Data Analysis of COVID-19 Patients in South Korea Midterm Report

1. Introduction

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. We want to make a contribution to prevent the spread of the disease and make predictions for the future. 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. 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.

2. Related Work Survey

2.1 Temperature, humidity, and latitude analysis to predict potential spread and seasonality for COVID-19[1] (Yiyuan Liu)

This article, by analyzing the temperature, latitude, and humidity of seasonal respiratory virus transmission behavior. Through modeling, it may be possible to predict a high-risk area that may become a serious spread of covid-19. They examined climate data from cities with significant community spread of COVID-19 using ERA-5 reanalysis, and compared to areas that are either not affected, or do not have significant community spread.
Pros:
- To data, The spread of covid-19 basically follows the predicted specific temperature (5-11°C), humidity (4-7 g /m³), At latitude (30-50° N’), showing the high accuracy of this prediction model.
Cons:
- Lack of significant community establishment in expected locations that are based only on population proximity and extensive population interaction through travel.

Possible extensions:
-Collect more relevant data to improve accuracy

2.2 Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions[2] (Yiyuan Liu)

By Integrated population migration data before and after January 23 and most updated COVID-19 epidemiological data into the Susceptible-Exposed-Infectious-Removed (SEIR) model to derive the epidemic curve. This paper also used an artificial intelligence (AI) approach, trained on the 2003 SARS data, to predict the epidemic. It finally proved that the implementation of control measures on January 23, 2020 is essential to reduce the scale of the final COVID-19 epidemic.

Pros:
-This paper modified the original SEIR-equation to account for a dynamic Susceptible [S] and Exposed [E] population state by introducing the move-in, In(t) and move-out, Out(t) parameters.
-They used the LSTM model, a type of recurrent neural network (RNN) that has been used to process and predict various time series problems to predict numbers of new infections over time.

Cons:
-The modeling method may be a bit insufficient

Possible extensions:
-This method can play a major role in data analysis in multiple locations.

2.3 Predicting COVID-19 in China Using Hybrid AI Model[3] (Yi Wang)

This article, which aims to predict the trend of the COVID-19, discovered that new daily confirmed cases at different time intervals have different contributions to susceptible infections. This paper employs the Hybrid AI Model to make predictions. Compared to traditional models, Hybrid AI model is more effective and can significantly reduce the errors of the prediction results.

Pros:
- The model has high prediction accuracy and the prediction results are highly consistent with actual epidemic cases.

Cons:

- This model only uses daily confirmed cases and infection rate to predict future infection rate and the number of confirmed cases. It did not combine with more complex factors like temperature and migration.

Possible extensions:

- Combine with more factors.

2.4 SEIR and Regression Model based COVID-19 outbreak predictions in India[4] (Yi Wang)

In this study, two machine learning models SEIR and Regression were used to analyse and predict the change in spread of COVID-19 disease. This paper aims at finding the rate of spread of the disease in India, developing a mathematical SEIR (Susceptible, Exposed, Infectious, Recovered) model to evaluate the spread of disease, using SEIR and Regression models to predict COVID-19 outbreak. Finally, two models both have good performance and the prediction result has high accuracy.

Pros:

- Algorithms used in models are easy to implement.
- High performance. The performance of the models was evaluated using RMSLE and achieved 1.52 for the SEIR model and 1.75 for the regression model.

Cons:

- The training data set is only based on confirmed cases, other factors may influence the result.

Possible extensions:

- This automated algorithm can be developed to fetch data in regular intervals and automatically predict the number of cases for weekly and biweekly data.

2.5 Correlation Between Temperature and COVID-19 (Suspected, Confirmed and Death) Cases based on Machine Learning Analysis[5] (Yuxuan Liu)

As COVID 2019 becomes more and more dangerous, Siddiqui M K [5] and his teammates decided to find the relationship between temperature and different COVID 2019 cases.

Siddiqui M K used one of the clustering methods: K-means to achieve his goal. Because clustering is a well-known unsupervised machine learning classifier which can do a great job on discovering inherent relationships on unlabeled data sets.

Finally, they found that temperature may have some relationship with COVID 2019 cases situation but the relationship is not significant which means temperature may not be the main factor.

Pros:
- Novelty on discovering relationship between temperature and COVID 2019 cases situation.
- Detailed analysis of the results.

