Feature Extraction Using Discriminant Graph Laplacian Principal Component Analysis with Application to Biomedical Datasets

Muhammad Aminu¹ and Noor Atinah Ahmad²*

¹,²School of Mathematical Sciences, Universiti Sains Malaysia, 11800, Penang, Malaysia

*nooratinah@usm.my

Abstract. In this paper, we propose a manifold learning method called discriminant graph Laplacian principal component analysis (DGLPCA) for feature extraction. The proposed method projects high dimensional data into a lower dimensional subspace while preserving much of the intrinsic structure of the data. Moreover, DGLPCA integrates maximum margin criterion into its objection function to improve class separability in the lower dimensional space. The effectiveness of the proposed method is demonstrated on two publicly available biomedical datasets taken from UCI machine learning repository. The results show that our proposed method provides more discriminative power compared to other similar approaches.

1. Introduction

Biomedical datasets are rapidly growing and becoming more and more complex [1]. These datasets are generally associated with high-dimensional features and multiple classes. The ill-structured nature of the biomedical data requires intelligent machine learning and data mining algorithms for automated analysis in order to make logical inferences from the stored raw data [2]. A diverse set of machine learning and data mining algorithms have been previously used to extract useful information from biomedical data [3]. The research in this area continues to grow partly because machine learning makes it possible to model data extremely well, without using strong assumptions about the modeled system.

Feature extraction is a machine learning technique that maps the original high dimensional data to a lower dimensional subspace while maximizing the variability of the new feature space. Several feature extraction methods have been used in the literature [4-6]. The most widely used feature extraction method is the principal component analysis (PCA). It is well-known that PCA preserves only global Euclidean structure of data. The method is not designed to detect the underlying manifold structure hidden in the high dimensional data. The efforts to include manifold learning techniques into PCA give rise to several variants of graph-Laplacian PCA [7-9]. However, this transformation matrix does not include class information and therefore may not be optimal for classification. When data comes with class labels, we would like to include this information into our feature extraction method so that the classification ability of the method is improved. Although it has been shown that the PgLPCA method in [9] is superior to the graph-Laplacian PCA (GLPCA) in [8] in terms of both clustering performance and noise reduction, neither method actually take classification explicitly into consideration. PgLPCA and GLPCA are completely unsupervised with regards to the class label
information, and this might sometimes lead to unsatisfactory performances of the methods in classification problems. Also, both PgLPCA and GLPCA use the \( k \)-nearest neighbor method to construct the graph. However, this approach is non-parameter free and sensitive to noise.

In this paper, we propose a novel algorithm for feature extraction and discriminant analysis called discriminant graph Laplacian PCA (DGLPCA). DGLPCA is essentially developed from PCA but has significant advantages over the ordinary PCA method. Benefiting from the objective of maximum margin criterion (MMC) [10], we integrate the objective function of MMC into GLPCA to improve its classification ability. We show how the solutions to this method can be obtained by solving a generalized eigenvalue problem.

The following points highlight the contributions of this paper:

- A new approach for PCA based on graph embedding is developed, which is called discriminant graph Laplacian PCA (DGLPCA). DGLPCA encodes discriminant information into the objective of graph Laplacian PCA to improve its classification ability.
- Instead of using \( k \)-nearest neighbour method for graph construction, in our proposed DGLPCA method, we construct a graph based on the class information of samples that capture as much of the discriminative information of data as possible. Our approach for constructing the graph is parameter free and can capture the discriminative information of data to a higher extent.
- We use the same approach as in maximum margin criterion [10] to maximize class separation in our proposed methods.
- The performance of DGLPCA is evaluated using two biomedical datasets publicly available from the UCI Machine Learning repository [11].

The remainder of the paper is organized as follows: In Section 2, we describe some related works from [8] and [9] followed a description of our newly proposed discriminant graph Laplacian PCA (DGLPCA) in Section 3. The extensive experimental results are presented in Section 4. Finally, we give some concluding remarks in Section 5.

