Project 3: Feature Selection

Saniya Naphade

March 15, 2020

Abstract

The objective of this project is to select a subset of features of a classification problem using an evaluation criteria using the sequential forward selection algorithm for feature selection.

Contents

1 Introduction

The fundamental problem of pattern recognition and machine learning is to approximate a function that relates the input X$\{x_i|i = 1...N\}$ to the output Y where X is a matrix made up of vectors[1]. However, many a times, Y is determined by a subset of features of X as some features of X may be irrelevant. Therefore, feature selection algorithms play an important function in approximating the relationship between the input and output[3]. This is done due to a number of factors:
1: Irrelevant input features can have higher computational cost.
2: These input features can also lead to overfitting of the data
3: These features usually have very small effect on the output therefore, we can keep the approximation model small by ignoring these features. This reduces the computational complexity as well as cost.
Feature selection strategy involves the implementation of a filter and a wrapper.[3]

1.1 Filter

Filters are generally used as preprocessing step. The selection of features is independent of any machine learning algorithm. Features are instead, selected on the basis of their scores in various statistical tests for their correctness with the outcome variable[3]. In this step, features are selected on the basis of their performance in Variance Ratio test. For a feature F with values S_F in a data set with C total classes, the variance ratio (VR) of between- to within-variance is calculated as

$$VR(F) = \frac{Var(S_F)}{1/C \sum_{k=1}^{C} Var_k(S_F)}$$

where $Var_k(S_F)$ is the variance of the subset of values from feature F which belongs to class c.

1.2 Wrapper

The wrapper function uses the subset of features obtained from the filter stage and trains a model using them. The predictive accuracy of the trained model is evaluated using cross validation or statistical re sampling. Based on the value of the predictive accuracy, a feature is either added or discarded for the final subset [3].
There are various selection algorithms that can be implemented in the wrapper function[3], such as:
1. Forward Selection: This is a simple greedy search algorithm which starts with an empty feature set and features are added to this set in each iteration based on whether they best improve the training model. The stopping condition for this method is reached when the addition of a new feature does not improve the model.
2. Backward Elimination: This is also a greedy algorithm in which the worst performing features are removed in each iteration. In each iteration, the model is then trained on the remaining features until all features are not exhausted.
3. Recursive Feature Elimination: It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the

best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.
In this project, forward selection algorithm is implemented.

2 Data

The feature selection module was implemented for two datsets, namely:
1. FACE 2. EEG

2.1 FACE

This dataset is obtained from the 3D scanning of the faces of 104 people. For every person, the Height difference and Orientation difference are computed which give 15500 feature. The dataset comprises of 2 classes.
Therefore, the dataset comprises of 104 datapoints, 15500 features and 2 classes.

