32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

# Mutual Information-Based Fisher Discriminant Analysis for Feature Extraction and Recognition with Applications to Medical Diagnosis

Ali Shadvar, *Student Member, IEEE*, and Abbas Erfanian, *Member, IEEE*

*Abstract*—**This paper presents a novel discriminant analysis (DA) for feature extraction using mutual information (MI) and Fisher discriminant analysis (MI-FDA). Most DA algorithms for feature extraction are based on a transformation which maximizes the between-class scatter and minimizes the within-class scatter. In contrast, the proposed method uses the Fisher's criterion to find a transformation that maximizes the MI between the transferred features and the target classes and minimizes the redundancy. The performance of the proposed method is evaluated using UCI databases and compared with the performance of some DA-based algorithms. The results indicate that MI-FDA provides a robust performance over different data sets with different characteristics. On average, an accuracy rate of 81.3% was achieved using MI-FDA.**

## I. Introduction

Dimensionality reduction of the raw input variable space is an essential preprocessing step in the classification process. There are two main reasons to keep the dimensionality of the input features as small as possible: computational cost and classification accuracy [1].

Reduction of the number of input variables can be done by selecting relevant features (i.e., feature selection) or extracting new features containing maximal information about the class label from the original ones (i.e., feature extraction) [2].

Linear discrimination analysis (LDA) is a well-known and popular linear dimensionally reduction algorithm for supervised feature extraction [3]. LDA computes a linear transformation by maximizing the ratio of between-class distance to within-class distance, thereby achieving maximal discrimination. In LDA, a transformation matrix from an n-dimensional feature space to a d-dimensional space is determined such that the Fisher criterion of between-class scatter over within-class scatter is maximized.

However, traditional LDA method is based on the restrictive assumption that the data are homoscedastic, *i.e,* data in which classes have equal covariance matrices. In particular, it is assumed that the probability density functions of all classes are Gaussian with identical covariance matrix but different means [4]. Moreover, traditional LDA can not solve the problem posed by nonlinearly separable classes. Hence, its performance is unsatisfactory for many classification problems in which nonlinear decision boundaries are necessary. To solve this, nonlinear extension of LDA has been proposed [5]-[6].

Moreover, LDA-based algorithms generally suffer from small sample size (SSS) problem when the number of training samples is less than the dimension of feature vectors [7]–[8]. A traditional solution to this problem is to apply PCA in conjunction with LDA [7]-[8]. Recently, more effective solutions have been proposed to solve the SSS [9]-[10].

Another problem that is common to most DA methods is that these methods can only extract $C$-1 features from the original feature space where $C$ is the number of classes. Recently, a method based on discriminant analysis (DA) was proposed, known as subclass discriminant analysis (SDA), for describing a large number of data distributions [11] and solve the limitation posed by the DA methods in the number of features that can be extracted.

One of the most effective approaches for optimal feature extraction is based on mutual information (MI). MI measures the mutual dependence of two or more variables. In this context, the feature extraction process is creating a feature set from the data which jointly have largest dependency on the target class and minimal redundancy among themselves. In [2], [12], a method was proposed, known as MRMI, for learning linear discriminative feature transform using an approximation of the mutual information between transformed features and class labels as a criterion. The approximation is inspired by the quadratic Renyi entropy which provides a nonparametric estimate of the mutual information. However, there is no general guarantee that maximizing the approximation of mutual information using Renyi's definition is equivalent to maximizing mutual information defined by Shannon.

In this paper, we propose a new method for feature extraction which is based on mutual information and Fisher-Rao's criterion. The proposed method is then evaluated by using seven databases. The results obtained using proposed method compare with that obtained using LDA [13], SDA [11], Kernel Gaussian LDA (KG-LDA) [14], Kernel Polynomial LDA (KP-LDA) [14], and MI-based feature extraction method proposed in [12].

