Dimensionality Reduction Analysis of Binary Classification Between Ripe Avocado and Eggplant

1st Nurfitri Anbarsanti
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore
nurfitri006@e.ntu.edu.sg

***Abstract*—In some applications, such as object detection and recognition, bioinformatics, and data mining, high data dimensions burden robust and accurate recognition due to a lack of data population knowledge and limited training samples. Machine learning methods for dimensionality reduction should aim to eliminate or adjust detrimental or unreliable dimensions for the classification task. Principal component analysis (PCA) and linear discriminant analysis (LDA) are rudimentary tools to reduce dimensionalities and extract features. In this study, we have performed sets of experiments that has used Kernel SVM and Mahalanobis Distance Classifier to analyze the role of PCA and LDA in the classification. This study will also discuss the results of their application on several machine learning classifications.**

***Index Terms*—PCA, LDA, SVM, Kernel SVM, Mahalanobis Distance Classifier**

I. Introduction

In some applications, such as object detection and recognition, bioinformatics, and data mining, high data dimensions burden robust and accurate recognition due to a lack of data population knowledge and limited training samples. Consequently, dimension reduction becomes perhaps the most important of these recognition systems.

Both principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [1] are rudimentary tools to reduce dimensionality and extract features. PCA is to maximize the data reconstruction capability of the features, and DA is to maximize the discriminatory power of the features.

Dimensionality reduction as a feature extraction has two objectives. One objective is to reduce the computational complexity of the classification process and minimize the loss of information. The second objective is to avoid the generalization problem and thus improve their accuracy and robustness. To achieve the first objective maximizing the information transmitted into the low-dimensional subspace. PCA maximizes the information structure of the data, and hence, it is optimal for data reconstruction; the discriminative information plays more roles in pattern recognition.

Although some approaches apply PCA for dimension reduction in the areas of the face recognition and object detection, many researchers turn to LDA for feature extraction [2], [3], due to the prevalent view that the discrimination of features is the most important for classification [2], [3] and its effectiveness in achieving the aforementioned first objective. The second objective of dimensionality reduction is, however, far from straightforward that it can be achieved by discriminant analysis [2].

PCA escalates the variances of the extracted features and thus downplays the reconstruction error and removes noise. The best data representation may not perform well from the classification because the total scatter matrix is contributed by both the within and between-class variations. To differentiate one class from the other class, the discrimination of the features is more important. LDA is assumed to be one of the efficient ways to extract the discriminative features as it tackles the within and between-class variations separately [5].

This study will help us understand both popular dimensionality reduction, i.e., PCA and LDA, and their pros and cons. This article will also discuss the results of their application on several machine learning classifications.

A. Principal Component Analysis

In PCA, an n-dimensional feature space will transform into an m-dimensional feature space, where the dimensions are orthogonal. PCA works on a process called Eigenvalue Decomposition of a covariance matrix (or it is also called the total scatter matrix) of a data set.

First, the data set of q n-dimensional training samples $x = [x_1, x_2, ..., x_q]$ are standardized by a data matrix by $\tilde{x}_i = x_i - \mu$ so that the dataset becomes $X = [\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_q]$.

Then, we calculate the covariance matrix representing the covariance of each pair of features in the data, which we call S^t, and

$$S^t = \frac{1}{q} \sum_{i=1}^{q} X^T X \tag{1}$$

Next, compute the eigenvalues and eigenvectors of S^T. The eigenvector of S^T represents directions in the data space, and eigenvalues represent the magnitude in those directions. The variance of the data projected onto the eigenvector is the eigenvalue of a covariance matrix.

We construct Φ, where $\Phi = [\phi_1, \phi_2, ..., \phi_m]$, where Φ consists of m number of largest eigenvalues of the total scatter matrix S^T.

Finally, the PCA transform the n-dimensional X into m-dimensional Y by:

$$Y = \Phi^T X \tag{2}$$

B. Linear Discriminant Analysis

The main idea of LDA is to find the linear combinations of the original variables (the axes or "discriminates") that best separate the classes. This is done by maximizing the distance between the means of the types while minimizing the spread (variance) within each category.

First, we define the data set of q n-dimensional training samples of c classes $x = [x_1, x_2, ..., x_q]$; and the number of training samples of type ω_j is q_j, where $j = 1, 2, ..., c$.

