Application of XGBoost Algorithm in The Detection of SARS-CoV-2 Using Raman Spectroscopy

Wandan Zeng1, Qi Wang1, Zhiping Xia2, Zhiping Li2 and Han Qu2

1 Shanghai Institute of Technology, Shanghai, 201418, China
2 Military Veterinary Institute, Jilin, Changchun, 130062, China
*Corresponding author’s e-mail: zengwd@sit.edu.cn

Abstract. The novel coronavirus (SARS-CoV-2), which was first discovered in late 2019 and rapidly spread to many countries around the world in a short period of time, is highly contagious and poses a significant threat to global public safety. How to quickly and efficiently detect whether a human is infected by the novel coronavirus is a crucial step in dealing with this public health emergency. Therefore, the collected Raman spectral data are preprocessed by normalization, smoothing denoising and feature extraction in this paper. A novel coronavirus detection method based on XGBoost and Raman spectroscopy is proposed. The experiments demonstrated the feasibility and accuracy of the method in the detection of novel coronavirus with an accuracy of 93.548%.

1. Introduction
COVID-19 is a novel coronavirus of the genus β, which is a single-stranded RNA virus that causes respiratory infections in humans [1]. Currently, the novel coronavirus nucleic acid detection can be used as a diagnostic basis [2]. However, due to the rapid development of the epidemic, more rapid and effective methods are needed to diagnose patients with varying degrees of clinical symptoms and suspected patients. The novel coronavirus detection method on the market is mainly PCR [3,4], but this method is time-consuming, expensive and has a high false-negative rate. Raman spectroscopy reflects the internal vibration and rotation energy level of molecules [5], is a material fingerprint, can be used to identify the presence of functional groups in molecules, with non-destructive, rapid, accurate and other characteristics, is a powerful tool for the identification of material composition. Raman spectroscopy method combined with machine learning algorithm for identification and classification is a commonly used method in spectral analysis. The use of Raman spectroscopy in combination with computer algorithms for identification and classification can shorten the pathogen detection cycle and greatly reduce the false positive rate of manual identification of Raman peaks. Saranjam Khan et al. [6] proposed Raman spectroscopy in combination with support vector machines (SVMs) for the classification of suspected dengue human blood serum and have achieved about 85% diagnostic accuracy, 90% precision, 73% Sensitivity and 93% specificity. Extreme Gradient Boosting (XGBoost) [7] is an Ensemble Learning method. Gertz et al. [8] used the XGBoost algorithm to evaluate sensor data and health records to predict "sick" and "healthy" cows. The method used shows great potential for the development of automatic detection tools in the future, which can continuously evaluate sports-related diseases.

This paper experimentally demonstrates the reliability of the XGBoost algorithm combined with Raman spectroscopy in the detection of novel coronaviruses. First, Raman spectroscopic data from
two populations collected using Raman spectroscopy. The spectral were normalized using Min-Max in the spectral pre-processing stage, and then the spectral data were de-noised using the Savitzky-Golay algorithm [9,10] in this experiment. After the pre-processing, the sample data are subjected to feature selection using Support Vector Machine Recursive Feature Elimination (SVM-RFE) [11]. In the model building stage, this paper uses grid search to adjust the global variables to find the optimal parameters. In the model evaluation phase, three evaluation indicators are used in this paper to evaluate the model.

2. Data collection and preprocessing

2.1. Experimental sample collection
The source of this dataset is collected by Yin Gang et al. [12] The experimental setup for the collection consisted of a Volume Phase holographic (VPH) spectrometer, a deep-cooled CCD camera and a Raman probe and laser. One hundred fifty-seven serum samples were collected from 53 diagnosed patient, 54 suspected cases, and 50 healthy controls. Corresponding 309 Raman spectral data were used through the spectrometer, including 150 Raman spectral data from healthy people and 159 Raman spectral data from people infected with SARS-CoV-2, with a more even distribution of data. Figure 1 shows the raw Raman spectral data.

