Final Project - Mid-stage Report

Yunhan Li
Stevens Institute of Technology
Jersey City, New Jersey, USA
yli330@stevens.edu

Yicong Pan
Stevens Institute of Technology
Jersey City, New Jersey, USA
ypan28@stevens.edu

Chengchen Zhao
Stevens Institute of Technology
Hoboken, New Jersey, USA
czhao36@stevens.edu

***Abstract*—The COVID-19 (coronavirus) is an ongoing pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The virus was first identified in mid-December 2019 in the Hubei province of Wuhan, China and by now has spread throughout the planet with more than 75.5 million confirmed cases and more than 1.67 million deaths. With limited number of COVID-19 test kits available in medical facilities, it is important to develop and implement an automatic detection system as an alternative diagnosis option for COVID-19 detection that can used on a commercial scale. Chest X-ray is the first imaging technique that plays an important role in the diagnosis of COVID-19 disease. In this paper, we will use three machine learning algorithms to analyze the pictures, namely Random Forest, SVM, Logistic Regression, and use the CNN algorithm in deep learning for analysis, and finally analyze the accuracy of the prediction by comparison.**

I. Introduction

The coronavirus (COVID-19) pandemic has affected billions of people since the time of its emergence from Wuhan, China in December 2019 [2]. The virus led to an outbreak at a very fast rate. A lot of research was conducted to identify the type of virus that caused COVID-19 disease and it was concluded that it belonged to a huge family of respiratory viruses that can cause diseases such as Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV). The new SARS-CoV-2 virus can develop viral pneumonia. The population has witnessed a very high mortality rate in some states. The death toll around the world is increasing day by day. Therefore, it is necessary to develop an accurate, fast and cost-effective tool for diagnosis of viral pneumonia. This will serve as the initial step for taking further preventive measures like isolation, contact tracing and treatment for stopping the outbreak.

One popular method to detect the virus is viral nucleic acid detection using real-time polymerase chain reaction, also known as RT-PCR test [3]. This test is very sensitive and has several limitations. For example, it cannot detect coronavirus developed before taking DNA sequence samples. Moreover, it takes 2-3 days to produce the result and requires many arrangements, public space. Many countries are not able to provide these conditions for testing of thousands of patients. Hence, continuing this method might slow down the process of controlling the pandemic.

In this scenario, medical imaging can prove to be a vital technique for diagnosis. Chest radiography plays an important role in the early diagnosis of pneumonia. It is commonly used because of its fast-imaging speed and low cost. However accurate and fast diagnosis of a X-Ray image is only possible with the help of expert knowledge [1].

We will use x-ray scans of human lungs collected. Since the Covid-19 virus acts on a person's lungs, we could theoretically distinguish whether a patient has Covid-19 based on a chest X-ray image of the patient. At the same time, our project also needs to consider common viral pneumonia, which can also cause abnormal images of people's lungs. Therefore, our final project aimed to classify chest X-rays into 3 classes: normal, Covid and viral pneumonia.First we will process the obtained image data to visualize the image data and generate training set,Then, the processed image data is judged by the selected three algorithms, and finally the accuracy score is obtained by comparing with the testing data set.

II. Related Work

In recent years, more and more machine learning methods have been used to solve COVID-19 detection problems, and neural network are the most dominant methods in solving the COVID-19 detection problem.

In 'A Deep Learning Approach for the Detection of COVID-19 from Chest X-Ray images using Convolutional Neural Networks' paper, they have proposed a deep convolutional neural network designed specifically for the detection of COVID-19 cases by implementing computer vision and image analysis on Chest X-Ray images gathered from five open access data repositories. they conducted experiments on COVID19 identification and compared it with four models: COVID-Net, ResNet18, ResNet and MobileNet-v2, and comparison experiment bettheyen the accuracy and loss generated on the validation set by each model. The experimental results show that the proposed model had the best performance accuracy on the validation set. Further, they investigated and applied different model parameters in order to gain deeper insights on the Chest X-Ray features critical for classifying Covid and non-Covid patients which can aid clinicians in improved screening as theyll as improve trust and transparency [1].