2. Related Works

Assuming we have a data matrix \( X = \begin{bmatrix} x_1, x_2, \ldots, x_n \end{bmatrix}^T \), where each row represents a sample (or a patient) and each column represents a feature. Each patient \( x_i \) belongs to one of the \( C \) classes \( \{ X_1, X_2, \ldots, X_C \} \), where \( X_c = \begin{bmatrix} x_{1c}, x_{2c}, \ldots, x_{nc} \end{bmatrix} \), \( c = 1, 2, \ldots, C \) and \( n_c \) denotes the number of patients in class \( c \). PCA aims to find a projection matrix \( U \) that minimizes the reconstruction error. The projection matrix \( U \) is obtained by solving the following minimization problem:

\[
\min_U \| X - UU^TX \|_F^2,
\]

subject to \( U^TU = I \), where \( I \) is an identity matrix.

To incorporate manifold learning into PCA, the graph Laplacian PCA (GLPCA) [8] is introduced and the modified minimization problem is given by:

\[
\min_U (-tr(U^TXX^TU) + \alpha tr(U^TXLX^TU)),
\]

subject to \( U^TXX^TU = I \), where \( \alpha \geq 0 \) controls the smoothness of the new representation. The matrix \( L = D - S \) is called the graph Laplacian [12] where \( S \) is an \( n \times n \) weight matrix defined by

\[
S_{ij} = \begin{cases} 
\exp \left( -\frac{x_i - x_j^2}{\sigma} \right); & \text{if } x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i) \\
0; & \text{otherwise}
\end{cases}
\]
such that $N_k(x_j)$ denotes the $k$ nearest neighbour set of $x_j$, $\sigma$ is a user specified parameter, and $D$ is a diagonal matrix whose entries are equal to the column (or row) sum of $S$. The optimal transformation matrix $U$ of GLPCA can be computed by applying an eigen decomposition on $XX^T - \alpha XLX^T$ and $XX^T$.

3. Proposed Methodology

3.1. Maximum Margin Criterion

In order to enhance the classification ability of GLPCA, we seek to capture the discriminatory information using the Maximum Margin Criterion (MMC) as proposed in [10]. Let $L = D - S$ be the graph Laplacian in (2) where $S$ is an $n \times n$ weight matrix defined as

$$ S = \text{diag} \left( S^{(1)}, S^{(2)}, \ldots, S^{(c)} \right), $$

such that $S^{(c)}$ is an $n_c \times n_c$ weight matrix having entries all equal to $1/n_c$ (i.e., $S^{(c)}_{ij} = 1/n_c$ for $i, j = 1, 2, \ldots, n_c$) and $D$ is a diagonal matrix whose entries are equal to the column (or row) sum of $S$.

With definitions $S_b = XX^T$ and $S_w = XLX^T$, the MMC finds the maximizing subspace of the following objective function,

$$ \max_U \text{tr} \left( U^T \left( S_b - S_w \right) U \right); $$

$$ \text{s.t.} \; U^T XX^TU = I $$

It can be shown that the matrices $S_b$ and $S_w$ are respectively the between class scatter matrix and the within class scatter matrix which are used to define the Linear Discriminant Analysis (LDA). The LDA method has long been a benchmark method for discrimination and classification. But it is well known that the objective function of LDA can easily lead to singularities in small sample-size problem. This problem is affectively avoided in MMC.

3.2. The proposed DGLPCA method

In PgLPACA and GLPCA, the graph Laplacian is constructed using the $k$-nearest neighbor method. The method defines $S$ as in (3) which requires user-defined parameters $k$ and $\sigma$ which are the number of nearest neighbors and the width of the heat kernel respectively. Although the performance of GLPCA (and PgLPACA) depends on the choices of $k$ and $\sigma$, it is not obvious how these parameters should be chosen for a particular problem. In MMC, $S$ is defined based on the information of the class labels and it is parameter free. Hence, by combining the objective function of GLPCA (2) and the objective function of MMC (5), we can improve discriminative ability of GLPCA while eliminating the need to adjust $k$ and $\sigma$.

