2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Reconstructable Generalized Maximum Scatter Difference Discriminant Analysis

Kai Huang
Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai 200240, China
Email: huangkai888888@aliyun.com

Liqing Zhang
Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai 200240, China
Email: zhang-lq@cs.sjtu.edu.cn

***Abstract*—Dimensionality reduction is a key preprocessing step for many applications. Until our knowledge, unsupervised approaches such as PCA and ICA do not take label information of the original data into account, so a supervised approach such as Linear discriminant analysis (LDA) performs better on many classification tasks. Unfortunately, the classical LDA approach has shortcomings, such as the well-known *small size problem*, the *heteroscedastic problem* and the *(C-1) low rank problem*. The (C-1) low rank problem greatly limits the dimension of the extracted features. In addition, the calculation of the between-class and within-class scatter matrices in the classical LDA approach actually only takes account of the Mahalanobis distance like covariance distance of data centers and each data class, so if the dataset has very few classes or the data distribution of each class is not Gaussian-like but has some spatial structure in the feature space instead, classical LDA does not work well. In this paper we propose a dimensionality reduction approach which avoids the limitations of classical LDA and improves handling of the between-class scatter matrix. Our approach approach takes the distribution of data in each class into consideration to calculate the projection matrix. It does not assume that the data distribution of each class approximates Gaussian; each can have its own spatial structure. Experiments show that our method can obtain better projection directions than the classical LDA approach and greatly improve the classification accuracy. In addition, our approach is able to reconstruct the original signal well, while the classical LDA approach ignores the reconstruction property.**

I. Introduction

Since in data mining, machine learning and bioinformatics applications raw data such as EEG, ECG, FMRI and CT is usually represented in a very high dimension space, dimensionality reduction is important to characterize the features of the data. This is sometimes called the curse of dimensionality problem[1][2]. A feature selection method selects a subset from the original feature vector which gives the most discrimination between classes with the extracted dimensions. Approaches include max-dependency, max-relevancy and min-redundancy[3].

There has also been considerable research on a more flexible feature transformation method which seeks a transformation to project the original feature set onto a subspace to get coefficients to use as the final feature vector. Approaches based on linear transformations include principal component analysis (PCA)[4], independent component analysis (ICA)[5] and linear discriminant analysis (LDA)[2]. The main principle of PCA is to find a projection with the least information loss during the dimension reduction process. PCA aims to find a set of projection vectors which minimize reconstruction error in the least squares sense. ICA is quite similar to PCA but it takes the independent properties of the transformed signal using the kurtosis and negative entropy to measure the degree to which the data is non-Gaussian to find independent signal components. The LDA approach aims to find the optimal discriminant vectors by maximizing the ratio of the between-class distance to the within-class distance to achieve the maximum class discrimination. These approaches are based on linear transformations. There are also some nonlinear approaches such as the extended version of PCA-Kernel PCA[6], and manifold learning techniques such as locally linear embedding (LLE)[7], Hessian LLE, Laplacian eigenmaps, and LTSA[8].

Some approaches, such as PCA and LDA, have been extended to a high dimensional version,for example 2dPCA[9], 2dLDA[10], tensor PCA[11] and tensor LDA[12]. These can effectively reduce the computing cost, restore original structure information and improve the performance.

For the classification problem, there are two main groups of methods, unsupervised approaches and supervised approaches. Unsupervised approaches such as PCA and ICA do not consider the class label of data while LDA, as a supervised approach, calculates the projection vector with the class label information to maximize the separation between classes and maintain the highly-aggregated characteristics of each class.

However the classical LDA approach has several shortcomings. The most critical issue is the under-sampling problem also known as the singularity problem and high-dimensional problem[13]. In this case, the within-class scatter matrix is singular, and the projection matrix cannot be calculated with the generalized eigenvalue method or Cholesky decomposition. Approaches to solve this problem include subspace LDA (PCA+LDA)[14], null space LDA[15] and regularized LDA[16]. Subspace LDA just adopts a PCA dimension reduction approach before the LDA is applied. Null space LDA limits the search to the null space of the within-class scatter matrix, which is effective when the null space contains enough information. Regularized LDA adds a diagonal matrix to the standard within-class scatter matrix making the matrix invertible.

Another well-known problem of classical LDA is the het-

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

eroscedastic problem[17]. The classical LDA approach assumes the data in each class has a Gaussian distribution with the same covariance matrix. Some heteroscedastic LDA[17] approaches have been studied. In addition, classical LDA can obtain only (C-1) optimal projections of multivariate data[18][19] which is the (C-1) low rank problem. Jose and Antonio proposed a complementary space LDA approach which searches for the next projection vector in the subspace orthogonal to the space spanned by the achieved projection vectors[18][19].

In this paper we propose an improved method for LDA which tackles the problems mentioned above. In addition, our approach improves the unreasonable design of the between-class scatter matrix. It takes the non-Gaussian distribution of the data for each class into consideration for calculating the projection matrix. Experiment shows that the projection matrix achieved by our method is better than that of the classical LDA approach. In real data classification, the classification accuracy of our method is greatly improved. In addition, our method can make a reconstruction process to get the original signal with the coefficients which classical LDA cannot do.

