UNIVERSITY OF IDAHO, MOSCOW, IDAHO

ANALYSIS OF TUMOR TYPES

By

ROSEBELLA CAPIO

A FINAL REPORT ON TUMOR TYPES USING MACHINE LEARNING ALGORITHMS

December 10, 2018

0.1 Problem Statement

Can gene expressions be used to determine what type of tumor a patient has? Using 5 different types and gene expressions, we wish to test and classify these types using machine learning and aid doctors with proper diagnosis tools.
A tumor, also known as neoplasm, is an abnormal mass of tissue that can be solid or fluid-like. That is, a tumor is a kind of lump or swelling that does not necessarily pose a health threat. A tumor is NOT the same as CANCER although some can develop into one.

0.2 Motivation

I decided to pursue this project due to my own tumor growth on my body. I was always told my growth will increase in size and cause me major health issues. This has been a great worry ever since growing up. Doctors in my home country, Ghana, were not helpful either. Through this project, I finally realized that it isn't a cancerous growth thankfully (because for as far back as I can remember, the growth has been the size and causes me zero pain). I believe tumor and or cancerous growth research works and classifications using genes will be great for health practitioners and ensure proper diagnosis especially in developing countries.

0.3 Objectives

- Principal Component Analysis
- Classification Analysis
- Clustering Analysis
- Association

0.4 Data Description

The data used in this project is a subset of the RNA-seq PANCRAN data set(that is, Pancreatic Cancer), which was obtained from the UCI Repository. It consists of random extraction of gene extractions of patients having different Pancreatic tumor types.

0.4.1 Preliminary Exploratory Data Analysis

a The data has 801 variables and $20,532$ features

b The feature names are gene1 to gene$20,532$ and the observations are labelled as the as of 5 different tumor types: BRA, KIRC, COAD, LUAD, PRAD.

c There are $20,532$ numerical features

d There is one categorical variable, that is the response (tumor classifier)

e There are no text variables

f No variables other than the ones mentioned above are mentioned

g The methods used to process/prepare this data are

h There are no zero missing values.

0.5 Literature Review & References

- Weinstein, John N., et al. 'The cancer genome atlas pan-cancer analysis project.' Nature genetics 45.10(2013) : $1113 - 1120$.
- Samuele Fiorini, samuele.fiorini '@' dibris.unige.it, University of Genoa, redistributed under Creative Commons license (http://creativecommons.org/licenses/by/3.0/legalcode) from https://www.synapse.org#Synapse:syn4301332

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

0.6 Modeling Process

0.6.1 Principal Component Analysis

Principal Component Analysis was used on the data set in efforts to reduce to the number of features. From the analysis, about 700 of the features can be used to explain almost all (if not all) of the variation. In fact, form the pca graph, the initial 200 features explains over 80% of the total variation. A graph of pca1 vs pca0 provided 5 distinct clusters.

0.6.2 Clustering

K Means Clustering was employed which gave the optimal number of clusters for both the pca and non-pca data to be 6. It was gave a Silhouette Score of 0.22469 for the non-pca and 0.22874 for the pca data. This signifies the inability to properly cluster the data.

Spectral Clustering, Gaussian Mixture Modeling, Agglomerative Clustering, DBSCAN and MiniBatchKmeans Clustering were all used on the pca data to provide a graphical presentation of the clusters. The Dendrogram produced many clusters which is quite difficult to explain which is due to the vast nature of the data and hence the clusters from pca0 vs pca1 is preferred.

0.6.3 Classification Models

Logistic Regression, K Nearest Neighbor, Random Forest, Decision Trees, Naive Bayes, Neural Networks and Support Vector Machine were the classification models used in this work. The table below gives the models versus their respective

Logistic Regression	0.99
KNN	0.99
Random Forest	0.26
Decision Trees	0.58
Naive Bayes	0.27
Neural Nets	0.28
SVM	0.28

test accuracies. Logistic Regression with no tuning parameter gave an accuracy on 0.99 on the training set which is pretty good. 10 fold cross validation also provided an accuracy of 0.998

K Nearest Neighbor gave a test accuracy of 0.99 for 5 neighbors and a 10 fold cross validation score of 0.998

Random Forest Classifier gave an accuracy of 0.26 on the training set

Decision Trees also provided a test accuracy score of 0.578 which isn't all that bad

Naive Bayes gave a test accuracy of 0.2704 signifying a poor job at classifying Neural Networks gave a test accuracy score of 0.28

Support Vector Machine also gave a test accuracy score of 0.28. Hence a bad choice as a classifier. This suggests that either Logistic Regression of KNN Classifier is a great choice for this data set.

0.6.4 Association

To ensure that Association Analysis could be performed, the data was recoded to transform the continuous data into $0's$ and $1's$. The median for each gene was found and values below the particular median was coded and 0 and those above the median 1. A random sample of size 2000 was used for the association because using all $20,532$ was taking over a day to run. Limiting the itemsets to 2, gene15994 provided the highest support with a value of 0.499. Tumor type BRCA appears to be the most common. Gene18905 also appears to be a common antecedent.For future work, a more powerful computer could be used to get more insight and ensure a thorough analysis.