A Hybrid Nonlinear-Discriminant Analysis Feature Projection Technique

Rami N. Khushaba, Ahmed Al-Ani, Adel Al-Jumaily, and Hung T. Nguyen

Faculty of Engineering, University of Technology, Sydney,
P.O.Box: 123, Broadway 2007, Sydney-Australia
{rkhushab, ahmed, adel, Hung.Nguyen}@eng.uts.edu.au

Abstract. Feature set dimensionality reduction via Discriminant Analysis (DA) is one of the most sought after approaches in many applications. In this paper, a novel nonlinear DA technique is presented based on a hybrid of Artificial Neural Networks (ANN) and the Uncorrelated Linear Discriminant Analysis (ULDA). Although dimensionality reduction via ULDA can present a set of statistically uncorrelated features, but similar to the existing DA's it assumes that the original data set is linearly separable, which is not the case with most real world problems. In order to overcome this problem, a one layer feed-forward ANN trained with a Differential Evolution (DE) optimization technique is combined with ULDA to implement a nonlinear feature projection technique. This combination acts as nonlinear discriminant analysis. The proposed approach is validated on a Brain Computer Interface (BCI) problem and compared with other techniques.

Key words: Feature projection, Nonlinear Discriminant Analysis

1 Introduction

Techniques that can introduce low-dimensional feature representation with enhanced discriminatory power are of paramount importance, because of the curse of dimensionality. Many methods have been proposed for dimensionality reduction and feature extraction, such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA)[1]. LDA, unlike other methods, is particularly suitable for solving classification problems. It aims to maximize the ratio of the determinant of the between-class scatter matrix of the projected samples to the determinant of the within-class scatter matrix of the projected samples. However, there are many problems with the classical LDA [2]. Classical LDA requires the scatter matrices to be nonsingular and fails when the scatter matrices are singular. Another limitation is that it pays no attention to the decorrelation of the data.

Uncorrelated features, are desirable in many applications, because they contain minimum redundancy. Recently, Jin et al [3] proposed the uncorrelated Linear Discriminant Analysis (ULDA), that can extract feature vectors having statistically uncorrelated attributes. Although being successful and enhanced version of the classical LDA, ULDA lacks the capacity to capture a nonlinearly clustered structure in the data because of its linear nature. Motivated by extracting nonlinear uncorrelated features, there were many attempts to solve this problem by employing kernel based approaches [4–6]. Due to the computational complexity associated with the kernel based approaches, especially for very large datasets, then it is a tempting task to search for alternative methods to perform the nonlinear mapping task.

In this paper, a two layer projection technique is presented. In the first layer a feed forward neural network layer is utilized as a nonlinear mapping stage for which the parameters are optimized with Differential Evolution (DE) [7]. The aim of using this layer is to nonlinearly map the input space to a high-dimensional feature space where different classes of objects are supposed to be linearly separable. This will prepare the scene for the second stage for further reducing the dimensionality by utilizing the ULDA, thus performing the linear mapping into a set of uncorrelated features.

This paper is organized a follows: Section 2 introduces the proposed projection technique and the DE optimization. The experiments and practical results are given in section 3. Finally a conclusion is given in Section 4.

2 Nonlinear Discriminant Analysis based Feature projection

An artificial neural network (ANN) model is an information processing paradigm consisting of many nonlinear computational elements operating in parallel and arranged in patterns reminiscent of biological neural nets [8]. Several studies were made to illustrate that ANN can perform well for pattern classification [9, 10]. These studies proved that within Multi-Layer Perceptrons (MLP), each layer of weights can be thought of as performing projections that try to separate as best as possible the different classes, so they can be linearly separable by the cells in the last layer. All of these studies suggest that the MLP actually consist of two projections: A Non-linear projection from input-to-hidden and from each hidden-to-hidden layer and a second projection being linear from the final hidden-to-output layer.

Studies in this field can be decomposed into two parts. The first focused on studies to enhance the functionality of multilayer feed-forward neural networks performing the nonlinear discriminant analysis [11, 12]. The second trend focused on Fisher's Discriminant Analysis itself as a statistical technique mixed with kernel functions to perform the nonlinear mapping [4–6]. Although many of these studies does actually perform well as a nonlinear discriminant analysis tool, but up to the authors knowledge there were no studies that combined neural networks with the statistical discriminant analysis for the specific purpose of feature projection. Thus the main focus of this paper is to combine these two techniques and compare the performance of the proposed nonlinear method with the existing techniques.

The basic structure proposed in this paper is shown in Fig.1 sharing similar architecture with the MLPs. However, we replaced the final linear layer of MLP with a ULDA implementation, and hence a linear discriminant analysis layer is incorporated. In addition, a Differential Evolution (DE) optimization technqiue is used to evolve the weights between the input and hidden layer. Thus, rather than optimizing the weights of many hidden layers, only the weights of the first hidden layer are optimized. Then ULDA acts upon the output of this hidden layer to perform the rest of the projection task.