Then, the covariance matrix of class ω_j is computed as:

$$\sum_j = \frac{1}{q_j} \sum_{X_i \in \omega_j} (x_i - \mu_j)(x_i - \mu_j)^T \tag{3}$$

where $\mu_j = \frac{1}{q_j} \sum_{X_i \in \omega_j} x_i$, the mean vector of class i. The within-class scatter matrix (S_{W_i}) is defined as:

$$S_W = \sum_{j=i}^{c} \frac{q_j}{q} \sum_j \tag{4}$$

While the within-class scatter matrix S_W is the sum of all the scatter matrices for each class so that $S_W = \sum_i S_{W_i}$. Then, the between-class scatter matrix of c classes is defined as:

$$S_B = \sum_{j=1}^{c} \frac{q_j}{q} (\mu_j - \mu)(\mu_j - \mu)^T \tag{5}$$

where μ_i is the mean vector of class i,a and μ is the overall mean of the data. The between-class scatter matrix S_B is the sum of all these matrices, so $S_B = \sum_i S_{B_i}$.

Then, we solve the generalized eigenvalue problem for the matrix $S_W^- 1 S_B$ to get λ as its eigenvalue and ϕ as its eigenvector. After that, we sort the eigenvectors by their corresponding eigenvalues in descending order by $\sum_{k=1}^{m} \lambda_k^{\frac{b}{w}}$

II. Experiment

The idea of this study is to perform different classification algorithms for a binary classification task, particularly SVM and Mahalanobis Distance Classifier. Then, PCA and LDA are performed to reduce the dimensionality of the dataset.

A. Programming Language

In this experiment, we use Python, which has become one of the most popular programming languages for machine learning and data science. Besides that, we use scikit-learn library to provide pre-implemented algorithms and tools, which significantly simplifies the process of developing this experiment. Besides using libraries like NumPy and pandas, Python supports multidimensional arrays, data frames, and various mathematical functions to perform numerical computations. All of the code for this experiment can be accessed in [8].

B. Dataset

The dataset contains 131 classes of different fruits and vegetables and 90483 total images from Kaggle [6][7]. Each size of the image is 100x100 pixels. We divide the dataset into 74.81% for the training data set and the rest of 25.19% for the testing data set. The proportion of the dataset can be seen in Figure 1. All parameters for dimensionality reduction and classification will be specified only by the training data. The testing data can only be used to generate the classification accuracy or error rate.

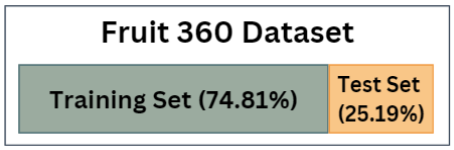

Fig. 1. The partition of the dataset

C. Class Choice

For binary classification in this study, we have decided to take Ripe Avocado and Eggplant classes because they almost look similar (they have nearly the same color), so the classification task will not be too easy. The chosen types can be seen in Figure 2 and Figure 3.

Fig. 2. EggPlant Dataset

D. Pre-processing of Image Data Set

In this study, StandarScaler is used for standardization in the pre-proecssing step. The purpose is to transform all features to the same scale by shifting the distribution of each attribute to have a mean of zero and a standard deviation of one (unit variance). This standardization of features can lead to improved model performance, faster convergence during training, and better interpretability of feature importance.

StandardScaler converts each imageinton a 100x100 numpy array for each RGB dimension. Then, the images have been flattened into one vector (Image Feature Vector. StandardScaler then subtracts the mean and divides by the standard deviation for each feature value.

Mathematically, the standardized value of feature index j for sample index i across n number of the samples is $z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$, where σ_j is the standard deviation of the

Fig. 3. Ripe Avocado Dataset

feature values, $\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \mu_j)^2}$; and μ_j is the mean of the feature values across all samples in the dataset, $\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$.

