Using Machine Learning Algorithms to Predict Efficiency of US's States Mask Policy During COVID-19

Jin Cui\textsuperscript{1,*}

\textsuperscript{1}School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China

* Corresponding author Email: jimcui@bupt.edu.cn

Abstract. Over the past year, the COVID-19 outbreak deeply and thoroughly changed the way the world is. However, the control policy's efficiency is still in dispute. Through the way of machine learning, now we are able to find and to probe into the data about corona-virus spreading patterns in a short period of time, suiting the remedy to the case, to launch targeted prevention policies, and minimize the economic loss under the premise of control the spread of the virus on a large scale. We directly use the LES algorithm and $K$-means clustering to make a comparison about the data feature. Therefore, the results are much more convincing than using any other recursive analyzing method alone. It is precise because of the ID3 algorithm, which we use for further analysis, to find the reason why those policies work.

1. Introduction

Since last year, human society had encountered one of the greatest obstacles in history: controlling the COVID-19 \cite{1} without showing down the economic growth. We wondered whether we could develop an evaluation system to demonstrate the efficiency of those control policies by only using small-sized data (like 10 or 20 days). If you only focused on the absolute value (each day cases) alone, the result might be confusing. Our goal is to find how state mask policies and various disease control policies influenced epidemic outbreaks in this project. To analyze the data set chosen before, we use the LES, $K$-means \cite{2}, and ID3 \cite{3} algorithm to fit the model we build specially in the project. For instance, the LES was used to showing the feature variation of the certain group data. Its vantage is easy to operate and has a relatively small amount of coefficient, which can be straightforward for further analysis. However, its shortcoming might be commonly under-fitting of the model. Furthermore, the $K$-means algorithm is the classifier when we need to sort the data in two different groups only due to its data feature in all dimensions. The advantage is that unsupervised learning simulates the control group environment, for its pattern is not defined before training, so no single label is weighted greater than others. Its disadvantage is the incapability to process scarce data. Moreover, the ID3 decision tree is defined to find the factor that most affects the effect of the Mask Mandates \cite{4}. Its merits and demerits are deemed to be its accurate presumption and the high chance of over-fitting. Our research work contains LES and $K$-means to evaluate the data feature and then compare each pair of data groups to conclude the final result. All data we collected from the internet are not intended for commercial use; therefore, there is no copyright issue. Precisely because of powerful machine learning tools, we can try to understand viruses and minimize their damage. The main contributions of this work can be summarized as follows:

1. The original data set is divided into separate groups in several different manners (whether based on the label) in order to obtain the data feature (such as variation or tendency), which would be used in...
the comparison.

2. Comparing and assessing the data feature to elucidate efficiency about the very label chose in the inchoate period.

3. Collecting other label efficiencies to make a new table about those policies and evaluate the factors which might influence this label.

The rest of this paper is organized as follows. Firstly, in Sect.2, we showed explicit methods we use to train the model, including LES, K-means, and ID3 algorithm. In Sect.3, we explained the data set features in charts and elucidated the results in tables and charts, which show the analyzing outcome explicitly. We used ACC as an evaluation of our classification results. Finally, Sect. 4 concludes this paper.

2. Methodology

2.1. Efficiency Evaluation of Mandating Masks

The original data set is two-dimensional, and the labels are time and the number of infected people, among which the time label needs to be regularized. Two groups of data are set: environmental group and control group. Since it is possible to get the actual date when the states announced blockades or imposed mandatory masks, the data can be classified in whether the date is before or after the policy promulgation and further processed using linear regression analysis (supervised learning). Next, the unsupervised learning k-means classifier is used to classify the original two-dimensional data set and further processed using gradient descent linear regression. Finally, the results of the two training sessions will be compared (data visualization) and made to test whether the government-mandated policies had a positive impact on epidemic control. If so, to what extent it had played its role.

