Feature Selection Methods for Cancer Subtype Classification

CONTENTS OF THIS FILE

Introduction

Differential diagnosis of Cancer sub-types and thus their treatments are one of the most challenging problems in clinical Medicine.. To choose the right treatment plan, a proper diagnosis is to be made.This project is intended to accurately classify a patient's cancer type based on gene expressions.
We have modelled a SVM classifier using different feature selection methods (Variance threshold,Selectkbest,selectmodel,RFE and SFS) to find the top features that would help in the accurate prediction of Cancer subclasses.The refined data set with quality samples makes it easier for differential diagnosis

Installation

- Download the zip folder
- Change Directory to G16-Source-code.
- Install stuff "**sudo apt -get install libfreetype6-dev libxft-dev**"
- Install Project dependencies "**pip install -r requirements.txt**"
- Run the different files named by the feature selection methods for results.

Data and Source Code Overview

DataSets :-
Both the training and testing Data with the different classes details are saved under Data folder
Inside Data folder:

1. **Training_cls.txt :**
   Contains the 14 subclasses of Cancer with their indices for training the data set
2. **Training_res.txt** :
   The file contains 144 samples of sub classes with 1084 features for training the model
3. **Testing_cls.txt :**

Contains the 14 subclasses details for testing the dataset

4. **Testing_res.txt** :
   The file contains 54 samples of subclasses with 1084 features for testing the model

Source Code :-

There are five different Feature Selection Methods implemented as a part of this project. Each of the methods and their results of classification are separated into different files.

Python files:

1. **selectkbest.py :**
   The file implements the selectkbest feature selection method on Training_cls.txt and Training_res.txt which then modelled using svm classifier to produce output on the test dataset
2. **variancethreshold.py:**
   Here, variance threshold feature selection method is used on the training data set.
3. **rfe.py**
   Here, Recursive feature Elimination method is used on the training data set.
4. **selectmodel.py**
   Here, the selectfrommodel method is used on the training data set.
5. **sfs.py**
   Here, the Sequential feature selection (SFS) method is used on the training data set(executed using python3 because SFS was added in the latest version).
6. **addon.py**
   This file contains the code for preprocessing of the raw data set which creates a dataframe file (traindata.txt) .This snippet is imported in all the feature selection method files.

Output

Output folder contains the screenshots of classification reports obtained from different feature selection methods .

Note: The results of SFS feature selection method is not shown due to the lack of processing support in PC but the code is shared.

References

- S. Ramaswamy, et al Multiclass cancer diagnosis using tumor gene expression signatures.
- https://towardsdatascience.com/5-feature-selection-method-from-scikit-learn-you-should-know-ed4d116e4172