Journal of Management Science and Statistical Decision
Vol.7 No1.(2010) p74-78

Regularized Least Squares LDA and Its Application in Text Classification

ZunXiong Liu
School of Information Engineering
East China Jiaotong University
Nanchang, China
e-mail:liuzunx@tom.com

LiHui Zeng
School of Information Engineering
East China Jiaotong University
Nanchang, China
e-mail:wssycmissyou@163.com

***Abstract*—Linear Discriminant Analysis (LDA) is a well-known technique for dimensionality reduction and classification, while the classical LDA formulation fails when the total scatter matrix is singular, encountered usually in undersampled problems. In this paper, regularized Least Squares LDA (RLS-LDA) based on the elastic net, is proposed to handle the problems, and the resulting models are robust and sparse. Firstly, the theories about linear regression and regularization are explored, and the equivalence relationship between the least squares formulation and LDA for multi-class classifications under a mild condition is summarized. Secondly, the construction of RLS-LDA is presented. Performance evaluations of these approaches are conducted on benchmark collection of text documents. Results demonstrate the effectiveness of the proposed RLS-LDA and it's the RLS-LDA based on the elastic net that is better than others.**

Keywords: LDA; linear regression; RLS-LDA

I. INTRODUCTION

This paper focuses on optimizing on the least squares LDA objective function with L1-norm, L2-norm and the elastic net on the parameters. There are currently significant interests in the related problems with high-dimensional and undersampled data. The sample sizes are much smaller than the data dimensionality for undersampled problems, such as face images, microarray expressions data and text documents. With them the classical LDA is not applicable because the corresponding total scatter matrices are singular. Many extensions of classical LDA have been proposed in the past to overcome the singularity problem, including subspace LDA[1 2], Uncorrelated LDA[3], Orthogonal LDA[3], regularized LDA[4 5], penalized LDA[6], and so on.

LDA can be applied for dimensionality reduction, in which each derived feature is a linear combination of all the original features, the equivalence relationship between the least squares formulation and LDA under a mild condition are put forward[7]. The coefficients stored in the transformation matrix are typically nonzero, i.e., the resulting models are often not sparse. However, sparsity often leads to easy interpretation and good generalization ability of the resulting model[8]. It's known that the linear regression with L1-norm, also named the Least Absolute Shrinkage and Selection Operator (Lasso)[9], can automatically select variables for the model, resulting in the sparse model. There are other regularized LS with different coefficient penalties, such as L2-norm and p-norm. Based on them, the elastic net comes true with some advantages. The elastic net algorithm is introduced into the least squares LDA, used for dimensionality reduction with high dimensional and sparse data. Through classification experiments on benchmark collection of text documents, the prediction performances with different regularized methods are compared. The results demonstrate the effectiveness of the proposed sparse least squares LDA.

II. LINEAR REGRESSION TECHNIQUES

A. Linear Discriminant Analysis(LDA)

The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible. Assume we have a set of d-dimensional samples $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \Re^d$ and $y_i \in \{1,2,...,k\}$ denotes the class label of the *i*-th sample, *n* is the sample size, *d* is the data dimensionality, and *k* is the number of classes. The data matrix $X = [x_1, x_2, ..., x_n]$ is partitioned into *k* classes as $X = [X_1, X_2, ..., X_k]$, where $X_i \in \Re^{d \times n_i}$, n_i is the size of the *i*-th class X_i and $\sum_{i=1}^{k} n_i = n$. For the k class problem, *(k-1)* projection vectors $w_i (i = 1, \cdots, k-1)$ should be found, arranged in columns of a projection matrix $W = [w_1, w_2, ..., w_{k-1}]$, so that any observation x_i can be represented a linear combination of the projection vectors, the encoding coefficients are $W^T x_i$. In Linear Dimensionality Analysis, three scatter matrices, called within-class, between-class and total scatter matrix are defined as follows:

$$S_w = \frac{1}{n}\sum_{i=1}^{k}\sum_{x \in X_i}(x-\mu_i)(x-\mu_i)^T \quad (1)$$

$$S_b = \sum_{i=1}^{k} n_i(\mu_i-\mu)(\mu_i-\mu)^T \quad (2)$$

$$S_t = \frac{1}{n}\sum_{j=1}^{n}(x_j-\mu)(x_j-\mu)^T \quad (3)$$

Where the centroid of the i-th class is $\mu_i = \frac{1}{n_i}\sum_{x \in X_i} x$, $\mu = \frac{1}{n}\sum_{\forall x} x = \frac{1}{n}\sum n_i \mu_i$ is the global centroid. It follows from the definition that $S_t = S_w + S_b$.

