Phan Bui

301325875

dbui@sfu.ca

Final Project Report:

Prediction on Number of Cases in Each State

And

Estimation on Number of COVID-19 Types

❖ Motivation and Background

- Many people say: "2020 is the year the world stops". As we all know, in the beginning of 2020 a new kind of virus named COVID-19 appears. This is a very dangerous kind of virus. It can "lay low" in the beginning, the host will not have any symptoms in the first two weeks, the virus can be passed by air or by touching the same surface, which makes it very contagious. COVID-19 is also dangerous because the host can die in short time without proper care. The world has never faced this kind of virus and has no treatment for it. Hence, the best thing we can do is to practice quarantine. This will impact the economy in a profoundly negative way.
- My project is a research on COVID-19 based on datasets available online. It will help us understand more about the virus and hopefully can deliver some hints to scientist about what we can do. Other people can also use my research to have a deeper intuition and knowledge about what is happening.
- In this research, I use techniques that I learned during studying CMPT 459 (Data Mining) in Simon Fraser University in Summer 2020 with the help of Professor Jian Pei and Teaching Assistant Madana Krishnan Vadakandara.

❖ Problem Statement

- There are many things that we want to know about the COVID-19, such as where does the virus come from, how contagious it is and what we can do to prevent spreading, how to make the vaccine, … In this project, I will use dataset available online, apply some technique I learned to figure out how many confirm cases and how many fatalities each state will have in the next week. Virus have many kinds and never stop evolving. In this project, I will also try my best to estimate how many kinds of COVID-19 are there.

❖ Datasets

- In this project, I will use 2 dataset in order to learn more information about COVID 19.
- The first data set is https://www.kaggle.com/c/covid19-global-forecasting-week-5. This dataset has a training file and a testing file. It has information about patients such as country, state, date confirmed, county, … I will use this data set to predict how many confirmed cases and fatalities will there be in the next week in each state.
- The second dataset is https://github.com/beoutbreakprepared/nCoV2019/tree/master/latest_data. This data set only have 1 file and it contains more specific information about patients like age, sex, city, symptoms, outcome, … I will use this dataset in order to estimate how many kinds of COVID-19 are there.

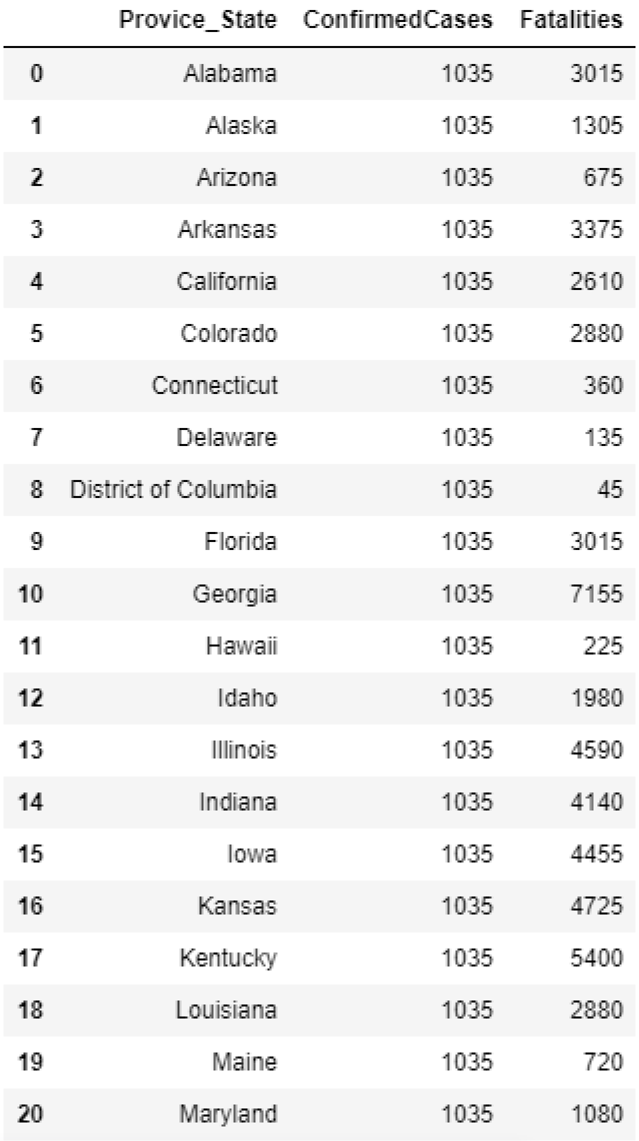
- After learning 2 datasets, hopefully I can deliver helpful information to readers so that we can understand more about this new dangerous kind of virus.