Cons:
- Only using K-means method.
- The factor chosen is too one-sided. Factors can include not only temperature.

Possible extensions:
- Analyzing more factors.
- Using more methods and making a comparison between each other.

2.6 Predictive Analytics of COVID-19 Using Information, Communication and Technologies[6] (Yuxuan Liu)

Mahalle P N [6] used machine learning methods to predict the trend and the confirmed cases of COVID 2019.

In the paper, Mahalle P N used many predictive analytics models, such as classification models, clustering model, forecast model, outlier models and time series model. They also utilized many predictive analytics algorithms which include random forest, gradient boosted models and K-means.

At last, they predicted there would be 1.6 million and 2.3 million CVOID-19 confirmed cases by the end of May and June respectively.

Pros:
- Detailed introducing many predictive methods.
- Did a lot of related works and wrote them in the paper in detail.

Cons:
- No comparison between different methods.
- Experiment part is too less.

Possible extensions:
- Add methods comparison part.
- Tell more about how to get the result.

2.7 A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models[7](Guoren Li)

Dianbo Liu's team is trying to provide stable and accurate COVID-19 forecasts ahead of days.

This paper presents a novel methodology which combines mechanistic model and machine learning methodologies to forecast COVID-19 2 days ahead of current time at the province level in China. In particular, it's based on and also improves the latest and most accurate infectious disease prediction model - ARGONet.
In this paper, authors assume the starting date of the epidemic is between Nov.15.2019 and Dec.1.2019, which matters the initial data. And they also assume the world has a detection of imported cases as low as 40%, which is kind of a conservative assumption according to other research works.
Because this is a real-time forecast model, authors evaluate it by performing prediction in a period and comparing the results with the ground truth. Specifically, they mainly take the dataset from China CDC official health reports, and evaluate the accuracy with the data between Feb.3.2020 and Feb.21.2020.
Pros:
- Forecasts at province level, which causes many noises, but still gets a good result. Prove their model is somewhat general.
- Utilizes various dataset, official reports from CDC, Internet search frequencies from Baidu, related news reports from media sources. Diverse data can make the model more robust.
Cons:
- The starting date of data is Nov.2020, but stable prediction starts from Feb.2020. Means the performance is not so good when the dataset is insufficient.
- The data processing methods of search engine dataset and media news reports dataset are not elaborated.
Possible extensions:
- Input more well-processed datasets to improve robust and accuracy.
- Implement in other countries and get a more general solution.

2.8 Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan[8](Guoren Li)

Yan Li's team, which is in Wuhan, is trying to predict the criticality of patients.
In this paper, authors present a novel approach based on the latest and hottest XGBoost ML algorithm. Their approach only takes two datasets, first is the epidemiological history of the patient, another one is clinical manifestations. With those data, they can also achieve 100% death prediction accuracy and 90% survival prediction accuracy. It's pretty a novel approach that somewhat minimum utilization of datasets can achieve such good results.

And also, the multi-tree XGBoost algorithm makes criteria on clinical manifestations be more explainable, which hugely helps doctors to analyze patients datasets.

Pros:

- A general operable formula to precisely and quickly quantify the risk of death.
- Their datasets, especially the most significant features (indicated in paper) are easily to collect by any hospital, means this work is more feasible and can be quick to verify.

Cons:

- Limitation datasets (only 375 patients), there also should be more criteria in clinical manifestations.
- Purely data driven, may vary from different datasets.
- Insufficient explanation on why choosing multi-tree XGBoost.

Possible extensions:

- Give more reference datasets, and result in more key features.
- Try on and compare with more models based on the same datasets.

2.9 Prediction and analysis of Coronavirus Disease 2019[9] (Songrui He)

Jia, Lin, et al. adopts three mathematical models(Logistic model, Bertalanffy model and Gompertz model) to predict the confirmed cases of COVID-19 in China. Firstly, the epidemic trends of SARS were fitted and analyzed in order to prove the validity of the three models. Based on the results, the models were then applied to the situation of COVID-19.

The regression coefficient (R^2) is used to evaluate the fitting ability of the three models. The result shows that generally the Logistic model has the best fitting effect while the Gompertz model may be better than the Bertalanffy model.

Pros:

- Provided a relative instructive prediction of COVID-19 confirmed cases number based on three mathematical models.
- Compared with SARS situation to confirm the validity.

