Predicting the COVID-19 Test Result

YUXIN YAO Student Number:19022267 2nd Year MEng Mathematical Computation
Internship Supervisor: Kevin Bryson
June 30th 2020 - August 25th 2020

Abstract

In this project, we created several models used to predict the result of individuals' COVID-19 tests with their clinical data. Initially, we used a simple model which predicts by checking the existence of some symptoms. After the improvements throughout the project, the model finally used was a random forest model with feature selection, pipeline, grid search and other optimization methods. The final result of the project reached the area under precision-recall curve 0.0977 and area under the receiver operating characteristic curve 0.7141.

I. Introduction

Nowadays, the coronavirus gave rise to millions infections and deaths. This internship is about a competition that requires the competitors to provide a model predicting if the person will have positive result in COVID-19 test, which will be accessed and ranked by the host – Dream Challenge [1]. The training data could only be accessed through UW EHR OMOP repository when the host runs the codes. During the internship, we focused on improving the model to score a higher value of area under the precision recall curve (AUPR) and area under receiver operating characteristic curve (AUROC). Besides, through this internship I first learnt and practised Machine Learning, including distinguishing the difference between models, choosing the most suitable model according to the question and data, improving the model with various methods and improving the time complexity.

II. Structure

A. Symptom-counting Method

In the beginning, the training of a model using data in the UW machine was not allowed. The specific symptoms and their thresholds are important to predict the COVID-19 test result. Several papers and official websites were used in this stage to select effective symptoms (conditions and measurement of a person like blood pressure or temperature.) that indicates positive tests in coronavirus. Papers [2] [3] indicated that the fever, cough, fatigue, lost of taste, age over 60 and etc. were statistically relative with coronavirus.

B. Linear regression Model

The training was allowed after an database update. The linear regression model was used with the default hyperparameters as a first trail. The data was cleaned before processed by model. All condition occurrences and measurement occurrences recorded with null values were deleted, and duplicated person ids were removed (unique key value for every person in data set) as well. Same hand-selected features used in symptom-counting method were selected for training.

C. Logistic regression Model

The logistic regression model was implemented after the linear regression one. All condition and measurement features were used in this model. Dependent features were pruned away after the data was loaded and cleaned. Moreover, vectorization was used when the loaded data frames(data type from pandas library) were processed to the format. Besides, the data type of the input data was converted to least space-occupying form. In-place operation prevented the duplication from storing the same data. Feature selection using χ^2 coefficient and random forest model was used. The weights for every feature were adjusted during the training. Grid search on l1 regression ratio with elastic net regression was applied. Error handling parts were added to avoid failure submission. In addition, garbage collection was employed.

D. Rondom Forest

Random Forest model was chosen, finally, all the improvement methods used in Logistic regression model were used. Pipeline built for selecting the value of the number of features that obtained the best f1-score was inserted.

III. Result Discussion

A. Symptom-counting Method

This method was only used when the host did not allow model training. The prediction on the probability was based on the number of occurrence of hand-selected features. Due to the simple method, the result for symptom-counting method was not good, with AUROC 0.4783 and AUPR 0.083.

B. Linear Regression Model

The features selected for training are same as the features used in symptom-counting method. The result of the linear regression was the worst, with AUROC 0.3696 and AUPR 0.0271, indicating the low accuracy. This is because predicting the coronavirus test is a classification problem which is not suitable for linear regression.

C. Logistic Regression Model

The logistic regression model is more appropriate than the linear regression model because it is used for classifications and it does not require a linear relationship between the input feature data and corresponding categorical data. Logistic regression model could deal with a large number of features and still performing good on accuracy and time spent on training.

Initially a default logistic regression model run with the hand-selected features and socred AUROC 0.6566 and AUPR 0.0793. According to the data published by news, the affection rate of coronavirus was about 2%, implying the database was not balanced. It lead to a biased model that poorly worked with the positive tests and predicted lots of positive cases negative. Through using weighted logistic regression model, the problem could be avoided, it automatically adjusts weights inversely proportional to class frequencies in the input data. The AUROC was 0.6604 and AUPR was 0.0817 after this improvement, means the true positive rate increased.