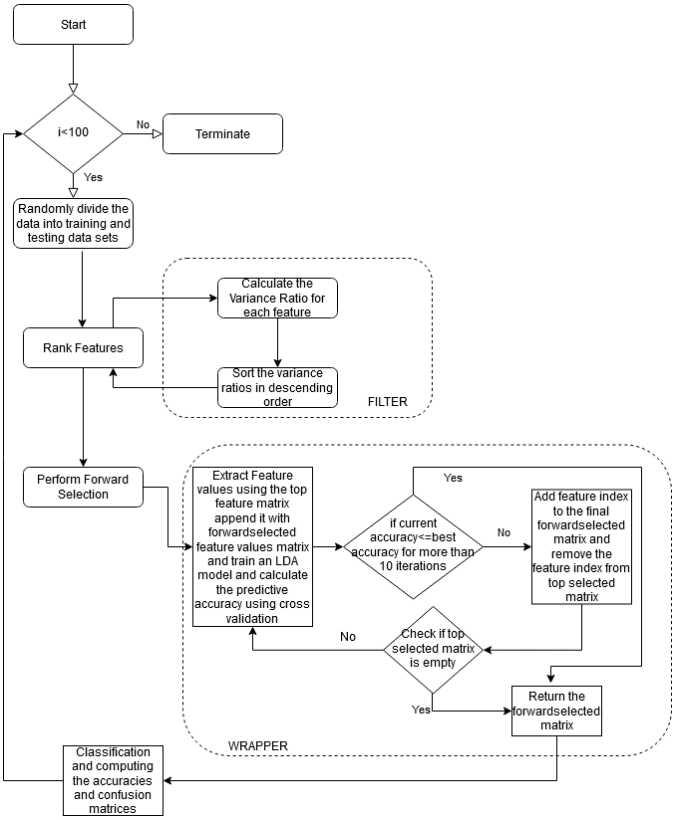
2.2 EEG

This dataset consists of the Fourier Coefficient Features and Temporal Features extracted from the original EEG dataset.
This dataset contains 1188 datapoints and around 49,920 features and 3 classes.

2.3 Extra Credit - Taiji

This dataset consists of individual footpressure frames from 10 different subjects performing Taiji. The dataset for each subject is a 4D matrix of dimensions [datapoints] x 60 x 21 x 2 which is reduced to a 2D matrix of dimensions [datapoints] x 2520. Therefore, the final dataset consisting of all the individual datasets giving the final dataset dimensions as 30,321 x 2520. The data is to be classified based on the gender i.e, male or female. Therefore, it has 2 classes: class 0 - Male; class 1 - Female.

3 Algorithm



As seen from the algorithm, Variance ratio is implemented for Filter and forward selection method is used for feature selection.

3.1 Workflow

1. the entire algorithm is ran 100 times.
2. In each iteraton, the dataset is divided and split into train data and test data (50-50 split for FACE and 80-20 split for EEG).
3. This top 1% features are extracted from the train data set using the variance ratio statistcal test. The variance ratios computed for all the features are sorted in descending order and the first 1% are returned.
4. The input for the wrapper function is the matrix consisting the indices and variance ratios of the top 1% features. Features are added one after the other and the prediction approximate of the model is computed. Based on the prediction accuracy, a decision is taken whether is the feature is selected or not.
5. The output of the wrapper function consists of the indices of the selected features. 6. A classification model is computed from these forwardselected features and the predictions are obtained for the test data.
7. Confusion matrices and the accuracy for each model is then computed which would give the overall accuracy of the feature selection module.

4 Results

This section described the results obtained for the two aforementioned data sets using the above described algorithm.

4.1 FACE

This result is obtained for 100 iterations and the top 1% ranked features are selected. The histogram is plotted for the number of times the features are in the top 10% features which are selected using the forward selection algorithm.

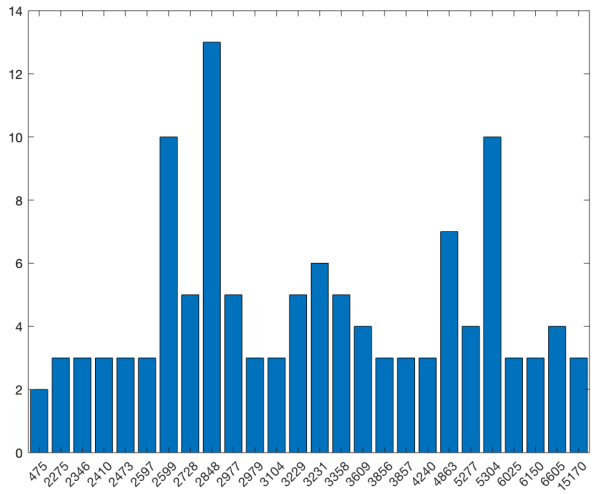
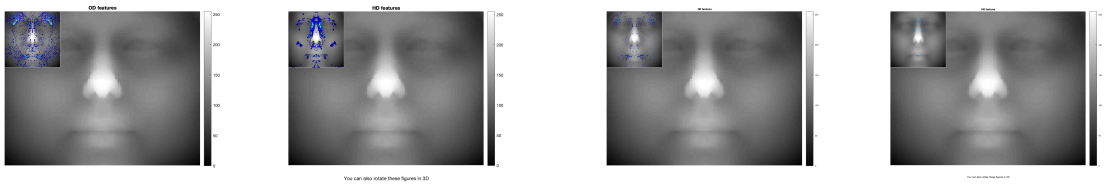


Figure 1: Histogram depicting the top 10% of the features selected via feature selection algorithm.