Figure 1. Raw Raman spectral

2.2. Data normalization
For the collected Raman spectral data, the intensity differentiation corresponding to different Raman offsets is relatively large. In this paper, normalization of the extracted feature vectors is used to ensure that each feature is treated equally by the classifier, allowing the data to be processed consistently. This paper uses Min-Max normalization:

\[ x_{\text{normalization}} = \frac{x - \text{Min}}{\text{Max} - \text{Min}} \]

where Max is the maximum value of the sample data, Min is the minimum value of the sample data, and \( x_{\text{normalization}} \) is the normalized sample value. The Raman spectral of the normalized processed samples are shown in Fig2.

2.3. Savitzky-Golay smooth denoising
When Raman spectrometer collects Raman spectral data, it will be affected by many factors such as the light of the collection environment and the purity of the sample itself. In this paper, Savitzky-Golay filtering algorithm is one of the commonly used denoising methods in Raman spectroscopy, and Savitzky-Golay denoising is also used.
Figure 2. Raman spectral after Min-Max normalization
The window width chosen for this paper is 35 and the polynomial order is 2. Fig.3 shows the Raman spectral image that has been smoothed and denoised by Savitzky-Golay, and it can be seen that some of the burrs on the Raman spectral have been smoothed to some extent.

Figure 3. Raman spectral after Savitzky-Golay smooth denoising

2.4. SVM-RFE feature selection
The SVM-RFE algorithm is a Wrapper method of SVM based on heuristic search strategy. SVM-RFE is also called a support vector machine-based recursive feature elimination algorithm. The SVM-RFE algorithm flow is shown in the table 1.

| SVM-RFE algorithm |
|-------------------|
| **Input:** Training data set $E(n$ samples, $m$ features), class label($n,1$). |
| **Step 1:** Initialize the current feature set $E_{\text{now}}$ as the original data set, the optimal feature set $E_{\text{best}}$ is empty, and the classification accuracy rate of the optimal feature subset $S_{\text{best}}$ is 0. |
| **Step 2:** Set the ratio of the number of features deleted each time $p$ ($0<p<1$). |
| **Step 3:** Repeat the following steps until the current feature set $E_{\text{now}}$ is empty: |
| - The SVM model is established by $E_{\text{now}}$, and the correct rate evaluation value $S_{\text{now}}$ is obtained; |
| - Sort the features in $E_{\text{now}}$ in descending order by the absolute value of feature weight $|w|$. |
| - Delete the lower $p\%$ features in the current subset $E_{\text{now}}$; |
| - If the correct rate $S_{\text{now}}$ of the current feature subset $E_{\text{now}}$ is greater than $S_{\text{best}}$: $E_{\text{best}} = E_{\text{now}}$. |
| **Output:** the best feature subset $E_{\text{best}}$. |

SVM-RFE feature selection is performed on 309 sets of Raman spectroscopy data after normalization and smoothing. The final number of selected features is 20, and the features are obtained at 407 cm$^{-1}$, 410 cm$^{-1}$, 412 cm$^{-1}$, 415 cm$^{-1}$, 442 cm$^{-1}$, 444 cm$^{-1}$, 447 cm$^{-1}$, 449 cm$^{-1}$, 452 cm$^{-1}$, 801 cm$^{-1}$, 804 cm$^{-1}$, 806 cm$^{-1}$, 808 cm$^{-1}$, 810 cm$^{-1}$, 835 cm$^{-1}$, 842 cm$^{-1}$, 1001 cm$^{-1}$, 1004 cm$^{-1}$, 1659 cm$^{-1}$ positions.
3. Experiment and Discussion

3.1. XGBoost Theory

XGBoost is an open-source machine learning project developed by Tianqi Chen et al. [13]. It is an integrated machine learning algorithm based on decision trees and framed by Gradient Boost. The objective function of XGBoost contains a regular penalty term as well as a loss function term to combine them for an overall optimal solution, which weighs the reduction of the loss function against the complexity of the model. The addition of the regularization term can reduce the variance of the model, making the model more simple to learn through the training set to prevent overfitting. The objective function is:

\[ L(\phi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{x} \Omega(f_x) \]  

Where \( \Omega(f) = \gamma T + \frac{1}{2} \lambda \| \omega \|^2 \)  \[ (3) \]

The first part of the right-hand side of Eq. (2) measures the loss function between the predicted and true values, and the second part represents the penalty term (regular term) on the model complexity. In the penalty term \( \gamma \) and \( \lambda \) represent the penalty coefficients, \( T \) represents the number of leaf nodes of a given tree, and \( \| \omega \|^2 \) represents the square of the output score on each tree leaf node (equivalent to the L2 regular). From the definition of the objective function, it can be seen that XGBoost takes into account the complexity of the model, the number of leaf nodes per tree, as well as the output score of each tree leaf node worth the squared sum.