❖ Methodology

- In this project, I use two main data mining methods which are decision tree and clustering. I use decision tree in dataset https://www.kaggle.com/c/covid19-global-forecasting-week-5 to figure out number of confirmed cases and fatalities in the next week in each state. I use clustering while working on dataset https://github.com/beoutbreakprepared/nCoV2019/tree/master/latest_data in order to know how many kind of COVID-19 are there.
- I also use some data cleaning to clean the data, get rid of tuples with "bad" attributes, prune the dataset to a smaller size which is easier to process.
- I use pandas to process .csv files into data frame and scikit-learn for decision tree and clustering. This is because these are tools that produce excellent results and convenient to use.
- In dataset https://www.kaggle.com/c/covid19-global-forecasting-week-5, I use pandas to read data from the .csv file. After that, I decide that I will use training attributes County, Province_State, Country_Region, Population, Weight to train and the target class is attribute Target (ConfirmedCases or Fatalities). For each training attribute, I count the number of rows with that attribute that is not NaN. I see that for each attribute, there a lot of row that satisfy the condition. Hence, I don't drop any training attribute. I drop every tuple with a NaN attribute and realize that only patient from the US are left. However, there are still plenty tuples to work on so I accept the outcome. After that, I have to digitalize training attributes that are not numbers. I use LabelEncoder() to perform this task. Then, I use a decision tree to train the training data. Finally, I use the same tree to guess the target class of the testing data and write the result to a table.
- In data set https://github.com/beoutbreakprepared/nCoV2019/tree/master/latest_data, I use pandas to read data from the .csv file. After that, I decide that I will use training attributes age, sex, city, sumptoms, lives_in_Wuhan, chronic_diseases_binary, outcome to divide the data into clusters. For each attribute, I count the number of rows with that attribute that is not NaN. I figure out that there is so little row with an "acceptable" symptoms and lives_in_Wuhan. Hence, I drop these two attributes so that I will have "large enough" data to use. I drop every tuple with a NaN attribute. After that, I have to digitalize training attributes that are not numbers. I use LabelEncoder() to perform this task. Then, I divide the data into 2 part, the first 1/9 of the data will be used for testing, the other 8/9 of the data will be used for training. For number of clusters range from 2 to 40, I divide training data and testing data into

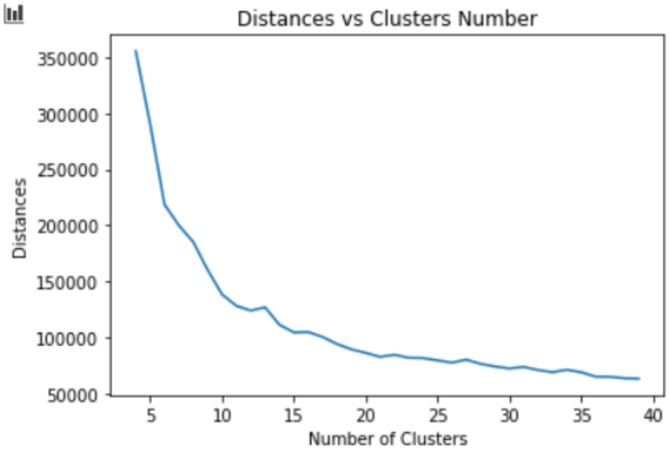
clusters. I calculate the kmeans of training data and use it to predict the cluster of testing data, then calculate the errors between testing cluster an training centroids. I also calculate kmeans of both training and testing data, then calculate the difference from points in clusters of testing data to cluster center in training data. The number of clusters which produce the minimum difference will most likely be the number of types of COVID-19.

