A New Multidimensional Feature Transformation for Linear Classifiers and Its Applications

EunSang Bak

Electrical and Computer Engineering Department
University of North Carolina
9201 University City Blvd, Charlotte, NC 28223, U.S.A.
bakeunsang@yahoo.com

Abstract. In this paper, a new feature transformation method is introduced to decrease misclassification rate. Linear classifiers in general are not able to classify feature vectors which lie in a high dimensional feature space. When the feature vectors from difference classes have underlying distributions which are severely overlapped, it is even more difficult to classify those feature vectors with desirable performance. In this case, data reduction or feature transformation typically finds a feature subspace in which feature vectors can be well separated. However, it is still not possible to overcome misclassifications which results from the overlapping area. The proposed feature transformation increases the dimension of a feature vector by combining other feature vectors in the same class and then follows typical data reduction process. Significantly improved separability in terms of linear classifiers is achieved through such a sequential process and is identified in the experimental results.

1 Introduction

The purpose of pattern classification is to decide the class of the data which is assumed to consist of (C, **x**) pairs where C is the class to which **x,** which is an *r*-dimensional feature vector, belongs. There are many traditional and modern approaches to estimate the underlying distributions. Some of them [1], [8] attempt to estimate the class conditional probability distribution of a feature vector **x** by assuming specific distributional form of underlying distributions. On the contrary, other approaches [2], [4], [5], which are typically called nonparametric methods, try to estimate the probability distributions without assuming any distributional form.

Once the distributions are estimated in either method, feature space is separated by a selected classifier. In fact, the selection of classifier is strongly related with the estimation method since the classifier is built upon the estimated underlying distributions.

In a high dimensional feature space, classification process may suffer from so called curse of dimensionality. Most of the approaches for feature classification involve a data reduction or feature transformation step. This step basically reduces the dimension of feature space so that feature vectors can be well separated in the new lower dimensional feature space.

Various methods have been proposed in literature for feature transformation/data reduction; Fisher's approach [3], [6], [12], [13], removal classification structures [15], adaptive linear dimensionality reduction [16], linear constrained distance-based classifier analysis [17] and others [7], [8], [9], [10], [11]. These approaches consistently try to transform feature vectors into a feature space of lower dimension.

In this paper, a new feature transformation is proposed and its effect is investigated in terms of linear classifiers. The paper is organized as follows: In Section 2, basic idea of the proposed feature transformation is introduced. In Section 3, a new feature transformation is explained in detail with mathematical derivation and its effect on the underlying distributions is analyzed. Unsupervised image classification process incorporating a new feature transformation is described in Section 4 as an application. The experimental results are presented in Section 5, which is followed by the conclusions in Section 6.

2 Basic Idea

For solving multidimensional classification problems, Fisher [3] suggested a method, which projected the data from d dimensions onto a line with specific direction for which the projected data were well separated. The projected value was obtained by linear combination of the components of $\mathbf{x}$, thus every multidimensional data was converted to a scalar by this linear combination.

Such a linear combination turned out to be the product of the difference of the means of classes and the common inverse covariance matrix. This process is equivalent to maximizing the ratio of between-class variation to within-class variation.

Fisher's main idea would rather be interpreted as how to linearly reduce the dimension of the feature space so that linear classifiers can give the best classification result. In this paper, the proposed method takes an opposite direction. Instead of reducing the dimension of the data, we first increase the dimension of the data by combining several feature vectors in the same class and make a new feature space and then reduce the dimension of the new feature space. By inserting the process of increasing data dimension, a very interesting fact is found with respect to feature classification.

Comparing the distributions between the original feature space and a new feature space, a new feature space gives much better separability for classification. That is, the distance between the means of existing probability distributions in the feature space gets longer compared to the change of their standard deviations. This in turn reduces the overlapping area between the probability distributions. Such an augmentation of the dimension of the feature space will be called a feature transformation hereafter.

Feature transformation requires a couple of assumptions. One of them is that the probability distributions of classes in the feature space should be normally distributed and the other is that those distributions have a common covariance.

In the following section, the proposed feature transformation will be derived mathematically and shows how the feature transformation changes the class conditional probability distributions so as to be suitable for linear classifiers.