Then the objective function is optimized using forward stepwise algorithm, and it is assumed to be the predicted value of the \( t \)-th sample in the \( t \)-th iteration (\( t \)-th tree). Then the loss function for the \( t \)-th time is:

\[ L^{(t)} = \sum_{i} l(\hat{y}_i, y_i^{(t-1)}) + \Omega(f_t) \]  

Do second-order Taylor expansion of equation (4), obtained equation (5).

\[ L(t) \approx \sum_{i} [l(\hat{y}_i, y_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \]  

In the generation of decision trees, we use metrics such as ID3, C4.5, and Gini index to select optimal split features and cut points, and XGBoost also defines metrics for feature selection and cut point selection.

\[ L_{\text{split}} = \frac{1}{2} \left[ \left( \sum_{i \in I_L} g_i \right)^2 + \left( \sum_{i \in I_R} g_i \right)^2 + \left( \sum_{i \in I_L} h_i \right)^2 + \left( \sum_{i \in I_R} h_i \right)^2 \right] - \gamma \]  

3.2. Model building and training

The experimental environment is configured as: Windows10 64-bit operating system; Development environment: JetBrains PyCharm 2019.1.2; Hardware configuration: memory 16.0GB; Processor: Intel(R) Core(TM) I5 9300H, CPU @ 2.40 GHz. Figure 4 shows the SARS-CoV-2 detection system. This paper combines XGBoost with Raman spectroscopy to develop a novel coronavirus detection model that includes the following core ideas.

a) Characteristic selection of Raman spectral data for the original two populations.

b) Integrate multiple decision tree models with low accuracy.

c) Using greedy strategy and quadratic optimization to determine the optimal node and the minimum loss function.

d) Use python to call XGBoost library to automatically run CPU multi-threaded.
In this paper, the XGBoost model mainly optimizes parameters such as `max_depth`, `learning_rate`, `n_estimators`. Among them, `max_depth` is used to control the maximum depth of the tree and control over-fitting. If the depth of the tree is too large, it will lead to over-fitting. `learning_rate` represents the step length of each iteration. If it is too large, the running accuracy is not high, and if it is too small, the training speed will be very slow. `n_estimators` represents the maximum number of trees generated and also the maximum number of iterations. The processed data are divided into test set and training set according to the ratio of 3:7. The `n_estimators` range is set to [0,200], the `max_depth` range is set to [0,15], and the `learning_rate` range is set to [0,0.5]. The above parameters are used as grid search parameters to train the model. The `max_depth` value selected in this article is 4, the `learning_rate` value is 0.25, and the `n_estimators` value is 200.

3.3. Model performance evaluation and results analysis

In this paper, the model performance is evaluated using three evaluation metrics. Since XGBoost is evolved from GBDT, we added the GBDT model in the evaluation phase. The performance of the final model in Raman spectral data for both populations is shown in Table 2.

| Model                  | Accuracy | Sensitivity | Specificity |
|------------------------|----------|-------------|-------------|
| GBDT                   | 87.097%  | 89.655%     | 84.848%     |
| XGBoost                | 88.710%  | 84.848%     | 93.103%     |
| PCA+XGBoost            | 91.935%  | 89.655%     | 93.939%     |
| RFE+XGBoost            | 93.548%  | 87.500%     | 96.774%     |
| SVM-RFE+XGBoost        | 93.548%  | 90.000%     | 96.875%     |