B. Linear Regression

Linear regression problems are usually coped with ordinary least squares(OLS) approximations, where the response variable y is approximated by the predictor in X(both the observations and the targets are centered), and the coefficients for each variable of X are contained in w, calculated by minimizing the following cost function:

$$L(w) = \|y - Xw^T\|^2 \tag{4}$$

Where $W = [w_1, w_2, ..., w_k]$ is the weight matrix, the solution for W can be given by

$$w_{LS} = (XX^T)^{-1} Xy \tag{5}$$

However, if some bias is allowed, estimators can be found with lowed mean square error than OLS when tested on an unseen set of observations. A common way to implement this is by introducing some constraints on the coefficients w. The described methods use constraints with the L1-norm, the L2-norm of w , or both.

The lasso is a penalized least squares method, imposing a constraint with L1-norm of the regression coefficients. Thus, the lasso estimates w_{lasso} are obtained by minimizing the lasso criterion

$$w_{lasso} = \arg\min_w \|y - Xw^T\|^2 + \theta \|w\|_1 \tag{6}$$

Replacing the L1-norm in the constraint with the L2-norm gives

$$w_{\text{ridge}} = \arg\min_w \|y - Xw^T\|^2 + \lambda \|w\|_2^2 \tag{7}$$

Lasso has proven to be a very powerful regression and variable selection technique, while it has a few limitations. If $d>n$, i.e., there are more variables than observations, lasso choose a maximum of n variables, which is clearly unsatisfactory. The elastic net regression method[10] was developed to overcome these drawbacks. The elastic net penalty is a convex combination of the constraints from the ridge and lasso penalties. For any non-negative λ and θ, the elastic net estimates w_{EN} are given as follows

$$w_{EN} = (1+\lambda)\left\{\arg\min_w \|y - Xw^T\|^2 + \lambda \|w\|_2^2 + \theta \|w\|_1\right\} \tag{8}$$

which is the lasso problem for $\lambda = 0$ .Given a value of λ, in the elastic net setting, Lars returns the solutions corresponding to all possible values of θ with the computation cost of a single least square fit.

III. Motivation and Details of Regularized LS-LDA

A. Relationship between ULDA and Multivariate Linear Regression

The classical LDA is not applicable to the text documents, where the total scatter matrix is singular. ULDA is a natural extension of classical LDA for undersampled problems. The optimal $\mathbf{W}_{\text{ULDA}}$ of ULDA is computed by solving the following optimization problem[3]:

$$\mathbf{W}_{\text{ULDA}} = \arg\min_W \{trace(\mathbf{S}_b^L (\mathbf{S}_t^L)^+)\} \tag{9}$$

The optimal $\mathbf{W}_{\text{ULDA}}$ consists of the top eigenvectors of S^+S_b corresponding to the nonzero eigenvalues[3], provided that the total scatter matrix S_t is singular.

In Least Squares Linear Discriminant Analysis[11] , $\mathbf{H}_b = \frac{1}{\sqrt{n}}[\sqrt{n_1}(\mu_1 - \mu), ..., \sqrt{n_k}(\mu_k - \mu)]$, $\mathbf{H}_t = \frac{1}{\sqrt{n}}(X - \mu e^T)$, μ_i is the centroid of the i-th class, μ is the global centroid and e is the vector of all ones of length n. Then $\mathbf{S}_b$ and $\mathbf{S}_t$ can be expressed as follows:

$$\mathbf{S}_b = \mathbf{H}_b \mathbf{H}_b^T \tag{10}$$

$$\mathbf{S}_t = \mathbf{H}_t \mathbf{H}_t^T \tag{11}$$

$\mathbf{H}_t$ can be computed via the singular value decomposition(SVD) of the matrix, i.e. $\mathbf{H}_t = \mathbf{U}\Sigma\mathbf{V}^T$, $\mathbf{U}$ and Vare orthogonal, $\Sigma = \begin{pmatrix} \Sigma_t & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, $\Sigma_t \in \Re^{t \times t}$ is diagonal, and t=rank($\mathbf{S}_t$).Then

$$\mathbf{S}_t = \mathbf{H}_t \mathbf{H}_t^T = \mathbf{U}\Sigma\Sigma^T\mathbf{U}^T = \mathbf{U}\begin{pmatrix} \Sigma_t^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T \tag{12}$$

