Received March 25, 2022; | Revised xx. xx., 2022; | Accepted xx. xx., 2022

DOI: xxx/xxxx

CLASS: UST PRML 2022-1

REPORT FOR THE STUDY

Classifying Bacteria Species with Dimension Reduction

Wonjun Lee[1] | **Junseong Bang**[2,3]

[1]Intelligence & Manufacturing Convergence Laboratory, ETRI, Daejun, Republic of Korea

[2]Police Science & Public Safety ICT Research Center, ETRI, Daejun, Republic of Korea

[3]Department of Computer Software, UST, Daejun, Republic of Korea

Abstract

Bacteria are microscopic, single-celled organisms. But There are only a few types of bacteria that cause the disease and There are only a few types of bacteria that cause the disease. The goal is to predict bacteria species based on repeated lossy measurements of DNA snippets. Using this dataset, after apply dimension reduction method(PCA, LDA), by Random Forest Classifier. It compares and analyzes to apply Naive, PCA, and LDA.

KEYWORDS:
Bacterias, PDA, LDA

1 | INTRODUCTION

Bacteria are microscopic, single-celled organisms. There are thousands of different kinds of bacteria. Many bacteria are present in various body parts, including human and animal skin, airway, mouth, digestive tract, reproductive organs, and urethra, without causing harm. Such bacteria are called bacteriophages or microbial communities.

There are at least as many bacteria in the bacteriophage as cells in the body. A lot of bivalve fungi actually help people.

There are only a few types of bacteria that cause the disease. These bacteria are called pathogens. Sometimes bacteriophage fungi cause disease under certain conditions.

Bacteria can produce harmful substances (toxins) or invade tissues and cause diseases. Some bacteria can trigger inflammation that can affect the heart, nervous system, kidneys, or gastrointestinal tract.

Some bacteria increase the risk of cancer.There are only a few types of bacteria that cause the disease. These include anthrax, botulinum poisoning, plague, and bacteria that cause Yato disease.

It is important to classify bacteria that cause diseases in advance and bacteria that can be used as weapons.

For Tabular Playground Series - Feb 2022 , the problem deals with classifying 10 different bacteria species using data from a genomic analysis technique that has some data compression and data loss. In this technique, 10-mer snippets of DNA are sampled and analyzed to give the histogram of base count.

The goal is to predict bacteria species based on repeated lossy measurements of DNA snippets.

2 | DATASET

This is Tabular Playground Series - Feb 2022 dataset, obtained from Kaggle.

the problem deals with classifying 10 different bacteria species using data from a genomic analysis technique that has some data compression and data loss.

Each row of data contains a spectrum of histograms generated by repeated measurements of a sample, each row containing the output of all 286 histogram possibilities , which then has a bias spectrum (of totally random ATGC) subtracted from the results.

The data (both train and test) also contains simulated measurement errors (of varying rates) for many of the samples, which makes the problem more challenging.

3 | LINEAR DISCRIMINANT ANALYSIS (LDA)

Linear Discriminant Analysis is a dimensionality-reduction technique in the pre-processing step for pattern-classification and machine learning applications. The goal is to project

a dataset onto a lower-dimensional space with good class-separability in order avoid overfitting and also reduce computational costs.

4 | PRINCIPAL COMPONENT ANALYSIS VS. LINEAR DISCRIMINANT ANALYSIS

Both Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are linear transformation techniques that are commonly used for dimensionality reduction. PCA can be described as an "unsupervised" algorithm, since it "ignores" class labels and its goal is to find the directions (the so-called principal components) that maximize the variance in a dataset. In contrast to PCA, LDA is "supervised" and computes the directions ("linear discriminants") that will represent the axes that that maximize the separation between multiple classes.

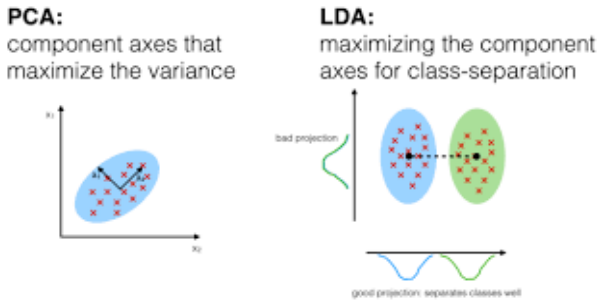

FIGURE 1 Comparison of PCA and LDA

5 | CODE

Import Library

```
#General Libraries
import numpy as np
import pandas as pd

# Machine Learning Libraries
from sklearn.preprocessing import scale
from sklearn.model_selection import
    train_test_split
from sklearn.metrics import confusion_matrix,
    accuracy_score, confusion_matrix,
    plot_confusion_matrix
from sklearn.metrics import roc_auc_score,
    roc_curve
import sklearn.metrics as metrics
from sklearn.preprocessing import StandardScaler

# Data Visualiztion Libraries
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

from warnings import filterwarnings
```

