CSE583 Project3: Feature Selection

Sree Sai Teja Lanka

March 12, 2019

Abstract

The goal of this project is to get you familiar with feature selection based on the forward selction methods involving filter and wrapper approches for the datasets Face and EEG. The objectives of this project involve: 1. Filter techniques assess the relevance of features by looking at the intrinsic propertiesof the data. In filter criteria, all the features are scored and ranked based on certainstatistical criteria. The features with the highest ranking values are selected and the lowscoring features are removed. Filter methods are fast and independent of the classifierbut ignore the feature dependencies and also ignores the interaction with the classifier. 2. Wrapper methods embed the model hypothesis search within within the feature subsetsearch. In this setup, a search procedure in the space of possible feature subsets is defined,and various subsets of features are generated and evaluated. The evaluation of a specificsubset of features is obtained by training and testing a specific classification model,rendering this approach tailored to a specific classification algorithm.

Contents

1 Introduction

We aim to select a subset of the features such that a criterion function is not aected much. Two important steps in feature selection are the search strategy and evaluation function. For the

lter I have chosen Augmented Variance Ratio as the selection criterion and performed ranking based on that. For the wrapper I have used greedy search strat- egy, speci

cally Sequential Forward Selection. The optimization criterion chosen is the classi

cation accuracy of the inbuilt Matlab linear discriminant classi

er. Both these ap- proaches are trained and tested on real world data. The performance has been evaluated using cross-validation.

2 Approach

2.1 Data

Two datasets have been used, they are as follows:
FACE:
This dataset is the result of quantified statistical facial asymmetry as a biometric under expression variations.
EEG: This dataset is the result of extraction of fourier coecient and temporal features extracted from EEG Signals. The total number of Fourier Coefficient Features are 48,384. There are 1,536 temporal features for each trial. The two sets of feature are concatenated together to create a total of 49,920 features for each observation.

2.2 Methods

Feature selection (also known as subset selection): Feature Selection is a process commonly used in machine learning, where a subset of features is selected from the available data for application of a learning algorithm [5]. So we prefer the model with the smallest possible number of parameters that adequately represent the data. Selecting the best feature subset is a NP complete problem. The task is challenging because first, the features which do not appear relevant singly may be highly relevant when taken with other features. Second, relevant features may be redundant so that omission of some of them will remove unnecessary complexity. An exhaustive search of all possible subsets of features will guarantee the best feature subset. The best subset contains the least number of features that most contribute towards accuracy. There are two approaches of feature selection:
Forward selection
1. Start with no variables.
2. Add the variables one by one, at each step adding the feature that has the minimum error.
3. Repeat the above step until any further addition does not signify any decrease in error.
Backward selection
1. Start with all variables.
2. Remove the variables one by one, at each step removing the feature that has the

highest error.
3. Repeat the above step until any further removal increases the error significantly Furthermore, The two broad categories of feature subset selection have been proposed:

- filter approach
- wrapper approach

2.2.1 Filter Approach

Filter techniques assess the relevance of features by looking at the intrinsic properties of the data. In filter criteria, all the features are scored and ranked based on certain statistical criteria. The features with the highest ranking values are selected and the low scoring features are removed. Filter methods are fast and independent of the classifier but ignore the feature dependencies and also ignores the interaction with the classifier. They also easily scale to very high-dimensional dataset. As a result feature selection need to be done only once and then different classifiers can be evaluated.
The common disadvantage of filter methods is that they ignore the interaction with the classifier and each feature is considered independently thus ignoring feature dependencies In addition, it is not clear how to determine the threshold point for rankings to select only the required features and exclude noise.

The selection criterion is the Augmented Variance Ratio which gives a score for discriminative power.

$$AVR(F) = \frac{(S_F)}{\frac{1}{C}\sum_{i=1}^{C} \frac{Vra_i(S_F)}{min_{i \neq j(|mean_i(S_F) - mean_j(S_F)}}}$$

The ranking is simply selecting the features that have the top 1 percent AVR values.