3 Multivariate Feature Transformation

Suppose there are two classes and the i^{th} class is represented as $\pi_i \sim \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$. It is assumed that all the classes have a common covariance matrix $\boldsymbol{\Sigma}$. We would like to assign a sample data $\mathbf{x}$, which is an r-dimensional random vector, to the most probable class among the two classes.

Let us define a new random vector $\mathbf{x}^*$ which consists of a certain number of random vectors in the same class. In other words, a new random vector is a composite random vector whose elements are random vectors from the same class. For example, take (s + 1) random vectors around a neighborhood of the centered random vector $\mathbf{x}_0$, $\left\{\mathbf{x}_j^{(i)} : j = 0, \ldots, s\right\}$ in the i^{th} class and make a new random vector $\mathbf{x}_0^{(i)*}$ which has $r(s$ + 1) dimension in proportion to the number of component vectors. The random vector $\mathbf{x}_0^{(i)*}$ is regarded as an extended random vector of $\mathbf{x}_0^{(i)}$ and is represented as follows:

$$\mathbf{x}^{(i)*} = \left[\mathbf{x}_0^{\mathrm{T}} \;\; \mathbf{x}_1^{\mathrm{T}} \cdots \mathbf{x}_s^{\mathrm{T}}\right]^{\mathrm{T}} \tag{1}$$

Note that the subscript of $\mathbf{x}^{(i)*}$ is omitted for brevity. Unless otherwise specified, the subscript of the extended random vector is the same as that of the first component vector. The mean vector of $\mathbf{x}^{(i)*}$ whose component random vectors are from the i^{th} class becomes

$$\boldsymbol{\mu}_i^* = \mathbf{1} \otimes \boldsymbol{\mu}_i \tag{2}$$

, where the operator $\otimes$ denotes a direct product in matrix operation. In addition, a covariance between two random vectors ($\mathbf{x}_j^{(i)}$, $\mathbf{x}_k^{(i)}$) can be defined as

$$\mathrm{Cov}\left(\mathbf{x}_j^{(i)}, \mathbf{x}_k^{(i)}\right) = \rho_{jk}\boldsymbol{\Sigma} \tag{3}$$

, which generate the covariance matrix of $\mathbf{x}^{(i)*}$ in (4). For simplicity, the superscript for labeling the class from which the random vectors come is omitted, therefore, $\mathbf{x}_j$ is substituted for $\mathbf{x}_j^{(i)}$ and $\mathbf{x}^*$ is substituted for $\mathbf{x}^{(i)*}$ as needed.

$$\mathrm{Cov}\left(\mathbf{x}^*\right) = \begin{bmatrix} \mathrm{Cov}\left(\mathbf{x}_0, \mathbf{x}_0\right) & \cdots & \mathrm{Cov}\left(\mathbf{x}_0, \mathbf{x}_s\right) \\ \vdots & & \vdots \\ \mathrm{Cov}\left(\mathbf{x}_s, \mathbf{x}_0\right) & \cdots & \mathrm{Cov}\left(\mathbf{x}_s, \mathbf{x}_s\right) \end{bmatrix} = \mathbf{C}^* \otimes \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^* \tag{4}$$

As a result, a new multivariate normal random vector $\mathbf{x}^*$ whose mean vector is $\boldsymbol{\mu}^*$ and covariance matrix $\boldsymbol{\Sigma}^*$ is obtained.

$$\mathbf{x}^* \sim \mathrm{N}\left(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*\right) \tag{5}$$

Now, discriminant function D_i^* [18] for the random vector $\mathbf{x}^{(i)*}$ is written compared to D_i for $\mathbf{x}^{(i)}$.

$$\mathrm{D}_i = \mathbf{L}_i^{\mathrm{T}}\mathbf{x}^{(i)} - \frac{1}{2}\mathbf{L}_i^{\mathrm{T}}\boldsymbol{\mu}_i, \qquad \mathbf{L}_i^{\mathrm{T}} = \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right)^{\mathrm{T}}$$
$$\mathrm{D}_i^* = \mathbf{L}_i^{*\mathrm{T}}\mathbf{x}^{(i)*} - \frac{1}{2}\mathbf{L}_i^{*\mathrm{T}}\boldsymbol{\mu}_i^*, \quad \mathbf{L}_i^{*\mathrm{T}} = \left(\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}_i^*\right)^{\mathrm{T}} \tag{6}$$