## II. Methods

### A. Mutual Information

Mutual information is a non-parametric measure of relevance between two variables. Shannon's information theory provides a suitable formalism for quantifying this concepts. Given two random variables $x$ and $y$, their mutual

Manuscript received April 23, 2010. This work was supported by Iran University of Science and Technology (IUST).

A. Erfanian and A. Shadvar are with the Department of Biomedical Engineering, Iran University of Science and Technology (IUST), Tehran, Iran (phone: 98-21-77240465; fax: 98-21-77240490; email: erfanian@iust.ac.ir).

information is defined in terms of their probabilistic density functions $p(x)$, $p(y)$, $p(x,y)$:

$$I(x;y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{1}$$

If the mutual information between two random variables is large, it means two variables are closely related. Indeed, MI is zero if and only if the two random variables are strictly independent.

*B. MI-Based Fisher-Rao's criterion*

Most DA methods defined so far are based on Fisher-Rao's criterion , which is given by

$$\mathbf{V}=\arg\max\frac{\left|\mathbf{V}^{\mathbf{T}}\mathbf{A}\mathbf{V}\right|}{\left|\mathbf{V}^{\mathbf{T}}\mathbf{B}\mathbf{V}\right|} \tag{2}$$

where the matrices $\mathbf{A}$ and $\mathbf{B}$, are assumed to be symmetric and positive-definite, so that they define a metric. LDA uses the between and within-class scatter matrices, $\mathbf{A}=\mathbf{S_B}$ and $\mathbf{B}=\mathbf{S_W}$, respectively, in (1); where

$$\mathbf{S_B}=\sum_{i=1}^{C}(\mu_i-\mu)(\mu_i-\mu)^T \tag{3}$$

$C$ is the number of classes, $\mu_i$ the sample mean of class $i$, $\mu$ the global mean,

$$\mathbf{S_W}=\frac{1}{n}\sum_{i=1}^{C}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)(x_{ij}-\mu_i)^T \tag{4}$$

$x_{ij}$ is the $j$th sample of class $i$, and $n_i$ the number of samples in that class. The objective is to find a linear transformation $\mathbf{V}$ which maximizes the between-class scatter matrix $\mathbf{S_B}$ and minimizes within-class scatter matrix $\mathbf{S_W}$ (Fisher's criterion).

A limitation of the Fisher's LDA is that it merely tries to separate class means as good as possible and it does not take the discriminatory information. Moreover, since LDA only makes use of second-order statistical information, the covariances, it is optimal for data where each class has a unimodal Gaussian density with well separated means. Furthermore, the maximum rank of $\mathbf{S_B}$ is $C-1$. Thus LDA cannot produce more than $C-1$ features.

In this work, we define a new information-theoretic criterion based on Fisher-Rao's criterion as

$$\mathbf{A}=\begin{bmatrix} I(f_1,c) & \frac{I(f_1,c)+I(f_2,c)}{2} & \cdots & \frac{I(f_1,c)+I(f_n,c)}{2} \\ \frac{I(f_2,c)+I(f_1,c)}{2} & I(f_2,c) & \cdots & \frac{I(f_2,c)+I(f_n,c)}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{I(f_n,c)+I(f_1,c)}{2} & \frac{I(f_n,c)+I(f_2,c)}{2} & \cdots & I(f_n,c) \end{bmatrix} \tag{5}$$

$$\mathbf{B}=\begin{bmatrix} 0 & I(f_1,f_2) & \cdots & I(f_1,f_n) \\ I(f_2,f_1) & 0 & \cdots & I(f_2,f_n) \\ \vdots & \vdots & \ddots & \vdots \\ I(f_n,f_1) & I(f_n,f_2) & \cdots & 0 \end{bmatrix} \tag{6}$$

where $I(f;c)$ is the mutual information between feature $f$ and class label $c$.

