PUNE INSTITUTE OF COMPUTER TECHNOLOGY,
DHANKAWADI PUNE-43.

A Project Report
On

Comparison of Classification algorithms on gender voice dataset with PCA

SUBMITTED BY

NAME: Akash Kale, Anuj Kanetkar, Shriniwas Nayak

ROLL NO: 4241, 4242, 4256

CLASS: BE-2

GUIDED BY
PROF. H.P. Channe

COMPUTER ENGINEERING DEPARTMENT
Academic Year: 2019-20

PUNE INSTITUTE OF COMPUTER TECHNOLOGY,
DHANKAWADI PUNE-43.

CERTIFICATE

This is to certify that Mr. ***Akash Kale, Anuj Kanetkar, Shriniwas Nayak*** *,* *R*oll No**. 4241, 4242, 4256** students of B.E. (Computer Engineering Department) Batch 2019-2020, have satisfactorily completed a mini-project report on "**Comparison of Classification algorithms on gender voice dataset with PCA"** under the guidance of **Prof H.P. Channe** towards the partial fulfillment of the fourth year Computer Engineering Semester VIII of Pune University.

Prof. H.P. Channe — Prof. M. S. Taklikar

Guide — **Head of Department,**

Computer Engineering

Date: **Place:**

Contents

Acknowledgement

We Akash Kale (4241), Anuj Kanetkar (4242), Shriniwas Nayak (4256) would like to extend our sincere gratitude towards our earnest gratitude towards our guide Prof. H.P. Channe.

It is because of her guidance and support that we have been able to complete the project in time and as per the requirements.

We take this opportunity to thank our college Pune Institute of Computer Technology (PICT) for providing all the required infrastructure and facilities required for the completion of the project.

We would also like to thank our family and friends for their love and support without which the project could not have been completed.

Problem Statement

Computers need to decide the gender of a person from his/her voice, the dataset for the same is available but contains many features and handling the same is not an easy task. There is a need for an efficient method to handle all the features without negatively affecting the ability of classifying the test dataset, also due the large number of features the response time of the system is slow which should be improved.

Abstract

Machine Learning is used widely today to provide solutions to many problems, increase profits, improve processes and at times for research and development. Machine Learning algorithms are particularly useful in this example as the presented data can be used for predicting the gender of the voice using supervised learning techniques.

Due to the high number of attributes at times it becomes extremely tedious to handle the data and also the attributes may be correlated which may violate some assumptions made by classification algorithms for example GNB, which assumes conditional independence amongst the attributes.

The project uses three classification algorithms and compares accuracy obtained from each of these and time required before and after the use Principal Component Analysis. Libraries used for the same include numpy, scikit learn, matplotlib amongst others.

The project is built in modular fashion to help in dynamic and plug and play programming, the project is capable of analysing any other similar file.

H/W & S/W Requirements

The project has certain hardware and software requirements as follows:

Hardware : 64 bit processor, 512 MB RAM, 4 GB HDD/SSD

Software : Fedora 20 OS (any other Open Source OS), python3 with scikit learn, numpy and pandas libraries, text editor (gedit).

Introduction

In many applications including robotics, chat bots, virtual assistants, e-commerce websites amongst others it is a required for the machine to understand the gender of the voice to make better suggestions to the user, for example, consider an e-commerce website will try to recommend apparels for women/men depending on the gender.

Even Though, the dataset for the same is available it has many attributes which are correlated, to reduce the dimensionality of the feature space and to make the attributes conditionally independent Principal Component Analysis is used.

Objective

The mini project aims to attain the following objectives:

1) Analyse provided data present data in readable format for any user

2) Reduce the dimensionality of the feature space.

3) Generate conditionally independent attributes.

4) Classify the data as per the labels and predict the outcome for similar data.

5) Compare the performance of different classification algorithms with respect to accuracy and time required for training them.

Scope

The project deals with data consisting of physical parameters of voice generated by a set of candidates, but names and unique ids of the candidates have not been used in the project and no group member had access to them during any time of the project.

The output of the project consists of these:

- Analysis Reports
- Feature space file after PCA
- Explained Covariance Ratio Graph
- Accuracy Comparison Graph for classification algorithms
- Time Comparison Graph for classification algorithms

The project can be further expanded into a full fledge system by adding following functionality:

1) Automated report mailing to users
2) Creating database and combining data from different sources
3) Developing an adequate front end for the system

Project Development Outline

For the successful implementation of the project we undertook the following activities:

1. Brainstorming for project field selection
2. Requirement gathering for the selected topic
3. Requirement analysis of the collected requirements
4. Finalizing requirements and problem statement.
5. Brainstorming for data analysis techniques
6. Incorporation of changes
7. Implementation of individual modules
8. Testing of individual modules
9. Exception handling in individual modules
10. Integration of modules
11. Integration testing
12. Integrated Exception handling
13. Final testing
14. Project Presentation
15. Project Report Submission

Functions

The project uses following functions to achieve the required objectives:

- analysis(data_frame) : Generates analysis reports
- pca(data_frame, threshold) : Returns reduced feature vector space
- GNB(data_frame,target_frame) : Performs Classification using Gaussian Naive Bayes Algorithm
- KNN(data_frame,target_frame) : Performs Classification using K Nearest Neighbor Algorithm
- SVM(data_frame,target_frame) : Performs Classification using Support Vector Machine Algorithm

Usage of Principal Component Analysis

Principal Component analysis is used to reduce the dimensionality of the feature space while using the effects of all the attributes. Consider a dataset with m data instances and n features represented as matrix Z_{mXn}. It is multiplied with its transpose $Z^T_{nXm} . Z'_{nXn} = Z^T_{nXm} X Z_{mXn.}$ This matrix Z' is represented as $P_{nXn} X D_{nxn} X P_{nXn}^T$, where P is the matrix of eigenvectors and D is a diagonal matrix of eigenvalues.