2.2.2 Wrapper Approach

Wrapper methods embed the model hypothesis search within within the feature subset search. In this setup, a search procedure in the space of possible feature subsets is defined, and various subsets of features are generated and evaluated. The evaluation of a specific subset of features is obtained by training and testing a specific classification model, rendering this approach tailored to a specific classification algorithm.
To search the space of all feature subsets, a search algorithm is then wrapped around the classification model.
However, as the space of feature subsets grows exponentially with the number of features, heuristic search methods are used to guide the search for an optimal subset. Thus feature selection is of considerable importance in classification as it
1. Reduces the effects of curse of dimensionality
2. Helps in learning the model
3. Minimizes cost of computation
4. Helps in achieving good accuracy The selection criterion I have used is the classification rate:

$$classificationrate = \frac{TruePositives + TrueNegatives}{Total}$$

Where $total = totalpositives + totalnegatives$

The search criterion is the Sequential Forward Selection whose algorithmic flowchart is shown below. The stopping condition is that when the addition of a new feature feature does not improve the classification accuracy or all the features have been selected.

Flow Chart:

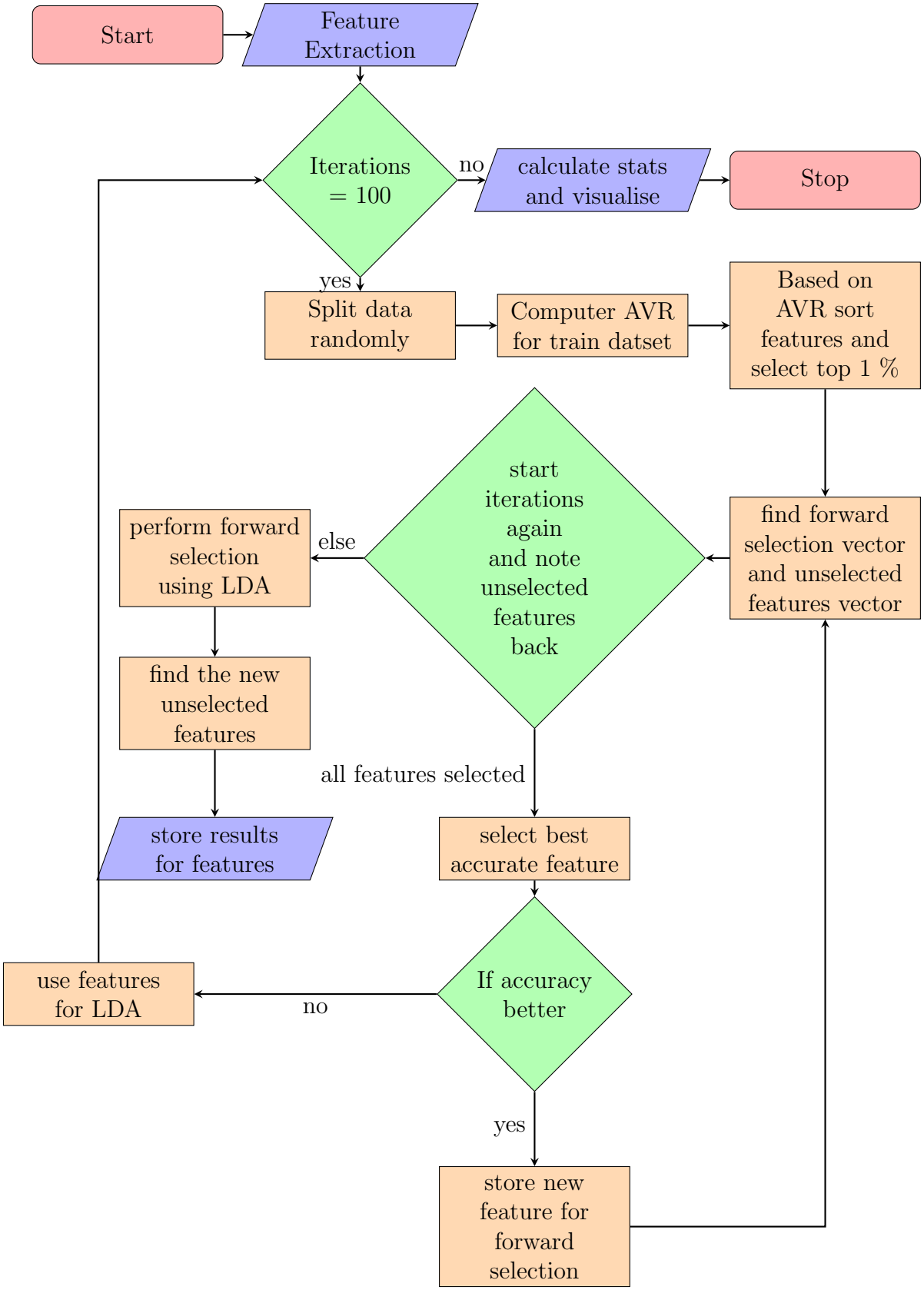
The flow chart explains the sequence of the feature selection algorithm using Filter and Wrapper approches.