Grid search which automatically pick up the best-performing model on a specific indicator with varied hyperparameter was applied to increase the AUPR. It took more time for training but it gave a much better result. The elastic net grid search varied the ratio of l1 regularization and selected the the model with best f1 score. The f1 score considers precision and recall. The AUROC was 0.7375 and the AUPR was 0.0979, which was the highest result in this project.

Although using the fixed features the model could get a good score, it is important to involve more features to improve the model. All of the condition features are used in this stage, which made the data processing process taking an unacceptably long time. Only 60 minutes could be used during the training according to the rule of the Challenge. The conditions and person data were recorded in different spreadsheets. Every condition occurring on a person at a time would be recorded as a unique condition occurrence, while every person had multiple condition occurrences. But the condition occurrences were not recorded with a regular pattern. The data needed to be processed into the pattern that shew the status of every condition for every individual. The process used simply for loops before vectorization was applied, which improved the speed of processing significantly. In contrast, the former took more than 3 hours and the latter took only 15 minutes on local synthetic data.

However, training the model with more than 900 features leads to long time grid searching, while the Challenge has a time limit for training. Feature selection was important to filter out those less relevant features before training. First, the Pearson correlation coefficient, measuring the degree of linear correlation between two variables, was chosen as the indicator. Whereas the features from input data did not necessarily have a linear correlation with the categorical value. After that, the χ^2 test, which could fit more general correlations between data, was selected. After the feature selection process, the AUROC was 0.6823 and AUPR was 0.0723, indicating the auto feature selection was worse than hand selection. However, we could not select all the features by hand, or a lots of important information would be missed. The model was developed further below.

All measurement occurrence were loaded after that. There was no indicative field of the abnormality of each measurement feature in the data set(with same pattern as condition data set). The measurement value for every measurement occurrence was compared with the corresponding upper-bounded value of the normal range.

Additionally, the polynomial features were formed after an initial feature selection, which would help the linear logistic regression model consider more features with interaction information. Considering the large number of features, a second feature selection was used to filter out less relative features before training. The filter used random forest, which filter out the features with less importance. This model scores AUROC 0.6866 and AUPR 0.0892 in the end.

D. Random Forest Model

The random forest was chosen at last. It is more suitable for non-linear parameters in data, which perform well on large number of data. Besides, the random forest is less impacted by noises. In addition, the pipeline was built in order to select the best number of features used for training. This model achieved the second highest AUROC 0.7141 and AUPR 0.0977.

IV. Conclusion

In this internship, I learnt the features of different models and variants of optimization method. In the best model that using random forest, it contains the data preprocessing, including the null value deletion, re-formatting of the original data with vectorizaton. Besides, the pipeline searching for the best number of feature for training, including feature selection by using χ^2 test and random forest model, together with grid search on the hyperparameters, were applied.

References

[1] "EHR Dream Challenge." http://dreamchallenges.org/project/covid-19-ehr-dream-challenge/, accessed 2020-04-29.

[2] L. Fu, B. Wang, T. Yuan, X. Chen, Y. Ao, T. Fitzpatrick, P. Li, Y. Zhou, Y.-F. Lin, Q. Duan, G. Luo, S. Fan, Y. Lu, A. Feng, Y. Zhan, B. Liang, W. Cai, L. Zhang, X. Du, L. Li, Y. Shu, and H. Zou, "Clinical characteristics of coronavirus disease 2019 (covid-19) in china: A systematic review and meta-analysis," *The Journal of infection*, vol. 80, pp. 656–665, Jun 2020. 32283155[pmid].

[3] Y. Zoabi and N. Shomron, "Covid-19 diagnosis prediction by symptoms of tested individuals: a machine learning approach," *medRxiv*, 2020.