2.1.1. LES Algorithm

Assuming that the independent variable is \( \overline{x}_i = [X_{i1}, \ldots, X_{im}] \) and the model's parameters are \( \overline{\beta} = [\beta_0, \ldots, \beta_n] \), then the model's prediction would be \( y_i = \beta_0 + \sum_{j=1}^{n} \beta_j \times X_{ij} \). If \( X_i \) is extended to \( \overline{x}_i = [1, X_{i1}, \ldots, X_{im}] \) then, \( Y_i \) would become a dot product of the parameter and the independent variable, i.e., \( y_i \approx \sum_{j=1}^{n} \beta_j \times X_{ij} = \overline{\beta} \overline{x}_i \). In the least-squares setting, the optimum parameter is defined as such that minimizes the sum of mean square loss:

\[
\overline{\beta} = \arg \min_{\beta} L(D, \overline{\beta}) = \arg \min_{\beta} \sum_{i=1}^{n} (\overline{\beta}, \overline{x}_i - y_i)^2. \tag{1}
\]

Now putting the independent and dependent variables in matrices \( X \) and \( Y \), respectively, the loss function can be rewritten as:

\[
L(D, \overline{\beta}) = \| X \overline{\beta} - Y \|^2 = (X \overline{\beta} - Y)^T (X \overline{\beta} - Y) = Y^T Y - Y^T X \overline{\beta} - \overline{\beta}^T X^T Y + \overline{\beta}^T X^T X \overline{\beta} \tag{2}
\]

As the loss is convex, the optimum solution lies at gradient zero. The gradient of the loss function is (using Denominator layout convention):

\[
\frac{\partial L(D, \overline{\beta})}{\partial \overline{\beta}} = -2X^T Y + 2X^T \overline{\beta} \tag{3}
\]

Setting the gradient to zero produces the optimum parameter:
\[
\frac{\partial L(D, \boldsymbol{\beta})}{\partial \beta} = 0 \Rightarrow -2X^T Y + 2X^T X \hat{\beta} = 0 \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T Y \tag{4}
\]

### 2.1.2. K-Means Clustering

The most common algorithm uses an iterative refinement technique. Due to its ubiquity, it is often called "the K-Means algorithm"; it is also referred to as Lloyd's algorithm, particularly in the computer science community. It is sometimes also referred to as "naive K-Means", because there exist much faster alternatives. Given an initial set of K-Means \( m_i^{(t)} \), the algorithm proceeds by alternating between two steps:

1. **Assignment step**: Assign each observation to the cluster with the nearest mean: that with the least squared Euclidean distance. (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means.)

\[
S_i^{(t)} = \{ \chi_p \mid \chi_p - m_i^{(t)} \leq || \chi_p - m_j^{(t)} ||^2 \forall j, 1 \leq j \leq k \}
\tag{5}
\]

where each \( \chi_p \) is assigned to exactly one \( S^{(t)} \) even if it could be assigned to two or more of them.

2. **Update step**: Recalculate means for observations assigned to each cluster.

\[
m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\chi_j \in S_i^{(t)}} \chi_j \tag{6}
\]

The algorithm has converged when the assignments no longer change. The algorithm is not guaranteed to find the optimum. The algorithm is often presented as assigning objects to the nearest cluster by distance. Using a different distance function other than (squared) Euclidean distance may prevent the algorithm from converging. Various modifications of k-means such as spherical k-means have been proposed to allow using other distance measures.

### 2.1.3 Information Gain

\[
IG(S, A) = H(S) - H(S \mid A) \tag{7}
\]

where \( H(S) \) is the entropy of the original set, which is the experimental group in our case. \( A \) is the subset from splitting set \( S \) by attribute \( A \), which is the third label in our data set-wearing mask or not. \( H(S \mid A) \) is the entropy of our control group.

Experimental group: add a label to the label list, such as before and after the promulgation of the mask policy, divide the original data into two groups, use linear fitting to reflect the features of the data set respectively, and compare the parameters of the fitting curve( \( K_1, K_2 \) ), and variation factor is \( \Delta K_e = |K_1 - K_2| \).

Control group: To exclude the influence of other factors, such as vaccination or large-scale sports events, the unsupervised classification method K-Means is adopted to directly classify the two-dimensional data. The original data is also divided into two groups this time. Linear fitting is used to reflect the features of the data set respectively, and the parameters of the fitting curve are compared ( \( K_1, K_2 \) ), and variation factor is \( \Delta K_c = |K_1 - K_2| \).