This chart explains the key points for strting the 100 iterations and moving towards the scenarios of feature selcetion for most commom fetures based on accuracy depending on the AVR.

This flow proposes to modify LDA for feature selection. The idea and method is not only novel but also simple and intuitive. LDA has been applied to many real-world problems such as face recognition and machine learning.

However, in the previous LDA, was always exploited as a feature extraction method and had never been used for feature selection. That is, the majority of the applications of LDA is to exploit it to transform samples into a new lower-dimensional space. As the feature extraction result is an integrated reflection of the original components of samples, it seems that LDA is not able to perform feature selection. Actually no one has ever made this attempt.

However, in this paper, our analysis shows that it is feasible to apply LDA to feature selection and to use LDA to select a number of components from all the components of original samples. We achieve this by viewing the LDA transform from a viewpoint of numerical analysis.


Start
Feature Extraction
Iterations = 100
no
calculate stats and visualise
Stop
yes
Split data randomly
Computer AVR for train datset
Based on AVR sort features and select top 1 %
find forward selection vector and unselected features vector
start iterations again and note unselected features back
else
perform forward selection using LDA
find the new unselected features
store results for features
all features selected
select best accurate feature
If accuracy better
no
use features for LDA
yes
store new feature for forward selection

3 Results

The above mentioned methods are performed on two different datasets namely: FACE and EEG.The results are analyzed using the confusion matrices, classification matrices as well as accuracy rate, standard deviations and the visualisation of top 1 % of the data predicted by training the model for the linear discriminant from the references of [1] and [2].

Confusion Matrix: A confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes.

Classification Matrix: A classification matrix sorts all cases from the model into categories, by determining whether the predicted value matched the actual value. All the cases in each category are then counted, and the totals are displayed in the matrix

3.1 Forward Selection using Filter and Wrapper

The least squares method is utilized to model a linear discriminant using one vs all classifier for predicting the classes and their accuracy with the visualization of plots for classification along with confusion matrices.

3.1.1 FACE Dataset

FACE dataset has nearly 15500 features for a data point and the classifier used is LDA. There are results based on plots for features of 1% and also the classification matrix and confusion matrix based on all features. From the figure 3, we can say that there are outliers for class 2, and class 1 has minimal outliers.

And from table 1 we can show that the confusion matrix after test is performed on model is decently not indented diagonally an hece more the result for over-fitting is not visualised much.

