2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)

2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA) | 978-1-7281-9004-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICPECA51329.2021.9362507

Global Epidemic Classification Based on K-Nearest Neighbor Algorithm

Jinhua Du
China University of Geosciences (Beijing)
School of Information Engineering
Beijing, 100083, China

***Abstract*—K-neighbor algorithm is an effective way to classify things according to their characteristics. Novel Coronavirus epidemic development and response situation of various countries around the world can be classified to simplify the difficulty of epidemic prevention and control and provide useful information for them. Based on the data of WHO and the existing research results at home and abroad, this paper analyzes the evaluation indexes of national epidemic situation, selects representative countries as training samples, and points out the epidemic situation types of corresponding countries. It aims to provide reference for relevant researchers in epidemic prevention and control and trend prediction through global epidemic classification, and hopes to improve relevant technologies in further research, so as to make the classification more scientific and accurate.**

Keywords—2019-nCoV, K-nearest neighbor algorithm, classification of the outbreak

I. Introduction

COVID-19 is an acute infectious pneumonia, and its pathogen is SARS-CoV-2 which has not been found in humans before. According to the current epidemiological investigation, COVID-19 has an incubation period of 1-14 days. During this period, novel coronavirus was highly contagious, mainly spread through respiratory droplets and close contact, and people were generally susceptible.

Since February 26, 2020, the number of newly confirmed cases in a single day in foreign countries exceeded that in China for the first time, the COVID-19 epidemic situation at home and abroad gradually showed different development trends. It can be seen from the right picture that, unlike the basic control of the epidemic situation in China, the epidemic situation in foreign countries generally presents an outbreak situation.

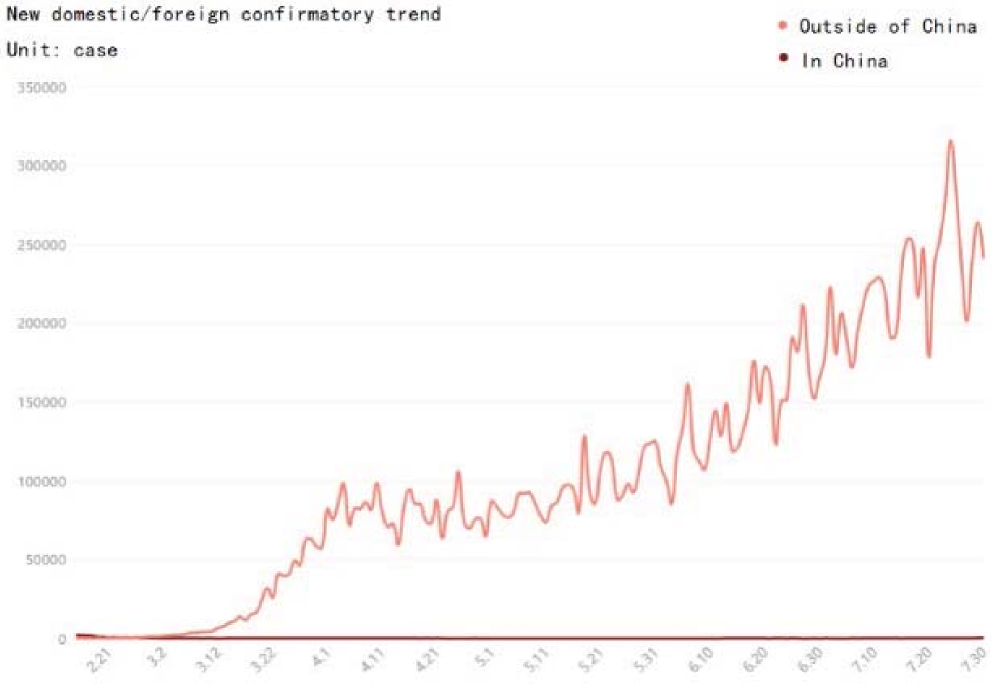


Fig. 1. Changes in the number of new confirmed cases at home and abroad from February 15 to July 31, 2020

Consider putting forward some feasible suggestions for countries in order to control the unhealthy development of the global epidemic as soon as possible. However, for 222 countries and regions in the world, due to the differences of subjective and objective factors, the epidemic development characteristics and the situation of fighting against the epidemic are different in different countries, so it is unrealistic to analyze and deal with each situation.