The difference of the fitting results of the experimental group and the difference of fitting results of the control group will be compared. Suppose the difference of the experimental group is greater than or far greater than the difference of the control group. In that case, it indicates that the information contained in this label, namely entropy reduction, is larger, and the policy corresponding to this label is more effective.

\[
IG(S, A) = H(S) - H(S \mid A) = \Delta K_e - \Delta K_c \tag{8}
\]
We use the rate of the relative change of information gain [5] in this case.

\[ RC = \frac{IG(S, A)}{H(S)} = \frac{\Delta K_e - \Delta K_s}{\Delta K_e} \]  \tag{9} 

![Fig. 1 The group structure we used in order to make a comparison about the data feature](image)

2.2. Assessment Analysis of Impact Factors for Policy Efficiency

Using the LES algorithm in the data before mask mandate and after mask mandate, we can obtain two slope \( K_2 \) and \( K_1 \), which means the increasing rate of new cases each day. To reduce \( K_2 \) by \( K_1 \), it becomes the efficiency parameter—the smaller the parameter, the better the mandate is. And the average efficiency parameter is -21.752030578117534, which means that mandates play positive roles in controlling the COVID-19 in general.

2.2.1. Decision Tree

The ID3 algorithm begins with the original set as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set \( S \). It calculates the entropy \( H(s) \) or the information gain \( IG(s) \) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set is then split or partitioned by the selected attribute to produce subsets of the data. (For example, a node can be split into child nodes based upon the subsets of the population whose ages are less than 50, between 50 and 100, and greater than 100.) The algorithm continues to re-curse on each subset, considering only attributes never selected before.

Recursion on a subset may stop in one of these cases:
(1) Every element in the subset belongs to the same class, in which case the node is turned into a leaf node and labeled with the class of the examples.
(2) There are no more attributes to be selected, but the examples still do not belong to the same class. In this case, the node is made a leaf node and labeled with the most common class of the examples in the subset.
(3) There are no examples in the subset, which happens when no example in the parent set matches a specific value of the selected attribute. An example could be the absence of a person among the population with age over 100 years. Then a leaf node is created and labeled with the most common class of the examples in the parent node's set.

Throughout the algorithm, the decision tree is constructed with each non-terminal node (internal node) representing the selected attribute. The data was split and terminal nodes (leaf nodes) representing the class label of the final subset of this branch.
3. Experiments

3.1. Datasets
In the experiments, we adapted 5 datasets (Mask, US pop, Density, Hospital, and Temperature) to build the model, which is summarized as follows:

1. Mask dataset: dataset contains 4 features, in which date and new cases per day are used in the original fitting and whether wearing the mask is included in the control group fitting.
2. US pop dataset: dataset contains 1 feature, including the population's proportion of each state.
3. Density dataset: dataset contains 1 feature, including the population density of each state.
4. Hospital dataset: dataset contains 2 features, including the number of hospitals and staffed beds in each state.
5. Temperature dataset: dataset contains the average temperature of each state and its ranking in the country.

| Dataset      | Element Indication                        |
|--------------|-------------------------------------------|
| Mask         | Date variation                            |
| US pop       | The population's proportion of each state  |
| Density      | The population density of each state       |
| Hospital     | The number of hospitals and staffed beds in each state |
| Temperature  | The average temperature of each state and its ranking in the country |

Note that all experiments are conducted on a computer with 1.99GHz Intel i7-8550U processor and 16GB RAM under Windows 10 operating system. The program codes of data processing and graphs modeling are written by Anaconda Python 3.8.0, which is available on https://anaconda.en.softonic.com/

3.2. Evaluation Metrics
In this study, ACC is adapted to measure the performance of experimental results, which can be calculated by:

\[
ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}
\]

where TP is the prediction classified as positive has been proved to be true, FN is the prediction classified as negative has been proved to be false. The prediction classified as positive has been proved to be falsely referred to as FP, and TN is the prediction classified as negative has been proved to be true.

3.3. Experimental Results and Analysis

3.3.1. Efficiency Evaluation of Mandating Masks
The example of evaluation result is summarized as follows:
Fig. 2 (a) G1 ($y_1 = 29.91868x + 793.6$), (b) G2 ($y_2 = -37.95165x + 878.82857$)

Firstly, we used the LES algorithm to calculate the fitting curve of each data set group. (The group classification method is shown in section 2.1.3.)

Then we computed the $K$, which is the variation factor of the mask label data set (experimental group), for further comparison.

$$\Delta K_e = H(S) = |K_1 - K^2| = 67.87032967032962$$  \hspace{1cm} (11)

Fig. 3 G3 ($y_3 = -11.13278816x + 853.98948598$)
Similarly, we used K-Means to divide the original data set group into two groups and then get the fitting curve of each group using LES.