Let $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$ be a partition of $\mathbf{U}$, where $\mathbf{U}_1 \in \Re^{n \times t}$ and $\mathbf{U}_2 \in \Re^{n \times (n-t)}$,so $\mathbf{U}_2$ lies in the null space of $\mathbf{S}_t$,i.e., $\mathbf{U}_2^T \mathbf{S}_t \mathbf{U}_2 = \mathbf{0}$ .For $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$ and $\mathbf{S}_w$ is positive semi-definite, we can obtain $\mathbf{U}_2^T \mathbf{S}_b \mathbf{U}_2 = \mathbf{0}$. So we have

$$\mathbf{U}^T \mathbf{S}_b \mathbf{U} = \begin{pmatrix} \mathbf{U}_1^T \mathbf{S}_b \mathbf{U}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{13}$$

Denote

$$\mathbf{B} = \Sigma_t^{-1} \mathbf{U}_1^T \mathbf{H}_b \in \Re^{t \times k} \tag{14}$$

B can be computed via the singular value decomposition(SVD) of the matrix, i.e., $\mathbf{B} = \mathbf{P}\dot{\Sigma}\mathbf{Q}^T$, $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal and $\dot{\Sigma} \in \Re^{t \times k}$ is diagonal. From Eq. (10), it follows that

$$\Sigma_t^{-1} \mathbf{U}_1^T \mathbf{S}_b \mathbf{U}_1 \Sigma_t^{-1} = \mathbf{B}\mathbf{B}^T = \mathbf{P}\dot{\Sigma}(\dot{\Sigma})^T \mathbf{P}^T = \mathbf{P}\Sigma_b \mathbf{P}^T \tag{15}$$

The multivariate regression model with the class label as the output following form:

$$y = xw^T \tag{16}$$

A popular approach for estimating w is the least squares, via the minimization of the following objective function:

$$L(w) = \left\|\tilde{Y} - \tilde{X}\mathbf{W}^T\right\|^2 \tag{17}$$

Where $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n] \in \Re^{d \times n}$ as the centered data matrix X and $\tilde{Y} = (\tilde{Y}_{ij})_{ij} \in \Re^{n \times k}$ as the centered data indicator matrix Y, respectively. $\tilde{x}_i = x_i - \bar{x}$, and $\tilde{Y}_{ij} = Y_{ij} - \bar{Y}_j$, where $\bar{x} = \frac{1}{n}\Sigma_{i=1}^n x_i$ and $\bar{Y}_j = \frac{1}{n}\Sigma_{i=1}^n Y_{ij}$

We have $\tilde{Y} = Y_3$ as the class indication matrix[11].That is $Y_3 = \frac{k-1}{\sqrt{k}} Y_2$, where $Y_2(ij) = \begin{cases} 1 & y_i = j, \\ -1/(k-1) & otherwise. \end{cases}$.

The optimal weight matrix W for multivariate linear regression in Eq. (5) becomes $\mathbf{W}_{LS} = \mathbf{S}_t^{+}\mathbf{H}_b$

Recall that in ULDA, the optimal transformation matrix $\mathbf{W}_{ULDA}$ consists of the top eigenvectors of $\mathbf{S}_t^{+}\mathbf{S}_b$ corresponding to the nonzero eigenvalue.The relationship between $\mathbf{W}_{ULDA}$ and $\mathbf{W}_{LS}$ is argued following. From Eq. (12), Eq. (13) and Eq. (14), the matrix $\mathbf{S}_t^{+}\mathbf{S}_b$ can be decomposed as follows:

$$\mathbf{S}_t^{+}\mathbf{S}_b = \mathbf{U}\begin{pmatrix}(\mathbf{\Sigma}_t^2)^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T\mathbf{H}_b\mathbf{H}_b^T$$

$$= \mathbf{U}\begin{pmatrix}(\mathbf{\Sigma}_t^2)^{-1}\mathbf{U}_1^T\mathbf{H}_b\mathbf{H}_b^T\mathbf{U}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T$$

$$= \mathbf{U}\begin{pmatrix}\mathbf{\Sigma}_t^{-1}\mathbf{B}\mathbf{B}^T\mathbf{\Sigma}_t & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T$$

$$= \mathbf{U}\begin{pmatrix}\mathbf{\Sigma}_t^{-1}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\Sigma_b & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}\mathbf{P}^T\mathbf{\Sigma}_t & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\mathbf{U}^T$$

