Study impact the latitude on Covid-19 spread virus by data mining algorithm

Ali Dawood Salman¹, Hasanain Ali Daway AL-farttoosi ², Abbas Jkhayyir Kadhim ³

¹Eotvos Lorand University- Hungary
²Halfaya Oil Field,Missan Oil Company, Missan, Iraq
³Amara technical institute
E-mail: abbasjk@stu.edu.iq; salman@inf.elte.hu; hasanain.daway@gmail.com

Abstract

Corona virus disease (COVID-19) is an infectious caused by a new virus, the virus causes respiratory disease with symptoms such as coughing and fever, causes pneumonia in more severe cases. problem statement: The new Corona virus, or what has become known as "COVID-19", has spread to more than 79 countries outside China, where the Wuhan region is the epicenter of the virus.Until now researchers not discovered vaccine COVID-19. For prevalence observation there are countries has spread the virus significantly, others in an average and other less, with countries until now where the virus hasn't spread. proposed solution: This research based on studying the impact geographical latitude on knowing the spread of the COVID 19 in confirmed cases. result: To help scientists and researchers whom still to work on discovery the vaccine, taking importance of the zones the spread of the Covid-19. Using Dataset from Kaggle by classification algorithms linear regression data mining to extract a knowledge from data beneficial.

Keywords: COVID-19, latitudes, Data mining, deep learning, Erlang Functional language.

1. Introduction

The outbreak of the epidemic novel of covid-19 in Wuhan , china has suddenly spread to the world since Dec 2019 [1]. the World Health Organization (WHO) has named The infectious disease which is caused by the virus as covid-19 on Feb 11, 2020 [2], all the suspected cases should be confirmed by the real-time polymerase chain reaction (RT-PCR) assay of the sputum [3], although it is the best way for diagnosis, confirming COVID-19 patients using RT-PCR it takes time and has been Advertised them suffer from high spurious negative rates. Many researchers have accomplished several studies about the effect of that virus on sex (a study has done by the Chinese Center for Disease Control, it says that men are more affected by Corona virus than women, the researchers saw that the death rate was 2.8 percent between the men while it was 1.7 percent between the women also Women has more
immune responses to infection, as the scientists see. And in China, men are smoking in much greater numbers than women) [4], as well the Relationship between the A, B and O Blood types people and the confirmed cases with Covid-19 [5].

The Research code applies have been written by Erlang language, Erlang is a functional programming language which provides runtime designed for extremely parallel, scalable software needs high uptime. Erlang is a strict, dynamically typed functional programming language that comes with built-in support for message-passing concurrency, interprocess communication, and distribution [6].

"Erlang started in the Computer Science Laboratory at Ericsson Telecom AB in 1986. In 1988, the Lab moved to another company called Ellemtel in a south-west of Stockholm. Ellemtel was owned by LM Ericsson and Televerket, the Swedish PTT, and was the fundamental centre for the development of new switching systems"[7]. Erlang is kind of similar to Java in case of using a virtual device and also supports multithreading. Its provides a very wide range of built-in types which include integer, float, atom, Logical, tuple, list, fun, string, pid, port, reference, bitstring, record, map [7].

Linear Regression is a machine learning algorithm based on supervised learning. LR is a linear approach to modeling based on independent forecasts of values, to create a clear relationship between variables to known the effect of values on the other for a specific goal. Regression methods are different because they depend on the number of independent Variables and the type of relationship between the Independent Variables and Dependent Variables. Regression is the study of dependence, used to analyze a central part of many research projects. It is used to answer questions such as Does changing class size affect success of students? Do changes in diet result in changes in cholesterol level, and if so, do the results depend on other characteristics such as age, sex, and amount of exercise? Do countries with higher per-person income have lower birth rates than countries with lower income? [8].

This research is based on studying the impact of longitude and latitude on knowing the spread of the COVID 19 in confirmed cases, The definition of latitude is the angular distance from some specified circle or plane of reference: such as. a: angular distance north or south from the earth's equator measured through 90 degrees an island located at 40 degrees north latitude. b: a region or locality as marked by its latitude [9].

2. Research methodology

The protocol of this retrospective is depending on study Chose and download Dataset from Kaggle site [10], by classification algorithms linear regression data mining to extract new knowledge from data beneficial.

This database contains 75 attributes.

(a) Province/State
(b) Country/Region;
(c) Latitudes;
(d) Longitude;
(e) Date series of COVID 19 confirmed, from 22/01/2020 until 01/04/2020.