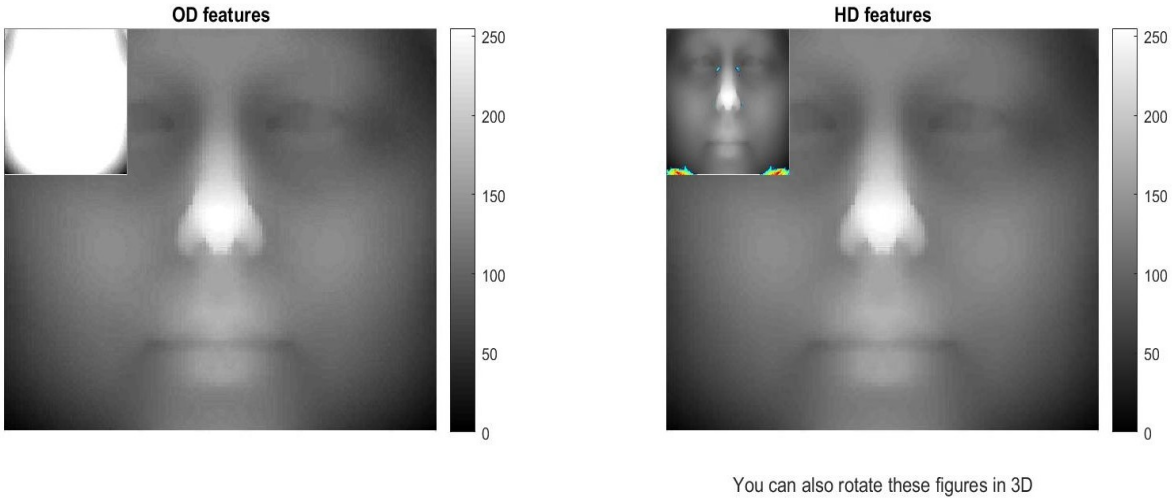


Figure 1: A visualization of variance ratio top 1% features for the FACE dataset.

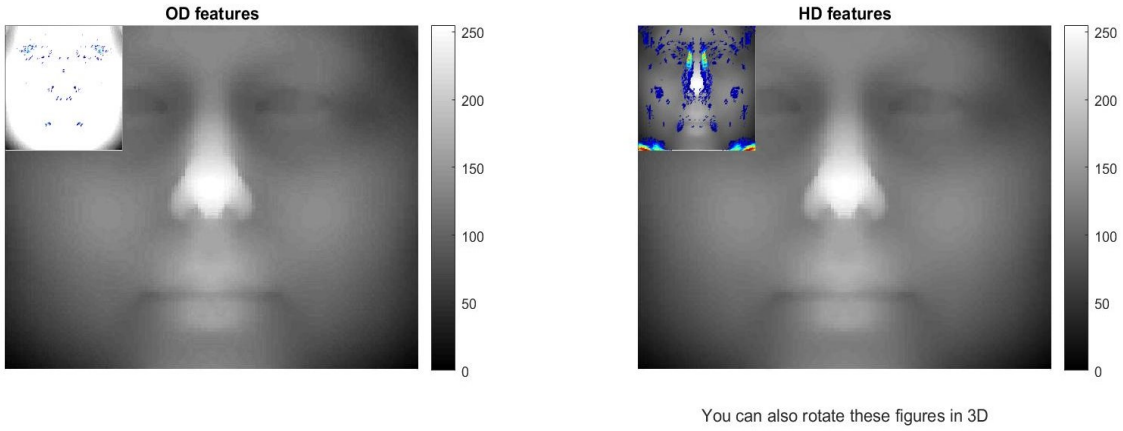


Figure 2: A visualization of top 1% features for the FACE dataset during 100 times of feature ranking.

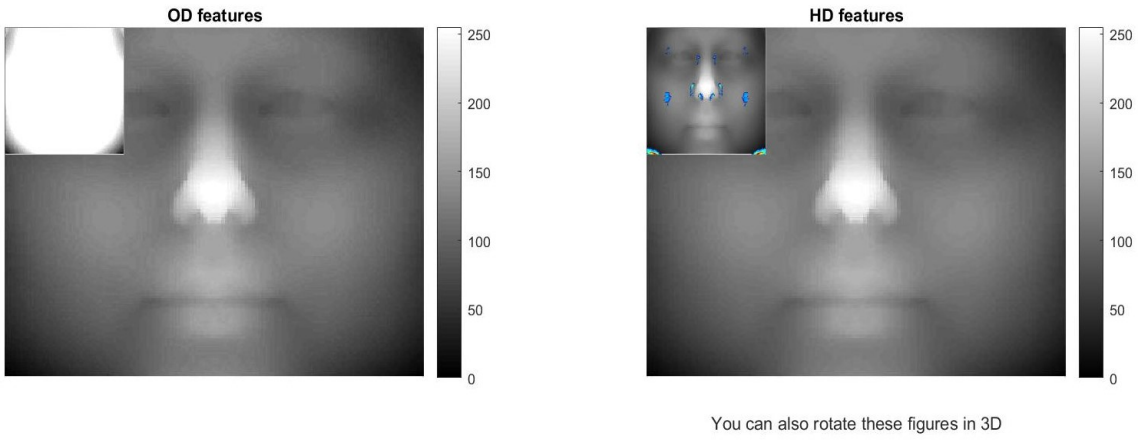


Figure 3: A visualization of top 10% or most features for the FACE dataset for 100 times of feature ranking.

