ISSN: 2508-7894 © 2020 KODISA & KAIA.
KJAI website: http://www.kjai.or.kr
doi: http://dx.doi.org/10.24225/kjai.2020.8.1.7

A Study on Methods to Prevent the Spread of COVID-19 Based on Machine Learning

Youngsang KWAK[1], Min Soo KANG[2]

Received: January 15, 2020. Revised: April 29, 2020. Accepted: June 05, 2020.

Abstract

In this paper, a study was conducted to find a self-diagnosis method to prevent the spread of COVID-19 based on machine learning. COVID-19 is an infectious disease caused by a newly discovered coronavirus. According to WHO(World Health Organization)'s situation report published on May 18th, 2020, COVID-19 has already affected 4,600,000 cases and 310,000 deaths globally and still increasing. The most severe problem of COVID-19 virus is that it spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes, which occurs in everyday life. And also, at this time, there are no specific vaccines or treatments for COVID-19. Because of the secure diffusion method and the absence of a vaccine, it is essential to self-diagnose or do a self-diagnosis questionnaire whenever possible. But self-diagnosing has too many questions, and ambiguous standards also take time. Therefore, in this study, using SVM(Support Vector Machine), Decision Tree and correlation analysis found two vital factors to predict the infection of the COVID-19 virus with an accuracy of 80%. Applying the result proposed in this paper, people can self-diagnose quickly to prevent COVID-19 and further prevent the spread of COVID-19.

Keywords: COVID-19, SVM, Correlation Analysis

Major classifications: Artificial Intelligence, Supervised Learning, Support Vector Machine

1. Introduction

Coronavirus is one of the major pathogens that primarily targets the human respiratory system. Previous outbreaks of coronaviruses (CoVs) include the severe acute respiratory syndrome (SARS)-CoV and the Middle East respiratory syndrome (MERS)-CoV, which have been previously characterized as agents that are a tremendous public health threat. The symptoms of COVID-19 infection appear after an incubation period of approximately 5.2 days (Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y.,Xing, X. et al, 2020). The period from the onset of COVID-19 symptoms to death ranged from 6 to 41 days, with a median of 14 days. This period is dependent on the age of the patient and the status of the patient's immune system. It was shorter among patients >70-years old compared with those under the age of 70. The most common symptoms at the onset of COVID-19 illness are fever, cough, and fatigue, while other symptoms include sputum production, headache, hemoptysis, diarrhea, dyspnea, and lymphopenia (Carlos, W. G., Dela Cruz, C. S., Cao, B., Pasnick, S., and Jamil, S, 2020).

For these reasons, it is necessary to study for finding self-diagnosis methods to prevent the spread of COVID-19. There are 21 attributes that contain a demographic feature, other diseases such as diabetes, heart disease, kidney disease, and corona result. First, correlation analysis was conducted to determine which attributes profoundly affect corona results, and then, corona results were predicted using SVM(Support Vector Machine) through variables that have a potent effect.

1 First Author, Student, Department of Medical IT Marketing, Eulji University, Email: ysk1188@naver.com
2 Corresponding Author, Professor, Department of Medical IT, Eulji University, Email: mskang@eulji.ac.kr

2. Supervised Learning

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object(typically a vector) and a desired output value(also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances.

2.1. SVM (Support Vector Machine)

In machine learning, Support-Vector Machine is supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, and an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the separate categories that are divided by a clear fap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall (Cortes, C., and Vapnik, V, 1995). The reason for using a two-class support vector machine is that technically it can be used in both classification and forecasting problems, and the second is less likely to be overfitted than neural network techniques, and thirdly, it is more accurate to predict, and lastly for its simplicity. For its experimental usage, the performance of two-class logistic regression and two-class neural networks were lower than the two-class support vector machine.

3. Experiment

3.1. Data Preprocessing

COVID-19 data utilized in this study were collected from the open source site, Kaggle specifically 128 rows and 21 columns, including demographic data, various symptoms, and diseases. Table 1 illustrates the data collected.

Table 1: COVID-19 DATA

Data Group	Example
Demographic data	Age, Gender, Body Temperature
Various Symptoms	Dry Cough, Sour Throat, Weakness, Breathing Problem, Drowsiness, Pain in chest, Travel History to infected countries, Change in Appetite, Loss of sense of smell
Various Diseases	Diabetes, Heart Disease, Lung Disease, Stroke or Reduced Immunity, Symptoms progressed, High Blood Pressure, Kidney Disease,
Result	Corona Result

Pre-processing of data collected was executed using correlation analysis. Figure 2 illustrates the pre-processing results of the data.