In our work, we want to use different algorithms to solve the problem, in order to see which algorithm is the best method to detect COVID-19 disease.

III. Our Solution

A. Description of Dataset

The dataset is from:

https:'//www.kaggle.com/pranavraikokte/covid19-image-dataset'.

We have a total of 251 images for training (70 viral pneumonia, 70 normal, 111 Covid) and 66 images for testing (20 viral pneumonia, 20 normal, 26 Covid).

Our dataset is actually those chest X-ray images. In order for the computer to be able to process these data, we need to convert them into two-dimensional matrices, that is, matrices of their grayscale values. The number of features for each data is the product of the rows and columns of the matrix.

These images are of different sizes, therefore we need to normalize them.

B. *Machine Learning Algorithms*

We basically use three Machine Learning models to solve this problem so far. Logistic Regression, SVM and Random Forest.

1. Logistic regression is one of the most commonly used machine learning algorithms. Logistic regression is a discriminative model with a lot of model regularization methods (L0, L1, L2, etc.), and we don't have to worry about whether our data features are related as we do with Naive Bayes. Compared to decision trees, SVMs, we can get decent probabilistic explanations, and we can even easily update the model with new data (using an online gradient descent algorithm). Therefore we train Logistic Regression model first to see how it works. But we think that Logistic Regression does not work well, because a. Logistic regression does not perform very well when the feature space is large; b. It is easy to under-fit, and the general accuracy is not very high; c. Cannot handle a large number of multi-class features or variables well.

2. SVM maps vectors into a higher-dimensional space, in which a maximum-margin hyperplane is established. Two parallel hyperplanes are built on both sides of the hyperplane separating the data. The separating hyperplane maximizes the distance between the two parallel hyperplanes. It is assumed that the larger the distance or gap between parallel hyperplanes, the smaller the total error of the classifier. We think SVM will work better than Logistic regression handling our dataset since a. Use the kernel function to map to a high-dimensional space; b. Use the kernel function to solve nonlinear classification; c. The idea of classification is very simple, that is, to maximize the interval between the sample and the decision surface. But a. SVM algorithm is difficult to implement for large-scale training samples; b. SVM is difficult to solve the multi-classification problem; c. SVM is Sensitive to missing data and the choice of parameters and kernel functions. We need to adjust the parameters and kernel functions many times to get the best classification result.

3. Random forest: a. can handle very high dimensional data (many features) without feature selection (because the subset of features is randomly selected); b. and after training, it can give which features are more important; c. The random forest model has strong generalization ability; d. The training speed is fast, because each decision tree is independent of each other during the training process, so it is easy to do a parallel method; e. For imbalanced data sets, the error can be balanced; f. If most of the features are missing, the accuracy can still be maintained. Since the number of features in the output data set is quite large ($over 10 \wedge 6$), we believe that random forests can perform well.

4. CNN: Next step in our time line form.

C. *Implementation Details*

1. Environment setup:

For our implementation, we use the following packages:

a. cv2 — reading images and data normalization

b. numpy — data argument and pre-processing

c. matplotlib.pyplot — visualization

d. sklearn — model training and model evaluation

2. Data argument part

As we stated in the proposal, we don't get many chest X-ray images. We have a total of 251 images for training and 66 images for testing. This is a relatively small dataset. Therefore we apply the following two methods to enlarge our dataset:

a.mirror: flip the image in horizontal direction.

b.crop: randomly extract a sub-image from the whole image and use it as a training sample.

eg. For an image of size $256 * 256$, by cropping the image to the size of $228 * 228$, we can enlarge the data set by $(256 - 227) * (256 - 227) = 841 times$. with data argument, we enlarge our training set from 251 images to 500 images and our testing set from 66 images to 120 images.