After some mathematical expansions, $\mathbf{L}_i^*$ can be described in terms of $\mathbf{L}_i$ in (7). The $\boldsymbol{\delta}$ in (7) is a vector whose elements are the sum of elements of each row vector of the inverse of correlation coefficient matrix $\mathbf{C}^*$ in (4).

$$\mathbf{L}_i^{*\mathrm{T}} = \boldsymbol{\delta}^{\mathrm{T}} \otimes \mathbf{L}_i^{\mathrm{T}}$$
$$\boldsymbol{\delta} = \left(\mathbf{1}^{\mathrm{T}}\mathbf{C}^{*-1}\right)^{\mathrm{T}} = \left[\delta_0, \delta_1, \cdots, \delta_s\right]^{\mathrm{T}} \tag{7}$$

The discriminant function D_i^* is finally represented in (8) and is described in terms of the random vector $\mathbf{x}$. In other words, the discriminant function of composite random vector $\mathbf{x}^*$ is characterized by the terms in the discriminant function of the random vector $\mathbf{x}$.

$$\mathrm{D}_i^* = \mathbf{L}_i^{\mathrm{T}}\left(\sum_{n=0}^{s}\delta_n\mathbf{x}_n^{(i)} - \frac{1}{2}\boldsymbol{\mu}_i\sum_{n=0}^{s}\delta_n\right) \tag{8}$$

Now, we define a new random vector $\mathbf{G}$ in (9) as a linear combination of (s + 1) number of random vector $\mathbf{x}$'s.

$$\mathbf{G} = \sum_{n=0}^{s}\delta_n\mathbf{x}_n \tag{9}$$

Eq. (9) corresponds to the proposed feature transformation. The coefficients of the linear combination are derived from the correlation coefficient matrix of random vector $\mathbf{x}$. Once the expectation $\boldsymbol{\mu}_\mathbf{G}$ and the covariance matrix $\boldsymbol{\Sigma}_\mathbf{G}$ are obtained, the discriminant function of $\mathbf{G}$ is described in (10). Comparing (10) with (8), it is recognized that the results are equivalent, which means that the discriminant function D_i^* can be considered as the discriminant function $\mathrm{D}_i(\mathbf{G})$.

$$\begin{aligned}\mathrm{D}_i\left(\mathbf{G}\right) &= \left(\boldsymbol{\Sigma}_\mathbf{G}^{-1}\boldsymbol{\mu}_\mathbf{G}\right)^{\mathrm{T}}\mathbf{G} - \frac{1}{2}\left(\boldsymbol{\Sigma}_\mathbf{G}^{-1}\boldsymbol{\mu}_\mathbf{G}\right)^{\mathrm{T}}\boldsymbol{\mu}_\mathbf{G} \\ &= \mathbf{L}_i^{\mathrm{T}}\left[\sum_{n=0}^{s}\delta_n\mathbf{x}_n - \frac{1}{2}\boldsymbol{\mu}_i\sum_{n=0}^{s}\delta_n\right]\end{aligned} \tag{10}$$

Since a linear combination of the normal random vectors also follows a normal distribution, the distribution of the random vector $\mathbf{G}$ will be

$$\mathbf{G} \sim \mathrm{N}\left(\boldsymbol{\mu}_\mathbf{G}, \boldsymbol{\Sigma}_\mathbf{G}\right)$$
$$\boldsymbol{\mu}_\mathbf{G} = k\cdot\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_\mathbf{G} = k\cdot\boldsymbol{\Sigma}, \quad k = \sum_n \delta_n \tag{11}$$

In summary, a multivariate random vector $\mathbf{x}$ is transformed into a multivariate random vector $\mathbf{G}$ by means of creating a composite random vector $\mathbf{x}^*$. The mean vector of $\mathbf{G}$ simply becomes k times larger than the mean vector of $\mathbf{x}$ and the covariance matrix becomes also k times larger than that of $\mathbf{x}$. This is a very important observation to indicate how feature transformation changes the class conditional probability distributions and achieves better separability. Since the distance between the means becomes k times larger while the spread of the distributions becomes $\sqrt{k}$ times larger, the distributions get farther away after feature transformation and it gives better separability.

4 Unsupervised Image Classification

In the previous section, we have seen that the linear combination of an extended random vector whose component random vectors came from the same class produced a new random vector and the separability of new random vectors between different classes was significantly improved in terms of linear classifiers.