Therefore, it is necessary to classify countries and select representative countries from each type as research objects. Further, from the perspective of types, this paper analyzes the current epidemic development and anti-epidemic situation of

Shenyang, China
January 22-24, 2021

specific countries, so as to facilitate the evaluation and prediction of epidemic prevention and control in different countries according to the types.

II. Problem analysis

In order to analyze the epidemic data of some major countries in the world in a certain period of time, it is necessary to visualize the data as a change curve. Considering that the fluctuation of data with time is too complicated, and this paper mainly studies the change trend of data, we can appropriately ignore the details of local data changes and use the least square fitting method to obtain the approximate change curve of source data.

Then, we can study the development of epidemic situation in China according to the tangent slope of the change curve at each point (i.e. the first derivative of the corresponding function at the corresponding point), and we can also study the fight against epidemic situation in China according to the concavity and convexity of the change curve (i.e. the second derivative of the corresponding function). Finally, the characteristics of epidemic development and anti-epidemic situation in different countries are regarded as the characteristics of corresponding countries, and countries can be further classified by clustering and other related algorithms.

Because the above classification of countries is based on the classification characteristics obtained according to the change curve, and the classification characteristics of some countries in the world are typical, it is easy to create an initial model training set, and then use this training set to classify other atypical countries, that is, use K- nearest neighbor algorithm to achieve classification.

III. Research object

In Chapter 1, what still needs to be clear is how to define the main countries in the world, that is, how to select the research object to make the model training set more typical, more comprehensive and more representative. Several irrelevant characteristic indexes can be obtained by principal component analysis [1], and seven representative countries from 193 countries can be selected as basic research objects.

TABLE I. Major countries in the world and their characteristic indicators

No.	Country name	Population/ten thousand	Location.	Medical level	Economic situation/USD	State system	Nucleic acid detection rate	Mask price/yuan
1	China	143981	East Asia	78	10261	Socialist Republic	5.27%	3
2	U.S.A	33117	North America	89	65280	Presidential Republic	3.25%	209
3	India	139120	South Asia	41	2104	parliamentary republic	0.15%	15
4	Germany	8380	Europe	92	49181	parliamentary republic	3.76%	72
5	Italy	6046	Europe	95	34671	parliamentary republic	4.86%	76.2
6	Iran	8408	West asia	72	5103	Unification of the state and the church	0.80%	18
7	Spain	4675	Europe	92	32023	Constitutional monarchy	4.11%	2200

Among them, the medical level parameter is the HAQ index of the corresponding country [2], which is used to reflect the comprehensive medical level of the country. The economic situation is the per capita GDP of the country in 2019, and the nucleic acid detection rate is the ratio of the number of people detected in the corresponding country to the total population. The price of masks is the price of masks of the same type in various countries, and its ratio with economic situation can reflect the shortage of personal protective materials in the country.

These indicators comprehensively cover five aspects of a country's economic level, political situation, medical conditions, material reserves and attitude to deal with the epidemic, ensuring that the selected research objects meet the research requirements.

IV. Model hypothesis

Hypothesis 1: The starting point of COVID-19 epidemic research around the world is February 26th, 2020.

Hypothesis 2: The newly added diagnosis data of countries around the world can effectively reflect the development of the epidemic situation in the country.

V. Symbol description

TABLE II. The notation used for the model

No.	Symbol	Significance
1	{x}	The arithmetic sequence corresponding to the time series starting from 1 and increasing by 1
2	S	The least squares fitting function of the data point corresponding function
3	dS	The first derivative of S
4	d^2S	Second derivative of S
5	C	K-nearest neighbor algorithm input feature vector
6	C_1	The first component of C is the number of points dS<0 in the specified domain
7	C_2	The second component of C is the number of points dS>=0 in the specified domain
8	C_3	The third component of C is the number of d^2S <0 points in the specified domain
9	C_4	The fourth component of C is the number of d^2S >=0 points in the specified domain
10	T	Corresponding types of feature vectors output by k-nearest neighbor algorithm

VI. Modeling

A. Overview of K-nearest Neighbor Algorithm

K-NearestNeighbor(kNN) algorithm is a basic classification and regression method. The input is the feature vector of the instance, which corresponds to the point of feature space. The output is the category of the instance, which can take multiple categories. K- Nearest Neighbor Algorithm assumes that given a training data set, the instance category in it has been determined. When classifying, a new example is predicted by majority voting according to the categories of its K nearest neighbor training examples.