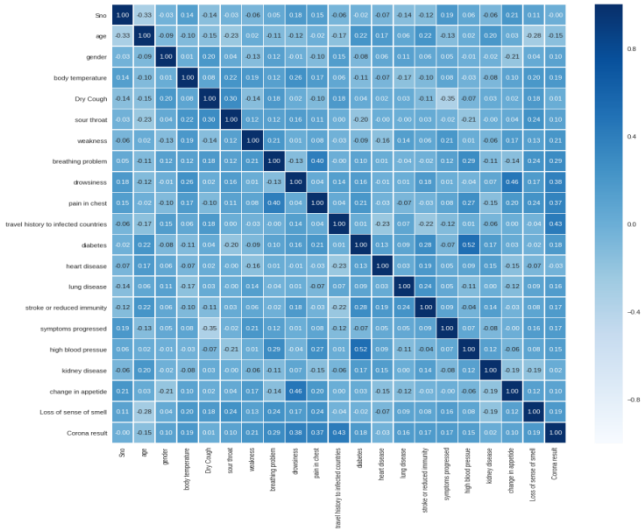

Figure 2: Correlation Analysis Results Using COVID-19 Data

Through the correlation result analysis result Drowsiness, Pain in chest, Travel history to infected countries attributes have 0.38, 0.37, 0.43 relation to the COVID-19 results. So, 1 experiment was executed using these three attributes applying SVM and Decision tree to predict COVID-19 results.

4. Results

The experiment with Drowsiness, Pain in Chest, Travel history to infected countries feature using SVM resulting in 0.806 accuracy, 0.737 precision, 0.933 recall, and finally 0.824 F1 score. Figure 3 visualizes how evaluate model came out. And with Decision Tree only three steps took to diagnose COVID-19 patients whether patients get COVID-19 or not.

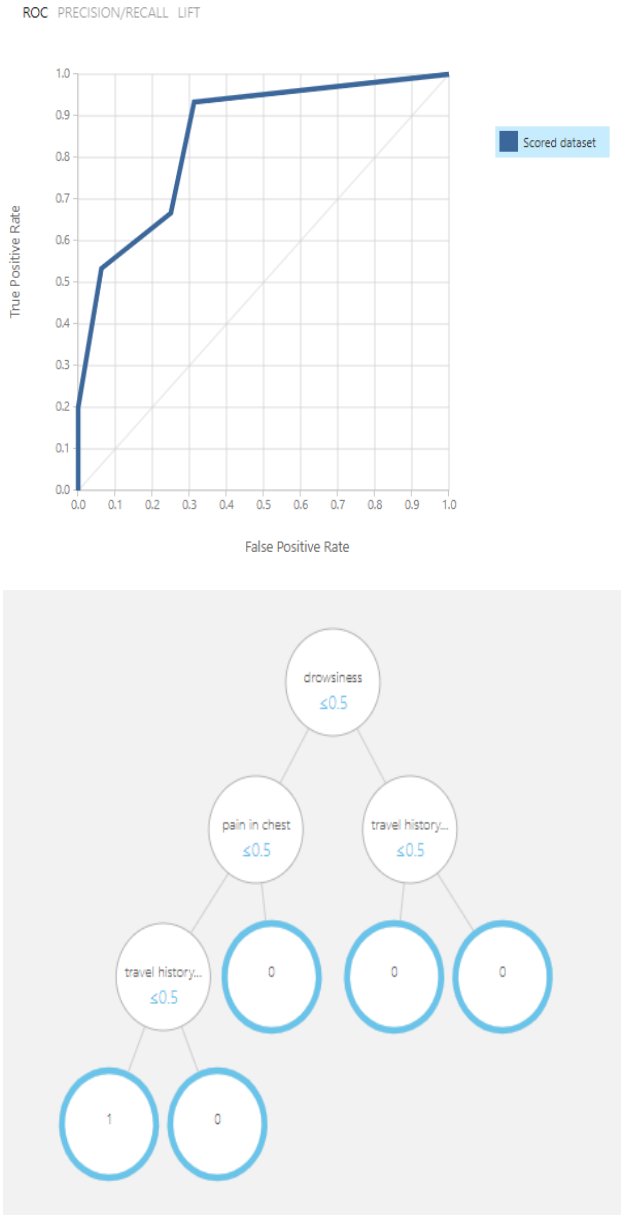

Figure 3: Evaluate Model

5. Conclusion

In this paper, 1 experiment was made to predict COVID-19 patients using two – class support vector machine and two-class boosted decision tree. The result showed means that when classifying COVID-19 patients, only checking the potential patient's drowsiness, pain in chest, and their travel history to infected countries is more efficient than doing self-diagnosis questionnaire, which takes time and has ambiguous standards. So, applying the result proposed in this paper, doctors could be able to reduce time when diagnosing COVID-19 patients and diagnose other potential COVID-19 patients more. Furthermore, doing this process could prevent the spread of COVID-19.

References

Carlos, W. G., Dela Cruz, C. S., Cao, B., Pasnick, S., & Jamil, S. (2020). COVID-19 Disease due to SARS-CoV-2 (Novel Coronavirus). *American Journal of Respiratory and Critical Care Medicine, 201*(4), 7-8.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Kang, M.S., Kang, H.J., Yoo, K.B., Ihm, C.H., & Choi., E. S. (2018). *Getting started Machine Learning with Microsoft AZURE ML*. Hanti Media

Kaggle (2020). COVID-19 Dataset, Retrieved April 20, 2020 from https://www.kaggle.com/bitsofishan/covid19-patientsymptoms

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y.,Xing, X. et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *New England Journal of Medicine,* 382, 1199-1207. DOI: 10.1056/NEJMoa2001316

Microsoft Azure Machine Learning(2020). Retrieved April 20, 2020, fromhttps://docs.microsoft.com/ko-kr/azure/machine-learning/studio/what-is-ml-studio