One of the applications for the proposed method would be image classification. In general, objective of image classification is the separation of the regions or objects in the image which have different characteristics. Due to the characteristics of image, most of the homogeneous objects or regions occupy a certain area in an image and feature vectors from the same object are located in the neighborhood. Thus, determination of classification of a feature vector can be made associated with the determinations of the neighboring feature vectors.

In our experiments, an iterative unsupervised classification is chosen. This classification process does not need a training process. Given that the number of regions (or classes) into which an image is supposed to be separated, a simple clustering algorithm is applied to classify the feature vectors and thus a provisional result of classification is obtained.

The purpose of the provisional result of classification by given clustering algorithm is to extract intermediate information about each class and to calculate the coefficient vectors $\boldsymbol{\delta}$ for each class for feature transformation, which compose an initial iteration. From the second iteration a selected linear classifier classifies the feature space resulted from the previous feature transformation.

As the proposed feature transformation has been mathematically proved in Section 3, the distributions of feature vectors of the classes becomes farther away each other, which results in a smaller misclassification rate from the smaller overlapping area between two probability distributions.

Until the conditions for terminating process are satisfied, the iterative procedure is continued. One of the conditions for termination is the size of the value k in (13). If k is not larger than one, such a transformation does not give a better separability in the feature space.

Fig. 1 shows the iterative procedure for image classification explained in the above. Note that since the main contribution of this paper is the method of feature

transformation, any other clustering method can be used depending on particular need, although, in our experiments, K-means clustering method is used.

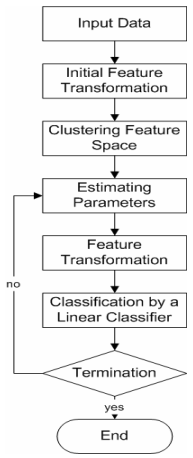


Fig. 1. Unsupervised classification procedure.

5 Experimental Results

We simulate the proposed method with a synthetic data set. The data set is generated from multivariate number generator with a mean vector and a covariance matrix. Each feature vector is located by given coordinate in an image as Fig. 2(e) according to its class. The size of the image is 26x26 so that the total number of feature vectors from two classes is 676. Synthetic feature vectors are generated with parameter sets whose mean vectors are [-1.0 1.0] and [1.0 -1.0] each, and the common covariance matrix whose diagonal elements are 10.0.

Fig. 2(a) shows the initial distributions of feature vectors from two classes. Since the distance from two mean vectors is much smaller than the size of variances in the covariance matrix, the feature vectors are heavily overlapped between classes. None of the linear classifiers seem to be adequate to classify feature vectors with desirable performance. Assuming that feature vectors are independently extracted, the first feature transformation is executed with the $\boldsymbol{\delta}$ vector having all ones. After first transformation, K-means clustering method is used to make a temporary classification map. Fig. 2(b) shows this classification map and the two classes in Fig. 2(b) look more separated than in Fig. 2(a). Fig. 2(b) simply represent temporary determined classes so that each class may contain feature vectors that are actually misclassified.

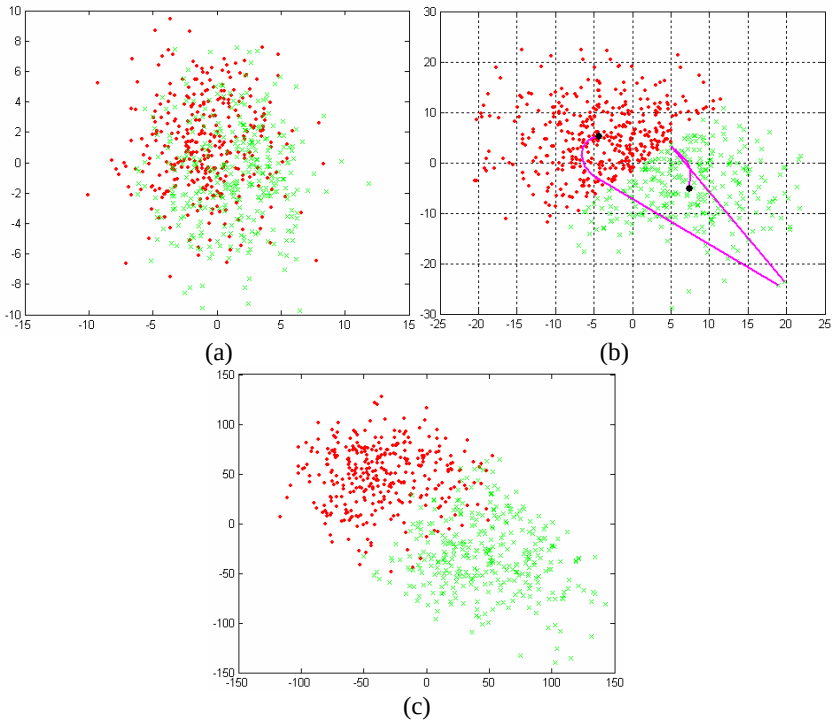
After four iterations, the distributions of feature vectors become more suitable for linear classification as in Fig. 2(c). Now, any linear classifiers can be selected to separate the transformed feature vectors. Fig. 2(d) shows the error rate at each iteration by linear discriminant function and the locations of misclassified feature vectors are

