PalArch's Journal of Archaeology of Egypt / Egyptology

Performance Comparison of PCA and LDA with Linear Regression and Random Forest for IRIS Flower Classification

[1]Debaraj Rana, [2]Swarna Prabha Jena, [3]Subrat Kumar Pradhan

[1,2,3] Department of Electronics and Communication Engineering, Centurion University of Technology and Management, Odisha, India

Email: [1]debaraj.rana@cutm.ac.in, [2]swarnaprabha@cutm.ac.in, [3]subrat.pradhan@cutm.ac.in

Debaraj Rana, Swarna Prabha Jena, Subrat Kumar Pradhan: Performance Comparison of PCA and LDA with Linear Regression and Random Forest for IRIS Flower Classification -- Palarch's Journal Of Archaeology Of Egypt/Egyptology 17(9). ISSN 1567-214x

Keywords: s- Machine Learning, Principal component analysis, linear discrimination analysis, logistic regression, Random forest

ABSTRACT

Classification is the mostly used machine learning problem now a day for varieties of application use in the field of security, agriculture, industry etc. This paper performs the classification of IRIS flower using Logistic regression and random forest algorithm. Principal component analysis (PCA) and linear discriminant analysis (LDA) used as feature extraction method for both case. In the second part a comparative study has been propose between the performance of both the machine learning method as well as both the dimensionality reduction method. The comparative study reveals that the LDA work far better than PCA, where as using LDA the logistic regression and random forest method gives nearly same result.

1. Introduction

Machine learning becomes a most interesting research topic [1]. Now a day's many people working in the field of Machine learning, even some researchers updating or finding new methods or algorithm for the machine learning. So gradually it becomes a vast area to do the research. Basically the machine learning is the basic process of making the machine to take decision just like human brain. This leads to the field of machine intelligence, which is some kind of incorporating intelligence inside the machine by making the machine to learn using some algorithm. The learning phase helps the machine to take decision from the learning experience.

The learning phase that's why classified as: Supervised learning, unsupervised learning and Reinforcement Learning [2, 3]. The supervised learning involved a presence of target output which help the system to learn, but in case of

unsupervised learning, no such target output there, but the system has to learn itself without any guidance. The third category learning is a method where an external agent gets involved and performs action by learning from the errors. Generally machine learning implementation consists of two phases, training and testing. Using some data sets called as training data, the system able to learn. The learning experience helps the system to take decision and predict the output during the testing phase where some testing data given to the system. Every machine perform both the training and testing with the help of some machine learning algorithm like logistic regression, support vector machine, Decision tree, Random Forest .

In this paper Iris flower classification has been done with the help of both logistic regression and random forest method using the data reduction and feature extraction technique PCA and LDA. Based on the classification accuracy a comparative study has been performed.

2. Principal Component Analysis

It is a statistical method which performs a transformation to produce an uncorrelated data set called as principal components from a large related data set [4]. This method helps to reduce the complexity of a problem by reducing the dimensionality of the large data set. The uncorrelated data set called the features set make the system work faster and accurate. The principal components are nothing but the Eigen vectors related to highest Eigen values of a covariance matrix derived from the standardized data set [5].

The steps of the required to extract the principal component as follows-

- Standardize the data set by removing the common features of the data set
- Determine the covariance matrix of the standardize dataset
- Calculate the Eigen vectors of the covariance matrix
- Create a new space using the Eigen vectors and project all the data on the feature space

3. Linear Discriminant Analysis

The Linear Discriminant Analysis is also used for dimensionality reduction technique [6]. LDA is used determine the linear combination of features sets which characterized two or more classes. It is a supervised dimensionality reduction technique to be used with continuous independent variable and a categorical dependent variable. The LDA aims to project the larger data set to a low dimensional space through discriminants features. Ronald A Fisher had formulated the linear discriminant method and he had demonstrated its use as a classifier [7].

Steps to solve LDA:

- Computing the d-dimensional mean vectors mi for i=1,2,3 (Three Class of IRIS Flower

$$m_i = \frac{1}{n_i} x_k \qquad (1)$$

- Compute the scatter matrices (in-between-class and within-class scatter matrix)
  - Within-class scatter matrix SW

$$S_W = \sum_{t=1}^{C} S_i \quad (2)$$

Where, C=number of class and $S_i = \sum_{x \in D_i}^{n} (x - m_i)(x - m_i)^T$

  - Between-class scatter matrix SB

$$S_B = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T \quad (3)$$

Where, m is the overall mean, and mi and Ni are the sample mean and sizes of the respective classes.

