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-CoV-2 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 (Zhao 2012) which usually takes upto 4-6 hours to reproduce the results which is not very optimal way to move forward as the rate of COVID-19 patients registeration grows exponentially. With this problem in scientific community, it motivated to aim of Data Sciene be brought to help. So, this leaded to a possibility of building classifiers which can diagonise patients as COVID-19 negative or positive based on their respective X-Ray images [citation of some similar experiments]. As this approach is can be less time and resource consuming and we hoped to achieve a competible performance compared to RT-PCR kits. Also in addition to a good prediction, we needed reasons that could justify what could be the features that are responsible in the diagonistic process(George 2012; Fenner 2012).

With this idea and motivation in hand, our work tries to experiment in building classifier with CXR (Chest X-Rays) as Ground Truth predicts whether an X-Ray image belongs to COVID-19 negative or positive. Along with, we try to come up with features that contributes to the detection of an image and also with an explaination delineating why was such a behaviour observed.

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.

Project Objective:

From the motivation to help flattening the curve of COVID-19 patient, we aim to build classifiers which achieves a base Accuracy, F_1-Score and AUC of 85% using X-Rays. Secondly, we aim to experiment on the classifiers which are intrinsically explainable as well as which aren't. With this experiments we would try to come up with set of features that would be responsible for the diagonistic process by Classifier as well as with an explaination of why such a set of features were significant in doing so.

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

In order to train the classifier, CT-Scan data would be gold-standard for us (Zhao 2012) and also the experiment by (**???**) suggest pretty good results in terms of accuracy and F1-Score. However, due to less in quantity, we would like to use X-rays as our dataset to train and test. Though, it won't be that competible in terms of quality w.r.t CT-Scans but as it is availbale in abundance for COVID-19 positives and negatives and as per (**???**) it has proved to be sufficient and comparable to diagonise w.r.t CT-Scans.

Our dataset consists of 1500 X-Ray images of so and so size. We have so and so COVID-19 positives and so and so negatives. We would like shuffle, batch and repeat to train 66% of data and the rest to test while building the model. [citing all the links from which we have collected our dataset]

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 the Dataset, Feature Extraction and Selection and then feeding descriptors (Trainable Vectors) to different classifiers to train and test and then finally 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:

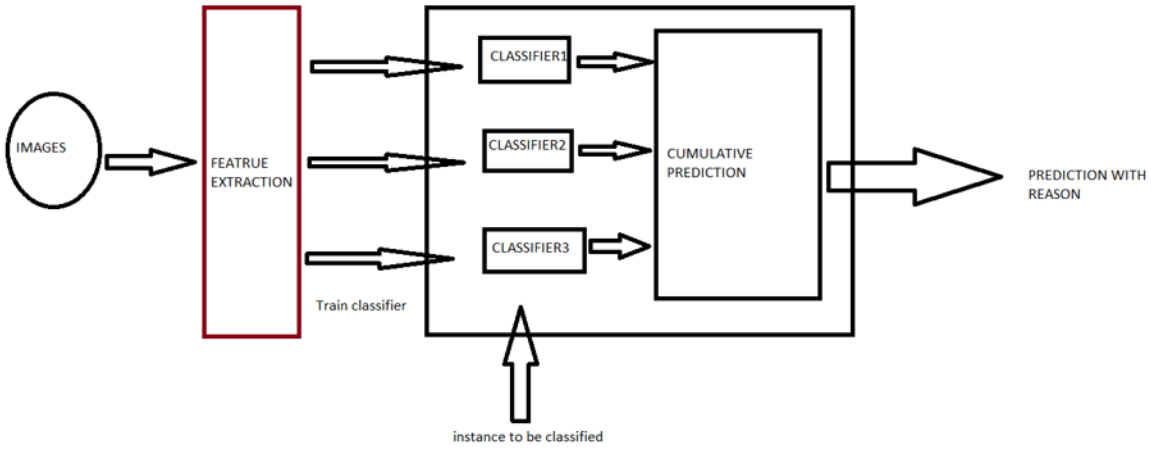
1. One that can be trained based on features to be extracted such as:

- 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-proccess 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.

$$Accuracy = Number\ of\ corrected\ predictions\ /\ Total\ number\ of\ predictions$$

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

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

$$Recall = TP/TP + FN$$

F_1-Score: It is a harmonic mean between Precision and Recall.

$$F_1 - Score = 2 * (Precision * Recall)/(Precision + Recall)$$

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, 25^{th} or lower quartile, 50^{th} or median quartile, 75^{th} or upper quartile and the maximum quartile, which summarizes the distribution of dataset. Our plan is to visualize our output in Box-plot diagram, for instance, we are gonna display probability of having COVID in quartiles. X-ray's without COVID as minimum quartile, X-ray's which has very less probability of getting COVID as the lower quartile and continues till X-ray's which has very strong probability of getting COVID as Maximum quartile.

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.

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

```
## [1] J. P. Cohen, P. Morrison, and L. Dao. "COVID-19 image data
## collection". In: _arXiv 2003.11597_ (2020). <URL:
## https://github.com/ieee8023/covid-chestxray-dataset>.
##
## [2] D. S. Kermany, M. Goldbaum, W. Cai, et al. "Identifying Medical
## Diagnoses and Treatable Diseases by Image-Based Deep Learning". In:
## _Cell_ 172.5 (2018), pp. 1122 - 1131.e9. ISSN: 0092-8674. DOI:
## https://doi.org/10.1016/j.cell.2018.02.010. <URL:
## http://www.sciencedirect.com/science/article/pii/S0092867418301545>.
##
## [3] C. D. Manning, P. Raghavan, and H. Schütze. _Introduction to
## Information Retrieval_. USA: Cambridge University Press, 2008. ISBN:
## 0521865719.
##
## [4] R Core Team. _R: A Language and Environment for Statistical
## Computing_. R Foundation for Statistical Computing. Vienna, Austria,
## 2019. <URL: https://www.R-project.org>.
##
## [5] J. Zhao, Y. Zhang, X. He, et al. "COVID-CT-Dataset: A CT Scan
## Dataset about COVID-19". In: _ArXiv_ abs/2003.13865 (2020).
```

World Health Organization (WHO). WHO Statement regarding cluster of pneumonia cases in Wuhan, China 2020 [14 Januray 2020]. Available from: https://www.who.int/china/news/detail/09-01-2020-who-statementregarding-cluster-of-pneumonia-cases-in-wuhan-china

World Health Organization (WHO). Novel Coronavirus – China 2020 [14 January 2020]. Available from: https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/

Fenner, Martin. 2012. “One-Click Science Marketing.” *Nature Materials* 11 (4): 261–63. https://doi.org/10.1038/nmat3283.

George, Martin. 2012. “One-Click Science Marketing.” *Nature Materials* 11 (4): 261–63. https://doi.org/10.1038/nmat3283.

Zhao, Martin. 2012. “One-Click Science Marketing.” *Nature Materials* 11 (4): 261–63. https://doi.org/10.1038/nmat3283.