The proposed approach is applied to ECG data analysis, although it is not limited to that, so related work on ECG feature extraction and classification will be discussed. There has been a lot of work on ECG feature extraction and classification. Some work decomposed the signal with a series of wavelet bases using projection coefficients as the feature vector[20]. Others decompose the ECG signal with the ICA components from the original signals with the FASTICA approach, which achieves quite high classification accuracy[21][22].

The paper is organized as follows, we first give a brief introduction to classical LDA based on the geometric meaning (Section 2.1). Then the limitations of classical LDA are introduced in Section 2.2-2.6. After that, we introduce some improvements based on the maximum scatter difference (MSD) corresponding to the several limitations of the classical LDA approach (Section 3). In Section 4, we show the effectiveness of our approach through several experiments. To validate the performance of the developed method, we compare with six other approaches. First, they are tested with UCI datasets (Section 4.1), and then they are applied to hospital ECG data analysis (Section 4.2). Finally the conclusion is given in Section 5.

mds

January 11, 2007

II. Limitations of Classical Linear Discriminant Analysis

In this section, we first briefly introduce the clear geometric meaning of LDA and then review its limitations, including the small size problem, the low rank problem, the heteroscedastic problem, unreasonable between-class scatter matrix, unconsidered distribution structure between classes and its inability to support reconstruction.

A. Geometric understanding of Classical LDA

Classical LDA computes a linear projection matrix G that maps each data point $a_j \in \mathbb{R}^m$ in the m-dimensional space to a vector y_i in l-dimensional space. It is actually the result of three equivalent optimization problems[1].

$$R(x) = \frac{x^T S_b x}{x^T S_w x}, R(x) = \frac{x^T S_b x}{x^T S_t x}, R(x) = \frac{x^T S_w x}{x^T S_t x} \quad (1)$$

S_w,S_b and S_t are the within-class scatter matrix, the between-class scatter matrix and the total scatter matrix. The matrix can be expressed as

$$\begin{aligned} S_w &= \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in A_i}(x-c^{(i)})(x-c^{(i)})^T \\ S_b &= \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in A_i}(c^{(i)}-c)(c^{(i)}-c)^T \\ &= \frac{1}{n}\sum_{i=1}^{k} n_i(c^{(i)}-c)(c^{(i)}-c)^T \\ S_t &= \frac{1}{n}\sum_{j=1}^{k}(x_j-c)^T(x_j-c) \end{aligned} \quad (2)$$

Here A_i means class i. n_i means the point count of class i. It is easy to understand that S_w is the sum of each class's covariance matrix. S_b is a weighted sum of the covariance matrix where the weight is the number in each data class. And S_t is the covariance matrix of all the points. It is easy to get the equation that $S_t = S_w + S_b$(notice the n_i in S_b). So S_t is named the total scatter matrix. The definition of Mahalanobis distance is as follows.

$$d(\overrightarrow{x}, \overrightarrow{y}) = \sqrt{(\overrightarrow{x}-\overrightarrow{y})^T S^{-1}(\overrightarrow{x}-\overrightarrow{y})} \quad (3)$$

It is easy to understand that inversion of the covariance matrix changes its eigenvalue to the reciprocal. The classical LDA approach uses the original covariance matrix so it gives a high weight to directions with high fluctuation. Actually the weight is the variance in this direction. The covariance matrix is a real symmetric matrix, and each real symmetric matrix has a orthogonal decomposition which can be expressed as the following form:

$$\begin{aligned} B &= P\begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} P^T \\ P &= [v_1, v_1, \cdots, v_n] \end{aligned} \quad (4)$$

In this equation, v_i is the eigenvector of the covariance matrix, and λ_i is the corresponding eigenvalue. It is obvious P is a orthogonal matrix. V_i forms a basis series. So each signal can be expressed as:

$$x = \sum_{i=1}^{n} a_i x_i \tag{5}$$

Through the above equation, the quadratic calculation of the covariance matrix can be expressed as the following form:

$$\begin{aligned} x^H B x &= \left(\sum_{i=1}^{n} a_i x_i\right)^H B \left(\sum_{j=1}^{n} a_j x_j\right) \\ &= \sum_{i,j=1}^{n} a_i a_j x_i^H B x_j \\ &= \sum_{i=1}^{n} \lambda_i |a_i|^2 \end{aligned} \tag{6}$$

So actually the main principle of the classical LDA approach is to maximize the ratio of the between-class covariance distance to the within-class covariance distance to achieve the maximum class discrimination.

B. Small size problem

In the classical LDA approach, the calculation of the ratio of the between-class covariance distance to the within-class covariance distance is actually a generalized Rayleigh quotient problem. Take the most widely used optimization problem which is the second optimization problem in Equation 1 as an example.

