Project Title: Cancer Subtype Multi-Class Classification in Gene Expression Data

Project Author: Ryan Urbanowicz – Cedars Sinai Medical Center - (Based on UCI Repository)

Short Description: This project deals with a multi-class classification task (5 tumor types) within RNA-seq gene expression data. This project will introduce participants to topics such as data cleaning, clustering, feature selection methods, and machine learning modeling methods.

Suggested Tags: multi-class, classification, machine learning, feature selection, automl

Long Description:

In this project, we will be looking at gene expression data first made available by the TCGA Pan Cancer analysis project, and collected by the UCI Machine Learning Repository. It has been sampled so that there is enough data for us to train models, without being too bulky to easily store locally or in the cloud.

We will focus on two different areas:

1. Unsupervised learning to explore relationships within the data.
2. Supervised learning to classify cancer type from the gene expression.

Abstract from the UCI Dataset Repository

This collection of data is part of the RNA-Seq (HiSeq) PANCAN data set, it is a random extraction of gene expressions of patients having different types of tumor: BRCA, KIRC, COAD, LUAD and PRAD.

Abstract from the Original Publication (Weinstein et. al. 2013):

The Cancer Genome Atlas (TCGA) Research Network has profiled and analyzed large numbers of human tumors to discover molecular aberrations at the DNA, RNA, protein and epigenetic levels. The resulting rich data provide a major opportunity to develop an integrated picture of commonalities, differences and emergent themes across tumor lineages. The Pan-Cancer initiative compares the first 12 tumor types profiled by TCGA. Analysis of the molecular aberrations and their functional roles across tumor types will teach us how to extend therapies effective in one cancer type to others with a similar genomic profile.

Dataset Sources:

Dataset files are included with this summary in the folder ‘Data’. These data files are also available from the UCI repository: https://archive.ics.uci.edu/dataset/401/gene+expression+cancer+rna+seq

The original data in its entirety is available from: https://www.synapse.org/#!Synapse:syn4301332

Basic Data Orientation/Preparation:

Dataset Name: TCGA-PANCAN-HiSeq-801x20531.tar

First extract the dataset from the '.tar' file. For example, you can unzip the .tar file, as well as secondary .tar file that is unpacked to get the two underlying data files ('data.csv', and 'labels.csv').

'data.csv' includes the 'X' part of the dataset (i.e. rows of instances, identified by sample identifier, and columns of gene expression features)

'lables' includes the 'y' part of the dataset (i.e. rows of class-outcome, also identified by the same sample identifier)

It may be useful to start by merging these two elements together into a single dataset or dataframe.

Cancer Subtypes in the Dataset (Multi-class outcomes)

- BRCA: Breast Invasive Carcinoma
- KIRC: Kidney Renal Clear Cell Carcinoma
- COAD: Colon Adenocarcinoma
- LUAD: Lung Adenocarcinoma
- PRAD: Prostate Adenocarcinoma

Dataset Information:

Samples (instances) are stored row-wise. Variables (features) of each sample are RNA-Seq gene expression levels measured by illumina HiSeq platform. There are no missing values.

A dummy name (gene_XX) is given to each feature. Check the original data from www.synapse.org or the platform specs for the complete list of probes name. The features are ordered consistently with the original data.

References and Suggested Reading:

Original Publication

- Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. and Stuart, J.M., 2013. The cancer genome atlas pan-cancer analysis project. Nature genetics, 45(10), pp.1113-1120.

Other Useful Publications

- Urbanowicz, R.J., Meeker, M., La Cava, W., Olson, R.S. and Moore, J.H., 2018. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, *85*, pp.189-203.

- Urbanowicz, R.J., Olson, R.S., Schmitt, P., Meeker, M. and Moore, J.H., 2018. Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of biomedical informatics*, *85*, pp.168-188.
- Olson, R.S. and Moore, J.H., 2016, December. TPOT: A tree-based pipeline optimization tool for automating machine learning. In Workshop on automatic machine learning (pp. 66-74). PMLR.
- Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P. and Poggio, T., 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences, 98(26), pp.15149-15154.
- Ooi, C.H. and Tan, P., 2003. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics, 19(1), pp.37-44.
- Senbagamalar, L. and Logeswari, S., 2024. Genetic Clustering Algorithm-Based Feature Selection and Divergent Random Forest for Multiclass Cancer Classification Using Gene Expression Data. International Journal of Computational Intelligence Systems, 17(1), p.23.