Thus, the optimal transformation matrix of ULDA is given by

$$\mathbf{W}_{ULDA} = \mathbf{U}_1\mathbf{\Sigma}_t^{-1}\mathbf{P}_q \tag{18}$$

Since only the first q diagonal entries of Σ_b is nonzero, $\mathbf{P}_q$ consists of the first q columns of p. On the other hand,

$$\mathbf{S}_t^{+}\mathbf{H}_b = \mathbf{U}\begin{pmatrix}(\mathbf{\Sigma}_t^2)^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T\mathbf{H}_b$$

$$= \mathbf{U}_1\mathbf{\Sigma}_t^{-1}(\mathbf{\Sigma}_t^{-1}\mathbf{U}_1^T\mathbf{H}_b)$$

$$= \mathbf{U}_1\mathbf{\Sigma}_t^{-1}\mathbf{P}\dot{\mathbf{\Sigma}}\mathbf{Q}^T$$

$$= \mathbf{U}_1\mathbf{\Sigma}_t^{-1}\mathbf{P}_q[\dot{\mathbf{\Sigma}}_q, \mathbf{0}]\mathbf{Q}^T$$

$$= \left[\mathbf{W}_{ULDA}\mathbf{\Sigma}_{bq}^{0.5}, \mathbf{0}\right]\mathbf{Q}^T$$

Where $\dot{\Sigma}_q, \Sigma_{bq} \in \Re^{rxq}$ consists of the first q rows and the first q columns of $\dot{\Sigma}, \Sigma_b$, respectively. It follows that

$$\mathbf{W}_{LS} = \left[\mathbf{W}_{ULDA}\mathbf{\Sigma}_{bq}^{0.5}, \mathbf{0}\right]\mathbf{Q}^T \tag{19}$$

Where Q is orthogonal. It is clear from Eq. (19) that the difference between W_{ULDA} and W_{LS} lies in the diagonal matrix $\mathbf{\Sigma}_{bq}^{0.5}$.

When the matrix Σ_{bq} is an identity matrix of size q, that is, W_{LS} and W_{JLDA} are essentially equivalent, under a mild conditionM1[7]

$$\text{rank}(S_b) = \text{rank}(S_b) + \text{rank}(S_w) \tag{20}$$

which holds in many applications involving high-dimensional and undersampled data.

B. Regularized LS-LDA

Based on the equivalence relationship established in the last section, ULDA(the natural extension of classical LDA) formulation can be extended using the regularization technique.

Regularization is commonly used to control the complexity of the model and improve the generalization performance. Linear regression using the L2-norm regularization, called ridge regression[11], minimizes the penalized sum-of-squares cost function. By using the class indicator matrix $\tilde{Y}$ in Eq. (17), we obtain the L2-norm regularized least squares LDA formulation (called "LS-LDA$_2$") by minimizing the following objective function:

$$L_2(W,\lambda) = \left\|\tilde{Y} - \tilde{X}W^T\right\|^2 + \lambda\|W\|_2^2 \tag{21}$$

Where $W = [w_1, w_2, ..., w_k]$,and $\lambda > 0$ is the regularization parameter.

In mathematical programming, it is known that sparseness can often be achieved by penalizing the L1-norm of the variables. It has been introduced into the least squares formulation and the resulting model is called lasso. Based on the established equivalence relationship between ULDA and least squares, we derive the L1-norm least squares LDA formulation (called"LS-LDA$_1$") by minimizing the following objective function:

$$L_1(W,\theta) = \left\|\tilde{Y} - \tilde{X}W^T\right\|^2 + \theta\|W\|_1 \tag{22}$$

Where $W = [w_1, w_2, ..., w_k]$,and θ is the regularization parameter.