Moreover, we calculate $\Delta K_c$, which is the variation factor of the control group.

$$\Delta K_c = H(S | A) = |K_1^2 - K_2^2| = 5.02087717$$ (12)

Finally, we got the rate of the relative change of information gain in this case.

$$RC = \frac{IG(S, A)}{H(S)} = \frac{\Delta K_c - \Delta K_e}{\Delta K_e} = \frac{67.87 - 5.02}{67.87} = 92.6\% >> 50\%$$ (13)

According to the final RC result, which is much greater than 50%, we can conclude that the policy corresponding to this label (mask mandatory) is indeed effective.

The table shows our judgement of each state's policy worked or not, only used about 10 days of data after the mask mandatory policy promulgation. In the first few days after the policy being issued, it is hard to tell whether the policy is efficacious or not. Because our data is short-term, or in other words, inadequate [6]. Suppose we only focus on the absolute number or trend of each state's daily increase of cases. In that case, we might get some really baffling conclusion. For instance, mask policy doesn't help to control the epidemic at all, or even, on the contrary, intensify it.

Our code final-LES.ipynb and final-kmeans generate these state results. ipynb, saved in the text file infection rate LES.txt and infection rate kmeans.txt. Using EXCEL for further processing the results and results are as follows.

states average: $\Delta K_{e,av} = 67.0925$

states average: $\Delta K_{c,av} = 22.7526$

$$RC_{average} = \frac{IG(S, A)}{H(S)} = \frac{\Delta K_c - \Delta K_e}{\Delta K_e} = \frac{67.0925 - 22.7576}{67.0925} = 66.08\% > 50\%$$ (14)

It indicates that mandating masks had a positive impact on epidemic control. On the whole, it had played its role pretty well. We can see that plurality of states' policy is effective, and we can sum up that based on the absolute number of cases per day of each state. An only small quantity of states prediction shows that this policy was effective. It might abate public confidence about this policy and reluctant to fully execute it, which might make the situation even worse [7].

In order to find out which method is more accurate to the real condition happened in each state after mask mandatory, we looked up the epidemic data of each state on New York Times six month after this policy committed, we then got the actual mask policy efficiency of each state based on the curve of the case [8]. All of the state's new cases per day curve had been saved in our project as a screenshot, which
is available for checking in our project file, or you might view it on the official website of the New York Times.

The accuracy of absolute number:

$$\text{ACC}_{ab} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{11+3}{23} = \frac{14}{23} = 60.87\%$$  \hspace{1cm} (15)

The accuracy of our method:

$$\text{ACC}_{our} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{18+3}{23} = \frac{21}{23} = 91.30\%$$  \hspace{1cm} (16)

To sum up, our method significantly improved the accuracy of judgement on the efficiency of mandating masks policy.

3.3.2. Find the factor that most affect the effect of the Mask Mandates

Using the LES algorithm in the data before mask mandate and after mask mandate, we can get two slopes $K_2$ and $K_1$, which means the increasing rate of new cases each day. To reduce $K_2$ by $K_1$, it becomes the efficiency parameter-the smaller the parameter, the better the mandate is. And the average of the efficiency parameter is -21.752030578117534, which means that mandates play positive roles in controlling the COVID-19 in general.

Efficiency parameter of each state:

Using the efficiency parameter, we set up a uniform standard to classify each state into good, normal, and bad three categories, representing the effect of the mask mandate in each state.

**Classification standard of the effect of mask mandate:**

Efficiency parameters $<=$-10: good
-10<$\text{Efficiency parameters}$<10: normal
Efficiency parameters $>$=10: bad

**Distribution of 3 categories:**

![Fig. 5 Efficiency parameter of each state](image)

![Fig. 6 Distribution of 3 categories](image)
To find out why the mask mandates have a different effect in a different state, we choose four relevant factors - temperature, staffed bed, infection rate, and population density as the attribute of the effect of mask mandates. We use the ID3 algorithm to draw a decision tree to find the most influential factor in the mask mandate.

**Reasons for choosing four factors:**

Temperature: Viruses can survive longer in cold temperatures, which increases the chance that people will be exposed to them.

Staffed bed: More staffed beds can contain more patients in the hospitals instead of isolating at home.

Infection rate: The higher the infection rate, the greater the chance that people will come into contact with a sick person.

Population density: High density means that it's more difficult for people to keep social distancing.

**Classification standard of four factors:**

**Temperature:**

To get the relative temperature difference in each state, we divide the states into three categories by their temperature ranks.