There are two approaches to solve the problem. The first is the method of Lagrange multipliers. It is clear from Equation 1 that x has infinite number of solutions. When the x is multiplied by a constant, $R(x)$ always keeps the same value (Offset the numerator and denominator). So the length of x is always restricted to make the denominator be 1. The restriction is used as a condition for the Lagrange method. The solution of this problem is to maximize the equation.

$$\begin{aligned} & c(x) = x^T S_b x - \lambda(x^T S_w x - 1) \\ & \Rightarrow \frac{dc}{dx} = 2S_b x - 2\lambda S_w x = 0 \\ & \Rightarrow S_b x = \lambda S_w x \end{aligned} \tag{7}$$

The equation above is a typical generalized eigenvalue problem. If the within-class scatter matrix S_w is invertible, this problem can be converted to a ordinary eigenvalue problem to calculate the result.

$$ {S_w}^{-1} S_b x = \lambda x \tag{8}$$

The within-class scatter matrix is the sum of each class's covariance matrix, and $rank(C) \leq rank(A) + rank(B)$, so the rank of the within-class scatter matrix is at most C (number of classes) less than the sample number.

For most application cases, such as image processing, video analysis, FMRI and CT, the dimension of the original signal is very high. Sometimes the training set is not that large and the number of cases is less than the dimension count. So in this case, the application of LDA is limited. Some modifications are needed in order to solve the problem.

C. (C-1) low rank problem

The classical LDA approach calculates the between-class scatter matrix by computing the covariance matrix of each class data's center. H_b can also be expressed in this form:

$$\begin{aligned} H_b &= \frac{1}{\sqrt{n}}[C^{(1)}, C^{(2)}, \cdots, C^{(i)}, \cdots, C^{(C)}] \\ C^{(i)} &= [(c_1^{(i)} - c), (c_2^{(i)} - c), \cdots, (c_{n_i}^{(i)} - c)] \end{aligned} \tag{9}$$

The vectors of each class are the same, so their rank is 1. Because each column of the matrix is reduced by the average of all columns, the weighted average is 0. In other words, they are linearly correlated.

$$\sum_{i=1}^{c} n_i (c^{(1)} - c) = 0 \tag{10}$$

Here n_i is the number of points in class i. Generally speaking, the centers of each class are linearly independent so the rank of H_b is $(C-1)$. Here C is the number of classes as before. Using the lemma below, it is easy to get that the rank of the between-class scatter matrix is (C-1).

lemma 1:for any m n matrix A $rank(A) = rank(A') = rank(AA') = rank(A'A)$.

Through the lemma above, it is easy to get the conclusion that the rank of S_b and H_b is the same, one less than the number of classes C. Although the C-1 is a upper bound, the actual value is quite close to this value. The rank of S_b is less than the upper bound $C-1$ only if the centers of different classes are on the same line linking to the origin.But for real data, this is rare so in most cases the rank of S_b is close to C-1.

So there is a paradox in classical LDA. If the number of classes is high, the accuracy of the results will be low, but if the number is comparatively low then the dimension extracted by LDA will be very low. Especially when using an SVM classifier, because multiclass LDA does not give good performance, the one to one, one to rest or directed acyclic graph strategy must be used. In this case, there are two classes in use for calculating the projection matrix. In other words, no matter how high the original data dimension, the calculated reduced dimension can only be 1. So in this condition, the classification result will be greatly affected.

D. Heteroscedastic Problem

Classical LDA assumes the distribution of data in each class is Gaussian. And here:

$$\begin{aligned} S_w &= \frac{1}{n} \sum_{i=1}^{k} S_{w_i} \\ S_{w_i} &= \sum_{x \in A_i} (x - c^{(i)})(x - c^{(i)})^T \end{aligned} \tag{11}$$

This approach assumes that the distribution of data in each class approximates a Gaussian distribution and the covariance matrices of all classes are similar. Actually for much real

data, this assumption cannot be the case. They may have some type of structure in the feature space or not have a good clustering structure. Data for different classes may be not linearly separable, or even not nonlinearly separable. Different class's data can overlap with each other.

E. Unreasonable between-class scatter matrix

In the traditional LDA approach, the design of the between-class scatter matrix is unreasonable as S_b in equation 2.

Section 2.1 explained the geometric meaning; it actually calculates the between-class covariance distance of class data centers. This approach seems simple and effective but has some problems. It assumes the centers of each class as a whole approximate a Gaussian distribution. This is often not the case for real data. On the other hand, if the number of classes is small, the data centers can be very sparse. In this situation, even if they actually have a Gaussian distribution, the calculated axis direction can have a large error because data is sparse. Besides this, it also makes the calculation difficult. Actually the low rank problem of dimension (C-1) is caused by this design.

F. Reconstructability problem

In general signal decomposition, the reconstructability of the signal can be very important. With wavelet decomposition or Fourier decomposition the original signal can be reconstructed from the decomposition coefficients. As for the PCA approach, because the basis achieved is normalized and orthogonal, the result of multiplying the projection matrix and the original signal is the coefficients in the new coordinate space which can be directly used for reconstruction. But for the classical LDA approach, the achieved basis is not orthogonal and cannot be directly used for reconstruction.