Similar to [13-15], we propose combining the objective functions in a multi-objective optimization problem of the form

$$ \begin{cases} 
\min_U \text{tr} \left( U^T \left( -XX^T + \alpha XLX^T \right) U \right); \\
\max_U \text{tr} \left( U^T \left( S_b - S_w \right) U \right); \\
\text{s.t.} \; U^T XX^TU = I . 
\end{cases} $$

Solving the multi-objective optimization problem (6) is an attempt to find a lower dimensional subspace that preserves the manifold structure of the data and maximizes the margin between different classes at the same time. These two properties are very important for clustering and classification tasks. The objective function (6) can be rewritten in the following equivalent form:

$$ \max_U \text{tr} \left( U^T \left( I - \alpha L + \eta (S - L) \right) X^T U \right), $$
\[ s.t. \quad U^T XX^T U = I, \]

where the regularization parameter \( \eta > 0 \) balances the contribution from the two parts. Using Lagrange multiplier method, the constraint optimization problem (7) can be transformed to a generalized eigenvalue equation of the form

\[ (X(I - \alpha L + \eta(S - L))X^T)U = \Lambda XX^T U. \]

where the solution to (8), namely \( U_{opt} \), is the principle eigenspace of \( U \). Let the columns of \( U_{opt} \) be \( u_0, u_2, \ldots, u_{d-1} \) ordered according to their eigenvalues \( \lambda_0 \geq \lambda_i \geq \cdots \geq \lambda_{d-1} \). For an arbitrary sample \( x_i \in \mathbb{R}^n \), the embedding is represented by:

\[ x_i \rightarrow y_i = U_{opt}^T x_i, \quad U_{opt} = [u_0, u_2, \ldots, u_{d-1}], \]

where \( y_i \) is a representation of \( x_i \) in a \( d \)-dimensional space \((d \ll m)\) and \( U_{opt} \) is a \( m \times d \) matrix.

Note that, in biomedical datasets, the number of features is very high compared to the number of the samples, i.e. the dimensionality of the data is usually higher than the number of samples. In such cases, the matrix \( XX^T \) in (9) can become singular. We apply the Tikhonov regularization technique \([16]\) to solve the singularity problem. The generalized eigenvalue problem (8) becomes:

\[ (X(I - \alpha L + \eta(S - L))X^T + \sigma I)U = \Lambda (XX^T + \sigma I) U, \]

where \( \sigma > 0 \) is the regularization parameter.

### 4. Experimental Results

In this section, we investigate the performance of our proposed discriminant graph Laplacian principal component analysis (DGLPCA) in extracting features capable of separating classes in data. All of our experiments have been performed on an Intel Core i7, 3.20GHz Windows PC with 8GB memory.

#### 4.1. Data and evaluation criteria

Two publicly available datasets taken from UCI machine learning repository are used in our experiments. They are Thyroid disease and Wisconsin diagnostic breast cancer data sets. The Wisconsin (diagnostic) data contains features computed from a digitized image of a fine needle aspirant (FNA) of a breast mass. This dataset consists of 569 patterns with 30 features categorized into two classes: benign and malignant. The original data was obtained by Wolberg and Mangasarian \([17]\).

The Thyroid disease dataset contains observations from 7200 patients characterized by 15 binary and 6 continuous patients attributes. Each of the patient is assigned to one of the three following classes: 1) Normal, not hyperthyroid, 2) Hyperfunction and 3) Subnormal functioning.

We used the \( K \)-NN classifier to identify the observations after feature extraction where \( K \) refers to the number of nearest neighbors. The classification accuracies are measured based on \( K = 5 \) for the breast cancer dataset, while for the Thyroid disease dataset, we set \( K = 20 \). The performance of our proposed DGLPCA is evaluated in comparison with the ordinary PCA and GLPCA methods.

