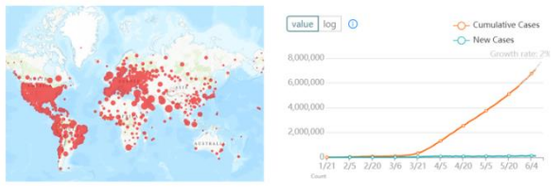
Data Analysis of COVID-19 Patients in South Korea

Songrui He, Yi Wang, Yiyuan Liu, Yuxuan Liu, Guoren Li

COVID-19



One of the most serious global pandemic.
188 countries/regions, more than 7 million cases!

High Fatality

What if 1) **new confirmed cases can be predicted** 2) **high death risk patients can be identified** and take some preventive actions?

Data Preprocessing

- Use OpenRefine
- Deleted some useless columns
- If missing too much values, delete the record.
- Fill in blanks:
  - According to the ratio of female and male.
  - Creating a new category
  - According to the province given.

Good data preprocessing is essential for some models!
Improve Accuracy!

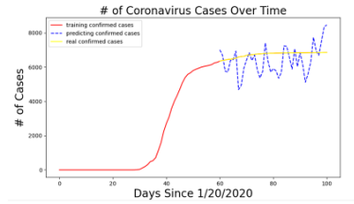
Linear Regression

Ridge Regression is used to mitigate the problem, providing improved efficiency in parameter estimation in exchange for a tolerable amount of bias.

$$L = \sum_{i=1}^{m}(y_i - w^T x_i)^2 + \lambda \underbrace{w^T w}_{\|w\|_2^2}$$

$$w_{ridge} = \arg\min_w L$$

$$w_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$$



Selected features: 1. "coronavirus" search trend 2. Temperature 3. Humidity

Training set: first 60% days
Test set: the rest 40%

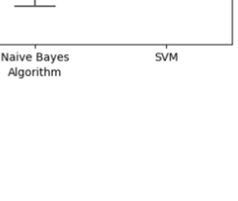
Data correlation analysis

To find the relationship quickly and efficiently between different random variables in these data, the most effective way is to perform correlation analysis on these random variables.

What we implemented

1. Data correlation analysis
2. Predict new confirmed cases: **Linear Regression**
3. Identify high death risk patients: **SVM, Naive Bayes, Logistic Regression**
4. Clustering analysis: **K-means**
5. Graph mining

SVM



The red line is a support vector which to classify data points as two classes. And the goal is to minimize ||w||.

Use linear kernel function and SMO algorithm to implement SVM, use OVO method to implement multi-classifier.

Randomly select 90% of dataset as the train set and remain 10% as the test set.

Selected features are:
1. Sex 2. Age
3. Country 4. Province
5. Infection case 6.Infection order

Accuracy: ~84.73 %

Naive Bayes

Implement Multinomial Naive Bayes algorithm to classify high risk patients. The Bayes posterior probability function is:

$$argmax_{c_k} logP(Y = c_k) + \sum_{j=1}^{d} x^{(j)} logP(w_j | Y = c_k)$$

Split dataset as 4:1, selected features: gender, age, country, province, infection case

Average accuracy: ~72%

Logistic Regression

Implement Logistic regression to classify high risk patients. Our prediction function is based on Sigmoid function:

$$h_\theta(x) = sigmoid(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

We also need the Cost function and J(θ) function to find the minimum value of J(θ)

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

$$cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & \text{if y=1} \\ -log(1 - h_\theta(x)) & \text{if y=0} \end{cases}$$

By trying different step size and iteration time, found 0.05 step size and 200 iteration time balances both efficiency and accuracy. **Accuracy: ~71.3%**

K-means

We use K-means algorithm to divide the patients into clusters and analysis the relations of each cluster.

K-means algorithm clusters data in n groups by minimizing the sum of squares(SSE).

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$$

Cluster result(K=3)

	Amount	Death rate
cluster1	1879	0.3%
cluster2	1142	0.8%
cluster3	283	14%

People of particular groups do have hidden relations.

People with high death risk:

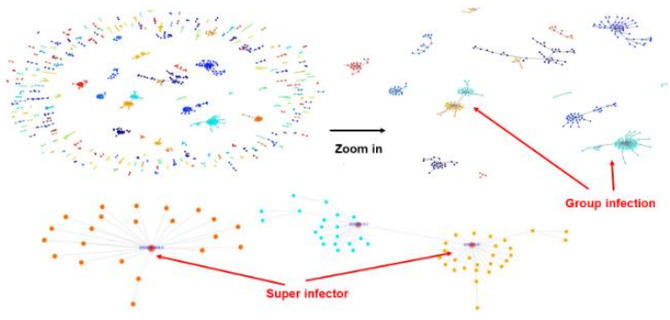
Came from Korea and US, having underlying disease, from age 60-80.

Graph Mining

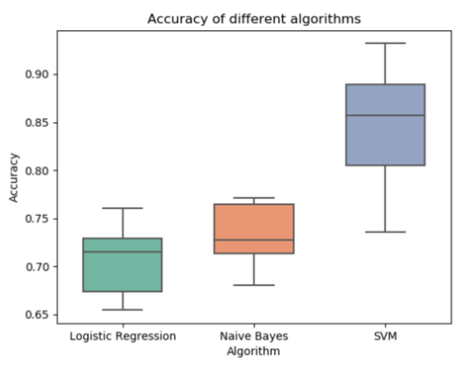
Based on the 'patient_id' and 'infected_by", construct a graph to connect people with "infection relation".

- **Group Infection Detection**: Use "Community Detection" to find potential groups of people having strong infection relations with each other.
- **Super Infector Detection**: Use "betweenness_centrality" to rank centrality, and identify patients who have spreading the virus to a large amount of people.

Graph Mining



Comparison



Conclusion

- Identifying the susceptible people and taking preventive measures accordingly is the key point to defeat the disease.
- Adapted and evaluated several machine learning algorithms to predict the growth trend of new confirmed cases and identify patients with high death risks.
- Used graph mining techniques to analyze group infection and super infector.
- Provided instructive information to help tackle the global problem of Coronavirus.