A sample of original cases confirmed dataset is shown in table1.
KDD is the process of determining useful information from a collection of data depending on CRISP-Data Mining process framework. KDD used data mining techniques is a process that includes data choose, Pre-processing data preparation and analyzing data through several processing stages in order to prepare the data without missing and proper data, data cleaning, incorporating prior knowledge on data sets and interpreting accurate solutions from the results and Evaluation so we will get the Knowledge, figure 1 will show us the steps (phases) of the CRISP-Data Mining process [11].
Regression is a data mining function that assigns items. The goal of regression is to accurately predict the target class for each case in the data, by Erlang functional language to grant the analyze of Dataset more flexible write code of the extract a knowledge. Figure 2 shows us the process steps of the algorithm Model life cycle.

```
Attributes types
- Nominal: 1, 2, 3, 4
- Ordinal: 5-75

Variable to be predicted:
- Latitudes
- Series days
```

3. Work steps:

3.1 Data Preparation:

Data preparation has been placed in the context of data exploration, in which the problem to be solved, rather than the technology, is paramount. Without identifying the problem to solve, it is hard to define how to extract value from the data mining activities that follow. In data, preparation does not mean that powerful techniques are not available or useful. Equally important is specifying the form of a solution. Without a firm idea of what success looks like, it is hard to determine if indeed the result found, and the form that it is delivered in, have actually succeeded. Having specified what a suitable solution looks like, and collected or discovered appropriate data, you can begin the process of data mining [12][13].

At the highest-level overview, the stages in the data exploration process are: [12]
1. Exploring the Problem Space
2. Exploring the Solution Space
3. Specifying the Implementation Method
4. Mining the Data (three parts)
   a. Preparing Data
   b. Surveying the Data
   c. Modeling the Data

sorting, arranging and analyzing data through several processing stages in order to prepare the data without missing and proper data, then mining on the data using linear classification method, by Erlang functional language to grant the analyze of Dataset more flexible write code of the extract a knowledge.

In this step, many process are done to prepare the data, as follow:
- Convert dataset from .xlsx extension to .txt file extension.
- Stop words are a division of natural language. The motive that stop-words should be removed from a text is that they make the text look heavier and less important for analysts. Removing stop words reduces the dimensionality of term space. The most common words in text documents are articles, prepositions, and pro-nouns, etc. that does not give the meaning of the documents. These words are treated as stop words. Example for stop words: the, in, a, an, with, ( ). Stop words are removed from documents because those words are not measured as keywords in text mining applications [59]. further in additional, every line in a text that ends Enter to a new line, the text automatically add \n, then should be delete \n from whole the text file.

- Aggregating each record of rows as list in order to handle it flexibly later.
- Filtering and nomination the important attribute that will be worked on as follows:

(a) Country/Region;
(c) Latitudes;
(d) Longitude;
(e) End Date series of COVID 19 confirmed 01/04/2020 that will be represent the acumolative amount

\[
\text{[["Germany"],["51.0"],["9.0"],["77872"]]}\\
\text{[["afghanistan"],["33.0"],["65.0"],["237"]]}\\
\text{[["Italy"],["43.0"],["12.0"],["110574"]]}
\]

3.2 Modelling & Evaluation Regression algorithm:

Logistic regression is an approach to prediction, with logistic regression, the researcher will be predicting a dichotomous result [15Noor]. The researcher has written and run the regression algorithm by sending values (country, latitude, longitude, total cases) thought of a list-comprehension to the regression function, to isolate all values that are equal or nearby together by grouping set.

\[
\text{[ to_Reg( Latitude, DataSet ) || [A,B,C,D] <- DataSet ]}
\]

Erlang block code below is the performs the regression algorithm:

```erlang
to_Reg(Latitude, [ ]) -> [ ];
to_Reg(Latitude, [H|T]) ->
[A,B,C,D] =H,
  case to_convert(Latitude) == to_convert(B) of
    true -> [ {B,A,D} | to_Reg(Latitude, T) ];
    false-> to_Reg(Latitude, T1)
    end.
```
7. The Results

When implementing the regression algorithm code, will get a result groups of countries that are equal or nearby together by grouping set to latitude as shown in Figure 3 below.