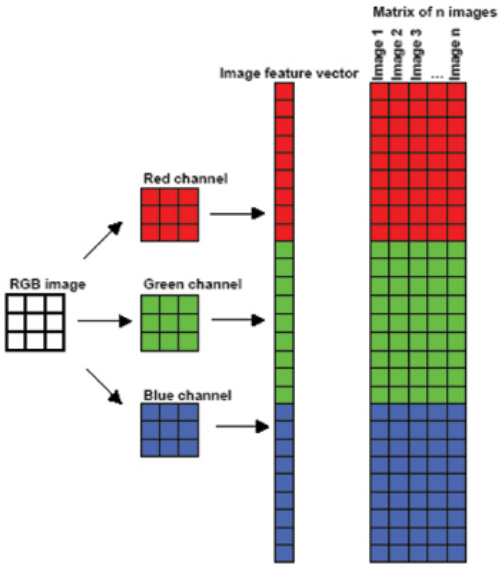

Fig. 4. Preprocessing of Image Data Set using StandardScaler Library

E. Principal Component Analysis (PCA)

This section will discuss how the dataset of Ripe Avocado compared with Eggplant appears in lower dimension. We use PCA from scikit-learn library to implement PCA in this experiment. PCA can reduce the training data dimension from (959, 30000) to (959, 2) or (959, 3), and reduce the test data dimension from (322, 30000) to (322, 2) or (322, 2). The first principal component of both datasets in two dimensions is shown in Figure 5, while in three dimension is shown in Figure 6.

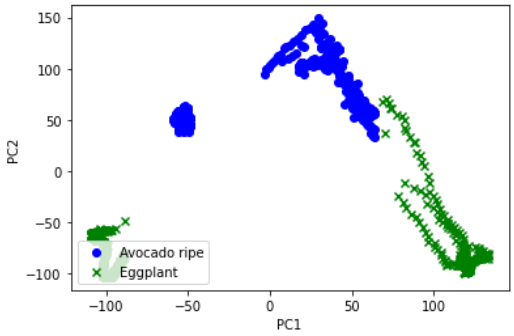

Fig. 5. First PC in 2D

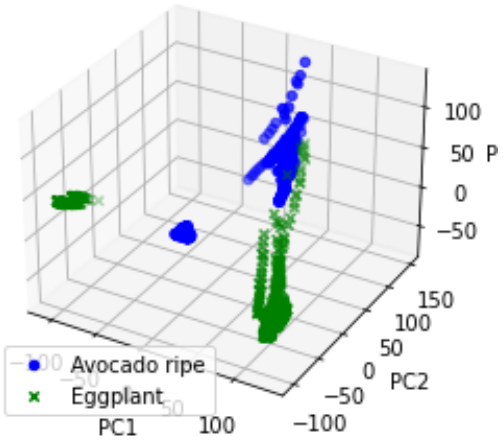

Fig. 6. First PC in 3D

F. Linear Discriminant Analysis (LDA)

Although LDA is able to be used as a classifier, in this study, we use LDA for dimensionality reduction before machine learning. We use LinearDiscriminantAnalysis from scikit-learn to implement LDA in this experiment. LDA can reduce the training data dimension from (959, 30000) to (959, 1) and reduce the test data dimension from (322, 30000) to (322, 1). The visualization of two dimension results of LDA as dimensionality reduction is shown in Figure 7 and the visualization of two dimension results of LDA as dimensionality reduction is shown in Figure 8.

G. Implementation of PCA and LDA on Kernel SVM Classifier

A Support Vector Machine (SVM) is a supervised classification method, that after a training phase can identify if a new point belongs to a class or another with the highest mathematical accuracy. In this study, we will use Kernel SVM as a classifier. After training the SVM classifier using the dataset, we get the accuracy of Kernel SVM without any dimensionality reduction is 51.55%. After implementing PCA, we get accuracy of Kernel SVM of 43.17%. Next, the accuracy of Kernel SVM with LDA for dimensionality reduction is increased, 88.82%. Kernel SVM boundaries after using PCA is