0.1011	0.8989
0.0383	0.9617

Train Classification Matrix

0.0812	0.9188
0.0576	0.9424

Test Classification Matrix

1.8200	16.1800
1.3400	33.6600

Train Confusion Matrix

1.3800	15.6200
1.9600	32.0400

Test Confusion Matrix

Table 1: Confusion and Classification matrices for the Face dataset

3.1.2 EEG Dataset

The dataset consists of two kinds of features extracted from the EEG dataset.

Fourier Coefficient Features: The feature vector of Fourier Coefficients consists of 126 coefficients sampled from twice the original data's sampling rate which was given as part of the dataset. The sampling rate of the original data is at 420 Hertz and each value of the discretized data represents 0.00238 ms blocks of time. These are saved in the original data in the form of the sin and cos elements. There is a vector of 126 coefficients for each of the 128 electrode trials. We combined each pair of sin and cos elements using the magnitude defined as $((cos + sin * i)^2) \cdot 5$ which is labeled abs features. The total number of Fourier Coefficient Features are $48,384 = 126 * 128 * 3$.

Temporal Features: The temporal features consist of the mean of 0.1 second intervals of the EEG responses. Within each trial, the control and stimuli are present for .6 seconds each for a combined total of 1.2 seconds. This leads to 12 temporal features for each electrode. Since there are 128 electrodes there are 1,536 = 128 * 12 temporal features for each trial.

The two sets of feature are concatenated together to create a total of 49,920 features for each observation.

This dataset has raw data so it contains NAN's and all 0 entries for features and observations. You will have to set up automated cleaning this data before running your experiments. Some useful matlab functions are nanmean, nanstd, and nanmin.

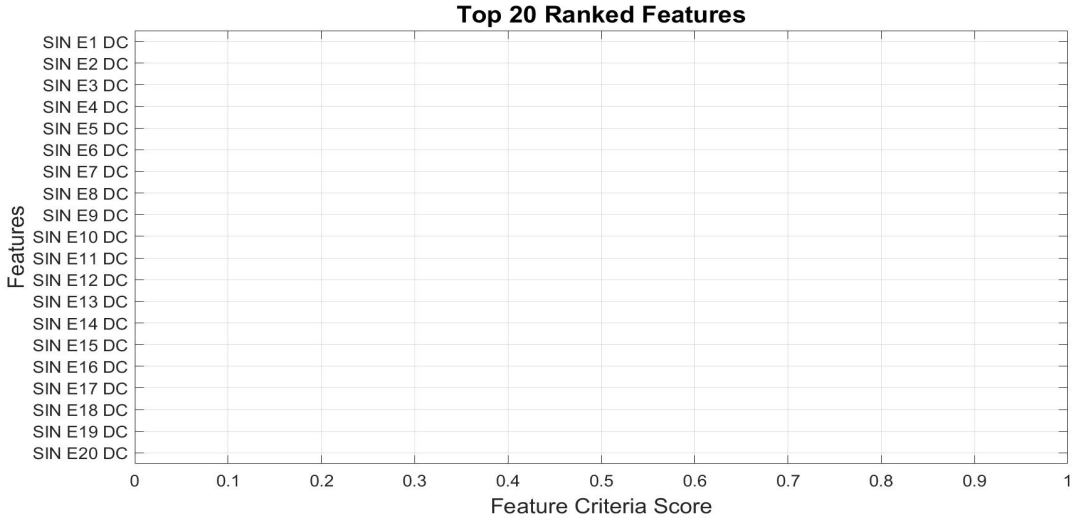


Figure 4: A visualization of variance ratio of features for the FACE dataset.

