Contents

<Update the table of contents by right-clicking on it and selecting Update Field… and then select page numbers only.>

Chapter 1 Introduction

Medical data plays a decisive role in disease diagnosis. The classification accuracy of high-dimensional datasets is often diminished by several redundant and irrelevant features. In this context, feature selection becomes an indispensable process. Feature selection primarily intends to identify a feature subspace which retains the classification accuracy while reducing the high computational cost of learning model as well as eliminating noise. The suitability of an appropriate feature selection approach heavily depends upon the capability of that approach to match the problem framework and to discover the intrinsic patterns within the data.

The problem domain involves predicting central neuropathic pain using brain EEG (Electroencephalogram) of 18 participants. Various pre-processing (including signal de-noising and normalization, temporal segmentation, frequency band power estimation) have already been applied on EEG data. The dataset includes data from 18 participants in blocks of 10 repetitions per participant (180 data rows in total). Each row contains 9 different features extracted from data recorded by 48 different electrodes (48x9=432 features ordered by electrode first, feature type second).

<u>Objective</u>:

Evaluate feature selection methods (at least 2 including one filter method) on the classifiers (at least two) measuring its performance in terms of accuracy, specificity and sensitivity by applying leave-one-subject-out cross-validation.

At last we need to suggest a technique which works well on unseen data but where all 3 performance measures are comparable (i.e. specificity is not too high than sensitivity or vice versa)

Chapter 2 Method

In our studies, we created evaluated pipelines using two feature selection techniques. F-stat based filter method and RFE wrapper method as features that look irrelevant in isolation (in filter method) may be relevant in combination. The detailed step-wise methodology is given below.

<u>Methodology:</u>

1) *Creation of Training and Testing dataset*
    a. *To simulate a situation where test didn't exist at the of feature extraction and classification pipeline*
2) *Creation of feature extraction and classification pipeline*
    a. *Two techniques (F-stat based filter method) & RFE Wrapper method*
    b. *Evaluated for 3 classifiers (Random Forest, Logistic Regression, SVC)*
    c. *Leave one group out cross validation (User defined function created)*
    d. *Confusion matrix calculated for each fold & summed at last*
    e. *Sensitivity and Specificity calculated on final confusion matrix*
    f. *Mean and standard deviation of accuracy is calculated*
3) *F-stat Filter method*
    a. *Sklearn selectKbest is used in loop to find the optimal number of features for each classifiers*
    b. *Number of features (**called k**) hyper parameter is tuned after exploration & Visualization (70 to 160 in the range of 5)*
    c. *Finally, optimum features are selected based on leave one group out cross validation measures*
    d. *At last Accuracy, sensitivity, specificity noted down for each classifiers*
4) *RFE wrapper method*
    a. *Sklearn RFE package is used to recursive feature elimination technique is used*
    b. *Hyper parameter **n_features_to_select** is tuned and varied in range to find the optimal features (including number of features)*
    c. *This hyper parameter is varied for each classifier depending on the results from filter method*
    d. *Finally, optimum features are selected based on leave one group out cross validation measures*
    e. *At last Accuracy, sensitivity, specificity noted down for each classifiers*
5) *Cross-validation on the whole dataset (180 records) for the final performance evaluation simulating unseen data in real world use of the model*

Chapter 3 Results

Dataset: The dataset includes data from 18 participants in blocks of 10 repetitions per participant (180 data rows in total). Dataset is divided into Train & test set as we want to separate the data into training and test at the very beginning to simulate a situation where the test set didn't even exist yet at the time you build your feature extraction and classification pipeline.

Training set: 120 records (12 participants, 10 repetition). Three participants (30 rows) from positive class and 3 (30 records) from negative class are randomly removed to make sure it does not affect the balance of the dataset.

The further evaluation using feature selection pipelines are performed on Training set.

Below is the leave one group out performance on the train datasets for the chosen feature selection technique and classifiers.

Cross-validation performance on train dataset

Techniques	Random Forest				Logistic Regression				SVC			
	# of Features	Accuracy*	Specificity	Sensitivity	# of Features	Accuracy*	Specificity	Sensitivity	# of Features	Accuracy*	Specificity	Sensitivity
F-Stat (Filter method)	145	75.83 (18.01)	75.61	75.95	115	85(15)	68.75	90.91	135	92.50 (9.24)	72	97.89
RFE (Wrapper method)	280	76.67 (10.27)	72	80	130	89.17 (11.15)	75.76	94.25	170	95 (11)	75	96.43