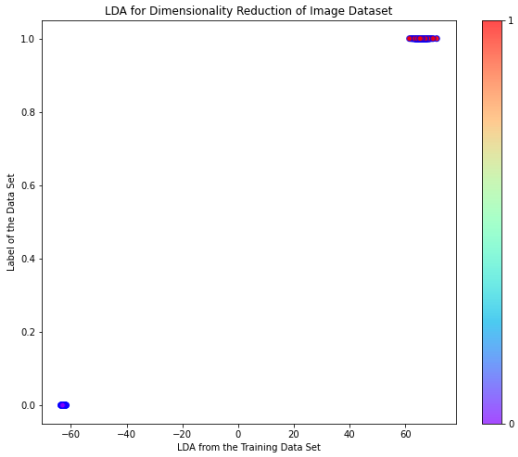

Fig. 7. Plot of LDA as dimensionality reduction result in two dimension

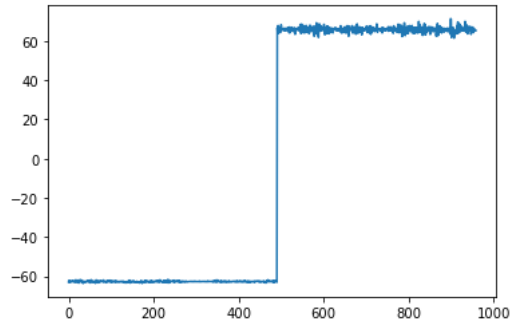

Fig. 8. Plot of LDA as dimensionality reduction result in one dimension

shown in Figure 9, while Kernel SVM boundaries after using PCA is shown in Figure 10. Recapitulation of this result can be shown in Table I.

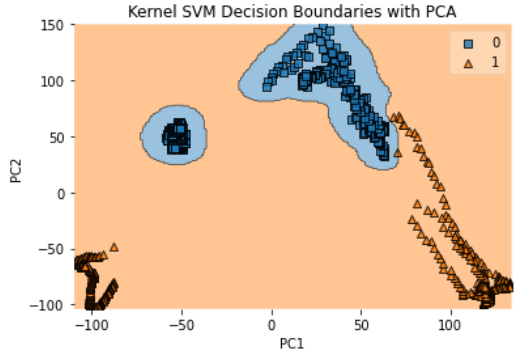

Fig. 9. Kernel SVM Decision Boundaries with PCA

H. Implementation of PCA and LDA on Mahalanobis Distance Classifier

The Mahalanobis Distance Classifier is a classification algorithm based on the concept of Mahalanobis distance. Mahalanobis distance is a measure of the distance between a point and a distribution that is defined as $D_M(x) = \sqrt{(x-\mu)^T \sum^{-1} (x-\mu)}$ where x is the point in question, μ is the mean vector of the distribution, $\sum$ is the covariance matrix of the distribution, and $\sum^{-1}$ is the inverse of the covariance matrix. In this study, Mahalanobis Distance Classifier is used with PCA.

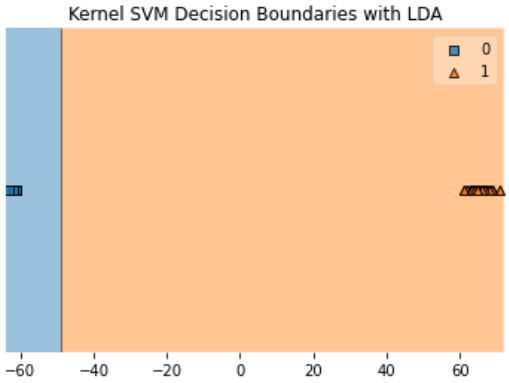

Fig. 10. Kernel SVM Decision Boundaries with LDA

TABLE I
ACCURACY OF DIFFERENT DR ON KERNEL SVM

Dimensionality Reduction	Kernel SVM
None	51.55%
PCA	43.17%
LDA	88.82%

However, the Mahalanobis Distance Classifier can not be used using original size of the available data set due to sparsity of data and volume of space. In high dimensions, data points tend to be far apart from each other, leading to sparsity. This sparsity makes it difficult to estimate the mean vector and covariance matrix accurately. The volume of the space increases exponentially with the number of dimensions, requiring exponentially more data to maintain the same level of statistical reliability.

Using PCA on Mahalanobis Distance Classifier, we get the accuracy result of 16.46% as shown in Table II. For visualizing results, we can use a scatter plot in the reduced feature space (after PCA) as shown in Figure 11. However, we are not successful in applying LDA on Mahalanobis Distance Classifier.

III. ANALYSIS

Tt is widely acknowledged that data dimensionality data is high; it can often lead to a deterioration in classification

TABLE II
ACCURACY OF DIFFERENT DR ON MAHALANOBIS DISTANCE CLASSIFIER

Dimensionality Reduction	Mahalanobis Distance Classifier
None	NaN
PCA	16.46%
LDA	NaN

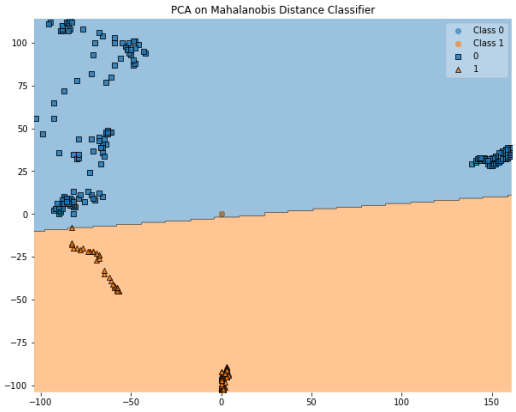

Fig. 11. Mahalanobis Distance Classifier Boundaries with PCA

performance in real-world applications, especially when the amount of training data remains constant. This phenomenon is referred to as the "curse of dimensionality."