B. An Overview of Epidemic Classification Models

To simplify the complexity of the model, the input of the model is the daily number of confirmed cases of COVID-19 in a country from February 26th, 2020, and the output is the epidemic type of the country.

In the model, the change curve corresponding to the input data set should be smoothed to obtain a smooth curve function, and the curve characteristics should be obtained by calculating the first and second derivatives of the curve function. According to the curve eigenvalue and the training data set of k- nearest neighbor algorithm, the model output is obtained by k- nearest neighbor algorithm.

C. Generating Training Data Sets

According to WHO Director-General Tedros Adhanom Ghebreyesus's criteria for the classification of global epidemic countries, countries around the world can be divided into four categories. The representative countries in these four categories are China, America, Germany and Iran, so the training data set can be composed of these four countries. Since the processing logic for each country is consistent, the following will take the data processing for the United States as an example.

In order to make the data continuous and smooth, we need to find a simple and easy-to-calculate function to approximate the changing trend of data points, that is, curve fitting, when we know the two-dimensional data points (date, newly diagnosed cases on the same day). In order to quantitatively explain that the fitting effect is the best in practice, it is necessary to introduce the criterion that the sum of squares of the difference between the new functions at known points is the smallest, and make this criterion reach the minimum value within the target range, that is, use least square fitting.

Because the purpose is to obtain the curve of the number of newly added cases with time, the emphasis is on change, not time. To simplify the model, consider preprocessing the abscissa date in the data points, and replace the specific time with an ordered sequence {x} such as {1, 2, 3...}.

In the least square fitting, if the mathematical model is selected as polynomial, the nth least square fitting polynomial of the corresponding function of the original data point can be defined as:

$$\Phi = H_n = \mathrm{span}\{1, x, \ldots, x^n\} \tag{1}$$

$$\varphi_i = x^i \tag{2}$$

The corresponding normal equation is:

$$\begin{bmatrix} \sum_{i=0}^{m}\omega_i & \sum_{i=0}^{m}\omega_i x_i & \cdots & \sum_{i=0}^{m}\omega_i x_i^n \\ \sum_{i=0}^{m}\omega_i x_i & \sum_{i=0}^{m}\omega_i x_i^2 & \cdots & \sum_{i=0}^{m}\omega_i x_i^{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=0}^{m}\omega_i x_i^n & \sum_{i=0}^{m}\omega_i x_i^{n+1} & \cdots & \sum_{i=0}^{m}\omega_i x_i^{2n} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{m}\omega_i f_i \\ \sum_{i=0}^{m}\omega_i x_i f_i \\ \vdots \\ \sum_{i=0}^{m}\omega_i x_i^n f_i \end{bmatrix} \tag{3}$$

The n-th least square fitting polynomial of the corresponding function of data points is:

$$S(x) = \sum_{i=1}^{n} a_k x^k \tag{4}$$

Take n=4. After calculation, the curve function of the number of newly confirmed cases in the United States every day is:

$$\mathrm{S}(x) = -0.0003x^4 + 0.1897x^3 - 30.6798x^2 + 1.8816x - 1.5529 \tag{5}$$

From its corresponding change curve, we can see the development characteristics of the epidemic situation and the situation of fighting against the epidemic situation in the United States:

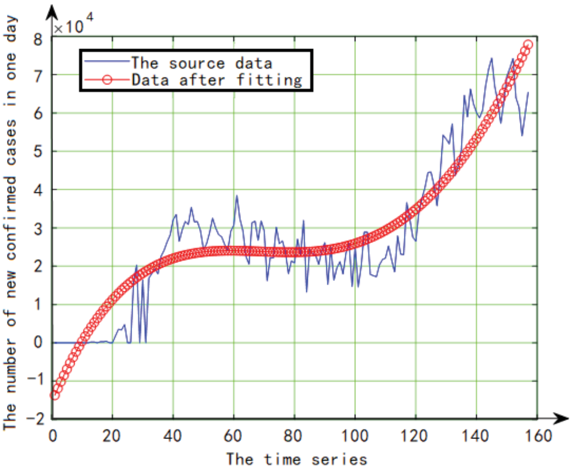


Fig. 2. Change in the number of new confirmed cases in the US from February 15 to July 31, 2020

The epidemic situation in the United States continues to rise, and has already ushered in the second peak. In the whole epidemic development process, the anti-epidemic effect was hardly reflected.

In order to quantify the above two characteristics of epidemic development and anti-epidemic effect, the first-order and second-order derivatives are taken from the curve function of the number of newly confirmed cases in the United States every day. In the definition domain, the specific characteristic information can be reflected by the magnitude of the first-order and second-order derivative values:

D. Overview of K-nearest Neighbor Algorithm

K-Nearest Neighbork- nearest neighbor (kNN) algorithm is a basic classification and regression method. The input is the feature vector of the instance, which corresponds to the point of feature space. The output is the category of the instance, which can take multiple categories. K- Nearest

Neighbor Algorithm assumes that given a training data set, the instance category in it has been determined. When classifying, a new example is predicted by majority voting according to the categories of its K nearest neighbor training examples.

TABLE III. COMPARISON TABLE OF RELATIONSHIP BETWEEN DERIVATIVE OF CHANGE CURVE AND EPIDEMIC CHARACTERISTIC VALUE

dS	d^2S	Characteristics of epidemic development	Characteristics of anti-epidemic effect	Type
Constant > 0	—	Increase in new quantity	No obvious effect	A
Constant < 0	—	Decrease in new quantity	The effect is very good	B
Have change	Constant < 0	The newly added quantity increases first and then decreases	Effective	C
Have change	Have change	The new quantity increases first, then decreases and then increases	Repetition	D

Among them, the number of newly confirmed cases decreased first and then increased from the beginning, and the two increases and decreases did not exist in practice so far, so it was not considered. At the same time, in order to facilitate the subsequent classification, these four cases are defined as A, B, C and D respectively.

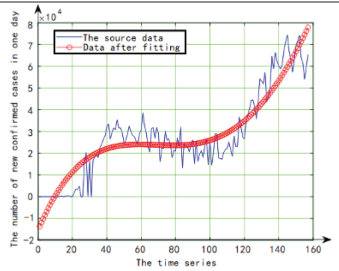
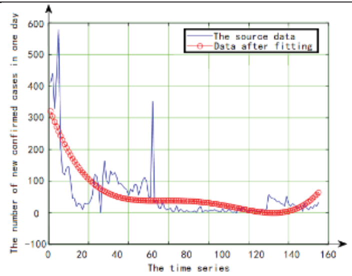
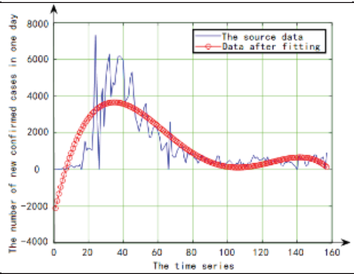
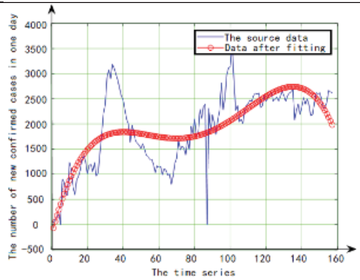
For the United States, the first derivative of the change curve is:

$$dS = - 0.0012x^3 + 0.5692x^2 - 61.3596x + 1881.5938 \quad (6)$$

Within the allowable range of error, dS>0 is always satisfied, that is, the overall development of the epidemic situation is continuously rising and the anti-epidemic failure, which belongs to Class A epidemic country. This is consistent with the qualitative description of the curve above.

By the same token, we can calculate and summarize the situations in China, Germany and Iran.