Cons:

- The prediction is only based on the number of past confirmed cases. When the policy and situation changes, the model will not fit anymore.

Possible extensions:

- Maybe more factors can be used in the prediction(e.g. weather, available hospital beds)

2.10 Inferring change points in the COVID-19 spreading reveals the effectiveness of interventions[10] (Songrui He)

Aiming at the challenge of the assessment of key epidemiological parameters and how they change when interventions show an effect, Dehning, Jonas, et al. proposed a method that combining an established epidemiological model, Susceptible-Infected-Recovered (SIR), with Bayesian inference to analyze the time dependence of the effective growth rate of new infections.

Focusing on COVID-19 spread in Germany, Bayesian inference is performed for the central epidemiological parameters of an SIR model using Markov-Chain Monte Carlo (MCMC) sampling. The results quantify the effect of interventions, and the corresponding change points can be incorporated into forecasts of future scenarios and case numbers.

Pros:

- The model provided a quantified effect of interventions using an novel and instructive model.
- The framework is scalable enough to be adapted to any other country or region.

Cons:

- The number of reported cases varies regularly over the course of a week and are especially low during weekends. This may affect the correctness of the model.

Possible extensions:

- Maybe some data transformation techniques will identify this periodic variation.

3. Project Progress

3.1 Data Preprocessing

We have done some data preprocessing work using Open Refine.

Data Reduction

Deleted Column	Reason
“latitude” and “longitude” in “Case.csv”	Province and city will be used as location information.
“global_num” in “PatientInfo”	It’s just a number given by KCDC.
“time” in “Time.csv”	Data update time is not useful for prediction

Dealing with missing data

"PatientInfo.csv"

- There are 77 records missing "sex", "age" and "infected_by" values. We treated them as "lacking too much information" and deleted them.
- For the records missing "sex"(there are only 9 records), since the rest of the records shows the ratio of "female" to "male" is 1.27:1, we manually filled in the missing values with 5 female and 4 male in random.
- There are 7 records missing both "age" and "birth_year". We filled in with a new category "unknown" in "age" for these records. For the rest of the records missing either "age" or "birth_year", we can calculate one of them based on the other value.
- For the records missing "city" values, since we know the province, we filled in "city" with existing cities in that province based on the proportion of existing cities.
- For the rest missing values, we filled in with a new category "unknown".

The rest of the dataset are processed based on similar strategies as listed below.

3.2 Work distribution update

Feature engineering and correlation analysis(Yiyuan Liu)
Linear Regression for prediction, Deep Learning for prediction(Yi Wang)
SVM and Decision Tree(Yuxuan Liu)
Bayes classification, Logistic regression(Guoren Li)
K-means clustering, Network analysis(Songrui He)

Reference

[1] Sajadi, Mohammad M., et al. "Temperature and latitude analysis to predict potential spread and seasonality for COVID-19." *Available at SSRN 3550308* (2020).
[2] Yang, Zifeng, et al. "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions." *Journal of Thoracic Disease 12.3* (2020): 165.
[3] Du, Shaoyi, et al. "Predicting COVID-19 Using Hybrid AI Model." (2020).
[4] Gaurav Pandey, Poonam Chaudhary, Rajan Gupta, Saibal Pal. "SEIR and Regression Model based COVID-19 outbreak predictions in India." *MedRxiv* (2020)
[5]Siddiqui, Mohammad Khubeb, et al. "Correlation Between Temperature and COVID-19 (Suspected, Confirmed and Death) Cases based on Machine Learning Analysis." *J. Pure Appl. Microbiol 14* (2020).
[6] Mahalle, Parikshit N., et al. "Predictive Analytics of COVID-19 Using Information, Communication and Technologies." (2020).
[7] Liu, Dianbo, et al. "A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models." *arXiv preprint arXiv:2004.04019* (2020).
[8] Yan, Li, et al. "Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan." *MedRxiv* (2020).
[9] Jia, Lin, et al. "Prediction and analysis of Coronavirus Disease 2019." *arXiv preprint arXiv:2003.05447* (2020).
[10] Dehning, Jonas, et al. "Inferring COVID-19 spreading rates and potential change points for case number forecasts." *arXiv preprint arXiv:2004.01105* (2020).