The elastic net proposed by Zou and Hastie solves a regression problem regularized by the L1-norm and L2-norm in a fast and effective manner. We derive the elastic net least squares LDA formulation (called "LS-LDA$_{EN}$") by minimizing the following objective function:

$$L_{EN}(W,\lambda,\theta) = \left\|\tilde{Y} - XW^T\right\|^2 + \lambda\|W\|_2^2 + \theta\|W\|_1 \tag{23}$$

The optimal w_j^* ,for $1 \le j \le k$, is given by

$$w_j^* = \arg\min_{w_j}\left(\left\|\tilde{Y} - Xw_j^T\right\|^2 + \lambda\|w\|_2^2 + \theta\|w\|_1\right) \tag{24}$$

IV. EXPERIMENT RESULTS

In this section, a collection of multi-label data sets is used for simulation experiments, showing the effectiveness of our proposed algorithm. In the experiments, five methods including ULDA, as well as LS-LDA and its regularization versions LS-LDA$_1$, LS-LDA$_2$ and LS-LDA$_{EN}$ are compared in performance. All these LDA methods are used to project the data into a lower-dimensional space where the K-Nearest-Neighbor(KNN) algorithm is employed for classification with the multi-label data, and the results are consistent with our the theoretical analyses. Here standard document collections, TDT2 are used as experimental data. The TDT2 corpus consists of data collected during the first half of 1998 and taken from six sources, includeing two newswires (APW, NYT), two radio programs (VOA, PRI) and two television programs (CNN, ABC). It consists of 11201 on-topic documents which are classified into 96 semantic categories. In this dataset, those documents appearing in two or more categories are removed, leaving us the largest 30 categories with 9394 documents in 30 categories, as described in Table 1.

TABLE I. 30 SEMANTIC CATEGORIES FROM TDT2 USED IN EXPERIMENTS

Category	Num of doc	Category	Num of doc	Category	Num of doc
20001	1844	20021	74	20056	66
20002	1222	20023	167	20065	63
20005	58	20026	72	20070	441
20008	71	20032	131	20071	238
20009	52	20033	145	20074	56

Category	Num of doc	Category	Num of doc	Category	Num of doc
20012	226	20037	65	20076	272
20013	811	20039	141	20077	120
20015	1828	20044	407	20086	140
20018	104	20048	160	20087	98
20019	123	20047	123	20096	76

The samples are high dimensional, whose categories are taken from 2class to 10class. These data are partitioned randomly, so the training set consists of two-thirds of the whole class samples, leaving the rest making the test dataset. The whole class samples are processed with the same dimensionality reduction techniques, one of the LDA approaches. Then KNN classifiers are applied to assign the testing samples to some category. The splitting was repeated 10 times. The resulting average accuracies of different algorithms are summarized in Figure 1.

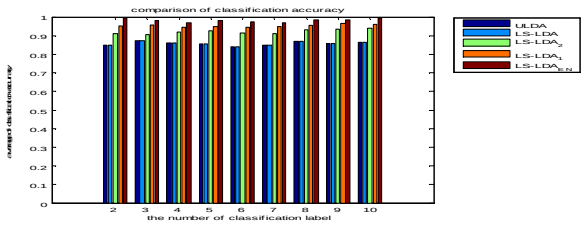

Figure 1. Comparison of all algorithms on text document datasets.

From figure 1, it can be observed that the regularized algorithms including LS-LDA$_1$, and LS-LDA$_2$, and LS-LDA$_{EN}$ perform much better than ULDA and LS-LDA without regularization, and the proposed LS-LDA$_{EN}$ performs the best in this data set. The figure 2 and 3 shows coefficient traces w for a linear model fit in the LS-LDA$_1$ and LS-LDA$_{EN}$.