3. Data pre-processing part

At data pre-processing part, We basically do 2 things: Normalization and Using PCA to reduce the dimension of features.

Using cv2 to read each image, we can get a 2-D matrix of that image i.e. its gray value. The size of each matrix is the size of that image. After reading some images, we find the sizes of all the images are different. For example one is (1400, 1648) and the other is (2953, 3604). We can not use these raw data to train our models. Therefore, we need to convert them into a same size (n, m). And the number of features of each X is $n * m$.

Using cv2 to read each image, we can get a 2D matrix of that image, its grayscale value. The size of each matrix is the size of that image. After reading some images, we found that all images are different sizes. For example, one is (1400, 1648) and the other is (2953, 3604). We cannot use this raw data to train our model directly. Therefore, we need to convert them to the same size (n, m). And the number of features for each data is $n * m$.

The choice of the converted size is important because usually the converted size is smaller than the original size. That is, some feature information will be lost during the normalization process. If the conversion size is too small, it may result in low precision. And if the transform size is too large, the dimensionality of the data is very high (more than $over 10 \wedge 6$), which may cause the training time to be too long and the model to fail to converge.

We initially set the size to (1000, 1000). Then the number of features is $over 10 \wedge 6$, which retains most of the feature information of the dataset, but is still high-dimensional. To reduce the dimensionality, we adopt the PCA method. The final data has 100 features.

4. Logistic Regression model

We use sklearn directly to implement the Logistic Regression model. If we use the data without PCA, it takes about 6 minutes to train the model. The hyper parameter I set is Logistic Regression(penalty="l2", C=1.0, random-state=None, solver="lbfgs", max-iter=3000, multi-class='ovr', verbose=0), the accuracy of the training set can reach 100%, and the accuracy of the test set is 86.36%, if we use PCA data, the accuracy of the training set can still reach 100%, but the accuracy of the test set will drop to 58.19%. However, the training time is very short, about 10 seconds.

The results still need to be improved.

5. SVM model Logistic Regression is not suitable enough for high dimensional data, then we also train SVM model. If we use the data without PCA, it takes about 3 minutes to train the model, the accuracy of the training set is 100%, and the accuracy of the test set is 87.88%, if we use PCA data, the accuracy of the training set can still reach 100%, but the accuracy of the test set will drop to 65.38The results still need to be improved.

6. Random Forest model

The random forest model is good at handling high-dimensional data, and it takes very little time compared to other machine learning algorithms. Even if we use data without PCA, it only takes 2 minutes to train the model. The hyper parameter I set is Random Forest Classifier(n-estimators=10, max-features=1000, max-depth=None, min-samples-split=2, bootstrap=True), the training set accuracy is 100%, the test set accuracy is 92.30%, if we use PCA data, training The accuracy on the set can still reach 98.41%, while the accuracy on the test set is 72.73%.

The results still need to be improved.

7. CNN model

To be added in next step...

IV. Comparison

Comparison of three Machine Learning models: By constructing and fitting the models separately, it is easy to identify their own strengths and weaknesses when compared to other models.Summarizing these strengths and weaknesses can be very helpful in choosing which model to use next time for a particular data prediction.

For logistic regression, it is a discriminative model, accompanied by many methods of model regularization (L0, L1, L2, etc.), and you don't have to worry about the relevance of your features as you do with plain Bayes. You also get a nice probabilistic interpretation compared to decision trees, SVM, and you can even easily update the model with new data (using online gradient descent algorithm - online gradient descent). If you need a probabilistic architecture (for example, to simply adjust the classification threshold, to specify uncertainty, or to get confidence intervals), or if you want to integrate more training data quickly into the model later, then do not hesitate to choose it.