![Figure 3. Latitude group set](image)

| Result          | Countries & Cities Up Equator | Countries & Cities Down Equator | percentage |
|-----------------|-------------------------------|---------------------------------|------------|
| Amount          | 61                            | 2151613                         | 96.34      |
| Cases confirmed | 31                            | 81565                           | 3.65       |

Table 2. percentage amount of confirmed cases the KDD
8. Conclusion & Recommendation:

The next table Table 3 shows data of COVID_19 confirmed cases in numerical patterns every 100,000 infections cases, linear regression by Latitude up and down the equator, indicating the number of countries with the number of infections confirmed cases and the highest cases in latitude, in addition below the code by Erlang language.

Figure 4. The percentage of confirmed cases in Up and Down Equator
Table 3. Data COVID_19 linear regression by Latitude up and down the equator

The diagram 1 below that show us the increase of cases confirmed for each 10,000 patterns according to accumulator cases informed Countries/Cities up and down of the equator.

Diagram 1. The increase of cases each 10,000 patterns

The diagram 2 below shows the focus and most spread Cases confirmed in the latitude at up part of the Equator
Diagram 2. Cases spread in latitude above the Equator

Diagram 3. Cases spread in latitude down the Equator

The diagram 3 below shows the focus and most spread Cases confirmed in the latitude at down part of the Equator.
The diagram 4 below shows the Regression line of focus and most spread Cases confirmed in the latitude at down part of the Equator.

8. Conclusion

This paper provides an applied study in the field of Knowledge Discovery and Data Mining, to help scientists and researchers whom still work on discover the vaccine, taking into account the importance of the geographical location in influencing the spread of the Covid-19, to extract new knowledge from data beneficial.

- According to the Regression Classification, the equator is the regression line.
- Countries, cities that above the equator have confirmed cases 26 twice the rate than countries, cities below the equator, as well in percentage 96.34 in above the equator compared with 3.65 down the equator.
- Zones {37,700130}, {40,193973}, {46,176526}, {43,174279}, {51,160232}, {55,109841}, of the latitude are the most affected and confirmed cases Covid-19 according the analysis and aggregate result of the dataset.
- Countries, cities below the Equator are hotter than hot above the Equator, which is recorded the lowest number of confirmed cases Covid-19 recorded compared with above the equator.
- One of most important notic, the countries that are above the equator are mostly is characterized by tourist cities such as Europe and America, which caused to spread the COVID-19 much more than countries that are below the equator according by touris characteriz.
- Most of the countries that are above the equator are characterized by the commercial exchange with various countries of the world more than those countries that are below the equator, which cause the movement of humanitarian for the purpose of trade in a huge way and this may also have helped spread the COVID-19, this confirms the effect of latitude 37,40,46,43,51,55, in the appearance of the most cases confirmed.
References

[1] Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine 2020.
[2] Gorbatenya AE. Severe acute respiratory syndrome-related coronavirus–The species and its viruses, a statement of the Coronavirus Study Group. BioRxiv 2020.
[3] Xie X, Zhong Z, Zhao W, Zheng C, Wang F, Liu J. Chest CT for typical 2019-nCoV pneumonia: relationship to negative RT-PCR testing. Radiology 2020: 200343
[4] https://www.nytimes.com/2020/02/20/health/coronavirus-men-women.html
[5] Zhao, J., Yang, Y., Huang, H. P., Li, D., Gu, D. F., Lu, X. F., ..., & He, Y. J. (2020). Relationship between the ABO Blood Group and the COVID-19 Susceptibility. medRxiv.
[6] J. Armstrong, R. Virding, C. Wikström, and M. Williams, Concurrent Programming in Erlang. Prentice Hall Europe, Herfordshire, Great Britain, second edition, 1996.
[7] Armstrong, Joe. "A history of Erlang." Proceedings of the third ACM SIGPLAN conference on History of programming languages. ACM, 2007.
[8] Weisberg, S. (2005). Applied linear regression (Vol. 528). John Wiley & Sons.
[9] Karasek Jr, R. A. (1979). Job demands, job decision latitude, and mental strain: Implications for job redesign. Administrative science quarterly, 285-308.
[10] https://www.kaggle.com/datasets
[11] Hipp, Jochen, and Ulrich Güntzer. "Is pushing constraints deeply into the mining algorithms really what we want?: an alternative approach for association rule mining." ACM SIGKDD Explorations Newsletter 4.1 (2002): 50-55
[12] Pyle, D. (1999). Data preparation for data mining. morgan kaufmann.
[13] Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. Knowledge and information systems, 1(1), 5-32.