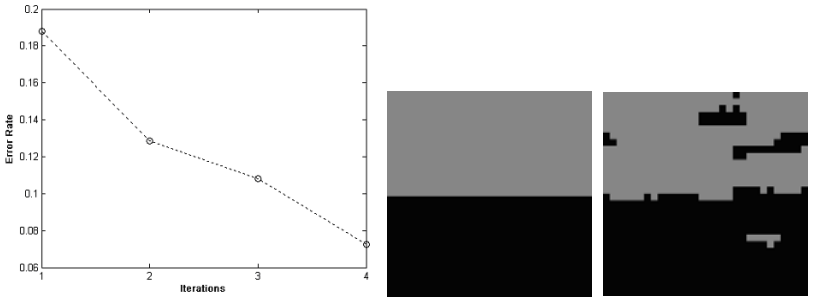
illustrated in Fig. 2(f) compared to the true classification map in Fig. 2(e). In the light of the above simulations, the proposed method indicates a new possibility for image classification.

Now, the proposed method is applied to a practical classification data. Original data set is from the UCI Machine Learning Repository [14]. Feature vectors in the original data set are extracted from an image which contains seven different regions, which are grass, path, cement, sky, brickface, foliage and window. A feature vector from each region is characterized by 19 feature components. In our experiments, for the sake of visualization, two regions (cement and foliage) are selected and two most significant feature components (intensity-mean and value-mean) are chosen in the experiments.

The data seems to be relatively linearly separable. However, as can be in Fig. 3(a) and Fig. 3(b), data distributions are not completely separable so that it would not be possible to separate the data without misclassification by any linear classifiers.

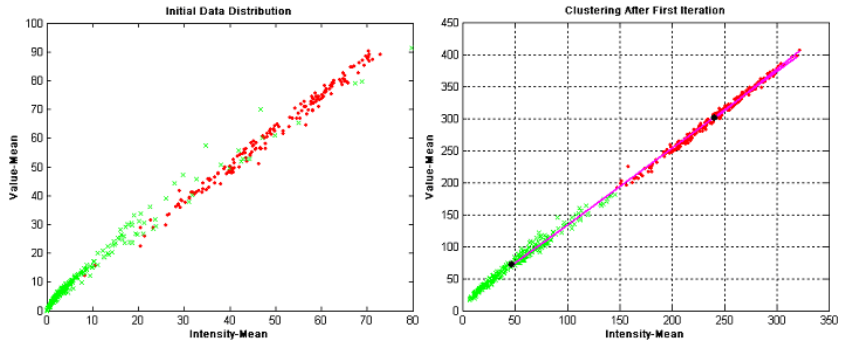
Surprisingly, Fig. 3(c)-(f) show the results of classification using feature transformation. Taking into account the features in the neighborhood, the proposed method changes the data distributions as much as it can be linearly separated. Fig. 3(c) shows the last data distributions on which linear discriminant function is to be applied. Feature vectors are separated without misclassifications and are illustrated in Fig. 3(f). As a result, two regions (cement-foliage) which have different natural characteristics are completely separated without misclassification through the proposed method.