#### 4.2. Discussion

To obtain reliable experimental results, we reshuffle all the datasets randomly in the experiments. For each of the datasets, \( p = (33, 50, 67) \) percent of samples are randomly selected as training sets and the rest are used as test sets. The training samples are used to determine the projection vectors. The test samples are then mapped into lower dimensional subspace using the obtained projection vectors. Classification is then performed in the lower dimensional subspace using the nearest neighbor classifier.

Table 1 presents the classification accuracies obtained on the breast cancer data using the PCA, GLPCA, DGLPCA and the baseline methods. For the baseline method, classification is performed in the original 30-dimensional space without any dimensionality reduction. The number of nearest
neighbors for constructing the nearest neighbor graph of GLPCA was set to $k = 3$ and the heat kernel width was set to $\sigma = 70$ for optimum performance (empirically determined). For DGLPCA, the regularization parameters $\alpha$ and $\eta$ are set to $1\times10^4$. As can be seen from Table 1, the performance of PCA, GLPCA and DGLPCA varies with different value of $p$. However, our proposed DGLPCA obtained the least classification error in all of the three cases.

Similar to the experiment on breast cancer data, we used the nearest neighbor classifier to classify the test samples of the Thyroid data using features extracted by the PCA, GLPCA, DGLPCA and the baseline method. The two parameters $k$ and $\sigma$ in GLPCA were set to 10 and 0.02 respectively (empirically optimized). While the regularization parameters $\alpha$ and $\eta$ in our proposed DGLPCA method were both set to $1\times10^4$. We extracted only 3 features using the different techniques and report the results in Table 2. As can be seen, our algorithm outperforms all the other methods in all cases. We observed that, even when the training data is small, the performance of DGLPCA is significantly higher than that of GLPCA and all other methods with large training data. This indicates that our method is able to learn from small training sets.

In summary, our proposed DGLPCA method obtained the best performance in all experiments. The experimental results further illustrate that, based on the datasets we used, the newly proposed DGLPCA method demonstrates more discriminating power than the ordinary GLPCA method. This results highlight the effectiveness of combining ideas from GLPCA and MMC in feature extraction.

**Table 1.** Classification results on the Wisconsin diagnostic breast cancer data.

| Method   | 33% Training        | 50% Training        | 67% Training        |
|----------|---------------------|---------------------|---------------------|
|          | error (%) | # of features | error (%) | # of features | error (%) | # of features |
| Baseline | 8.16      | 30             | 7.72      | 30             | 7.89      | 30             |
| PCA      | 8.42      | 2              | 8.77      | 2              | 7.89      | 2              |
| GLPCA    | 13.68     | 2              | 9.47      | 2              | 6.84      | 2              |
| DGLPCA   | 3.68      | 2              | 4.56      | 2              | 2.11      | 2              |

**Table 2.** Classification results on the Thyroid disease data.

| Method   | 33% Training        | 50% Training        | 67% Training        |
|----------|---------------------|---------------------|---------------------|
|          | error (%) | # of features | error (%) | # of features | error (%) | # of features |
| Baseline | 7.04      | 21             | 6.80      | 21             | 6.75      | 21             |
| PCA      | 7.40      | 3              | 7.33      | 3              | 7.58      | 3              |
| GLPCA    | 7.38      | 3              | 6.92      | 3              | 7.63      | 3              |
| DGLPCA   | 5.00      | 3              | 3.75      | 3              | 3.75      | 3              |

5. Conclusions

In this paper, we have introduced an efficient feature extraction technique called discriminant graph Laplacian PCA. Our approach combines ideas from spectral graph analysis and maximum margin criterion to provide an efficient and effective approach for feature extraction. We showed empirically that the newly proposed approach performed better than the previous approaches in extracting features capable of discriminating the classes in a data. Other than improving the discriminative power of graph-Laplacian PCA, the proposed DGLPCA also provides a parameter-free graph construction method so that its performance is independent on the graph parameters.