For the generalized eigenvalue problem in equation 11, S_w and S_b are real symmetric positive definite matrices because $(A^{(-1)})' = (A')^{(-1)} = A^{(-1)}$. so ${S_w}^{-1}$ is a symmetric matrix. When A and B are both symmetric, the necessary and sufficient condition for AB to be a symmetric matrix is $AB = BA$. So generally ${S_w}^{-1}S_b$ is not a symmetric matrix so the eigenvector obtained by calculating the Eigenvalue problem is not orthogonal.

Or consider this issue with another solution, Cholesky decomposition. Here, $G^{-1}S_b(G^{-1})^H$ is real and symmetric. Its eigenvectors are orthogonal. Orthogonal eigenvectors should be adopted with a transformation $G^H x = y$. $S_w = GG^H$ actually the G is:

$$[v_1, v_2, \cdots, v_n] \begin{pmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_n} \end{pmatrix} \tag{12}$$

So it is not an orthogonal matrix. The projection vector x after the transformation is not an orthogonal projection vector.

Reconstuctability is important, because the LDA approach also can be used to automatically find the characteristic of each type, or find the most discriminative characteristic between different types. In the application of ECG analysis, it can be used to find the discriminative waveform between confusing diseases and the typical waveforms of each heart disease.

III. Improved methods

For the above-mentioned problems, we propose a series of improvements. Each of the improvements removes at least one limitation of the classical LDA approach. Each individual improvement as a method can be used separately.

A. Pairs of classes as a whole

As mentioned above, classical LDA regards the centers of classes as a whole to calculate the axis direction while calculating the between-class scatter matrix. It has not taken the distribution of each class into consideration. It is quite possible that the centers of classes are scattered or have high variance but the data overlaps with other classes. In this case, the projection matrix cannot achieve good dimension reduction performance. Here we propose a method to take data for each pair of classes to extract the projection direction.

$$R(\mathbf{x}) = \frac{x^T S_b x + w x^T S_{11} x}{x^T S_w x}$$

$$S_{11} = \sum_{i,j} w_{ij} \sum_{x \in \{c_i, c_j\}} (x - c^{(ij)})(x - c^{(ij)})^T \qquad i \neq j \tag{13}$$

here $c^{(ij)}$ is the average of two centers c_i, c_j. This approach also deals with the (C-1) low rank problem. So a better method is to take all the data of each class into consideration but not the connection of the two classes' centers. So here we consider not only the centers of each class as in classical LDA but also each point of each pair of classes.

$$S_{11} = \sum_{i,j} w_{ij} \sum_{x \in A_i A_j} (x - c^{(ij)})(x - c^{(ij)})^T \qquad i \neq j \tag{14}$$

here $c^{(ij)}$ is the mean of all the data of class $A_i A_j$. Here w_{ij} is the weight. It can be determined by the total number of points in each class pair, the experimental similarity or importance. Such as the following equation:

$$w_{ij} = \frac{n_i + n_j}{n} \tag{15}$$

Here n_i is points number of class i, and n means points number of all classes. A data class with more data points than the others or a class vulnerable to confusion in classification or a class with significant importance, all these cases should be assigned a higher weight than the others.

The geometric explanation of these two approaches is as follows. The first improved method is similar to the original LDA with only two classes of data. The calculation of the between-class scatter matrix with only two classes is as follows:

$$B = P \begin{pmatrix} \lambda & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} P^T \tag{16}$$

$$\begin{aligned} P &= [v_1, v_2, \cdots, v_n] \\ x^H Bx &= |x| \cos\theta \lambda \cos\theta |x| \\ &= |x|^2 \cos\theta^2 \lambda \end{aligned} \tag{17}$$

Whether the projection direction for the between class scatter matrix is good depends on the angle between the projection direction and the connection of the two class centers.

And our first improved method is quite similar to this condition. It takes the direction of the center connection of each class pair into consideration. So its projection direction depends on the sum of all the center connections of each class pair. Its equation is as follows:

$$\begin{aligned} x^H S_{11} x &= \sum_{i \neq j, ij = c_i\ c_j} |x| \cos\theta_{ij}\ \lambda_{ij} \cos\theta_{ij}\ |x| \\ &= \sum_{i \neq j, ij = c_i\ c_j} |x|^2 \cos\theta_{ij}^2\ \lambda_{ij} \end{aligned} \tag{18}$$

As for the first improved method, the rank of every singular matrix is 1. According to the lemma: matrix A and matrix B are all $S * N$ matrix, $A + B$ come to matrix C, $rank(C) \leq rank(A) + rank(B)$. There are C_n^2 class pairs, so the final rank is less than the number of combinations C_n^2.

$$rank(S_{11}) \leq C_n^2 \tag{19}$$

In this case, the low rank problem is eased, but not that much. This problem still exists.

