. If we take the output class as an additional input feature, it forms an *(N+1)*-dimensional space, and applying ICA to this dataset gives us *(N+1)*-dimensional vectors. These vectors are made to well describe the data distribution over the extended *(N+1)*-dimensional feature space and can be useful for classification problems

if we project the vectors onto the original N-dimensional feature space. In the problem above, the vector shown as a thick arrow in Fig. 1 will surely indicate the apparent discontinuity in the data distribution over the extended feature space. This vector can provide us with a great deal of information about the problem and projecting it onto the (x_1, x_2) feature space gives us a new feature relevant to the output class.

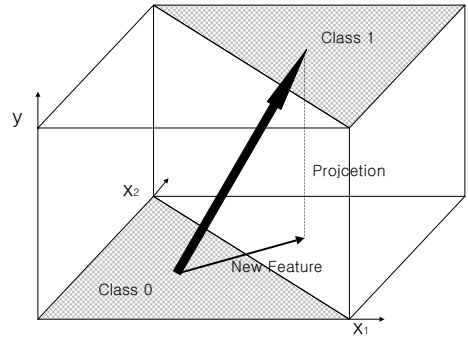

Fig. 1. Concept of ICA based feature extraction

Based on this idea, a feature extraction algorithm for classification problem is proposed as follows:

1. **Preparing data**
   (a) With N input features $\mathbf{x}= [x_1, \cdots, x_N]^T$ and one output class c, compose a new *(N+1)*-dimensional input dataset $\{(\mathbf{x}^T, c)^T\}$.
   (b) Normalize each feature f_i by $(f_i - m_i)/2\sigma_i$ where m_i and σ_i are the mean and the standard deviation of f_i, respectively.
2. **Performing ICA**
   Apply ICA to the new dataset, and store the resulting weight matrix $\mathbf{W}$ of dimension *(N+1)* $\times$ *(N+1)*.
3. **Shrinking small weights**
   (a) For each *N+1* independent row vector W_i of $\mathbf{W}$, compute the absolute mean $a_i = \frac{1}{N+1}\sum_{j=1}^{N+1} |w_{ij}|$.
   (b) For all w_{ij} in $\mathbf{W}$, if $|w_{ij}| < \alpha \cdot a_i$, then shrink $|w_{ij}|$ to zero. Here, α is a small positive number [1].
4. **Extracting feature candidates**
   (a) For each weight vector W_i, project it onto the original input feature space, i.e., delete weights $w_{ic} (= w_{i,N+1})$ corresponding to the output class, resulting in new N-dimensional row weight vector W_i'.
   (b) By mutiplying new weight matrix $\mathbf{W}'$ of dimension *(N+1)* $\times N$ to the original input data $\mathbf{x}$, construct a *(N+1)*-dimensional vector whose components f_i's are new feature candidates.

[1] In the simulations, we set $\alpha = 0.2$ or 0.5.

5. **Deleting unappropriate features**
   (a) Form a set of feature candidates $F = \{f_i = W_i'\mathbf{x}, \quad i \in 1 \cdots N{+}1\}$. Then set $F_S = F$.
   (b) For each feature candidate f_i, if the corresponding weight for class w_{ic} is zero, then exclude it from F_S.
   (c) For each feature candidate f_i, if corresponding weights $w_{ij} = 0$ for all $j \in 1 \cdots N$, then exclude f_i from F_S.
   (d) Resulting F_S contains final N' extracted features.

Step 3 is useful for excluding the effect of features which do not contribute significantly to the new features. The effect of the exclusion is controlled by a suitable choice of α.

When we applied the algorithm to the problem illustrated in Fig. 1, with 1,000 samples, we obtained 3 raw weight vectors $W = \{(w_1, w_2, w_c)\} = \{(7.55, 1.74, -2.24), (2.10, 7.88, -2.48), (-0.75, -0.78, 3.53)\}$. Note that the last one leads to a new feature $-0.75x_1 - 0.78x_2$ which is nearly equivalent to the optimal feature $x_1 + x_2$.

Though we consider only two-class problems in the algorithm, the algorithm can easily be extended to multi-class problems which can be set up as a series of two-class problems, where in each two-class problem, one class is distinguished from the others.

3 Experimental Results

In this section we apply the proposed algorithm on UCI dataset [8] and will present some experimental results which show the characteristics of the proposed algorithm. For the classification problems in this section, to show the effectiveness of the proposed algorithm, we selected appropriate numbers of original features and extracted features by MIFS-U [9] and compared the classification performance. We tested the proposed algorithm for several problems, but present the results of only two cases for space limitation. The other cases show the similar enhancement in performance.