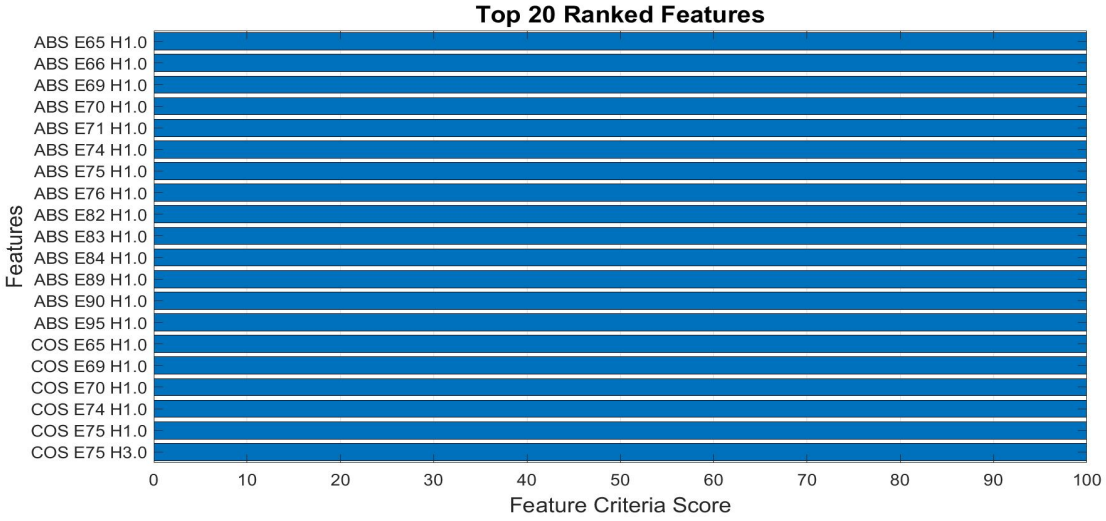


Figure 5: A visualization of top 1% features for the FACE dataset during 100 times of feature ranking.

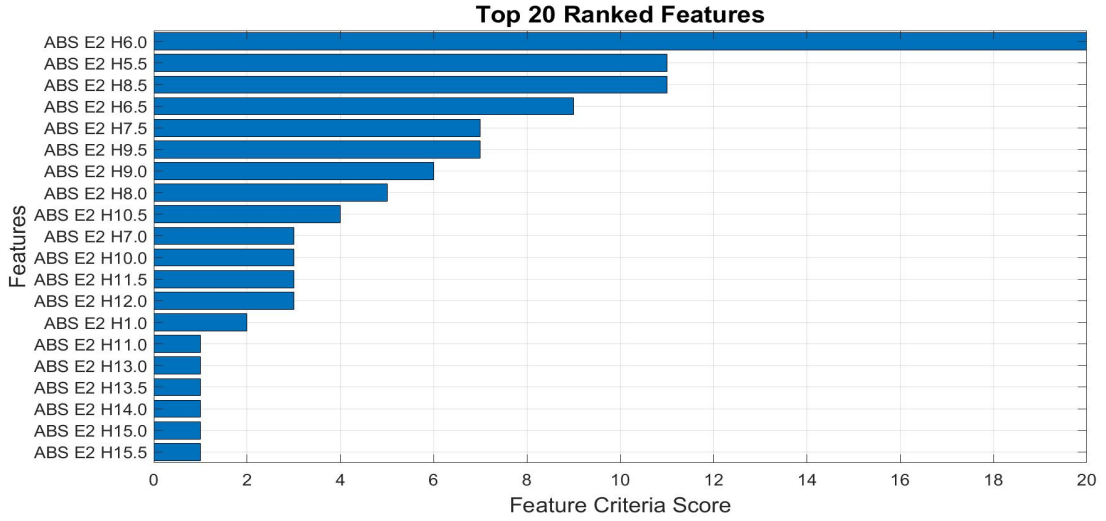


Figure 6: A visualization of top 10% or most features for the FACE dataset for 100 times of feature ranking.

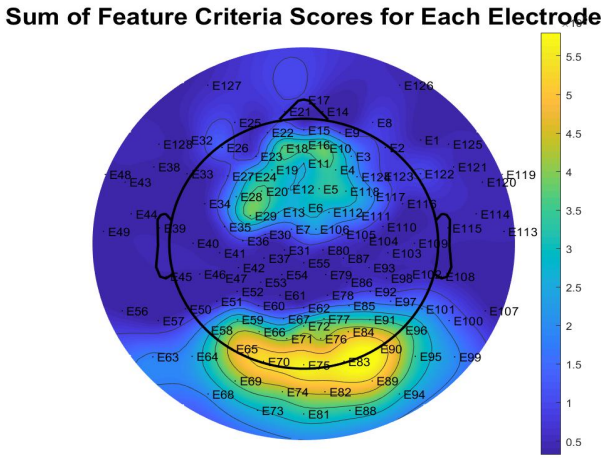


Figure 7: A visualization of top features for the FACE dataset for forward selection