It can be seen from the results that XGBoost has higher accuracy and specificity compared to GBDT. This is due to the fact that XGBoost explicitly adds the complexity of the tree model as a regular term to the optimization target and uses a greedy strategy and quadratic optimization to determine the optimal node and the minimum loss function. Due to the large number of features collected, we combine the XGBoost algorithm with three feature extraction methods to obtain their detection results. The SVM-RFE binding XGBoost algorithm showed the best performance in Sensitivity and Specificity, respectively 90.000% and 96.875%. It can be seen that XGBoost has a better robustness than the other three models on the Raman spectral of two different populations.
4. Conclusion
In view of the characteristics of Raman spectral, the data pre-processing operations such as normalization, smoothing and denoising, feature selection, etc. are performed on the raw Raman spectral. In the model construction stage, the log loss function is used as the error function in the objective function to make the model fully fit the training data set, reduce the bias of the model, while adding regularization penalties in the objective function, so that the model is not easy to produce over-fitting, in order to improve the stability of the model. After the construction of the detection model, the model was evaluated using three model evaluation indicators, which proved that the model has good generalization ability. This provides a reference for novel coronavirus detection, which can be used as an auxiliary diagnostic method.

References
[1] Amirfakhryan, H., and Safari, F. (2020) Outbreak of SARS-CoV-2: Pathogenesis of infection and cardiovascular involvement. Hellenic Journal of Cardiology, https://doi.org/10.1016/j.hjc.2020.05.007
[2] Ayouba, A., Thaurignac, G., Morquin, D., Tuallon, E., Raulino, R., Nkuba, A., Lacroix, A., Vidal, N., Foulongne, V., Le Moing, V., Reyes, J., Delaporte, E., and Peeters, M. (2020) Multiplex detection and dynamics of IgG antibodies to SARS-CoV-2 and the highly pathogenic human coronaviruses SARS-CoV and MERS-CoV. J Clin Virol 129, 104521.
[3] Choi, Jun Yong , and M. Joo . "The pathogenesis and alternative treatment of SARS-CoV-2." Integrative Medicine Research 9.3(2020):100421.
[4] Bouiller, K., Humbert, S., Payet-Revest, C., Brunel, A. S., Mareshal, A., Lepiller, Q., Grillet, F., and Chirouze, C. (2020) Utility of CT scan in patients with initial negative PCR for SARS-CoV-2: a report of three cases. Infection.
[5] Uusitalo, S., Popov, A., Ryabchikov, Y. V., Bibikova, O., Alakomi, H.-L., Juvonen, R., Kontturi, V., Siitonen, S., Kabashin, A., Meglinski, I., Hiltunen, J., and Laitila, A. (2017) Surface-enhanced Raman spectroscopy for identification and discrimination of beverage spoilage yeasts using patterned substrates and gold nanoparticles. Journal of Food Engineering 212, 47-54.
[6] Khan, S., Ullah, R., Khan, A., Wahab, N., Bilal, M., and Ahmed, M. (2016) Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM). Biomed Opt Express 7, 2249-2256.
[7] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, 13-17 August 2016, 785-794.
[8] Gertz M, Große-Butenuth K, Junge W, et al. Using the XGBoost algorithm to classify neck and leg activity sensor data using on-farm health recordings for locomotor-associated diseases. Computers and Electronics in Agriculture. 2020;173.
[9] Liu Y, Dang B, Li Y, Lin H, Ma H. Applications of Savitzky-Golay Filter for Seismic Random Noise Reduction. Acta Geophysica. 2016;64(1):101-124.
[10] Zimmermann B, Kohler A. Optimizing Savitzky-Golay parameters for improving spectral resolution and quantification in infrared spectroscopy. Appl Spectrosc. 2013;67(8):892-902.
[11] Duan KB, Rajapakse JC, Wang H, Azuaje F. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans Nanobioscience. 2005;4(3):228-234.
[12] Yin L, Li S, Lu Y, Yin Y, Su Y, Zeng M, Luo M, Ma H, Zhou D, Yao G, Liu J, and J. Lang, Data and code on serum Raman spectroscopy as an efficient primary screening of coronavirus disease in 2019 (COVID-19), (2020). doi:10.6084/m9.figshare.12159924.v1.
[13] Wang S, Dong P, Tian Y. A Novel Method of Statistical Line Loss Estimation for Distribution Feeders Based on Feeder Cluster and Modified XGBoost. Energies. 2017;10(12).