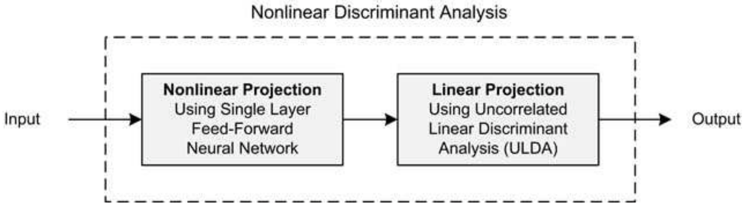

Nonlinear Discriminant Analysis
Input
Nonlinear Projection Using Single Layer Feed-Forward Neural Network
Linear Projection Using Uncorrelated Linear Discriminant Analysis (ULDA)
Output

Fig. 1. Block Diagram of the proposed projection technique

2.1 Differential Evolution based Weight Optimization

Differential Evolution (DE) is simple, parallel, direct search optimization method having good convergence, and fast implementation properties[7]. The crucial idea behind DE is based on generating trial parameter vectors by adding the weighted difference of randomly chosen two population members ($X_{r1,g}$ and $X_{r2,g}$) to a third member ($X_{r0,g}$) to create a mutant vector, $V_{i,g}$ from the current generation g, as shown in Eq.1 below:

$$V_{i,g} = X_{r0,g} + F \times (X_{r1,g} - X_{r2,g}) \tag{1}$$

where $F \in (0, 1)$ is a scale factor that controls the rate at which the population evolves.

In addition, DE employs uniform crossover, also known as discrete recombination, in order to build trial vectors out of parameter values that have been copied from two different vectors. In particular, DE crosses each vector with a mutant vector, as given in Eq. (2):

$$U_{j,i,g} = \begin{cases} V_{j,i,g} & \text{if rand(0,1)} \leq C_r \text{ or} \\ X_{j,i,g} & \text{Otherwise} \end{cases} \tag{2}$$

where $U_{j,i,g}$ is the j^{th} trial vector along i^{th} dimension from the current population g. The crossover probability $C_r \in [0, 1]$ is a user defined value that controls the fraction of parameter values that are copied from the mutant. If the newly generated vector results in a lower objective function value (better fitness) than the predetermined population member, then the resulting vector replaces the vector with which it was compared.

Each member of the population acts as one possible representation for the weights attached to each connection in the network. A population of 100 members was initially randomly generated. In order to bound the search space, the weight values were limited to a range between -1 and +1. This constraint also helps reduce the chance that the evolutionary process will produce a forced model with extreme weight values. The evolution process starts after initialization according to DE equations mentioned above as shown in Fig.2 (A modified version of the one published by [7]). The output of each node will be computed according to the following equation:

$$\mu_j(0) = f_t\left(\sum_{i=0}^{n-1} w_{ij}x_i - \theta_j\right) \tag{3}$$

where, $\mu_j(t)$ is the output of node j at time t, x_i is the element i of the input, and f_t is the nonlinear transfer function chosen as the sigmoid function in this paper. θ_j is the threshold value associated with each neuron, that can also be included in the genome linear representation.