By solving the optimization problem (2) using (5) and (6), the projection vector set consists of the eigenvectors corresponding to nonzero eigenvalues of $\mathbf{B^{-1}A}$ as

$$\mathbf{A}\mathbf{V}=\mathbf{B}\mathbf{V}\Lambda \tag{7}$$

The transformation matrix $\mathbf{W}$ must be constituted from the largest eigenvectors $\mathbf{V}$ as

$$\mathbf{W}=\left[v_1,v_2,v_3,\ldots v_i\right] \tag{8}$$

where $v_1,v_2,...,v_i$ are the largest eigenvectors that satisfy $\lambda_i\geq\lambda_{\max}/\alpha$.

The optimal feature set is obtained by projecting the original feature set on the projection matrix as

$$\mathbf{Y}=\mathbf{W}^{\mathbf{T}}\mathbf{X} \tag{9}$$

where $\mathbf{X}$ is the original feature set and $\mathbf{Y}$ is the optimal feature set.

By solving the optimization problem (2) using (5) and (6), new features are extracted from the original features which jointly have largest dependency on the target class with minimal redundancy. When $\mathbf{A}$ is maximized, features are extracted with largest dependency on the target class, and when $\mathbf{B}$ is minimized, features extracted with minimum redundancy.

## III. RESULTS

In this section, we investigate the performance of the proposed method using several UCI data sets [15] and compare the results obtained with the well-known feature extraction methods: LDA, SDA, KG-LDA, KP-LDA, and MI-based feature extraction method proposed in [12] known as MRMI-SIG. The UCI machine learning repository contains many real-world data sets that have been used by a large variety of investigators [1]-[3], [11], [12]. Table I shows brief information of the data sets used in this paper.

TABLE I
BRIEF INFORMATION ON UCI DATA SETS

| *Data set* | *Class* | *Instances* | *Attributes* | *Train* | *Test* |
|---|---|---|---|---|---|
| Breast Cancer Wisconsin Diagnostic | 2 | 568 | 30 | 284 | 285 |
| Haber man's Survival | 2 | 306 | 3 | 153 | 153 |
| Parkinson | 2 | 197 | 22 | 99 | 98 |
| Pima Indians Diabetes | 2 | 768 | 8 | 500 | 268 |
| MUSK1 | 2 | 166 | 476 | 300 | 176 |
| Lung Cancer | 3 | 32 | 56 | 16 | 16 |
| SPECTF Heart | 2 | 267 | 44 | 80 | 187 |

The discriminant methods mentioned above are first used to find a low dimensional representation of the data and, then, the k-nearest-neighbors (KNN) classifier with $K$=1 is used to classify each of the testing samples according to the class label. This process is carried out ten times and the average is calculated. In this work, we used a two-dimensional mutual information estimation based on histogram method [16].

### *A. Wisconsin Diagnostic Breast Cancer Data Set*

The first database used in this study is the "Wisconsin Diagnostic Breast Cancer" (WDBC) set. In this database, we have 30 features decribing the shape and texture of the nucleus of the cells to be analyzed. The task is to discriminate between harmfull and harmless cells. The 569 samples available are randomly divided into a training set of 285 samples and a testing set of 284 instances. This assures that the sample-to-dimension ratio is appropriate. The results of classification are summarized in Table II. It is observed that average classification accuracy 95.2% is obtained using MI-FDA method which is comparable with that obtained using LDA, SDA, and KG-LDA. The accuracy obtained using KP-LDA is 90.3%.

### *B. Haberman's Survival Data Set*

The second databases utilized is "Haberman's Survival". In this database, there are cases from the study that was carried out between 1958 and 1970 at the University of Chicago's Billing Hospital considering the survival of patients who had undergone surgery for breast cancer. The purpose of this study is to discriminate between the patient survived 5 years or longer and died within 5 years. The 306 samples are randomly divided into a training set of 153 samples and a testing set of 153 instances. The average accuracy obtained using MI-FDA method is 70.5% which is comparable with the results obtained using KP-LDA and KG-LDA. The accuracies obtained using LDA and SDA are 66.9% and 65.6%, respectively.