As for the second improved method, every matrix is not like the classical LDA, It consider each point of each pair of classes as a whole. Now the rank of a single matrix is one less than the sum of the number of points in the two classes. The rank of the sum of the C_n^2 matrices is less than sum of the rank of each matrix:

$$rank(S_{11}) \leq \sum_{i \neq j} num(data_i) + num(data_j) - 1 \tag{20}$$

The condition here is similar to the problem discussed in the section about the C-1 low rank problem. Although it is an upper bound, the actual value is close to this upper bound. The relation of these three upper bounds is as follows:

$$C - 1 \leq C_n^2 \leq \sum_{i \neq j} num(data_i) + num(data_j) - 1 \tag{21}$$

The second improved method further eases the C-1 low rank problem but this approach still has a limitation; it assumes that the distribution of each pair of classes as a whole is Gaussian.

To summarize, the first approach can enhance the rationality but the number of total projection vectors cannot be increased that much. The second approach not only takes the distribution of each class into consideration but also solves the (C-1) limitation of dimension. If the dataset is large enough, the signal can have enough dimensions to be reconstructed.

B. Pairs of points in pairs of classes

Although the two methods mentioned above, to some extent, solve two limitations of LDA, they still have some problems. They calculate the between-class covariance distance, regarding each class as a whole as having a Gaussian distribution and calculating its axis direction. But each class of data has some spatial structure. The direction with big variance of the two classes of data may not be a good projection direction. It may not give a good separation plane. We make another improvement based on the previous approach.

$$\begin{aligned} &R(\mathbf{x}) = \tfrac{x^T S_b x + w_1 x^T S_{11} x + w_2 x^T S_{oo} x}{x^T S_w x} \\ &S_{oo} = \textstyle\sum_{i,j} w_{ij} \sum_{x \in A_i, y \in A_j} S_{xy} \qquad i \neq j \\ &S_{xy} = (x - c^{(ij)})(x - c^{(ij)})^T + (y - c^{(ij)})(y - c^{(ij)})^T \end{aligned} \tag{22}$$

The matrix used here with only two vectors is quite similar to the special case with only two classes of data centers. Its rank is one. Its eigenvector corresponding to the nonzero eigenvalue is the connection of two points.

This approach takes each pair of points from the two classes into consideration. The final projection direction is the sum of the angle between the direction and the connection for each pair of points. This approach also solves the (C-1) dimension limitation. But here each point has the same impact on the final result. It may not be the best choice.

To give a geometric explanation, it determines the projection direction according to the angle between it and each pair of points of different classes. It is as follows:

$$\begin{aligned} x^H S_{oo} x &= \sum_{i \neq j, ij \in A_i\ A_j} |x| \cos\theta_{ij}\ \lambda_{ij} \cos\theta_{ij}\ |x| \\ &= \sum_{i \neq j, ij \in A_i\ A_j} |x|^2 \cos\theta_{ij}^2\ \lambda_{ij} \end{aligned} \tag{23}$$

The rank of it is as follows:

$$rank(S_{oo}) \leq \sum_{i \neq j} num(data_i) * num(data_j) \tag{24}$$

The relation of different improved methods' rank is as follows:

$$\begin{aligned} &C - 1 \leq C_n^2 \leq \sum_{i \neq j} num(data_i) + num(data_j) - 1 \\ &\leq \sum_{i \neq j} num(data_i) * num(data_j) \end{aligned} \tag{25}$$

For the low rank problem, this is better than the previous two approaches.

Here the rank is still a upper bound. But only when data points are on the same line passing the origin will the rank be affected. For real data, this condition is rare so the rank of the matrix is close to the upper bound.

This approach fully considers the distribution of each class or between classes. At the same time, the C-1 low rank problem is also eased. Also this approach has flexibility for the further extension of adding weight to each data point pair which have significant importance for performance improvement.

C. Considering distribution structure between classes

It is obvious, features in the feature space have a distribution which may be very scattered. The above method which calculates pairs of points from pairs of classes considers the spatial distribution of each class. It is better than the LDA approach which only considers the centers of each class. On the other hand, a better classification plane is determined more by the closer point pairs. So a natural extension of the above method is to assign a weight to each point depending on its position.

$$S_{oo} = \sum_{i,j} w_{ij} \sum_{x \in A_i, y \in A_j} w(d_{xy}) S_{xy} \qquad i \neq j \tag{26}$$

The easiest way of determining a weight is to use the reciprocal of distance ($\frac{1}{d_{xy}}$),

$$\mathrm{w}(d_{\mathrm{xy}}) = {d_{\mathrm{xy}}}^{-n}$$

or in the following form:

$$\mathrm{w}(d_{\mathrm{xy}}) = \begin{cases} = 1 & \text{if} d_{\mathrm{xy}} \in (N_{\%} \sim M_{\%})(Max(d_{x_i y_j}) - Min(d_{x_i y_j})) \\ = 0 & \text{if} d_{\mathrm{xy}} \notin (N_{\%} \sim M_{\%})(Max(d_{x_i y_j}) - Min(d_{x_i y_j})) \end{cases} \tag{27}$$

For example, we can assign 1 or ${d_{xy}}^{-1}$ to the point pairs whose distance is between the range $(2_{\%} \sim 10_{\%})(Max(d_{x_i y_j}) - Min(d_{x_i y_j}))$. Because data points of different classes may overlap with each other to some extent we add a weight added on the a interval which depends on the distance between the data pair of different type. So our approach adds nonzero weight to data pairs close to each other but not the most near to each other and excludes the data pairs far from each other. In the experiment, this approach gives extremely good performance. In this way, a more reasonable classification plane and projection direction can be achieved.