- Compute the eigenvectors e1, e2 …and corresponding eigenvalues λ1, λ2, for the scatter matrices by solving

$$S_W^{-1} S_B w = \lambda w \quad (4)$$

- Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a low dimensional matrix W
- Transforming the samples onto the new subspace W by projecting with the Y=X.W where X is the n sample and Y is k sample data set

4. Regression and Random Forest

The logistic regression [8] is a technique which predicts the output from a set of features set generally used for binary classification, but it can be extended to multi class logistic regression by subdividing the problem in to n binary classification and again for each binary classification it determine whether true of false. In this ways the logistic regression can be applied to multi class classification problem. The regression uses a sigmoid signal defined as y = 1 / (1 + e^-x) as a logistic function to map the classification output [9].

Random Forests is a supervised machine learning algorithm used for both Regression and Classification [10]. This algorithm creates forest of trees by randomly selecting decision trees. Forest which means collection of trees (here Decision trees) and these trees are trained on subsets (equal to the size of training set) selected at random. The decision of majority of trees is chosen as the final decision. It is the most powerful Supervised Machine learning algorithm. In general the more trees in the forest the more robust the prediction and thus higher accuracy.

Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree. [11, 12]. Bagging method is used to build the forest which is the collection of Decision trees. Multiple trees are built, to classify a new object based on attributes, each tree gives a classification and the class which has higher votes

is chosen for the final classification, in case of regression average of all outputs of different is considered.

5. Result Analysis

The classification of IRIS flower and the comparative study have been performed using python 3.6 and the Scikit-learn library [13]. For the simulation purpose data set of IRIS flower has been collected from Machine Learning IRIS Data set [14]. The data set contain three types of iris flower namely Setosa, Verginica and Versicolor. The data set contain 50 set of features (Sepal Length, Sepal Width, Petal length, Petal Width) from each class of flower. A sample of set has shown in table 1and figure 1.

Figure 1. IRIS flower types

Table 1. IRIS flower feature set

	sepal length	sepal width	petal length	petal width	Class Label
0	0.613030136	0.108501054	0.947517828	0.736039671	Iris-virginica
1	-0.567766273	-0.124001205	0.384914467	0.348083182	Iris-versicolor
3	-0.213527351	1.736016872	-1.190374946	-1.20374277	Iris-setosa
...	...	...	...	...	...

Before performing the accuracy analysis the PCA and LDA has been applied to select the principal components and discriminant components respectively from the feature data set. The significant components and their distribution as a scatter plot which is the results of PCA and LDA have shown in figure 2 and 3.

Two Principal Components		
S.N.	1st Principal Component	2nd Principal Component
0	2.264542	0.505704
1	2.086426	0.655405
2	2.36795	0.318477
3	2.304197	0.575368
4	2.388777	0.674767
...	...	...

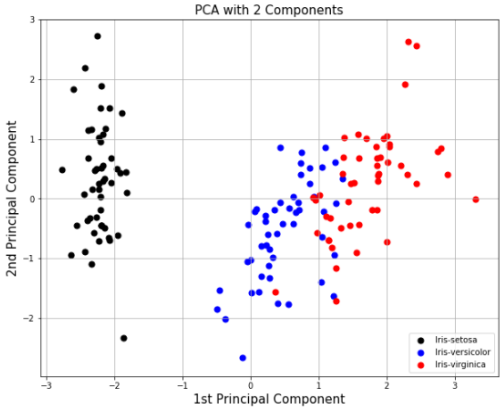

Figure 2. Principal Component Analysis

Two Linear Discriminant Components		
S.N.	1st LD Component	2nd LD Component
0	8.084953	0.328454
1	7.147163	0.755473
2	7.511378	0.238078
.....	.....	.....

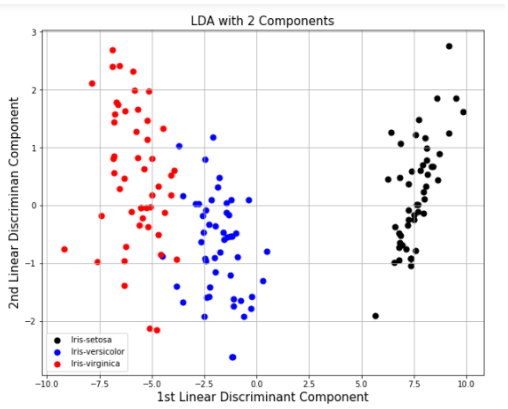

Figure 3. Linear Discriminant Analysis

For the performance study of the technique implemented the data set has been randomly divided into training set and testing set. As discussed earlier the paper mainly focus on a comparative effect of PCA and LDA on IRIS flower classification using Logistic regression and random forest classification method.