### *C. Parkinson Data Set*

This database composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). The number of attributes is 23. The 197 samples available are randomly divided into a training set of 99 samples and a testing set of 98 instances. The average accuracy obtained is 91.8% which is higher than that obtained by other methods.

### *D. Pima Diabetes Data Set*

In this database, all patients are females at least 21 years old of Pima Indian heritage. The number of attributes is 8. The 768 samples available are randomly split into a test and a train set, each with size of 384 samples. The average accuracies obtained using MI-FDA method and KG-LDA, are 72.1% and 75.2%, respectively. The rates obtained by LDA, SDA, KP-LDA, and MRMI-SIG are about 69.0%.

### *E. Musk Data Set*

This data set describes a set of 92 molecules of which 47 are judged by human experts to be musks and the remaining 45 molecules are judged to be non-musks. The 166 features that describe these molecules depend upon the exact shape, or conformation, of the molecule. The goal is to learn to predict whether the new molecules will be musks or non-musks. The 476 samples available are randomly divided into a training set of 300 samples and a testing set of 176 testing instances. The accuracies obtained using MI-FDA method and KG-LDA, are 94.6% and 94.2%, respectively, while the rates obtained by KP-LDA and MRMI-SIG are 86.5% and 81.8%, respectively. In contrast, LDA and SDA achieve a 100% accuracy rate on this data set.

### *F. Lung Cancer Data Set*

This database describes 3 types of pathological lung cancers. Te aim was to learn to predict the type of pathological lung cancers. The 32 samples available are randomly divided into a training set of 16 samples and a testing set of 16 testing instances. The precision is 64.6% for MI-FDA method, 38.8% for LDA, 43.1% for SDA, 59.4% for KP-LDA, and 52.5% for KG-LDA.

### *G. SPECTF Heart Data Set*

This database describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images.

TABLE II
THE AVERAGE CLASSIFICATION RATE S (%) USING DIFFERENT FEATURE EXTRACTION METHODS

| *UCI Data Sets* | *Original set* | *LDA* | *SDA* | *KP-LDA* | *KG-LDA* | *MRMI-SIG* | *MI-FDA* |
|---|---|---|---|---|---|---|---|
| Breast Cancer | 91.8 ± 1.3 | 94.5 ± 1.2 | 96.5 ± 0.8 | 90.3 ± 2.1 | 95.7 ± 0.9 | 82.6 ±3.2 | 95.2 ± 0.8 |
| Haber man's Survival | 68.8 ± 3.5 | 66.9 ± 4.3 | 65.6 ± 3.1 | 71.4 ± 1.8 | 72.8 ± 1.6 | 67.6 ± 3.2 | 70.5 ± 1.2 |
| Parkinson | 85.8 ± 2.2 | 82.6 ± 1.7 | 81.2 ± 4.6 | 82.5 ± 2.7 | 87.4 ± 1.3 | 87.0 ± 4.7 | 91.8 ± 1.4 |
| Pima Indians Diabetes | 67.9 ± 1.0 | 69.1 ± 2.9 | 69.0 ± 2.2 | 69.2 ± 2.3 | 75.2 ± 1.2 | 69.7 ± 1.7 | 72.1 ± 1.2 |
| MUSK1 | 86.0 ± 1.7 | 100.0 ± 0.0 | 100.0 ± 0.0 | 86.5 ± 2.7 | 94.2 ± 1.4 | 81.8 ± 3.5 | 94.6 ± 1.6 |
| Lung Cancer | 67.2 ± 10.4 | 38.8 ±4.9 | 43.1 ± 9.0 | 59.4 ± 6.8 | 52.5 ± 11.1 | - | 64.6 ± 6.6 |
| SPECTF Heart | 75.5 ± 1.7 | 67.7 ± 3.4 | 71.6 ± 2.4 | 79.8 ± 2.0 | 80.6 ± 1.0 | 73.7 ± 3.2 | 80.0 ± 1.3 |
| MEAN | 77.6 ± 3.1 | 74.2 ± 2.6 | 75.3 ± 3.2 | 77.0 ± 2.9 | 79.8 ± 2.6 | 77.0 ± 3.3 | 81.3 ± 2.0 |

Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) has been processed to extract features that summarize the original SPECT images. As a result, 44 continuous feature patterns have been created for each patient. The 267 samples available are randomly divided into a training set of 80 samples and a testing set of 187 instances. The average recognition rate is 80.0% for our method, 67.7% for LDA, 71.6% for SDA, 79.8% for KP-LDA, and 80.6% for KG-LDA.

## IV. CONCLUSIONS

Feature extraction plays an important role in classification systems. In this paper a novel method for feature extraction based on mutual information and Fisher discriminant analysis (MI-FDA) was proposed. The goal of MI-FDA is to create new features from transforming the original features so that maximizes the mutual information between the transformed features and the class labels and minimizes the redundancy. In contrast to LDA-based algorithms which are based on second-order statistics, the proposed method is based on information-theoretic which is able to compares the nonlinear relationships between random variables (i.e., between a vector of features and the class label). The proposed method was evaluated using seven databases of UCI. On average, an accuracy rate of 81.3% was achieved using MI-FDA. The improvement in MI-FDA's classification rate over KG-LDA, KP-LDA, SDA, LDA, and MRMI-SIG are 1.9%, 5.6%, 8.0%, 9.6%, and 5.6%, respectively.

## REFERENCES


[1] J. M. Leiva-Murillo and A. Artes-Rodriguez, "Maximization of mutual information for supervised linear feature extraction," *IEEE Transactions on Neural Networks*, vol. 18, no. 5, pp. 1433–1441, 2007.

[2] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *The Journal of Machine Learning Research*, vol. 3, no. 7-8, pp. 1415–1438, 2003.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, second edition, 2001.

[4] R. P. W. Duin and M. Loog, "Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 732–739, 2004.

[5] S. Mika, G. Rätsch, J Weston, B. Schölkopf, A. Smola, and K.-R. Müller, "Constructing descriptive and discriminative nonlinear features: rayleigh coefficients in kernel feature spaces," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 623-628, May 2003.

[6] C. Yang, L. Wang, J. Feng, "On feature extraction via kernels," *IEEE Transactions on System, Man, and Cybernetics*, vol.38, no. 2, pp. 553-557, 2008.

[7] J. Yang, A.F. Frangi, J. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230-244, Feb. 2005.

[8] M. Kyperountas, A. Tefas and I. Pitas, "Weighted piecewise LDA for solving the small sample size problem in face verification," *IEEE Transactions on Neural Networks*, vol. 18, no. 2, March 2007.

[9] H. Yu and J. Yang, "A direct LDA algorithm for high dimensional data—with application to face recognition," *Pattern Recognition*, vol. 34, no. 10, pp. 2067–2070, 2001.

[10] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Transactions Neural Netw.*, vol. 17, no. 1, pp. 157–165, Jan. 2006

[11] M. Zhu and A. M. Martinez, "Subclass discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1274–1286, 2006.

[12] K. E. Hild II, D. Erdogmus, K. Torkkola, and J. C. Principe, "Feature extraction using information-theoretic learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1385–1392, 2006.

[13] C.R. Rao, Linear *Statistical Inference and Its Applications*, second ed. Wiley Interscience, 2002.

[14] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, pp. 2385-2404, 2000.

[15] C.L. Blake and C.J. Merz, "UCI Repository of Machine Learning Databases, " Dept. of Information and Computer Sciences, University of California, Irvine, http://www.ics.uci.edu/mlearn.

[16] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distribution," *Signal Processing*, vol. 16, no, 3, pp. 233-246, 1989.