The geometric meaning of this approach is to fully consider the distribution of data points of different classes which are close to each other and exclude the influence of data points which are far from the classification plane.

D. Extension to maximum scatter difference problem

The above method has solved the (C-1) low rank problem, heteroscedastic problem, the problem of the unreasonable between-class scatter matrix, and the problem of the data distribution between classes not being considered. But the small size problem still has not been solved. Actually classical LDA is a multi-objective optimization problem. It is about simultaneously optimizing the following equation:

$$\arg\max_x \{x^T S_b x\} \qquad \arg\min_x \{x^T S_w x\} \tag{28}$$

A multi-objective optimization problem can be solved by optimizing several equations simultaneously, or by combining multiple optimization targets into one. Classical LDA combines the maximization and minimization problem into one by dividing the first one by the second one, which actually leads to the small size problem. So the S_w must be full rank. But if we just change the division to subtraction:

$$\arg\max_x \{x^T\, S_b\, x + w_1 x^T\, S_{11}\, x + w_2 x^T\, S_{11o}\, x - w_3 x^T\, S_w\, x\} \tag{29}$$

In this way, the problem can be easily solved. We can use the similar method for classical LDA and Maximum Scatter Difference Discriminant Analysis. We just do Eigenvalue decomposition for the following matrix.

$$S_b + w_1 S_{11} + w_2 S_{11o} - w_3 S_w \tag{30}$$

Then we order the eigenvectors by its corresponding eigenvalues. Then the ordered eigenvectors is the projection vector we wanted which actually is our solution for the objective function. Some work has been done with a similar idea[23][24].

E. Reconstructability

The first three improvements above still retain a form similar to classical LDA. From the above discussion, it is clear multiple projection vectors are not orthogonal to each other. If using this basis to transform the original dataset, the coefficients cannot be used to reconstruct the signal, so the coefficients should be adjusted.

For example, we set the projection vector$\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, to get a set of coefficients $\mathbf{p} = \alpha_1 \mathbf{x}_1 + \cdots + \alpha_k \mathbf{x}_k$, α_i is determined by the orthogonality condition: $\mathbf{x}_i^T(\mathbf{z} - \mathbf{p}) = 0, i = 1, \ldots, k$. let $n \times k$ Order matrix $X = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_k \end{bmatrix}$, The $\mathbf{p}$ written as $\mathbf{p} = X\alpha, \alpha = (\alpha_1, \ldots, \alpha_k)$, the k equations can be expressed as a matrix form, as follows:

$$\begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_k^T \end{bmatrix} (\mathbf{z} - X\alpha) = X^T(\mathbf{z} - X\alpha) = \mathbf{0} \tag{31}$$

To get $X^T X\alpha = X^T \mathbf{z}$. Row vector of X is linearly independent, $\mathrm{rank}X = k$, use nature of matrix rank $\mathrm{rank}(X^T X) = \mathrm{rank}X = k$, To know $X^T X$ is a k-order reversible square matrix, α has a unique solution:

$$\alpha = (X^T X)^{-1} X^T \mathbf{z} \tag{32}$$

So the orthogonal projection amount is obtained

$$\mathbf{p} = X\alpha = X(X^T X)^{-1} X^T \mathbf{z} \tag{33}$$

The projection vectors for the extended maximum scatter difference problem are orthogonal to each other. In other words, $(X^T X) = I$ so

$$\alpha = (X^T X)^{-1} X^T \mathbf{z} = X^T \mathbf{z} \tag{34}$$

This signal can be used to find the discriminative characteristic of an ECG waveform.

IV. Experiments

In our experiments, to demonstrate the applicability of our approach, we compare the performance of our method with other methods on 11 datasets from the UCI database. Finally, we use real data, specifically ECG signals, to test the effectiveness of our approach. We compare our method with other dimension reduction methods. ICA, PCA, classical LDA and complement space LDA approaches are tested with the same dataset.