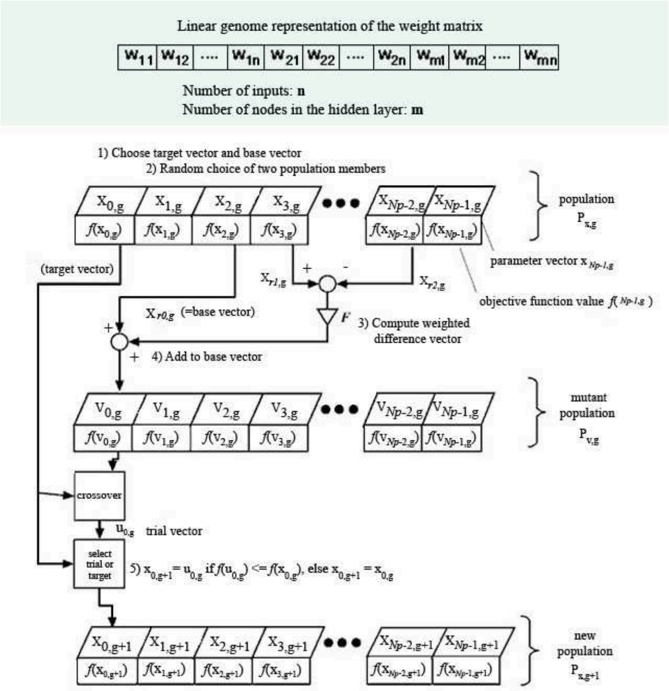

Linear genome representation of the weight matrix
w11 w12 .... w1n w21 w22 .... w2n wm1 wm2 .... wmn
Number of inputs: n
Number of nodes in the hidden layer: m
1) Choose target vector and base vector
2) Random choice of two population members
x0,g x1,g x2,g x3,g ••• xNp-2,g xNp-1,g
f(x0,g) f(x1,g) f(x2,g) f(x3,g) f(xNp-2,g) f(xNp-1,g)
population Px,g
parameter vector xNp-1,g
objective function value f(Np-1,g)
(target vector)
xr1,g
xr2,g
+
-
F
3) Compute weighted difference vector
xr0,g (=base vector)
+
+ 4) Add to base vector
v0,g v1,g v2,g v3,g ••• vNp-2,g vNp-1,g
f(v0,g) f(v1,g) f(v2,g) f(v3,g) f(vNp-2,g) f(vNp-1,g)
mutant population Pv,g
crossover
u0,g trial vector
select trial or target
5) x0,g+1 = u0,g if f(u0,g) <= f(x0,g), else x0,g+1 = x0,g
x0,g+1 x1,g+1 x2,g+1 x3,g+1 ••• xNp-2,g+1 xNp-1,g+1
f(x0,g+1) f(x1,g+1) f(x2,g+1) f(x3,g+1) f(xNp-2,g+1) f(xNp-1,g+1)
new population Px,g+1

Fig. 2. DE based weight optimization technique

Since the weights of the proposed neural network are evolved using DE optimization technique, then there is a need for a fitness function in order for the

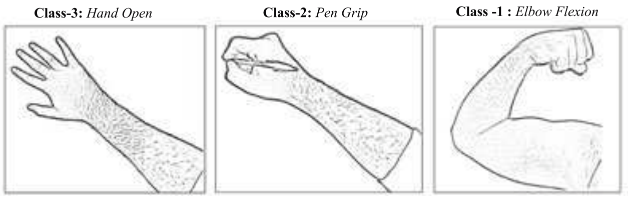


Fig. 3. Different classes of hand movements that the user imagined

DE technqiue to function. The classification accuracy was used as a fitness function of the DE. An LDA classifier was used for this purpose. The advantage of this classifier is that it does not require iterative training, avoiding the potential for under- or over-training. Finally, due to space limitiation we omit the ULDA details and refer the reader to [1, 3] for more details about ULDA.

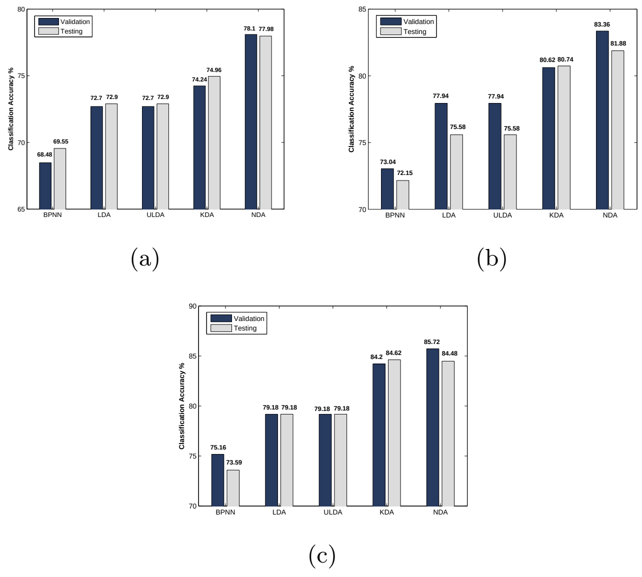
3 Experiments and Practical Results

A brain Computer Interface (BCI) probelm is considered in this paper to prove the effecteviness of the proposed technqiue. This problem was chosen due to the fact that the classification of the multichannel Electroencephalogram (EEG) signal proved to be quite challenging. The EEG dataset was recorded using two EEG channels and processed by the ProComp2 encoder from Thought Technology Ltd. Five subjects participated in the experiments. Electrodes were placed on the C3 and C4 locations that are known to show the most prominent changes for motor imagery data. Each user was instructed to imagine three different classes of the arm movement, these are: Elbow Flexion, Pen Grip, and Hand Open as shown in Fig. 3. The user was asked to perform around 12 trials of imagining each of these classes. Within each trial, a total of 30 seconds of data were recorded at 256 Hz sampling rate.

Different window lengths (128, 256, and 384 samples) were adopted to test the effectiveness of the proposed technique under various situations. These windows were incremented by 64 samples each time. The extracted feature set included a combination of autoregressive (AR) features with additional time domain features like skewness (SKEW), mean average value (MAV), waveform length (WL), and root mean square (RMS). The reason for selecting such a combination of features is that it does not need large computational power [13], while at the same time being an effective feature set. The total number of extracted features were 10 from each channel, thus 20 features were extracted from the two channels (10 features/channel = 6 AR + SKEW + MAV + WL + RMS).