To enhance classification and accuracy, machine learning methods for dimensionality reduction should aim to eliminate or adjust detrimental or unreliable dimensions for the classification task. This can be achieved by removing this harmful dimension or correcting its inconsistent statistical properties.

At first, we highly expect that PCA helps improve the classification accuracy, not because it minimizes the data reconstruction error, but because it has some role in removing the unreliable dimension. The dimensionality reduction by the supervised principal component analysis (including PCA) that plays the most vital role in boosting the classification accuracy while the discriminative method can greatly reduce the dimensionality with the minimum loss of the discriminative information [2].

The important role of PCA in the classification is to remove unreliable dimensions due to insufficient or unrepresentative training data. However, this work demonstrates that implementing PCA on the data total scatter matrix does not effectively remove the unreliable dimensions if one class is represented by its training data much better or worse than the class. In other words, PCA has deteriorated the classification performance.

From the Kernel SVM experiment in this study, we can see that LDA greatly enhances the classification accuracy.

While the primary goal of all components in a recognition system is to extract information that most effectively distinguishes between different categories, it is crucial to focus on obtaining this discriminative information from the entire data population rather than relying solely on a specific set of training data. We need discriminant analysis to maximize the discriminatory power of the extracted features [3][4]

Mahalanobis distance, and many classifiers are proven optimal only under Gaussian assumption [2]. Because of in this work, Mahalanobis Distance Classifier does not work under Gaussian data, the classifier result is quite low.

IV. Conclusion

We have performed sets of experiments that has used Kernel SVM and Mahalanobis Distance Classifier to analyze the role of PCA and LDA in the classification.

Mahalanobis Distance Classifier can not be used using original size of the available data set due to sparsity of data and volume of space. In this work, the PCA has been incorporated into Mahalanobis Distance Classifier to get unexpected result.

However, the performance of PCA and LDA can be seen in Kernel SVM part. PCA brings the most representative information in the most minor square error. Besides, LDA brings the most discriminative information in the linear constraint.

The PCA approach shows reduced performance. Otherwise, LDA outperforms all dimensionality reduction method in this study. The LCA achieves the best performance in all experiments.

As discriminative information is used for the research classification and based on the experimental results, we prefer LDA to PCA.

Mahalanobis Distance Classifier can not be used using original size of the available data set due to sparsity of data and volume of space. In this work, the PCA has been incorporated into Mahalanobis Distance Classifier to get unexpected result.

References

[1] R. O. Duda, P.E. Hart, and D.G Stork, Pattern Classification. New York Wiley, 2001.
[2] X. D. Jiang, Linear Subspace Learning-Based Dimensionality Reduction. IEEE Signal Processing Magazine, March 2011.
[3] X. D. Jiang, Asymmetric principal component and discriminant analyses for pattern classification. IEEE Transaction of Pattern Classification, vol 31, no. 5, pp 931-937, May 2009.
[4] D. L. Swets and J. Weng. Using Discriminant Eigenfeatures for Image Retrieval. IEEE Transaction of Pattern Analysis and Machine Intelligence. Vol 18, no. 8, pp. 831-836, Aug 1996.
[5] X. D. Jiang, B. Mandal, and A. Kot. Eigenfeature Regularization and Extraction in Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, No. 3, March 2008.
[6] M. Oltain, "Fruit 360", Available: Kaggle. [Online]. Available: https://www.kaggle.com/datasets/moltean/fruits. [Accessed: October 23, 2023].
[7] H. Muresan, M. Oltean. Fruit recognition from images using deep learning. Acta Univ. Sapientiae Informatica Vol 10, Issue 1, pp. 26-42, 2018.
[8] N. Anbarsanti. "Dimensionality Reduction Implementation of Binary Classification Between Ripe Avocado and Eggplant", Available: Github. [Online]. Available: https://github.com/anbarsanti/PCA_LDA_SVM_Mahalanobis/tree/main. [Accessed: October 27, 2023]