A. UCI datasets analysis

Although we mainly focus on ECG analysis, and compare our work with other approaches for ECG feature extraction, our approach is not confined to the field of ECG analysis. To demonstrate the effectiveness of our approach, we test on eleven datasets from the UCI database and compare the classification accuracy of projected features with different approaches. The selected datasets are Car Evaluation, Connect-4, Image Segmentation, Letter Recognition, Nursery, Statlog (Shuttle), Steel Plates Faults, MiniBooNE particle identification, Wall-Following Robot Navigation Data, and a Wine Quality and Yeast dataset. Each dataset has more than 4 types. We divide large datasets into training data and test data with a ratio of 1:10, and small datasets with a ratio of 1:2. The dimension of the original data is reduced to about half of the original dimension to be tested. The final results are listed in Table 1. The best two results for each dataset are in italics and bold. It is very clear that our approach is significantly superior to other methods for almost all datasets. This is enough to prove the effectiveness of our method-Reconstructable Generalized Maximum Scatter Difference Discriminant Analysis(RGMSDDA or in short RGM).

TABLE I
Comparisons of different approaches' classification accuracy for 11 UCI databases

apps(%)	*ICA*	*PCA*	*LDA*	*LDAc*	*RGM*
car	75.75	74.54	***91.15***	90.91	***93.40***
connect	65.84	***66.39***	66.33	66.07	***67.32***
image	***94.48***	89.05	65.86	85.00	***94.67***
letter	93.69	92.10	95.26	***96.29***	***95.65***
nursery	86.77	79.39	***93.81***	92.88	***95.08***
statlog	98.89	98.29	97.76	***99.29***	***99.74***
steel	71.25	***78.26***	73.88	64.40	***100***
mini	83.40	83.01	***89.43***	***90.20***	89.01
wall	73.35	80.21	***87.07***	84.40	***87.15***
wine	***79.21***	75.99	79.20	53.84	***79.50***
yeast	50.54	***60.92***	57.82	57.41	***61.59***

B. Hospital ECG data analysis

To evaluate the performance of our method, we test it on a large dataset which we get from our local hospital. The dataset consists of 3000 pieces high quality 12-leads ECG records. Each piece includes about 10 to 25 beats and there are 65,716 beats in total. These records are detected from a wide range of people: men and women, young and old, healthy and unhealthy. The doctors' conclusion are taken as the label for the beats, which is one of the following 6 types(Symbol in parenthesis): Normal beat(N), Left bundle branch block beat(L), Right bundle branch block beat(R), Left ventricular hypertrophy(V), Sinus bradycardia(S),Electrical axis left side(E). After the preprocess step for the raw ECG signal, we get the following single heartbeat segments: 19400 of N type, 7056 of L type, 10080 of R type, 6720 of V type, 14540 of S type, 7920 of E type. Next we split the dataset into two parts: the train part and the test part. We use the train part to get the projection vectors. And then we use the train part for SVM model training. The model are then use for the classification of testing dataset. We split the original dataset into two parts randomly: the train part consists of 20000 beats, the test part consists of 45716 beats.

TABLE II
ECG dataset used for experiment(train and test set)

Type	N	L	R	V	S	E
Train	5397	2445	3164	2256	4694	2044
Test	14003	4611	6916	4464	9846	5876

The detail of the dataset we use is listed in Table 2. At first, the divided ECG beats is aligned by the R peak and then normalized to same length. Then each beats is processed with short time Fourier transform to transform the original signal into time and frequency domain. Because valuable diagnosis feature may be obvious in the time and frequency domain but not in the time domain only. Then different dimension reduction methods which are shown in Fig 1 are adopted to reduce the dimension of the original signal and extract valuable features. At last, we use SVM with RBF kernel to compare the classification accuracy of these features. Figure 1 compares the classification accuracy of different feature dimensions. It is obvious that the classification accuracy increase smoothly with dimension for our method RGMSDDA. On the whole, the RGMSDDA is with best performance compared with other methods.

The Table 3 has listed out all the accuracy value for each class of ECG beats. It is shown that Our method is the best for each ECG beat type compared with other dimension reduction methods.

V. Conclusions

From the geometric meaning of the LDA, our paper analyses problems and design flaws of classical LDA and suggests a series of improvements. Our approach solves the under sampling problem, the (C-1) low rank problem, and the heteroscedastic problem of classical LDA. At the same time, the problem that

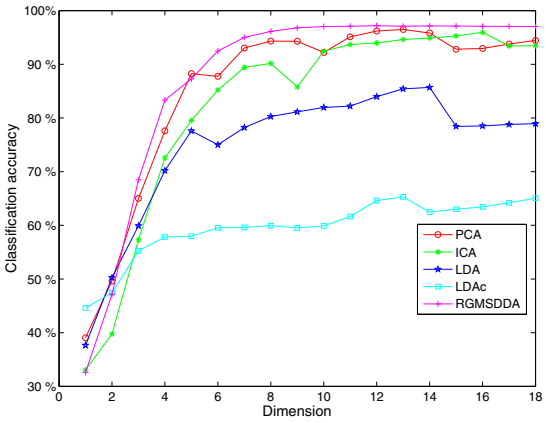

Fig. 1. Cross validation comparison for different models

TABLE III
COMPARISON WITH OTHER VECTOR-BASED ALGORITHM

Model	N	L	R	V	S	E
ICA	96.5	85.5	94.5	94.1	95.6	96.1
PCA	97.4	93.3	96.0	97.2	97.6	95.7
LDA	91.7	72.7	78.5	90.4	82.9	82.7
LDAc	86.0	10.8	48.7	81.5	97.1	61.3
RGM	99.5	94.3	97.6	94.5	98.0	96.5

