Proposal on Visualizing COVID-19 and Classifying Severe Cases with Immediate Medical Need

April 27, 2020

Project Members:	Yun Ki, LEUNG	1155109801
	King Yeung, CHAN	1155119394
	Tsz Chun, LAI	1155125208

Problem Statement

We may hear "YOU STAY AT HOME FOR US" so many times because of the tremendous pressure on the medical industry. We want to give doctors a hand on making a decision based on the urgency of patients from the novel coronavirus. To understand the current circumstances quickly, we are going to visualize the overall trend and classify the issue of death from the novel coronavirus. We anticipate the outcome can assist professionals to decide a better judgement on healing the patients.

Motivation

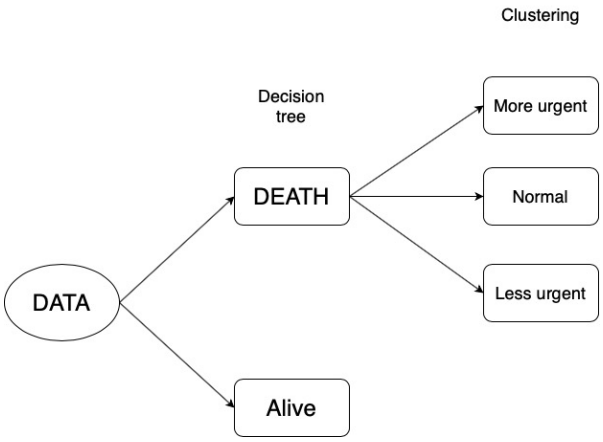
At the late end of 2019, the novel coronavirus, also known as COVID-19, has become an epidemic and communicable diseases around the world. The novel coronavirus has not only influenced the economy over the world but also put unprecedented pressure on medical personnel around the world. Due to the lack of information, a lot of medical personnel was not able to save lives from the novel coronavirus. For a general citizen without the knowledge of professional medicine, we do not have such abilities to take off their shoulders. We understand that they are fighting for us in the front line, therefore, we want to try our best to relieve their pressure in the view of making decisions. We expect to provide essential information, such as the classification of confirmed patients with the potential death and the urgency of a patient who suffers from the novel coronavirus. We hope the results can help doctors to decide the treatment of each patient, such that they can put more resources and effort into those classified patients and save their lives. In the project, consequently, we aim to relieve the stress on medical personnel.

Data Sources

To establish the model we are concerning, we are going to grab the data set *Novel Corona Virus 2019 Dataset* which published from Kaggle. Kaggle is famous for many competitions about Machine Learning and Data Science. The data set that we have chosen contains daily information on the novel coronavirus. The sources of data are organized on a daily basis from the international organization, such as the World Health Organization. We believe the data are reliable with such reason. The model building relies on the training data and the testing data. As a result, the outcomes of the project are biased on the data. Thus, we assume the provided data are dependable.

Data Processing Approaches

The blueprint of the project is the visualization of the overall trend and the classification for the issue of deaths caused by the coronavirus. The graph on the right briefly illustrates the ideas of the project. We propose to visualize the overall trend of the current situation of the novel coronavirus. As we are investigating the relationship between confirmed patients and the urgency of death, we would concentrate on studying the tendency of the number of mortalities. Subsequently, we are going to use both supervised and unsupervised learning to inspect the behavior of the data. Decision tree (supervised learning) is adopted to evaluate whether a patient confirmed with the novel coronavirus would face the matter of death. Clustering along with k-means (unsupervised learning), afterwards, is going to help us to classify the urgency of a confirmed patient with a high death rate to receive treatments by only studying the historical data on patients who have passed their lives.



Planned Analysis/ Implementation/ Experimentation

For the visualization, we are going to plot the overall trend, including the number of confirmed, death and recovered cases, of the novel coronavirus. We plan to use a time series model, known as regression against time, to capture the behavior of the data. R software is being favored for the visualization since it is a robust language to deal with a huge amount of data.

For the classification, we intend to divide patients into 2 groups in the decision tree which are “alive” and “dead”. The data set provides us with a variety of attributes, such as age, travelling record, and symptoms, which are valuable to construct multiple levels for the tree. We are planning to use about 800 observations to construct the decision tree using Weka. We would divide the data into a training data set and a testing data set for the model verification. Moreover, we are considering using k-means to discover the Euclidean distance between each patient, such that we can classify different clusters for patients. We label those clusters based on the severity of the patients from the time of death. The number of clusters k we use depend on the fitness of the data. We consider k with the SSE no longer drop rapidly.

Potential Conclusions

We expect the visualization gives us an insight into the current situation about the novel coronavirus so that we can easily understand the behavior of the data. From the decision tree, we hope that it can classify the characteristics of the patients facing the issue of death. Then, we can make use of k-means to classify the urgency of treatments among severe cases. With our common sense that the severe cases with a longer infection period tend to have a higher risk of death and they should have immediate medical needs. Thus, our ultimate goal is the medical stuff can use our results to detect severe cases efficiently in order to have immediate medical support.