Data Analysis of COVID-19 Patients in South Korea

Project Type: Software
Dataset: https://www.kaggle.com/kimjihoo/coronavirusdataset
Labor division: Correlation Analysis(Yiyuan Liu), Regression(Yi Wang), Clustering(Yuxuan Liu), Classification(Guoren Li), Network Analysis and Graph Mining(Songrui He)

Background and motivation

Coronavirus disease 2019 (COVID-19) is the infectious disease caused by the most recently discovered coronavirus. The disease has spread globally from the ongoing 2019–20 coronavirus pandemic. As of 13 April 2020, more than 1.91 million cases have been reported in 210 countries and territories, resulting in more than 118,000 deaths. More than 448,000 people have recovered. Governments of many countries and CDC recommended people to stay at home to prevent COVID-19.

We want to make a contribution. We decided to use the structured dataset based on the report materials of KCDC and local governments because this dataset is detailed enough for us to analyze and mine. It is known that identifying the death risk of all patients is critical for a country to control the disease and allocate the medical resources. According to existing researches, there are some relationship between the patient's condition(e.g. Age, sex, underlying disease, country/region) and their death risks. Besides, the spread of the virus is influenced by the region, population, migration etc. based on a hidden model.

Overall, our goal is to use data mining knowledge to identify the hidden relationship between the patient's condition and the death risks, and gain a better understanding of how the virus spread.

What we are planning to do:

- Predict the number of confirmed cases and deaths
- Find hidden correlations(e.g. patients' states vs condition, growth rate vs region/weather)
- Predict death risk of a particular patient
- Visualizations

Method

1. Association and correlation analysis
   The dataset about COVID-19 that we use is huge. By using some techniques of association and correlation analysis, we will analyze the correlation of different items, including: weather, death risk, patient's physical condition, regional case growth rate, patient's region, traveling route and number of contacts, to find some hidden correlations.

2. Regression
   We will use some regression model(e.g. Linear regression) to predict the future confirmed cases and death cases in each region based on previous data. This is not an accurate prediction model based on virology and migration. It’s just a projection based on previous data. We can use it to determine the worst case if no precautions are taken.

3. Clustering (unsupervised learning)
   According to the correlation analysis, we may use some clustering algorithms(e.g. K-means, kNN) to find some similarities of infected patients, such as physical conditions and regions. Then, we are able to get internal relationships among data by clustering.

4. Classification (supervised learning)
   By applying some classification algorithms such as naive Bayes, Decision Tree and Logistic regression, we can build a classification model to determine the final state (isolated/released/deceased) of a particular patient based on the patient’s condition. In this way, we can predict the death risk of a new patient and take some reasonable precautions.

5. Network Analysis and Graph Mining
   The spread and infection of the virus among people will form a network, such as the “Who is infected by whom” graph and the patients’ route. Thus, with the help of network analysis and graph mining techniques, we may find some interesting patterns(e.g. Super infector) and gain a better insight into the spread mode of the virus.

Evaluation Criteria

1. Descriptive and predictive

We will show what happened so far, and how our methods and results explain the current situation. Besides, we will try to use our methods to predict the things that will happen.

2. Coverage and accuracy

For some methods and specific purposes, we may need to filter out some data. We will show the coverage of the data we used, and explain the reason why we pick those data. Besides, we will show how accurate our methods can be.

3. Performance

We will give main performance metrics of our methods (e.g. time consuming, computing resources, etc).