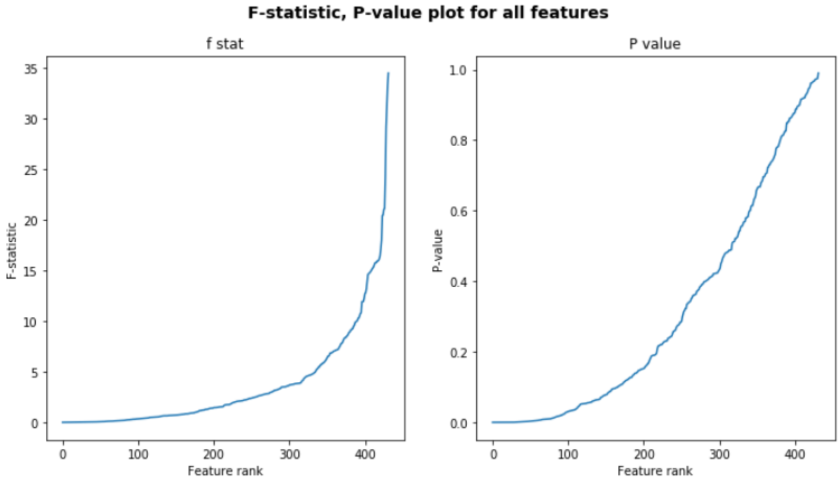
The best technique and classifier chosen after train set performance is also evaluated on the whole dataset (180 records). It will also include test data on which was never available during testing.

Cross-validation performance on whole dataset

Techniques	Random Forest				Logistic Regression				SVC			
	# of Features	Accuracy*	Specificity	Sensitivity	# of Features	Accuracy*	Specificity	Sensitivity	# of Features	Accuracy*	Specificity	Sensitivity
F-Stat (Filter method)	145	78.33 (15.37)	82.54	76.07	115	81.11 (14.87)	77.27	83.33	135	85.56 (11.65)	75.76	91.23
RFE (Wrapper method)	280	78.89 (12.86)	80.88	77.68	130	84.44 (13.83)	75.81	88.98	170	85.00 (8.33)	84.42	85.44

Rationale behind hyper parameter of Filter method (F-stat):

After visualizing the f-stat against feature rank and feature rank against p-value, we decided to keep a range of (70 to 160) for number of features based on significant low p-value.



Rationale behind hyper parameter of RFE wrapper method**:**

RFE wrapper method is used to include the features that look irrelevant in isolation (in filter method) may be relevant in combination. The lower range if ***n_features_to_select*** is set based on filter method final values of number of features of each classifiers (Like 145 for random forest, 115 for Logistic regression and 135 for SVC).

Rationale behind evaluation the final performance on whole dataset.

All classification pipelines are evaluated on the whole dataset - using all data from all subjects as test data; it is therefore necessary, again, to perform leave-one-subject-out cross-validation here. So, to be clear, there are two levels of cross-validation:

1. cross-validation on the training data to compare feature selection methods, classifiers, and hyper-parameters, and

2. cross-validation on the whole dataset for final performance evaluation.

Chapter 4 Discussion

Conclusions:

Following conclusion can be drawn from the analysis:

1) Filter method provide a great way to select features based on their individual relevance. Visualization would help in setting up the hyper-parameter to shorten the run-time of the pipeline
2) Wrapper method provides a way to include the features that look irrelevant in isolation (in filter method) may be relevant in combination.
3) It is always preferable to setup wrapper-method hyper-parameter based on the results on previous tried methods. In our case we have used filter method output to setup the lower range.
4) The performance of wrapper method is in general better than filter method for all classifier both on train set and on whole dataset
5) On Unseen data (Whole data set):
    a. Performance decreases a bit from train data which is what expected when we have small dataset to train on
    b. Performance on **RFE wrapper method using SVC classifier** is better than others and that model is finally recommended
        i. as it has low standard deviation of accuracy (8%) with a mean accuracy of 85%
        ii. Both specificity and sensitivity are comparable (around 85%) while in other cases sensitivity is quiet high

At last we want to select a model where all performance measures (accuracy, sensitivity & specificity) are comparable i.e. sensitivity should not be high at the cost of specificity or vice versa. It is so because treatment can't be given to wrong persons as preventive medication have strong side effects.

Future work:

One can also try other hybrid methods involving wrapper method (+filter method) where features are arranged based on their importance or relevance from the already executed filter methods.

References

[1] Aleksandra Vuckovic, Vicente Jose Ferrer Gallardo, Mohammed Jarjees, Mathew Fraser, Mariel Purcell, Clinical Neurophysiology 129 (2018) 1605-1617, https://doi.org/10.1016/j.clinph.2018.04.750

[2] https://doi.org/10.1016/j.clinph.2018.04.750.