Wiscosin Breast Cancer Diagonosis. The dataset consists of nine numerical attributes and two classes which are benign and malignant. It contains 699 instances with 458 benign and 241 malignant. There are 16 missing values and in our experiment we replaced these with average values of corresponding attributes.

The proposed algorithm extracted seven new features. We selected appropriate numbers of original features and extracted features by MIFS-U [9] and trained the data with C4.5 [10] and multilayer perceptron with standard backpropagation(BP). MLP was trained for 500 iterations with three hidden nodes, zero momentum and learning rate of 0.2. For verification, 10-fold cross validation is used. The classification results are shown in Table 1 and the numbers in parentheses are the tree sizes of C4.5.

Table 1. Classification performance for Breast Cancer data (%)

no. of features selected	Original Features (C4.5/MLP)	Extracted Features (C4.5/MLP)
1	91.13(3)/92.42	95.42(3)/95.42
2	94.71(13)/96.28	95.57(9)/95.42
3	95.85(23)/96.14	95.71(11)/96.57
9	94.56(31)/96.42	–

Table 2. Classification performance for Chess data (%)

no. of features selected	Original Features (MLP)	Extracted Features (MLP)
1	67.43	96.23
2	74.52	97.98
3	91.08	98.80
36	98.54	–

It shows that with only one extracted feature, we can get near the maximum classification performance that can be achieved with at least two or three original features. Also, note that the tree size does not expand much for the case of extracted features, which is desirable for generating simple rule.

Chess End-Game Data. There are 36 attributes with two classes, *white-can-win ('won')/white-cannot-win ('nowin')*. There are 3,196 instances of which 1,669 are *'won'* and the other 1,527 are *'nowin'* with no missing values.

In conducting feature extraction algorithm, we set $\alpha = 0.5$ and obtained six additional features. After feature selection process with MIFS-U, we trained the data for 500 iterations by the standard BP with 0.2 learning rate and zero momentum. We divided the data into two sets, one half for training and the other half for testing. Table 2 shows the classification performances on the test data. The proposed method shows significant improvement in performance over the original features.

4 Conclusions

In this paper, we have proposed an algorithm for feature extraction. The proposed algorithm is based on ICA and have the ability to generate appropriate features for classification problems.

As the name indicates, ICA is characterized by its ability of identifying statistically independent components based on input distribution. Although ICA can be directly used for feature extraction, it does not generate useful information because of its unsupervised learning nature. In the proposed algorithm, we added class information in training ICA. The added class information plays a critical role in the extraction of useful features for classification. For problems

with original N features, we get *N+1* weight vectors in *(N+1)*-dimensional extended feature space. Projecting them onto original feature space leads to new feature candidates and by deleting unappraite ones which have little correlation with the class information, we can extract better features.

Since it uses the standard feed-forward structure and learning algorithm of ICA, it is easy to implement and train. If we use the proposed algorithm with the conventional feature selection algorithm, we can get simpler classifier with better performance. Experimental results for several datasets show that the proposed algorithm generates good features which outperforms the original features for classification problems. Though the algorithm is proposed for two-class problems, it is also useful for multi-class problems with simple modification.

Acknowledgments

This project is partially supported by the Brain Science and Engineering Program of the Korea Ministry of Science and Technology and the Brain Korea 21 Program of the Korea Ministry of Education.

References

1. V. Cherkassky and F. Mulier, *Learning from Data : Concepts, Theory, and Methods*, pp. 60-65, John Wiley & Sons, 1998.
2. H. Liu and H. Motoda, "Less is more," *Feature Extraction Construction and Selection*, pp. 3-11, Boston: Kluwer Academic Publishers, 1998.
3. A.J. Bell and T.J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol 7, pp. 1129-1159, 1995.
4. A. Hyvarinen, E. Oja, P. Hoyer and J. Hurri, "Image feature extraction by sparse coding and independent component analysis," *Fourteenth International Conference on Pattern Recognition*, vol. 2, pp. 1268 -1273, 1998.
5. A.D. Back and T.P. Trappenberg, "Input variable selection using independent component analysis," *The 1999 Int'l Joint Conf. on Neural Networks*, July 1999.
6. H.H. Yang and J. Moody, "Data visualization and feature selection: new algorithms for nongaussian data," *Advances in Neural Information Processing Systems*, vol 12, 2000.
7. T.M. Cover, and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.
8. P.M. Murphy and D.W. Aha, UCI repository of machine learning databases, 1994. For information contact ml-repository@ics.uci.edu or http://www.cs.toronto.edu/∼delve/.
9. N. Kwak and C.-H. Choi, "Improved mutual information feature selector for neural networks in supervised learning," *The 1999 Int'l Joint Conf. on Neural Networks*, July 1999.
10. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA., 1993.