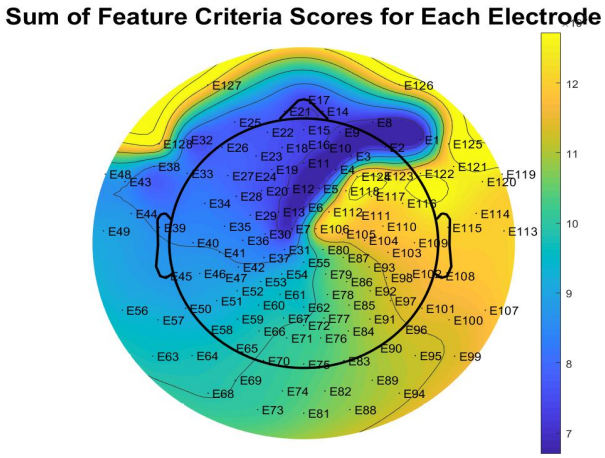


Figure 8: A visualization of top features for the FACE dataset for filter approach

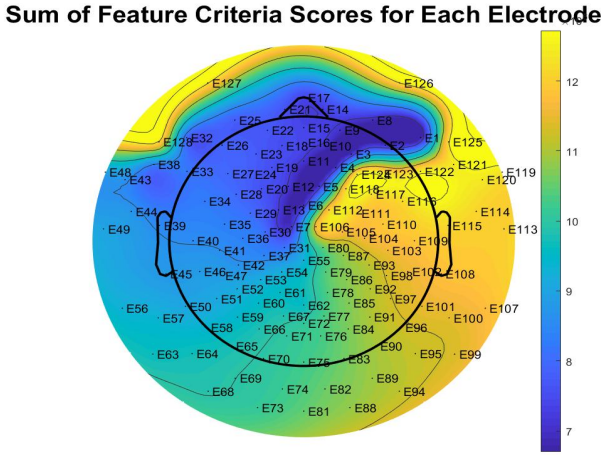


Figure 9: A visualization of top features for the FACE dataset for wrapper approach

0.3399	0.3665	0.2936
0.3220	0.3932	0.2848
0.3231	0.3706	0.3063

Train Classification Matrix

0.3361	0.3691	0.2948
0.3197	0.3856	0.2947
0.3266	0.3684	0.3051

Test Classification Matrix

Table 2: Confusion and Classification matrices for the EEG dataset

107.7600	116.1800	93.0600
102.0800	124.6500	90.2700
102.4100	117.4800	97.1100

Train Confusion Matrix

26.5500	29.1600	23.2900
25.2600	30.4600	23.2800
25.8000	29.1000	24.1000

Test Confusion Matrix

Table 3: Confusion and Classification matrices for the EEG dataset

3.1.3 Comparision for two Datasets

When the three datasets are compared the test accuracy of Face, EEG are unique and having the difference in accuracy for train ad test. For the face the difference is about nearly eual to 50% whereas, the EEG dataset also has the difference of 50%.
However, the comparison of both confusion and classification matrices are not that different as for both the datasets the diagonal representation of the averages is not that different but based on EEG dataset the values in confMat and classMat are huge and very large numbers compared to Face dataset.

Dataset	Train Acc	Train Std	Test Acc	Test Std
Face	1.0628	0	0.5118	0.6089
EEG	0.6930	0	0.3422	0.2914

Table 4: Train and Test accuracy and standard deviations for WINE, Wallpaper and Taiji datasets using Least Squares method for features 1, 7.

4 Conclusion

- This project would be a better scope to feature selection for a huge datasets.
- Based on the classification results the data in the matrices perform well and compared to previous project it s a way different and not far better performing in accuracy as I observed.

References

[1] C. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006. 6

[2] A. K. V. K. Pang-Ning Tan, Michael Steinbach, *Introduction to data mining*. New York, 2019. 6