TABLE IV. ANALYSIS OF THE CURVE AND CHARACTERISTICS OF THE EPIDEMIC IN THE UNITED STATES, GERMANY AND IRAN

Country	Change curve	Features	Type
U.S.A		In the whole process, it has been in the intensive outbreak period, and no obvious effect of anti-epidemic has been seen.	Class A
China		At the beginning of the global outbreak, it showed good anti-epidemic effect and effectively restrained the development of COVID-19 virus.	Class B
Germany		National compulsory prevention and control measures, WHO's experience in epidemic prevention and control shared around the world, and people's trust in the government ensure that the epidemic is controlled after the outbreak.	Class C
Iran		Relaxation after the first peak of the epidemic, led to the arrival of the second peak, and even made the epidemic grow more rapidly.	Class D

Obviously, the criteria in Table II are too idealistic and difficult to achieve in practice. In the allowable range of error, if we take the first derivative and second derivative values less than 0 as statistical basis, we can obtain the number of data points with the first derivative less than 0, the first derivative not less than 0, the second derivative not less than 0 and the second derivative not less than 0 among all corresponding data points in the epidemic situation curve of a certain country. It can be used as an index to judge the type of anti-epidemic countries.

TABLE V. QUANTITATIVE DATA AND MATHEMATICAL CHARACTERISTICS OF COVID-19 CHANGE CURVES IN THE UNITED STATES, GERMANY AND IRAN

Country	C1	C2	C3	C4	Features	T
	dS<0	dS>=0	d^2S <0	d^2S >=0		Type
U.S.A	20	137	68	89	Most of the first derivative is > =0, the second derivative is flat, and the curve increases monotonously	Class A
China	107	50	38	119	The first derivative is mostly < 0, the curve decreases monotonously, the second derivative is mostly > =0, and the curve decreases monotonously	Class B
Germany	87	70	94	63	The first derivative is flat and single peak, the second derivative is mostly < 0, and the curve increases first and then decreases	Class C
Iran	52	105	101	56	The first derivative is mostly > =0, the curve increases monotonously, the second derivative is mostly < 0, the curve increases monotonously and the curve rises twice	Class D

The above table is the training data set of k- nearest neighbor algorithm.

E. KNN Classification

According to the basic principle of KNN algorithm, after the model training set has been constructed by some samples of known classes (represented by all feature vectors), for a new sample of unknown class, k neighbor samples with the smallest distance can be found. By looking at which class these k neighbors belong to, we can think that the new sample also belongs to that class.

Among them, the solution method of feature vector c of new samples of unknown class is the same as that in Section 4.4, but the calculation method of distance and the value of k still need to be determined.

There are many ways to calculate the distance. To simplify the model, we use Euclidean distance, that is, the spatial distance between two points is L2 norm of vector difference between two points. Therefore, the euclidean distance between two n-dimensional vector sums can be expressed as:

$$d_{12} = \sqrt{\sum_{k=1}^{n} (x_{1k} - x_{2k})^2} \tag{7}$$

However, the selection of k value is complicated: because there is only one feature vector corresponding to each type in the training data set, k can only take 1 now. When the KNN algorithm is implemented, it only needs to find the category corresponding to the nearest feature vector, which means that the category corresponding to the new sample is the same.

After using this model to process certain samples again, the corresponding types of these samples can be modified according to the actual situation and added to the training data set of this model.

When the sample size in the training data set of this model exceeds 2/3 of the countries in the world, cross-validation can be used to determine the most suitable K value, that is, Holdout validation method: take the available K values one by one, use this K value to establish KNN model, and predict the types of the remaining 1/3 countries. If it is inconsistent with the facts, sum the squares of errors (i.e., if the type is inconsistent, the error is 1, and if the type is consistent, the error is 0), and finally take k value to minimize the square sum of errors. The k value at this time can ensure that the model is neither under-fitted nor over-fitted. The following description assumes that the optimal k value has been obtained.

Let the number of samples in the existing model training set be n, and the feature vector of the new sample be NC. The distance vector can be obtained by substituting the feature vector of each existing sample and the feature vector of the new sample into the Euclidean distance calculation formula, where each component represents the Euclidean distance between the old and new samples.