The system gone through a training process using the training data set by implementing the machine learning technique. After that the testing data have been given to the trained model and Classification result using both Logistic regression and Random forest has been achieved. The PCA with logistic regression gave an accuracy of 86% with 80% of training data and 20% testing data as shown in table 2. In contrast to that the LDA with random forest method gave 100% accuracy rate. The result shown in figure uses PCA for logistic regression and LDA for Random forest for a case study and the result shown below figure 5.

Table 2. Classification Result

Predicted Class Label			
Iris-virginica	Iris-versicolor	Iris-setosa	Iris-virginica
Iris-setosa	Iris-virginica	Iris-virginica	Iris-versicolor
Iris-versicolor	Iris-virginica	Iris-versicolor	Iris-setosa
Iris-setosa	Iris-setosa	Iris-versicolor	Iris-versicolor
Iris-virginica	Iris-setosa	Iris-setosa	Iris-versicolor
Iris-setosa	Iris-virginica	Iris-virginica	Iris-versicolor
Iris-versicolor	Iris-versicolor	Iris-setosa	Iris-setosa
Iris-versicolor	Iris-setosa		

Target Class Label			
Iris-virginica	Iris-versicolor	Iris-setosa	Iris-virginica
Iris-setosa	Iris-versicolor	Iris-versicolor	Iris-versicolor
Iris-versicolor	Iris-versicolor	Iris-versicolor	Iris-setosa
Iris-setosa	Iris-setosa	Iris-virginica	Iris-versicolor
Iris-virginica	Iris-setosa	Iris-setosa	Iris-versicolor
Iris-setosa	Iris-virginica	Iris-virginica	Iris-versicolor
Iris-versicolor	Iris-versicolor	Iris-setosa	Iris-setosa
Iris-versicolor	Iris-setosa		

PCA: In the analysis of only PCA in the classification it can be observed that the accuracy level increases with the increase of training data size at a fixed set of principal components, even with the same training data size it can be seen that the accuracy level is more when the principal component taken will be more as per performance graph shown in figure 4.

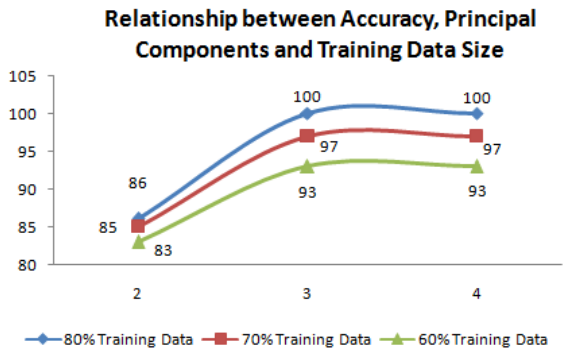

Figure 4. Performance of PCA

LDA: In case of PCA when the number of principal components increases then it increases the accuracy rate from 86% to 100% (shown in figure 5). This characteristics also seen in case of PCA but the accuracy reach to 100% with 3 principal components but same has been achieved with 2 discriminant components in the case of LDA. It means LDA give more accuracy as compared to PCA at the earliest. That percentage of training and testing data set was taken as 80:20 ratios.

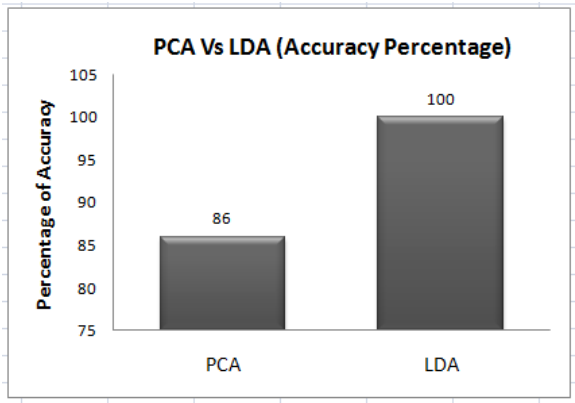

Figure 5 Comparisons of PCA and LDA

LDA LR-PCA-LR: This graph shown in figure 6 reveals that the accuracy level get increase with the increase of percentage of training data set. The more data used for training make the system more robust to perform which increase the percentage of accuracy. When the comparison done between PCA and LDA for classification using Logistic regression method, the first one achieved 81% accuracy with 50% training data, and increase to 86% of accuracy at most when increase the training set to 80%, where as in case of LDA the accuracy level achieved was 96% with 50% of training data and increased to 100% with 80% training data with two significant components taken in both cases. LDA achieved the highest accuracy faster than LDA even with less training set data as compared to PCA.