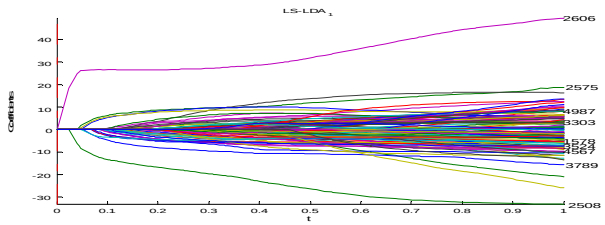

Figure 2. Coefficient Traces for LS-LDA$_1$

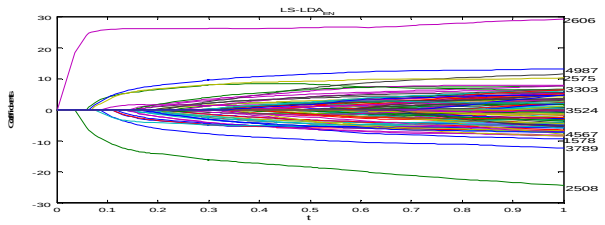

Figure 3. Coefficient Traces for LS-LDA$_{EN}$

Figure 2 and 3 tell that the coefficients for LS-LDA$_{EN}$ are smaller than that for LS-LDA$_1$, and the LS-LDA$_1$ coefficients are unstable, while LS-LDA$_{EN}$ coefficient traces are good. Moreover, the LS-LDA$_{EN}$ results in sparse and stable model with variable selection technique, variable with coefficients of zero are effectively omitted from the model.

In summary, the experiments above show that

- condition(M1) is more likely to hold for high-dimensional data;
- LS-LDA is equivalent to ULDA when condition(M1) holds;
- LS-LDA and ULDA achieve similar classification perfomance even when condition(M1) does not hold.

Thus, LS-LDA can be applied as a general least squares formulation for LDA for multi-class classifications. From figure 1, we can observe from the figure that the regularized methods perform much better than ULDA and LS-LDA, and LS-LDA$_1$ is comparable to LS-LDA$_{EN}$. The sparse formulation LS-LDA$_{EN}$ performs the best for this data set.

V. CONCLUSION

ULDA for multi-label classifications can be formulated as a least squares problem under a mild condition ,which tends to hold for undersampled problem. Based on the equivalence relationship extensions, the Regularized Least Squares LDA is proposed in this paper. The experiments on a collection of multi-label datasets are conducted to validate the effectiveness with the presented algorithm. Experimental results show that the performance of the proposed the regularized Least Squares LDA effectively omitted the variables from the model. The proposed regularized Least Squares LDA performs well for the text document data set. As the extension of the research, the effectiveness of this the regularized Least Squares LDA model for learning from labeled and unlabeled data, will be examined. The regularized Least Squares LDA model will be generalized into feature extraction in the web page relevance analysis.

ACKNOWLEDGMENT

This research is sponsored in part by funds from the Jiangxi Provincial Department of Education Research Foundation under Grant No. GJJ09507.

REFERENCES

[1] P.N. Belhumeour, J.P. Hespanha, and D.J. Kriegman,"Eigenfaces vs Fisherfaces: Recognition using class specific linear projection," IEEE Trans Pattern Analysisand Machine Intelligence, vol 19,pp. 711-720,Jul 1997.

[2] D. L. Swets and J.Y. Weng. "Using discriminant eigenfeatures for image retrieval." IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18,pp. 831-836, Aug 1996.

[3] J. Ye. "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems." The Journal of Machine Learning Research,vol. 6, pp. 483-502, 2005.

[4] J.H. Friedman. "Regularized discriminant analysis." Journal of the American Statistical Association, vol. 84, pp.165-175, Mar 1989.

[5] Y. Guo, T. Hastie, and R. Tibshirani. "Regularized discriminant analysis and its application in microarrays."Technical report, Stanford University, 2003.

[6] T. Hastie, A. Buja, and R. Tibshirani. "Penalized discriminant analysis." Annals of Statistics, vol. 23, pp.73-102, Feb 1995.

[7] J. Ye and T. Xiong. "Computational and theoretical analysis of null space based and orthogonal linear discriminant analysis." Department of

Computer Science and Engineering, Arizona State University,vol. 7, pp. 1183-1204, 2006.

[8] T. Hastie, R. Tibshirani, and J.H. Friedman. "The elements of statistical learning : data mining, inference, and prediction." The Mathematical Intelligencer, Springer New York,vol. 27, pp. 83-85, June 2005.

[9] R. Tibshirani,"Regression Shrinkage and Selection via the Lasso."*Journal of the Royal Statistical Society. Series B,vol. 58,pp. 267-288,1996*.

[10] H. Zou and T. Hastie "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society Series B.vol. 67, pp. 768, 2005.

[11] J. Ye. "Least squares linear discriminant analysis." Proceedings of the 24th international conference on Machine learning, USA. Vol. 227, pp. 1087-1093,2007.