(a) (b)

Figure 2: Visualization of Fig. (a) top 10% features based on the number of times they appeared within the top 1%. Fig. (b) top 10% of the features based on the number of times they were chosen by the forward selection algorithm.

Table 3 gives the Average Confusion Matrix obtained for the feature module in which the features where pre-processed using Variance Ratio and selected using forward selection

algorithm.
The module has *average accuracy of 61.47% and standard deviation of 10.36* as shown in Table 5 for test data.

Table 1: Th Confusion Matrix of train data for 100th iterations

	0	1
0	12	6
1	0	35

Table 2: Th Confusion Matrix of test data for 100th iterations

	0	1
0	7	10
1	7	27

Table 3: The average Confusion Matrix of test data for 100 iterations

	0	1
0	6.19	10.81
1	6.29	27.710

Table 4: Classifier Results for train data

Overall Accuracy(%)	Standard Deviation
93.28	4.46

Table 5: Classifier Results for test data

Overall Accuracy(%)	Standard Deviation
61.47	10.36

4.2 EEG

This result is obtained for 50 iterations and top 1% ranked features are selected.

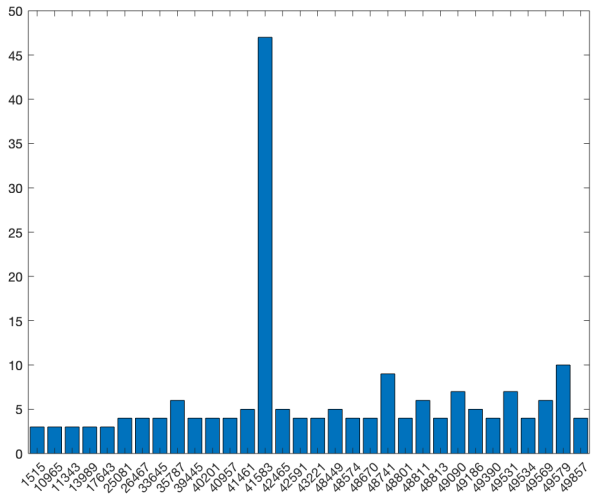
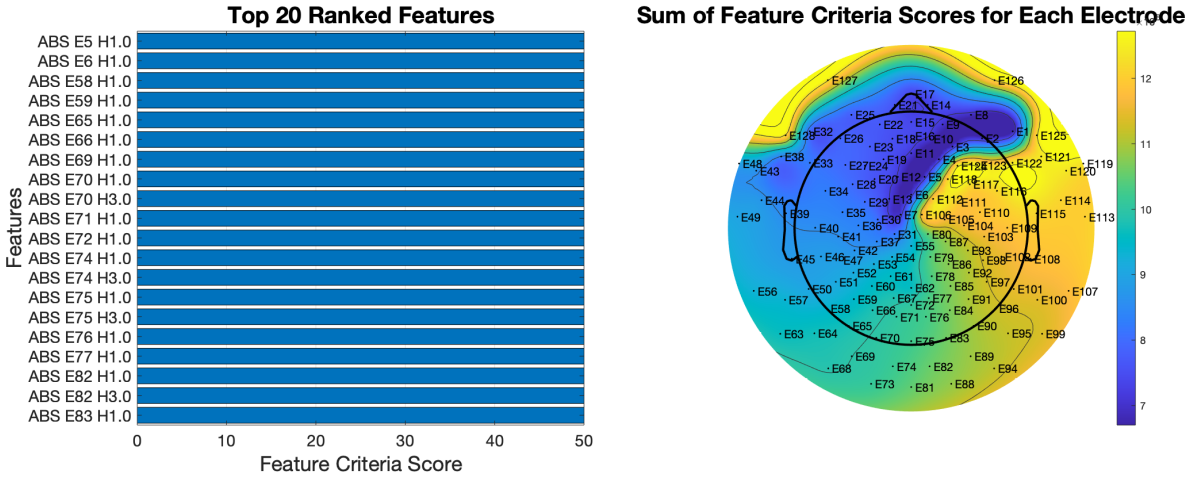
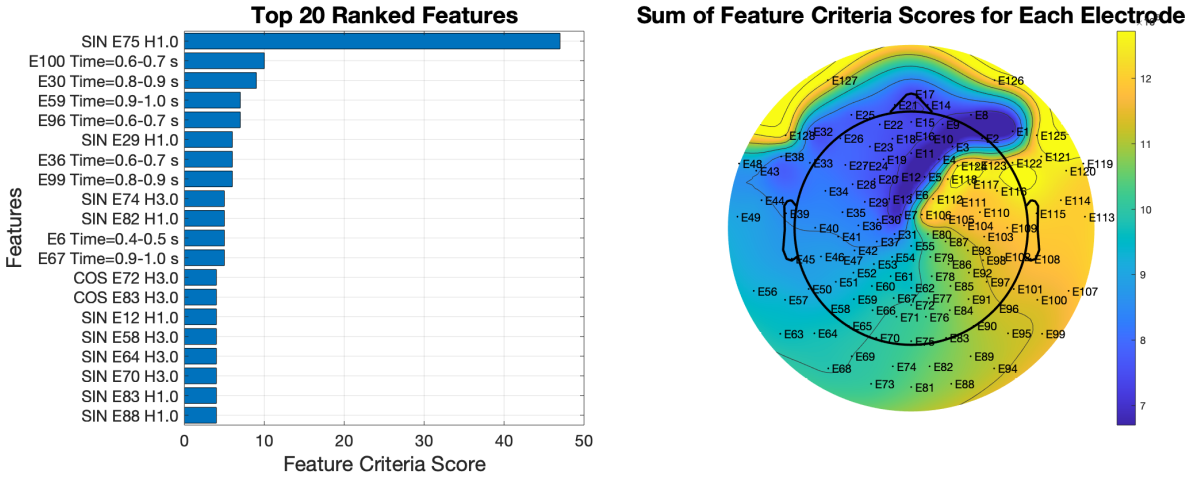