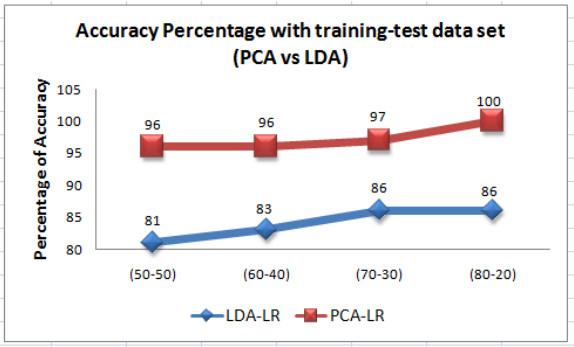

Figure 6. Performance of Linear Regression

LDA-LR-LDA-RF: The performance graph of LDA (shown in figure 7) for classification using linear regression method express that , the accuracy achieved was 96% with 50% of training data size which increases to 100% with 80% data size. The implementation of random forest also impact quiet similar effect which gives result with 94% to 100% with increase of 50% training data size to 80% training data size. But when it compared to each other at each level then linear regression perform well with less training data where as the random forest gives better result with large training data size. The significant components taken were fixed in all the cases. So comparatively it can be seen both are performing quiet same in the context of IRIS flower classification.

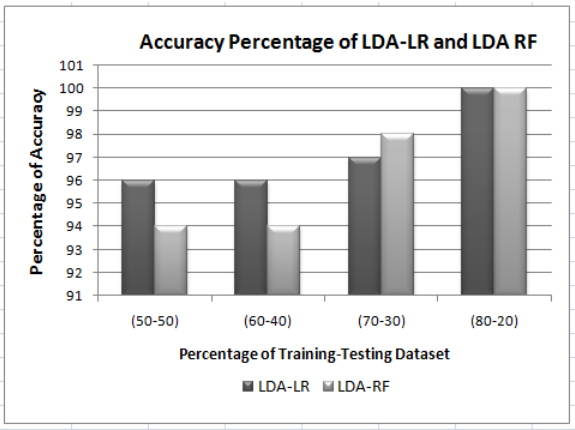

Figure 7. Performance of LDA with LR and FR

6. Conclusion

The main aim of the study was to derive a conclusion regarding the performance of PCA and LDA feature extraction method in IRIS flower classification. The study reveals that both methods giving good result in the classification, even the accuracy depends on the number of principal component to be selected. But with a fixed set of principal components the performance of LDA is better than PCA. The study also show that the increase in percentage of training data also increase the accuracy level. For the classification Logistic regression and random forest has been used. The

Random Both the method performance are quiet similar with PCA or LDA. The LDA performing much better by giving 100% accuracy as compare to 86 % result produced using PCA.

References

Alpaydin, Ethem. Introduction to machine learning. MIT press, 2020.

Zhao, Z. and Liu, H., 2007, June. Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th international conference on Machine learning (pp. 1151-1157).

Love, B.C., 2002. Comparing supervised and unsupervised category learning. Psychonomic bulletin & review, 9(4), pp.829-835.

Shlens, J., 2014. A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100

Turk, M. and Pentland, A., 1991, January. Face recognition using eigenfaces. In Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition (pp. 586-587)

Raschka, Sebastian. "Linear discriminant analysis bit by bit." Disponible en:

Linear Discriminant Analysis (LDA) is a classification method originally developed in 1936 by R. A. Fisher

Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M. and Klein, M., 2002. Logistic regression. New York: Springer-Verlag.

https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html#introduction

Machine Learning: The Ultimate Beginners Guide for Neural Networks, Algorithms, Random Forests and Decision Trees Made Simple by Ryan Roberts, CreateSpace Independent Publishing Platform, 2017

Smith, Chris. Decision trees and random forests: a visual introduction for beginners. Blue Windmill Media, 2017.

https://www.edureka.co/blog/random-forest- classifier/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, pp.2825-2830.

http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data