- Rank <=16: Hot
- 17<Rank<=34: Warm
- Rank>34: Cold

**Staffed bed:**

Use the state's total cases in a month to divide the summation of staffed beds [9] in the state, then get the cases per bed. Then use all the cases per bed in a different state to classify each state. Cases per bed<=1.5: Vacant

- 1.5<Cases per bed<=3.5: Normal
- Infection>3.5: Redundant

**Infection rate:**

Use the state's total cases in a month to divide the state's population, then get the infection rate. Then use all the infection rates in different states to do a normalization. Finally, use the updated infection rate to classify each state.

- Infection rate <=0.2: Low
- 0.2<Infection rate<=0.45: Medium
- Infection>0.45: High

**Population density:**

To get the relative population density difference in each state [10], we divide the states into three categories by their number of residents per square mile.

- Number <=100: Low Density
- 100< Number <=250: Normal Density
- Number >250: High Density

Use all of the attribute and the effect of mask mandates to produce a table:

**Decision tree:**
Fig. 7 Illustration of decision tree

Through this decision tree, we can find out that infection rate is the most influential factor to the effect of mask mandate. Besides, we can also see the influence of temperature, population density, and cases per bed on the effect of mask mandate. According to the final decision tree, we can help the government develop more reasonable solutions to control the COVID-19.

4. Conclusion

This paper proposed a method containing several models to fit the original data and make the comparison. For each part of our data set, we use a suitable model to fit the model. Among which the coefficient and the purpose of each learning model are combined delicately. We set the experimental and control group during the analysis phase to minimize the noise (control variate method). It might seem like a rudimentary model of ensemble learning. Because our data demonstrate only a month before and after the mandating of mask policy, we could easily conclude that plurality of state's policy is inefficient based on absolute cases tendency. According to the final result, some states we originally reckoned as a total failure based on the absolute cases per day implementing mask policy have a silver lining in our final evaluation. Our predictions show that even though the new cases per day are still increasing in some states, the policy is still highly efficient, which suggests that the trend is going downward in the future. After checking up the following data of those states, we confirmed our prediction in most cases. We can conclude that the model we select is proved to be robust and slightly under-fitting.

In the future, we could load the data set more efficiently. For instance, data can be processed by a normal distribution model to conduct variance analysis, test for homogeneity function of variance, and student’s T test. Furthermore, regression analysis that is closer to the biological curves model can be used for data curve fitting, such as J-shaped curves and S-shaped curves, which may have a more significant correlation in parametric (feature) analysis.

References

[1] Donsimoni, J. R., et al. "Projecting the Spread of COVID-19 for Germany," IZA Discussion Papers, 2014

[2] X. Huang, Y. Ye, and H. Zhang, "Extensions of Kmeans-Type Algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercaluster Separation[J]," IEEE Transactions on Neural Networks & Learning Systems, vol. 25, no. 8, pp. 1433-1446, 2014

[3] Gaddam, S. R., V. V. Phoha, and K. S. Balagani. "K-means+id3: A Novel Method for Supervised Anomaly Detection by Cascading k-Means Clustering and id3 Decision Tree Learning"
Methods." IEEE Trans KNOWL DATA ENG, vol. 19, no. 3, pp. 345-354, 2007

[4] R Gatter, and S. Mohapatra. "COVID-19 and the Conundrum of Mask Requirements." Social Science Electronic Publishing.

[5] Kent, JOHN T. "Information gain and a general measure of correlation." Biometrika 1(1983):163-173.

[6] Terejanu, G. "Active data collection for inadequate models. "International Conference on Information Fusion IEEE, 2015.

[7] Guglielmi, S., et al. "Public Acceptability of Containment Measures During the COVID-19 Pandemic in Italy: How Institutional Confidence and Specific Political Support Matter." Social Science Electronic Publishing.

[8] Basch, C. H, A. Kecojevic, and V. H. Wagner. "Coverage of the COVID-19 Pandemic in the Online Versions of Highly Circulated U.S. Daily Newspapers." Journal of Community Health (2020):1-9.

[9] Hassan, and Jenifa. "Monte Carlo simulation model of availability of staffed beds in urban hospitals during pandemic periods." Journal of Historical Pragmatics 14.2(2010):328-333(6).

[10] Griffith, D. A., and D. W. Wong. "Modeling population density across major US cities: a polycentric spatial regression approach." Journal of Geographical Systems 9.1(2007):53-75.