Figure 3: Histogram depicting the top 10% of the features selected via feature selection algorithm.



(a) (b)

Figure 4: Fig. (a)Histogram depicting the top 20 Ranked features based on the number of times they appeared within the top 1%. Fig. (b) Visualization of top 1% feature ranked based on variance ratio.



(a) (b)

Figure 5: Fig. (a)Histogram depicting the top 20 Ranked features based on the number of times they were chosen by the forward selection algorithm. Fig. (b) Visualization of top 20 Ranked features based on the number of times they were chosen by the forward selection algorithm.

Table 8 gives the Average Confusion Matrix obtained for the feature module in which the features where pre-processed using Variance Ratio and selected using forward selection algorithm.
The module has *average accuracy of 52.08% and standard deviation of 3.53* as shown in Table 10 for test data.

Table 6: The Confusion Matrix of train data for 50th iterations

	P2	P4M	PMG
P2	247	39	31
P4M	100	141	76
PMG	59	45	213

Table 7: The Confusion Matrix of test data for 50th iterations

	P2	P4M	PMG
P2	48	17	14
P4M	30	26	23
PMG	15	15	49

Table 8: The average Confusion Matrix for 50 iterations

	P2	P4M	PMG
P2	55.24	15.24	8.52
P4M	27.28	24.86	26.86
PMG	15.52	17.54	45.94

Table 9: Classifier Results for train data

Overall Accuracy(%)	Standard Deviation
61.50	1.68

Table 10: Classifier Results for test data

Overall Accuracy(%)	Standard Deviation
52.08	3.53

5 Conclusion

Through this project we understood the concept of feature selection for machine learning problem. We learnt various methods used for feature selection such as, Filter and Wrapper methods. This project helped us to understand the importance if feature selection in machine learning problems. We also observed that by the elimination of the irrelevant feature, the computation time reduced significantly. However, for the EEG dataset we

could see that this approach was not very fruitful. As the number of features increased, the time taken by the forward selection algorithm to run also increased significantly.
The model was trained using Linear Discriminant Analysis classifier. We observed that by reducing the number of features in the FACE dataset, we were able to reduce the over-fitting of data points. However, the model did not perform well for the EEG dataset.

6 Extra Credit

The program was run for 10 iterations and the top 5% ranked features are selected and forwarded to the forward selection algorithm.