the original LDA method does not fully consider the spatial distribution of each class and the spatial relations between classes has been improved by our approach. Besides this, a weight is assigned to each point on the distance between points of different classes to increase the impact of closer points which greatly improves the performance. In comparisons between cardiology features, PCA, ICA, LDA, complementary space LDA, RGMSDDA, our approaches RGMSDDA outperform others. They find more effective projection vectors and greatly improve the classification accuracy. Besides this, our approach also solves the reconstructability problem. It can reconstruct the original signal from the coefficients after dimension reduction. Good performance on 11 UCI datasets proves that our method is not limited to ECG analysis but can also be used for other types of data.

VI. ACKNOWLEDGMENTS

The work was supported by the National Natural Science Foundation of China (Grant No. 91120305 and 61272251).

REFERENCES

[1] Keinosuke Fukunaga. *Introduction to statistical pattern recognition (2nd ed.)*. Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[2] Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. New York: Springer-Verlag, 2001.

[3] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[4] Timmerman M.E. Principal component analysis (2nd ed.). i. t. jolliffe. *Journal of the American Statistical Association*, 98:1082–1083, January 2003.

[5] Aapo Hyvrinen. Independent component analysis. *Neural Computing Surveys*, 2, 2001.

[6] Bernhard Schlkopf, Alexander Smola, and Klaus-Robert Mller. Kernel principal component analysis. In Wulfram Gerstner, Alain Germond, Martin Hasler, and Jean-Daniel Nicoud, editors, *Artificial Neural Networks ł ICANN'97*, volume 1327 of *Lecture Notes in Computer Science*, pages 583–588. Springer Berlin / Heidelberg, 1997. 10.1007/BFb0020217.

[7] David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10):5591–5596, 2003.

[8] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal of Scientific Computing*, 26:313–338, 2002.

[9] Jian Yang, David Zhang, Alejandro F. Frangi, and Jing-yu Yang. Two-dimensional pca: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26:131–137, January 2004.

[10] Ming Li and Baozong Yuan. 2d-lda: A statistical linear discriminant analysis for image matrix. *Pattern Recogn. Lett.*, 26:527–532, April 2005.

[11] Haiping Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Multilinear principal component analysis of tensor objects for recognition. In *in Proc. Int. Conf. on Pattern Recognition*, pages 776–779, 2006.

[12] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J. Maybank. General tensor discriminant analysis and gabor featuresforgaitrecognition,ieee trans. *Pattern Anal. Mach. Intell*, page 1700, 2007.

[13] Yunhui He. Solving undersampled problem of lda using gram-schmidt orthogonalization procedure in difference space. In *Proceedings of the 2009 International Conference on Advanced Computer Control*, pages 153–157, Washington, DC, USA, 2009. IEEE Computer Society.

[14] Peter N. Belhumeur, Jo?o P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, 1997.

[15] Jieping Ye and Tao Xiong. Computational and theoretical analysis of null space and orthogonal linear discriminant analysis. *J. Mach. Learn. Res.*, 7:1183–1204, December 2006.

[16] Jerome H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[17] Marco Loog and Robert P. W. Duin. Linear dimensionality reduction via a heteroscedastic extension of lda: The chernoff criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:732–739, 2004.

[18] José M. Leiva-Murillo and Antonio Artés-Rodríguez. Maximization of mutual information for supervised linear feature extraction. *IEEE Transactions on Neural Networks*, 18(5):1433–1441, 2007.

[19] Chandra Dhir and Soo Lee. Discriminant independent component analysis. In Emilio Corchado and Hujun Yin, editors, *Intelligent Data Engineering and Automated Learning - IDEAL 2009*, volume 5788 of *Lecture Notes in Computer Science*, pages 219–225. Springer Berlin / Heidelberg, 2009.

[20] Qibin Zhao and Liqing Zhang. Ecg feature extraction and classification using wavelet transform and support vector machines. In *Neural Networks and Brain, 2005. ICNN B '05. International Conference on*, volume 2, pages 1089 –1092, oct. 2005.

[21] Xing Jiang, Liqing Zhang, Qibin Zhao, and S. Albayrak. Ecg arrhythmias recognition system based on independent component analysis feature extraction. In *TENCON 2006. 2006 IEEE Region 10 Conference*, pages 1 –4, nov. 2006.

[22] Yang Wu and Liqing Zhang. Ecg classification using ica features and support vector machines. In *ICONIP (1)*, pages 146–154, 2011.

[23] Fengxi Song, David Zhang, Dayong Mei, and Zhongwei Guo. A multiple maximum scatter difference discriminant criterion for facial feature extraction. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(6):1599–1606, 2007.

[24] Jun Gao and Lili Xiang. Laplacian maximum scatter difference discriminant criterion. In *ICAIC (1)*, pages 691–697, 2011.