However, more work needs to be done in evaluating the performance of the proposed method on other datasets. Also, the performance of the proposed method using different classification techniques such as support vector machine, random forest and neural networks should further be investigated.
Acknowledgements
The authors would like to acknowledge UNIVERSITI SAINS MALAYSIA for the financial support through Research University Grant (RUI) (acc. No. 1001/PMATHS/8011040) and the USM Bridging Grant (acc. No. 304/PMATHS/6316049).

References
[1] Wasan S, Bhatnagar V and Kaur H 2006 The impact of data mining techniques on medical diagnostics, *Data Science Journal*, 5, pp. 119–126.
[2] Huang S, Zhou J, Wang Z, Ling Q and Shen Y 2017 Biomedical informatics with optimization and machine learning, *EURASIP J Bioinform Syst Biol*, 2017:4.
[3] Foster K R, Koprowski R and Skufca J D 2014 Machine learning, medical diagnosis, and biomedical engineering research: Commentary, *BioMedical Engineering Online*, 13–94.
[4] Boulesteix A -L 2004 PLS dimension reduction for classification with microarray data, *Statistical applications in genetics and molecular biology*, 3(1), 1–30.
[5] Liu J -X, Wang D, Gao Y -L, Zheng C -H, Shang J -L, Liu F and Xu Y 2017 A joint-L2,1-norm-constraint-based semi-supervised feature extraction for RNA-Seq data analysis, *Neurocomputing*, 228, 263–269.
[6] Liu H, Zhang Z, Zhang X, Yang Y, Zhang Z, Liu X, Wang F, Han Y and Zhang C 2018 Dimensionality Reduction for Identification of Hepatic Tumor Samples Based on Terahertz Time-Domain Spectroscopy, *IEEE Transactions on Terahertz Science and Technology*, 8(3), 271-277.
[7] Shahid N, Kalofolias V, Bresson X, Bronstein M, and Vandergheynst P 2015 Robust principal component analysis on graphs, in *Proceedings of the IEEE International Conference on Computer Vision*, 2812-2820.
[8] Jiang B, Ding C, Luo B, and Tang J 2013 Graph-Laplacian PCA: Closed-form solution and robustness, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3492-3498.
[9] Feng C -M, Gao Y -L, Liu J -X, Zheng C -H, and Yu J 2017 PCA based on graph laplacian regularization and P-norm for gene selection and clustering, *IEEE transactions on nanobioscience*, 16(4), 257-265.
[10] Li H, Jiang T, and Zhang K 2004 Efficient and robust feature extraction by maximum margin criterion, in *Advances in neural information processing systems*, 97-104.
[11] UCI repository of machine learning databases, University of California-Irvine, Department of Information and Computer Science.
[12] He X and Niyogi P 2004 Locality preserving projections, in *Advances in neural information processing systems*, 153-160.
[13] Gui J, Jia W, Zhu L, Wang S -L, and Huang D -S 2010 Locality preserving discriminant projections for face and palmprint recognition, *Neurocomputing*, 73(13-15), 2696-2707.
[14] Gui J, Sun Z, Jia W, Hu R, Lei Y, and Ji S 2012 Discriminant sparse neighborhood preserving embedding for face recognition, *Pattern Recognition*, 45, (8), 2884-2893.
[15] Zhang J, Wang J, and Cai X 2017 Sparse locality preserving discriminative projections for face recognition, *Neurocomputing*, 260, 321-330.
[16] Wang Q, Meng Z, and Li X 2017 Locality adaptive discriminant analysis for spectral–spatial classification of hyperspectral images, *IEEE Geoscience and Remote Sensing Letters*, 14(11), 2077-2081.
[17] Wolberg W H and Mangasarian O L 1990 Multisurface method of pattern separation for medical diagnosis applied to breast cytology, *Proceedings of the national academy of sciences*, 87(23) 9193-9196.