The eigenvalues in D are sorted in descending order and accordingly P is changed to P^*_{nXn} and then Z is pre multiplied to this matrix to obtain the reduced feature space.

$$Z^*_{mXn} = Z_{mXn} X P^*_{nXn}$$

There are mainly three methods used for deciding the number of components to be retained in the feature space. They are :

- ➢ Deciding static number of attributes to be retained
- ➢ Deciding a threshold on explained variance percentage
- ➢ Deciding the number of attributes using cumulative explained variance graph

Test Cases

The following table encapsulates the test and their results for the "gender_voice_dataset.csv" file.

Algorithm	Accuracy (%)	Time (ms)	Result
GNB	90.53	7.41	Pass
KNN	74.49	36.39	Pass
SVM	92.93	3139.153	Pass
GNB after PCA	93.43	8.22	Pass
KNN after PCA	96.71	50.92	Pass
SVM after PCA	96.72	50.15	Pass

Table 1

Results

The results obtained indicate that in the majority of the cases performing PCA helps to increase accuracy, also the time required for training is reduced.

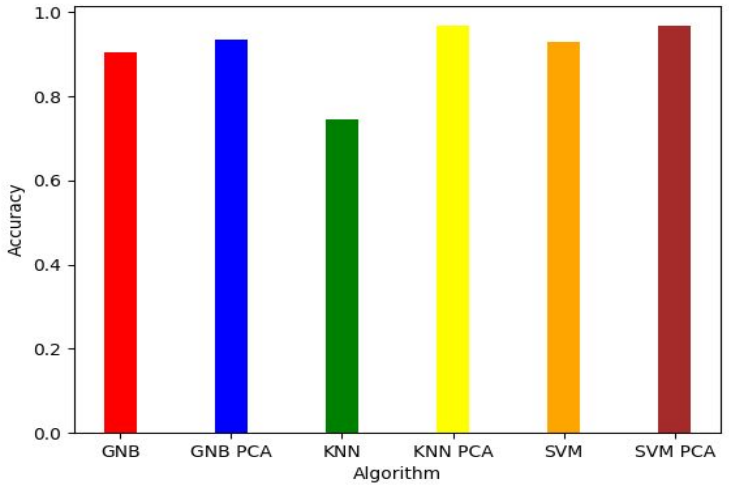

Image 1

The above image helps to understand that for every algorithm there is an increase in accuracy after implementing PCA.

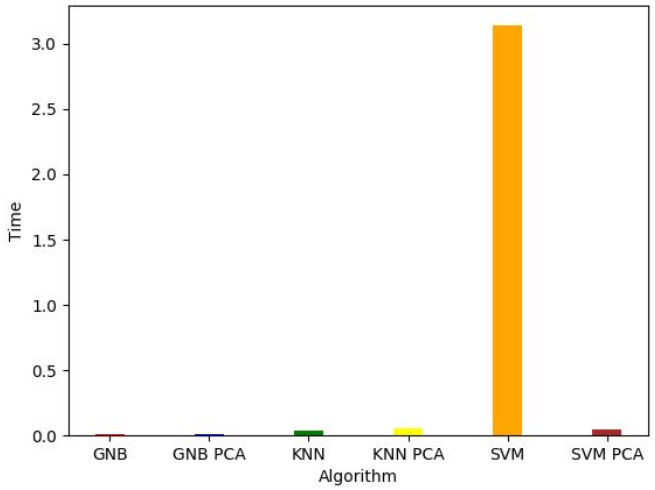

Image 2

This graph indicates that there is significant reduction for the time required to train a model using SVM after using PCA technique.

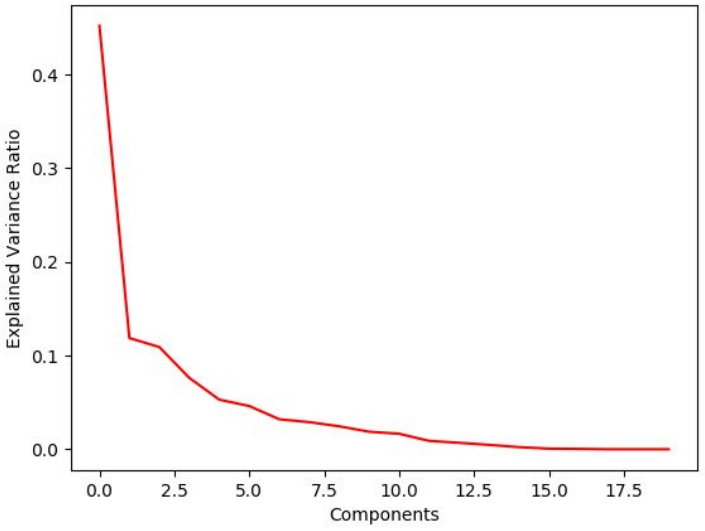

Image 3

This graph indicates that 6th component onwards the explained variance ratio decreases almost linearly.

Conclusion

We have successfully completed the project with the required functionalities. We have also studied the future prospects of the project. Created required visualizations for any user and generated analysis reports.

References

[1] https://scikit-learn.org/ (Date visited : 5/4/20 time : 15:30)

[2] https://matplotlib.org/ (Date visited : 5/4/20 time : 15:30)

[3] https://en.wikipedia.org/wiki/Principal_component_analysis (Date visited : 3/4/20 time : 17:30)

[4]https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c (Date visited : 3/4/20 time : 17:45)

[5] https://www.kaggle.com/primaryobjects/voicegender (Date visited : 1/4/20 time : 15:30)