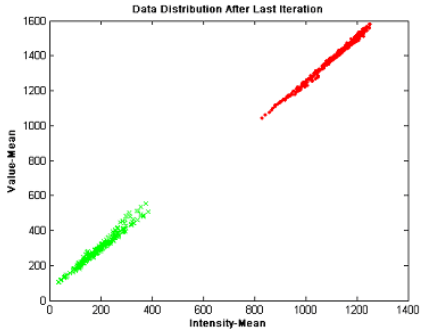


(d) (e) (f)

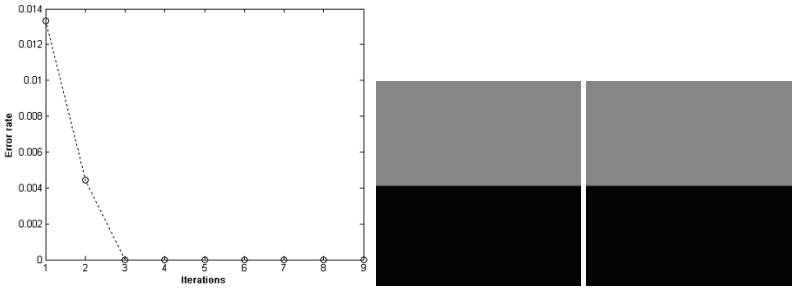
Fig. 2. Result from simulation. (a) Distributions of feature vectors of two classes. (b) Temporary classification by K-means algorithm. (c) Distributions of feature vectors of two classes after the last feature transformation. (d) Error rates on every iteration. (e) True classification map. (f) Classification map from the proposed method.



(a) (b)



(c)



(d) (e) (f)

Fig. 3. Results from a real data set. (a) Distributions of feature vectors of two classes. (b) Temporary classification by K-means algorithm. (c) Distributions of feature vectors of two classes after the last feature transformation. (d) Error rates on every iteration. (e) True classification map. (f) Classification map from the proposed method.

6 Conclusions

A new feature transformation method is introduced. It increases the dimension of a feature vector by combining other feature vectors in the same class and then follows a typical data reduction process. The proposed method eventually gives significantly improved separability in feature space in terms of linear classifiers and the promising experimental results are presented.

References

1. W. Chou, "Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition," *Proc. IEEE*, vol. 88, no. 8, pp. 1201-1223, Aug. 2000.

2. T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 6, pp. 607-616, June 1996.

3. R.A. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics*, vol. 8, pp. 376-386, 1938.

4. L.J. Buturovic, "Towards Bayes-optimal linear dimension reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, pp. 420-424, 1994.

5. T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *J. Royal Statistics Soc., B*, vol. 58, pp. 155-176, 1996.

6. C.R. Rao, "The utilization of multiple measurements in problems of biological classification," *J. Royal Statistical Soc., B*, vol. 10, pp. 159-203, 1948.

7. M. Hubert and K.V. Driessen, “Fast and robust discriminant analysis,” *Computational Statistics & Data Analysis*, vol. 45, Issue 2, pp. 301-320 March 2004.

8. W.L. Poston and D.J. Marchette, “Recursive dimensionality reduction using Fisher's linear discriminant,” *Pattern Recognition*, vol. 31, no. 7, pp. 881-888, 1998.

9. M. Loog, R.P.W. Duin and R. Haeb-Umbach, “Multiclass linear dimension reduction by weighted pairwise Fisher criteria,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 762-766, 2001.

10. L. Rueda and B.J. Oommen, “On optimal pairwise linear classifiers for normal distributions: The two-dimensional case,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 274-280, 2002.

11. H. Brunzell and J. Eriksson, “Feature reduction for classification of multidimensional data,” *Pattern Recognition*, vol. 33, pp. 1741-1748, 2000.

12. Duda, R.O., P.E Hart, and D.G.. Stork, *pattern Classification*, 2ed. John Wiley & Sons, New York, Jan. 2000.

13. A.K. Jain, R.P.W. Duin and J. Mao, “Statistical pattern recognition: A review,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4-37, 2000.

14. UCI Repository of Machine Learning Databases, www.ics.uci.edu/mlearn/mlrepository.html, 2004

15. M. Aladjem, “Linear discriminant analysis for two classes via removal of classification structure,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 187-192, 1997

16. R. Lotlikar and R. Kothari, “Adaptive linear dimensionality reduction for classification,” *Pattern Recognition*, vol. 33, Issue 2, pp. 177-350 2000

17. Q. Du and C.-I. Chang, “A linear constrained distance-based discriminant analysis for hyperspectral image classification,” *Pattern Recognition*, vol. 34, Issue 2, pp. 361-373, Feb. 2001

18. A.M. Kshirsagar, *Multivariate Analysis*, M. Dekker, New York, 1972.