Advantages of logistic regression, the most basic and simple model: simplicity of implementation and wide application to industrial problems. very low computational effort for classification, high speed and low storage resources. Convenient observation of sample probability scores. multicollinearity is not a problem for logistic regression, which can be combined with L2 regularization to solve the problem. computationally inexpensive and easy to understand and implement.

Disadvantages of logistic regression: Low prediction accuracy. Difficulty in representing nonlinear data or highly complex data well, and difficulty in modeling regression when there are correlation polynomials between data features. Tendency to overfit because the model is very simple (the use of regularized models can counteract this effect). poor separation of signal and noise, requiring the removal of irrelevant features before use. Not understanding the feature interactions in the dataset. Logistic regression may be unstable if the dataset has redundant features. It can only handle two classification problems (softmax derived on this basis can be used for multi-classification) and must be "linearly separable". for non-linear features, transformation is required.

logistic regression application areas: Used in the field of dichotomous classification to derive probability values, applicable to the field of ranking based on classification probabilities, such as search ranking. The extended softmax of logistic regression can be applied to multi-classification domains, such as handwriting recognition, etc. Credit evaluation Measuring the success of marketing Predicting the revenue of a particular product Whether an earthquake will occur on a specific day

For SVM, Support vector machines, an enduring algorithm with high accuracy, provide a good theoretical guarantee to avoid overfitting, and work well even if the data is linearly indistinguishable in the original feature space, given a suitable kernel function. It is particularly popular in text classification problems that are often ultra-high-dimensional. Unfortunately, memory consumption is large, difficult to interpret, and running and tuning the reference is a bit annoying, while random forest just avoids these disadvantages and is more practical.

Advantages of SVM:

SVMs can use kernel functions to map to higher dimensional spaces Nonlinear mapping is the theoretical basis of SVM method, and SVM can use kernel function to solve the classification of nonlinear problems. The principle of classification is very simple, easy to understand and conceptualize, and the classification effect is good. It only requires training to obtain the support vector, which plays an intermediate role in the SVM classification decision. SVM is a novel small sample learning method with a solid theoretical foundation. It is different from existing statistical methods because it basically does not involve probability measures and the law of large numbers, etc. In essence, it avoids the traditional process from induction to deduction and achieves efficient "transductive inference" from training samples to forecast samples, which greatly simplifies the usual problems of classification and regression. The final decision function of SVM is determined by only a small number of support vectors, and the complexity of the computation depends on the number of support vectors rather than the dimensionality of the sample space, which in a sense avoids the ”dimensionality disaster”. A small number of support vectors determines the final result, which not only helps us to catch the key samples and ”eliminate” a large number of redundant samples, but also predestines the method to be not only simple, but also has a good ”robustness”. This ”robustness” is mainly reflected in the following three points:

1) The addition and deletion of non-support vector samples have no effect on the model;

2) The support vector sample set has a certain level of robustness;

3) In some successful applications, the SVM method is insensitive to the selection of the kernel

Disadvantages of SVM:

SVM algorithm is difficult to implement for large training samples.Since SVM solves the support vector with the help of quadratic programming, and solving quadratic programming will involve the computation of a matrix of order m (m is the number of samples), the storage and computation of this matrix will consume a lot of machine memory and computing time when the number of m is large. The main improvements for the above problem are J. Platt’s SMO algorithm, T. Joachims’ SVM, C.J.C. Burges et al.’s PCGC, Xue-Gong Zhang’s CSVM, and O.L. Mangasarian et al.’s SOR algorithm

Difficulties in solving multi-classification problems with SVMs.The classical support vector machine algorithms only give algorithms for two-class classification, while in practical applications of data mining, multi-class classification problems are generally to be solved. It can be solved by the combination of multiple two-class support vector machines. There are mainly one-to-many combinatorial models, one-to-one combinatorial models and SVM decision trees; then it is solved by constructing combinations of multiple classifiers. The main principle is to overcome the inherent shortcomings of SVM and combine the advantages of other algorithms to solve the classification accuracy of multi-class problems. For example, it is combined with rough set theory to form a combined classifier with complementary advantages for multi-class problems.

sensitive to missing data.There are also techniques for the choice of kernels (libsvm comes with four kernel functions: linear kernel, polynomial kernel, RBF, and sigmoid kernel).