Load Dataset, Split Data

```
test = pd.read_csv("./test.csv", index_col="
    row_id")
train = pd.read_csv("./train.csv", index_col="
    row_id")

TARGET = 'target'
FEATURES = [col for col in train.columns if col
    not in [TARGET]]

print(f'Number of observations in TRAIN:{len(
    train)}')
print(f'Number of observations in TEST:{len(test
    )}')
train.target.value_counts()

print(train[FEATURES].duplicated().sum())
print(test[FEATURES].duplicated().sum()) #

train.drop_duplicates(keep='first', inplace=True
    ) #

X = train.drop("target", axis=1).astype(np.
    float32)

from sklearn.preprocessing import LabelEncoder


target_encoder = LabelEncoder()
y = pd.Series(target_encoder.fit_transform(train
    ["target"]))


X_train, X_test, y_train, y_test =
    train_test_split(train[FEATURES].values, y,
    random_state=42)
```

Standard Scale

```
from sklearn.preprocessing import StandardScaler

std_scale = StandardScaler()
std_scale.fit(X_train)
X_train_std = std_scale.transform(X_train)
X_test_std  = std_scale.transform(X_test)
```

LDA

```
from sklearn.discriminant_analysis import
    LinearDiscriminantAnalysis
lda = LinearDiscriminantAnalysis()
lda.fit(X_train_std, y_train)
X_train_lda = lda.transform(X_train_std)
X_test_lda = lda.transform(X_test_std)
```

LDA Visualization

```
from sklearn.discriminant_analysis import
    LinearDiscriminantAnalysis as LDA

train_sub = train.sample(10000, random_state=
    42)
lda_data = LDA(n_components=2).fit_transform(
    train_sub.drop(columns='target'),train_sub.
    target)
plt.figure(figsize=(10,10))
sns.scatterplot(x = lda_data[:, 0], y = lda_data
    [:, 1], hue = 'target', data=train_sub)
```

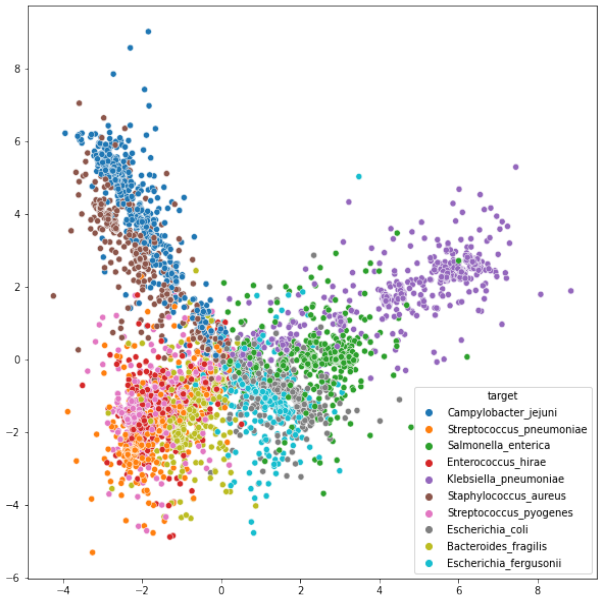

FIGURE 2 LDA Visualization

Random Forest Classifier

```
from sklearn.ensemble import
    RandomForestClassifier
# learning..
clf_rf_lda = RandomForestClassifier(max_depth=2,
     random_state=42)
clf_rf_lda.fit(X_train_lda, y_train)

# prediction
pred_rf_lda = clf_rf_lda.predict(X_test_lda)

from sklearn.metrics import accuracy_score
accuracy_lda = accuracy_score(y_test,
    pred_rf_lda)
print('# Prediction with LDA: ')
print(accuracy_lda)
```

6 | RESULT

This is the result of applying Random Forest Classifier to data applied with Naive which is version that have not been processed, PCA, and LDA.

	Naive	PCA	LDA
Testing Score	0.74	0.48	0.82

TABLE 1 Result of Naive, PCA, LDA

7 | CONCLUSION

In order to analyze classifying 10 different bacteria specie, the dimension was reduced to PCA and LDA, and then a model was created by applying a Random Forest Classifier .

this dataset, Random Forest Classifier showed the highest performance after applying LDA, and PCA showed the lowest performance.

Future research plans aim to increase accuracy and performance by using more diverse dimensional reduction techniques.