Classification of Cancer based on Gene Expression RNA-Seq Data Set

Zhanyang Zhu Zhanyang.Zhu@Gmail.com 10/18/2022 for UCSD Machine Learning Bootcamp Capstone Project

Goals:

1. **Train multi-class classification models to determine a cancer type given gene expression data of a patient**
2. **Analyze the importance genes (features) that can distinguish the cancers**
3. **Perform feature dimension reduction to determine a minimal set features (i.e., genes) that can be for testing**

In this exercise, the cancer types are limited to BRCA, KIRC, COAD, LUAD and PRAD, SKCM, THCA, LGG. More cancer types can be included.

TCGA Study Abbreviations https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations

Data Selection:

For this UCSD Data Bootcamp projects, I downloaded eight data sets from https://www.synapse.org/#!Synapse:syn2812961. All eight data sets are Illumina HiSeq RNASeq V2 data collected by unc.edu (data set name and cancer type)

1. **unc.edu_BRCA_IlluminaHiSeq_RNASeqV2.geneExp.tsv - BRCA: Breast invasive carcinoma**
2. **unc.edu_COAD_IlluminaHiSeq_RNASeqV2.geneExp.tsv - COAD: Colon adenocarcinoma**
3. **unc.edu_KIRC_IlluminaHiSeq_RNASeqV2.geneExp.tsv - KIRC: Kidney renal clear cell carcinoma**

4. **unc.edu_LUAD_IlluminaHiSeq_RNASeqV2.geneExp.tsv - LUAD: Lung adenocarcinoma**
5. **unc.edu_PRAD_IlluminaHiSeq_RNASeqV2.geneExp.tsv - PRAD: Prostate adenocarcinoma**
6. **unc.edu_SKCM_IlluminaHiSeq_RNASeqV2.geneExp.tsv - SKCM: Skin Cutaneous Melanoma**
7. **unc.edu_THCA_IlluminaHiSeq_RNASeqV2.geneExp.tsv - THCA: Thyroid carcinoma**
8. **unc.edu_LGG_IlluminaHiSeq_RNASeqV2.geneExp.tsv - LGG: Brain Lower Grade Glioma**

Samples (instances) are stored column-wise. Variables (attributes in rows) of each sample are RNA-Seq gene expression levels measured by illumina HiSeqV2 platform.

- **4859 samples with 20530 features**
- **For more data summary see https://github.com/ZhanyangZhuSD/UCSDMLCapstone/blob/main/GeneExpressionCancerRNA-Seq.ipynb**

Classification Model Selection

For explorative analysis with 40% of the total data, the following classification models are trained and the best model is selected. For details, see https://github.com/ZhanyangZhuSD/UCSDMLCapstone/blob/main/GeneExpressionCancerRNA-Seq.ipynb

1. **K-Nearest Neighbours,**
2. **Logistic Regression,**
3. **Naïve Bayes, Stochastic Gradient Descent,**
4. **Decision Tree,**
5. **Random Forest, Support Vector Machine and**
6. **XGBoost**

Results

- **Best classification model is XGBoost based on the explorative analysis.**

(https://github.com/ZhanyangZhuSD/UCSDMLCapstone/blob/main/GeneExpressionCancerRNA-Seq.ipynb)

- **The accuracy is 0.99918 when XGBoost model is applied to the full data set with all the features are used.**
- **With feature importance cutoff at 0.0005, the # of features (i.e. genes) is 105. The classification accuracy remains at 0.99918.**
- **Given the number of genes reduced to 105 from 20530, the cancer test will be cheaper and easier to run.¶**

For detail, see
https://github.com/ZhanyangZhuSD/UCSDMLCapstone/blob/main/GeneExpressionCancerRNA-Seq_FullData_XGBoost.ipynb

Others Notes:

This project was inspired by a similar UCI data set - a random extraction of gene expressions of patients having different types of tumor: BRCA, KIRC, COAD, LUAD and PRAD

https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

Relevant Papers:

Weinstein, John N., et al. 'The cancer genome atlas pan-cancer analysis project.' Nature genetics 45.10 (2013): 1113-1120.