

Data Science with R Project Proposal

Team: COVID-19 Predictor

20.05.2020

Project Title: COVID-19 Prediction using Explainable Machine Learning

Background and Motivation:

The COVID-19 or the SARS-COV2 originated from the district of Wuhan, China has transpired to be a pandemic worldwide [WHO reference]. Research on the COVID-19 is a hot topic among the Artificial Intelligence community recently. Due to shortage and limited efficiency of current testing mechanism of COVID-19 tests, i.e. through RT-PCR kits @Zhao2020COVIDCTDatasetAC. These kits usually takes upto 4-6 hours to reproduce the results which, statistically speaking not very optimal way to move forward as the rate of COVID-19 patients registration grows exponentially. with this problem in scientific community, it motivated to aim of Data Sciene can be brought to help. So, it leaded to a possibility of building a classifier which can diagnose patients as COVID-19 negative or positive based on their respective X-Ray images. As this approach seems to be less time and resource consuming and we hoped to achieve a competible performance w.r.t RT-PCR kits.Also in addition to a good prediction, we needed reasons that could justify what are the features that may be responsible to the diagnostic process. In addition to current research, our work tries to finds insights of COVID-19 detection using simpler explainable models and effectively use the research surrounding the virus. Aim of our project is to come up with an explanation that why the model chosen to work with by researchers behaves in the fashion it did. Specifically, we are suppose to work on finding out how and what features are contributing in Deep Learning Approach and in other explainable classifiers. Secondly, We would also try other simpler classifiers and would compare the performance and also want to know whether we could achieve similar, close or even better performance on the same experiment using other classifiers. [-@R-base]

Project Objective:

Ground Truth and Technology Stack:

Technology Stack

The project will be built in R with usage of API's like `magick`, `opencv` for image processing and `tidyverse` packages like `dplyr` and `tidyr` for data manipulation, `ggplot2` for data visualization, `rmarkdown` and `knitr` for reproducible & automated reporting, `shiny` for interactive web applications, and `tidymodels` for inferential and predictive modeling.

Dataset

Our approach to learn would be from chest X-Rays however, as literature surveys suggest, CT-scans transfer much knowledge and thus is better suitable Deep Learning based approaches, X-Rays on the other hand are publicly available and abundant source than CT-scans and gives good results on the fronts of vision tasks. Both Deep Learning based approaches and features extraction using computer vision will be done from them. We will for reasons of abundance of data stick to X-Ray images. In particular we will use the Dataset as our Ground Truth.

Data Science Pipeline Design Overview:

We would be following a typical Data Science pipeline starting with Pre-Processing of Data, Feature Extraction and Selection and then feeding descriptors (trainable vectors) to different classifiers to train and then evaluation would be done based on predictor's results. The details are delineated in the following sections.

Feature Extraction and Selection

For the part of feature extraction, we will try texture-based descriptors. There exist several texture-based vision algorithms. We will try to combine features before training and train our model on a combined feature set. Or else we can train models on individual features, and then combine prediction results might be combined and thus one feature might only not be selected but multiple features can be selected. Literature survey tells us Local Binary Patterns shall be a good choice for texture-based descriptor. We will also try to use pretrained networks to gain texture descriptors or vision API's for the extraction part. Moreover, there are several Neural-Nets we faced in literature survey, that are carefully curated for the purpose of the COVID-19, which requires the image to directly fed to the net, and thereby auto-encodes the parameters.

Model Selection and Modelling

The problem in hand is a classification problem where we would be classifying whether an image as positive or negative w.r.t COVID-19. Here, we would like to emphasize that the model won't predict presence or absence or pneumonia, which is a result not only of COVID-19 but other kind of reasons also affect this.

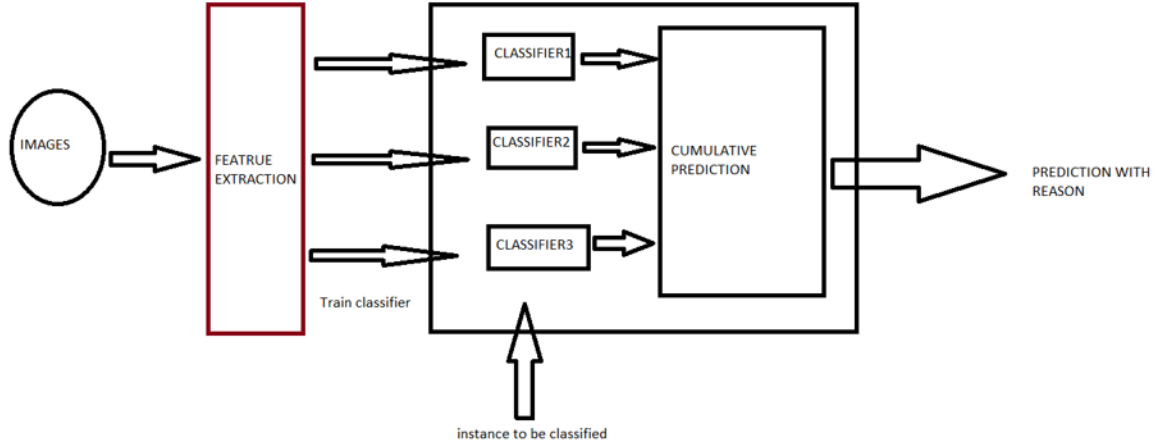
We intend to work on two kinds of algorithms:

1. One that can be trained based on features to be extracted such as:
 - Clustering Algorithms
 - k-Means
 - kNN
 - Support Vector Machines
 - Binary Classifiers
 - Decision Trees
 - Naive Bayes

Mostly because the models are intrinsically explainable.

2. Neural Network based approaches, where the model is a black box model and we will use tools like saliency maps for description.

Overview of Design



The above figure suggests the overall overview of an architecture of the system we would like to develop. We will have images as data and we would pre-process by cropping of images and extract features from them. Then after, the features or preprocessed images are fed to the classifier for training. Once trained unknown instance is supplied for classification. We will use late fusion; hence we accumulate the prediction of each of the classifier with certain confidence. The confidence shall be extracted by calculating the MAP score. The Map score can be calculated on a test set that shall be segregated from the overall dataset before the training phase begins. The ratio of the MAP score shall give the confidence contribution of each of the systems.

Evaluation and Visualization

Evaluation Methodology

Evaluation Metrics are used to calculate the performance of the model. We have different type of evaluation methods but selecting a metric is an important step in the project. Most commonly used metrics are Precision and Recall. Precision and Recall are also used with other metrics like Accuracy, F1-Score, Area under ROC curve, MAP Score. The higher the metric value the better the performance.

Accuracy: Best and mostly used metric. Easily suited for binary as well as multiclass classification problem.

$$Accuracy = \text{Number of corrected predictions} / \text{Total number of predictions}$$

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Precision: It is a best choice when we want to be very sure of our prediction.

$$Precision = TP / TP + FP$$

Recall: Captures as many positives as possible

$$Recall = TP / TP + FN$$

F₁-Score: It is a harmonic mean between Precision and Recall.

$$F_1 \text{ Score} = 2 * (Precision * Recall) / (Precision + Recall)$$

MAP score: Quantifies how good our model at performing the query. First we calculate the average of the precision for each query and then the mean of all these AP scores.

Q – No.of queries in the set

AveP(q) – Average precision for a each query q

$$\sum_{q=1}^Q AveP(q)/Q$$

AUC & ROC: Indicates how well the probabilities from the positive classes are separated from the negative classes. Mostly used to check or visualize the performance of the multi-class classification models. ROC is the probability curve and AUC represents degree or measure of separability.

Visualization

Visualization is a computer generated image using a computer representation of data as primary source and a human as its primary targets. It's an abstract of information. Box Plot easily display data and we can see outliers as well. It graphically displays the data in five statistics Minimum quartile, 25th or lower quartile, 50th or median quartile, 75th or upper quartile and the maximum quartile, which summarizes the distribution of dataset.

Time Plan:

2 Meetings per week. Tuesday and Friday at 17:30 Sharp!

Individual topics have been assigned in a fashion where each pair or individual would research and implement the functionality in the given time period. Any hindrances and issues would be discussed and will be tried at best to resolve within a weeks time. As the components of the pipeline depend on each other, research of the methodology would be done in parallel whereas the implementation part in an incremental approach.

Table 1: Time-Line

Sr.Nr.	Work_Slot	Time_Period	Tasks	Member_Name
1	1	22.05.2020 - 05.06.2020	Data Pre-processing & Feature Engineering	Jalaj, Subhankar
2	2	05.06.2020 - 19.06.2020	Model Selection and Modelling	Subhajit, Shivam
3	3	19.06.2020 - 03.07.2020	Visualization and Evaluation	Roshmitha

Team:

Jalaj, Vora *M.Sc. Digital Engineering*

Shivam, Singh *M.Sc. Digital Engineering*

Subhankar, Patra *M.Sc. Data and Knowledge Engineering*

Subhajit, Mondal *M.Sc. Data and Knowledge Engineering*

Roshmitha, Thummala *M.Sc. Data and Knowledge Engineering*

Supervised by: M.Sc. Uli Niemann

References