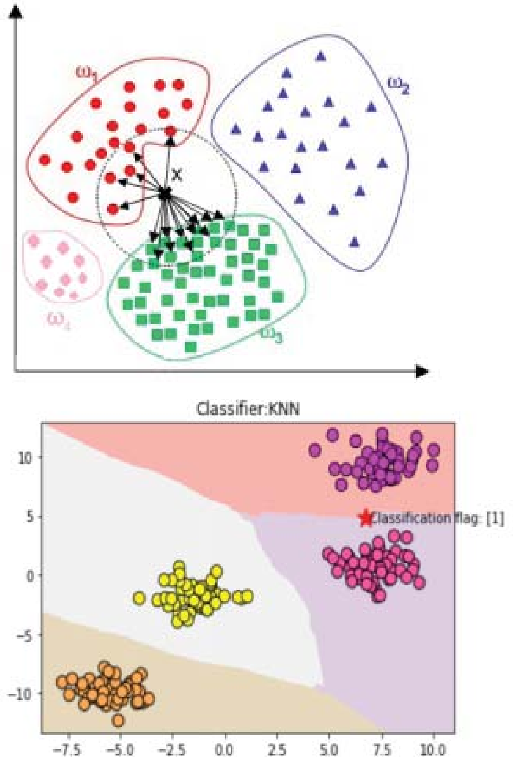


Fig. 3. KNN algorithm results schematic diagram [4]

On the premise of not changing the training set order of the original model and the components of the distance vector, the components, that is, the distance between the samples in the training set and the samples to be measured, are sorted from small to large, and the sorting results are recorded in the new vector.

Take the original samples corresponding to the first k components, that is, k neighbor samples with the smallest distance from the current sample, and determine the frequency f of each category in these k samples. The class with the highest frequency f is used as the prediction classification of the sample.

VII. Model Checking

A. Quote

Among the major countries in the world selected for this modeling, India, Italy and Spain will be used to test the accuracy of this model, except the United States, China, Germany and Iran as training sets of model data.

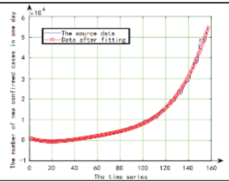
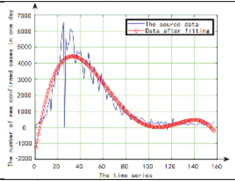
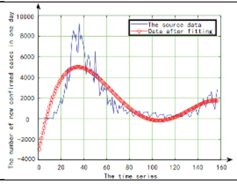
The data of newly diagnosed cases in these three countries from February 26 to July 31, 2020 are respectively substituted into the least square fitting algorithm for four times, but the corresponding feature vectors are:

TABLE VI. Corresponding feature vectors of the epidemic change curves of India, Italy and Spain

Country	Feature vector C
India	[19,138,0,157]
Italy	[93,64,93,64]
Spain	[74,83,87,70]

Substituting the feature vectors into KNN classification implementation code one by one, the corresponding prediction classification types can be obtained. Compared with the type reflected by the actual change curve, it can test the accuracy of the national classification model of epidemic situation established in Question 1.

TABLE VII. Quantitative data and prediction of epidemic change curves in India, Italy and Spain

Country	Change curve	Features	Actual type	Forecast type	Error
India		In the whole process, it has been in the intensive outbreak period, and no obvious effect of anti-epidemic has been seen.	Class A	Class A	0
Italy		National compulsory prevention and control measures, WHO's experience in epidemic prevention and control shared around the world, and people's trust in the government ensure that the epidemic is controlled after the outbreak.	Class C	Class C	0
Spain		Relaxation after the first peak of the epidemic, led to the arrival of the second peak, and even made the epidemic grow more rapidly.	Class D	Class C	1

In the above tests, the accuracy of the model is 67%, and there is still room for improvement. However, a careful analysis of the sample of "Spain" causing the error shows that the variation curve can be classified into category C within the allowable error range. However, the actual definition of the secondary outbreak peak is not considered in the model accurately, which leads to the deviation between the predicted value and the real value. Generally speaking, this model has achieved the expected goal of modeling.