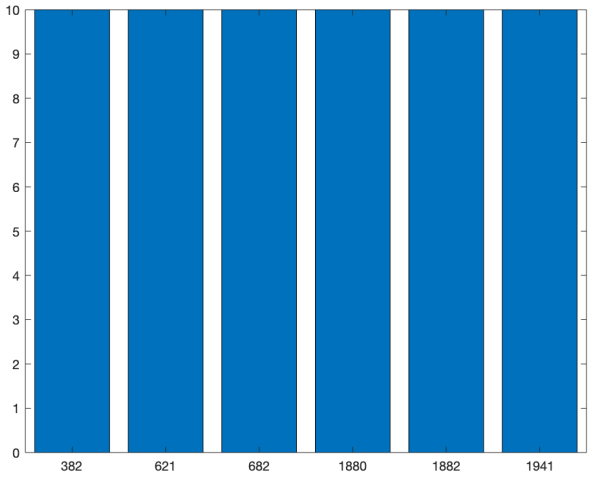
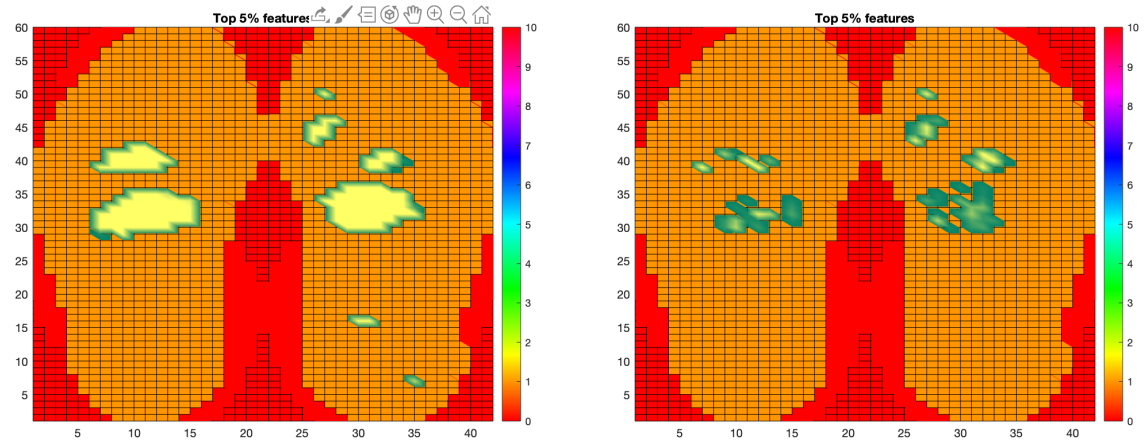


Figure 6: Histogram depicting the top 10% of the features selected via feature selection algorithm from the top 5%.

Table 13 gives the Average Confusion Matrix obtained for the feature module in which the features where pre-processed using Variance Ratio and selected using forward selection algorithm.
The module has *average accuracy of 98.296% and standard deviation of 0.0013* as shown in Table 15 for test data.



(a) (b)

Figure 7: Visualization of Fig. (a) top 10% features based on the number of times they appeared within the top 5%. Fig. (b) top 10% of the features based on the number of times they were chosen by the forward selection algorithm.

Table 11: The Confusion Matrix of Train Data for 10th iterations

	0	1
0	12435	20
1	361	11442

Table 12: The Confusion Matrix for Test Data for 10th iterations

	0	1
0	3102	11
1	87	2863

Table 13: The average Confusion Matrix of Test Data for 10 iterations

	0	1
0	3106.800	6.200
1	102.400	2847.600

Table 14: Classifier Results for train data

Overall Accuracy(%)	Standard Deviation
98.464	0.00098

Table 15: Classifier Results for test data

Overall Accuracy(%)	Standard Deviation
98.296	0.0013

References

[1] Feature Selection,
http://research.cs.tamu.edu/prism/lectures/pr/pr_l11.pdf 2

[2] Bishop,Pattern Recognition and Machine Learning,
http://users.isr.ist.utl.pt/ wurmd/Livros/school/Bishop

[3] https://www.analyticsvidhya.com/blog/2016/12/introduction 2