Breast Cancer Detection

March, 2019

Problem Description

Breast cancer is the leading type of cancer in women worldwide. Molecular subtyping of breast cancer has become common practice to understand prognosis of disease, and to design a treatment plan [1]. The subtype indicates the severity of the cancer and influences the treatment plan formulation. This task is to develop an automated method to classify the molecular subtype of breast cancer based on ultrasound images and clinical diagnostic data.

Dataset

The dataset can be downloaded from the following link (token: 65yf):
https://pan.baidu.com/s/1saEeST34iLkn7nSN3fNVPQ
In this dataset, you are given hundreds of medical records of breast cancer patients. Each medical record is associated with several ultrasound images and some clinical diagnostic data. The clinical diagnostic data contains the following fields:

Field	Type	Meaning
id	str	Patient ID
age	int	Age of the patient
HER2	int	Scale (0-3) of how strongly HER2 (marker for genetic predisposition for breast cancer) is detected
P53	bool	Whether P53 (marker for genetic predisposition for cancer) is positive
molecular_subtype	int	Molecular subtype of breast cancer, there are four types of molecular subtypes in the dataset (1: Luminal A, 2: Luminal B, 3: HER2-Enriched, 4:Triple Negative)

Submission & Evaluation

To evaluate your work, you are asked to submit a csv file contains the predication results of ALL the patient IDs in the test set. The submission file should consist N lines where N is the number of patients in the test set. Each line contains a patient id and your prediction of the patient separated by comma. A sample of the submission file can be found in the dataset.
PLEASE NOTE:

- No additional lines (e.g., header lines, comment lines) are allowed in the submission file.
- The order the the patient ids in your submission file is NOT necessary to be consistent with that in the test set.

Submissions are scored on *Accuracy* and *Macro-F1*:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{y}_i = y_i) \tag{1}$$

$$Macro\text{-}F1 = \frac{1}{C}\sum_{c=1}^{C}F1_c = \frac{1}{C}\sum_{c=1}^{C}\frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c} \tag{2}$$

where:

- N is the number of patients in the test set,
- $\hat{y}_i$ and y_i are the predicated and truth label of the i-th patient, respectively,
- $\mathbb{1}(\cdot)$ denotes the indicator function[1],
- C is the number of possible labels,
- $Precision_c$, $Recall_c$ and $F1_c$ are the precision, recall and F1 score of class c, respectively.

We will provide the evaluation script later.
In addition to the submission of the prediction results of test set, you will be required to submit an experiment report and the code used in this task. The experiment report should include, but is not limit to, methodologies, experiment results, related work and references.

References

[1] Celebi, Filiz et al. 2015. The role of ultrasonographic findings to predict molecular subtype, histologic grade, and hormone receptor status of breast cancer. *Diagnostic and Interventional Radiology.* 21(6), 448.

[1] https://en.wikipedia.org/wiki/Indicator_function