❖ Results

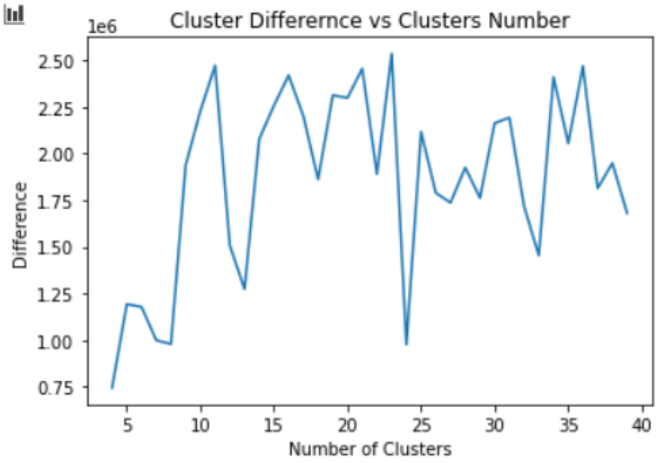


	Provice_State	ConfirmedCases	Fatalities
0	Alabama	1035	3015
1	Alaska	1035	1305
2	Arizona	1035	675
3	Arkansas	1035	3375
4	California	1035	2610
5	Colorado	1035	2880
6	Connecticut	1035	360
7	Delaware	1035	135
8	District of Columbia	1035	45
9	Florida	1035	3015
10	Georgia	1035	7155
11	Hawaii	1035	225
12	Idaho	1035	1980
13	Illinois	1035	4590
14	Indiana	1035	4140
15	Iowa	1035	4455
16	Kansas	1035	4725
17	Kentucky	1035	5400
18	Louisiana	1035	2880
19	Maine	1035	720
20	Maryland	1035	1080

This is the result from processing the dataset https://www.kaggle.com/c/covid19-global-forecasting-week-5 with decision tree. Only the results of the first 20 states are shown. The rest of the result can be viewed in the Kaggle notebook. I see that the states with bigger population often have more fatalities. This can be the result of more interactions between people. Some small states like Kansas also have pretty high fatalities. This might be because the people are not well informed about the virus and does not take it seriously.



This is the result of dataset https://github.com/beoutbreakprepared/nCoV2019/tree/master/latest_data when I predict cluster of testing data and calculate the distance between that cluster and centroids of training data. The bigger the number of clusters, the smaller the error. This is because each point can have more choices to find centroids that fit it the most.



This is the result of dataset https://github.com/beoutbreakprepared/nCoV2019/tree/master/latest_data when I calculate the kmeans of both training and testing data. After that, I try to find at which point 2 kmeans are the most similar. It can be seen on the graph that at number of clusters equals 7 or 8, the difference is the smallest. I estimated that there are about 8 types of COVID-19.

- **Evaluation**

  I am confident that my result is quite accurate. This is because I employed techniques and algorithms that have been developed for a long time and have been proven to often produce great result. Before running any algorithm, I do data cleaning to assure that the inputs of each algorithm are “clean” and “acceptable”. The clustering result from dataset https://github.com/beoutbreakprepared/nCoV2019/tree/master/latest_data turns out to be surprisingly quite encouraging in term of accuracy.

- **Summary**

  After employing decision tree and clustering method to process two datasets, I have learned more about COVID-19. The result is not perfect but is good in many aspects. I have predicted the number of confirmed cases and fatalities in the near future in each state. I have also estimated that there are 3 types of COVID-19. However, there are still some flaws in my result. After I run decision tree on https://www.kaggle.com/c/covid19-global-forecasting-week-5, I realize that the prediction is 100% correct on the testing data, which is “too good to be true”. I suspect that my model is quite simple and there are not enough data to feed in the algorithm. If I have a more complex model and larger data size, the result will become more accurate.