VIII. Sensitivity analysis of model

Among the parameters of the model, the training data set can be measured by the actual situation, so the certainty is large. However, K value, that is, the number of neighbors selected by KNN algorithm can not be accurately defined, so it is reasonable to doubt whether the model is applicable to all k values.

Due to the lack of data in the training data set, the K value can only be taken as 1. However, after using the model to make a certain prediction and expand the size of the training data set, it will be found that the prediction classification results change obviously, so it can be considered that the prediction type is sensitive to K [5].

In order to quantify the sensitivity relationship between prediction type and k, the relationship between prediction type error and k can be actually plotted. Because there is no definite quantitative relationship between them, differential cannot be used to describe the relationship. However, by observing the relationship diagram, it can still be concluded that the K value which minimizes the classification prediction error is the best realization of the model, which has achieved the goal of modeling.

IX. Model evaluation

The whole idea of national classification model of epidemic situation is to use least squares to fit the data change curve, obtain the feature vector according to the curve characteristics, and use KNN data training set to find the most possible type corresponding to the newly obtained feature vector.

The whole model has clear logic, reasonable structure, scientific usage and good performance in practical application. There are two innovative features: first, preprocessing the initial data, making the data continuous and smooth, and cutting off unnecessary complexity. Second, KNN is used for classification. The data training set of the model can be continuously expanded during the gradual use of the model, which makes the classification results more accurate and the model has good self-optimization function.

However, the model still has many disadvantages: firstly, the data training set needs to be verified manually to ensure the accuracy of the data, and to a certain extent, the model needs to invest more labor costs. Secondly, the model has less data in the initial training set, which makes the error of classification results unable to reach the best level. Third, the selection of each component of feature vector in the model can not restore the epidemic characteristics of corresponding countries to the greatest extent, which makes the accuracy of the results of model components questioned.

X. Model improvement

In view of the disadvantages of the above model, there are two ways to improve: first, expand the dimension of feature vector, and find the variable that can better describe the characteristics of the change curve as the new feature vector component. The second is to expand the size of the model training set, which can be expanded by using the model for 3-5 rounds of classification, each round of classification of 4-8 countries, and adding effective classification results to the training set. It is believed that the accuracy and reliability of the model will be greatly improved after further improvement.

XI. Conclusion

In order to cope with the severe situation of global epidemic prevention and control, reduce the difficulty of work. In this paper, the data of countries provided by WHO are processed, and the effective numerical analysis methods such as principal component analysis, least square fitting and K-nearest neighbor algorithm are used to establish the national classification model of epidemic situation. Taking the number of newly added countries in a single day in a period of time as the model input, the corresponding epidemic country type of the country is exported. It is hoped that it can help the relevant personnel to think about prevention and control schemes according to the corresponding types first, and apply them to many countries of the corresponding types, so as to improve work efficiency and reduce cost input.

References

[1] Wang Xuemin. Applied Multivariate Statistical Analysis (5th edition) [M]. Shanghai: Shanghai University of Finance and Economics Press, 2017:201-224. (Wang Xuemin. Applied Multivariate Statistical Analysis (5th edition) [M]. Shanghai: Shanghai University of Finance and Economics Press, 2017:2017-224.)

[2] Prof Rafael Lozano, Nancy Fullman ,Jamal Yearwood. Measuring performance on the Healthcare Access and Quality Index for 195 countries and territories and selected subnational locations: a systematic analysis from the Global Burden of Disease Study 2016 [J]. The Lancet, 2018, 391 (10136):2236-2271.

[3] China.com.cn. Who Director-General TEdros Adhanom Ghebrev: Countries around the world are divided into four epidemic states [EB/OL].[2020-07-14].http://med.china.com . cn/content/pid/191152/tid/3

[4] pengjunlee. The KNN adjacent classification algorithm of machine learning [EB/OL]. [2018-09-15]. https://blog.csdn. net/pengjunlee/article/details/82713047.

[5] Meerschaert, Mathematical Modeling Method and Analysis [M]. Beijing: Machinery Industry Press, 2014:12. (Meerschaert, Mathematical Modeling Method and Analysis [M]. Beijing: Machinery Industry Press, 2014:1