In the dimensionality reduction stage, different techniques were employed to present a fair comparison. These included: LDA, ULDA, and the Kernel Discriminant Analysis (KDA)[4]. Also included was the MLP trained with back propagation algorithm, referred to as BPNN. The BPNN was added as it employs a nonlinear mapping internally within its hidden layers. All of these methods were compared with the one proposed in this paper, referred to as NDA. The testing scheme employed included a three way data split in which the total data was divided into training (≈ 2500 sample), validation(≈ 1000 sample), and testing(≈ 1000 sample). The objective function was to minimize both the training and validation errors. Then the network was tested with the completely unseen testing set to measure the generalization capability of the system. An important note to mention here is the number of neurons utilized within the hidden layer, which was roughly set to three times the number of features, as this proved to present a resonable coice for this problem.

The results of the comparison are shown in Fig. 4. These results indicate that the performance of both LDA and ULDA is the same. This is expected as both perform the same task but the latter also considers the redundancy and singularity within the scatter matrices, if such a problem exists. Also shown is that the performance of BPNN was the worst in all the cases. This is justified by the fact the back propagation algorithm cannot escape a local minima. When comparing the results achieved by KDA and NDA, it is clear that NDA almost always achieved better results than KDA. One important note here is that the NDA was more powerful than KDA when dealing with smaller window size, while for larger window size, the performance of both methods was very close. Initial results were very encouraging, achieving a maximum of 81.88% with a 1 second window lentgh that was incremented by 0.25 second each time.



(a) (b)

(c)

Fig. 4. Classification accuracies averaged across 5 subjects with different dimensionality reduction techniques (a) Window Length =128 and (b) Window Length =256 and (c) Window Length =384

4 Conclusion

In this paper, a new nonlinear discriminant analysis based feature projection technique was proposed. It included a hybrid of neural networks and Fisher's discriminant analysis. The theory and justification behind this technique was explained. The algorithm was compared with other statistical techniques and multilayer perceptron, in a BCI problem with three classes of imagination, achieving better results than all other methods even the kernel based discriminant analysis (81.88% for NDA and 80.74% for KDA). The results indicate that the proposed technique is a powerful combination for feature projection purposes. More experiments will be conducted in the future as we are currently extending this technqiue to have a self tuning capability.

References

1. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press (2006)
2. Ye, J., Janardan, R., Li, Q., Park, C.H., Park, H.: An optimization criterion for generalized discriminant analysis on undersampled problems. IEEE Transactions on Pattern Analysis and Machine Intellogence **26**(8) (2004) 982–994
3. Jin, Z., Yang, J.Y., Hu, Z., Lou, Z.: Face recognition based on the uncorrelated discriminant transformation. Pattern Recognition **34**(7) (2001) 1405–1416
4. Lu, J., Plataniotis, K.N., Venetsanopoulos, A.N.: Face recognition using kernel direct discriminant analysis algorithms. IEEE Transactions on Neural Networks **14**(1) (2003) 117–126
5. Xiong, T., Ye, J., Cherkassky, V.: Kernel uncorrelated and orthogonal discriminant analysis: A unified approach. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). (2006) 125–131
6. Liang, Z., Shi, P.: Uncorrelated discriminant vectors using kernel method. Pattern Recognition **38**(2) (2005) 307–310
7. Price, K., Storn, R., Lampinen, J.: Differential Evolution: A Practical Approach to Global Optimization. Springer (2005)
8. Lippmann, R.: An introduction to computing with neural nets. IEEE ASSP Magazine **4**(2) (1987) 4–22
9. Webb, A.R., Lowe, D.: The optimised internal representation of multilayer classifier networks performs nonlinear discriminant analysis. Neural Networks **3**(4) (1990) 367–375
10. Gallinari, P., Thiria, S., Badran, F., Fogelman-Foulie, F.: On the relations between discriminant analysis multilayer perceptrons. Neural Networks **4**(3) (1991) 349–360
11. Casasent, D., Chen, X.: Radial basis function neural networks for nonlinear fisher discrimination and neyman-pearson classification. Neural Networks **16**(5-6) (2003) 529–535
12. Kwon, Y., Moon, B.: Nonlinear feature extraction using a neuro genetic hybird. In: Proceedings of the 2005 conference on Genetic and evolutionary computation. (2005) 2089–2096
13. Dharwarkar, G.S., Basir, O.: Enhancing temporal classification of aar parameters in eeg single-trial analysis for brain-computer interfacing. In: Proceedings of The 28th IEEE EMBS Annual International Conference, New York City, USA. (2006) 2171–2174