First, if the number of samples is smaller than the number of features, then there is no need to choose a nonlinear kernel; simply using a linear kernel will do. second, if the number of samples is greater than the number of features, then a nonlinear kernel can be used to map the samples to higher dimensions, which generally yields better results. Third, if the number of samples and the number of features are equal, a nonlinear kernel can be used in this case, and the principle is the same as the second one. For the first case, it is also possible to first reduce the dimensionality of the data and then use a nonlinear kernel, which is also a method.

Application areas of SVM:

Text classification, image recognition (mainly in the field of binary classification, after all, conventional SVM can only solve binary classification problems)

For random forest(decision tree), One of the great advantages of decision trees is that they are easy to interpret. It can handle interactions between features without stress and is non-parametric, so you don’t have to worry about outliers or whether the data is linearly separable (for example, a decision tree can easily handle the case where category A is at the end of some feature dimension x, category B is in the middle, and then category A appears again at the front of feature dimension x). One of its drawbacks is that it does not support online learning, and so the decision tree needs to be rebuilt in its entirety when new samples arrive. Another disadvantage is that it is prone to overfitting, but this is the entry point for integrated methods such as random forest RF (or boosted treeboosted tree). Also, random forests are often the winner of many classification problems (usually a little better than support vector machines), they are fast and adjustable, and you don’t have to worry about tuning a bunch of parameters like support vector machines, so they have always been popular in the past.

Advantages:

Decision trees are easy to understand and interpret, can be analyzed visually, and rules can be easily extracted. can handle both nominal and numerical data. being more suitable for handling samples with missing attributes. the ability to handle uncorrelated features. relatively fast operation when testing datasets. the ability to produce feasible and effective results for large data sources in a relatively short period of time

Disadvantages:

Decision tree models are prone to overfitting, but random forests can reduce overfitting to a large extent. Tend to ignore the interconnection of attributes in the dataset. For those data with inconsistent sample sizes in each category, different decision criteria bring about different attribute selection preferences when performing attribute partitioning in decision trees; the information gain criterion has a preference for more desirable attributes (typically represented by the ID3 algorithm), while the gain rate criterion (CART) has a preference for less desirable attributes, but CART no longer simply (instead of directly using the gain rate due diligence division, CART uses a heuristic rule) (whenever information gain is used, it has this drawback, e.g., RF).

V. Future Directions

We will continue to complete several other models to fit and predict the data based on our time line form. Finally, we will do an overall comparison of the prediction results of the models, as well as their respective characteristics and advantages and disadvantages.

VI. Conclusion

Although it is too early to do a summary for our current process, by the few models we have completed now, they satisfy our function to some extent, but their performance does not satisfy us completely. We will continue to complete the remaining more complex models and observe their predictive performance. However, before starting the prediction of the neural network model, after a brief comparison of Assignment 4, I guess the prediction performance of the neural network will be better than all the three previous models.

Different algorithms have different characteristics and strengths and weaknesses, and some models may be able to compensate for each other. We may consider whether to try to optimize the predictive performance of the data by fusion of results after we have completed the construction and fitting of individual models.

References

[1] Saxena, A. and Singh, S.P., 2022. A Deep Learning Approach for the Detection of COVID-19 from Chest X-Ray Images using Convolutional Neural Networks. arXiv preprint arXiv:2201.09952.

[2] Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, and et al. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infectious Disease Modelling, 5:256-263, 2020

[3] V. M. Corman et al., “Detection of 2019 novel coronavirus (2019-ncov) by real-time rt-pcr,” Eurosurveillance, vol. 25, no. 3, p. 2000045, 2020.