RESEARCH ARTICLE | OCTOBER 03 2017

Predictive analysis effectiveness in determining the epidemic disease infected area FREE

Najihah Ibrahim; Nur Shazwani Md. Akhir; Fadratul Hafinaz Hassan

AIP Conf. Proc. 1891, 020064 (2017)
https://doi.org/10.1063/1.5005397

Predictive Analysis Effectiveness in Determining the Epidemic Disease Infected Area

Najihah Ibrahim [1, a)], Nur Shazwani Md. Akhir [1, b)] and Fadratul Hafinaz Hassan [1, c)]

[1]*School of Computer Sciences,*
Universiti Sains Malaysia, Malaysia.

[a)] Corresponding author: najihah.ib@gmail.com
[b)] nur_shazwani_mdakhir@student.usm.my
[c)] fadratul@usm.my

Abstract. Epidemic disease outbreak had caused nowadays community to raise their great concern over the infectious disease controlling, preventing and handling methods to diminish the disease dissemination percentage and infected area. Backpropagation method was used for the counter measure and prediction analysis of the epidemic disease. The predictive analysis based on the backpropagation method can be determine via machine learning process that promotes the artificial intelligent in pattern recognition, statistics and features selection. This computational learning process will be integrated with data mining by measuring the score output as the classifier to the given set of input features through classification technique. The classification technique is the features selection of the disease dissemination factors that likely have strong interconnection between each other in causing infectious disease outbreaks. The predictive analysis of epidemic disease in determining the infected area was introduced in this preliminary study by using the backpropagation method in observation of other's findings. This study will classify the epidemic disease dissemination factors as the features for weight adjustment on the prediction of epidemic disease outbreaks. Through this preliminary study, the predictive analysis is proven to be effective method in determining the epidemic disease infected area by minimizing the error value through the features classification.

INTRODUCTION

Epidemic diseases are the contagious diseases that are possible to be spread into the entire nation if the contagion measurement had reached the outbreak level and manage to wipe out the entire population. There are some well-known epidemic outbreaks that were happened in the entire world such as dengue, yellow fever, cholera, diphtheria, influenza, bird flu and many more [1-8]. This contagious diseases had caused a major world health issues and was believed to be the one of the major factors that had caused the 43% of life lost globally [7]. Malaysia also had experienced some of epidemic diseases outbreak such as dengue, hepatitis, chikungunya and many more. However, recently from 2013 to 2016, there were reports on the outbreak of diphtheria in Malaysia [9]. Diphtheria may cause the inflammation of the nerves, paralysis and bleeding problem to the host. Even though diphtheria involves the sole host, the infection may spread in many ways and by multiple types of agents. The transmission usually happened via direct contact and contaminated air. However, this disease also was believed had been cause by the irregular of vaccination.

The infection of epidemic disease can be spread vigorously by the active mobilization of the pathogen and the rapid production and stimulation of the pathogens [6]. This disease outbreak was stimulated by several factors that can be identified as the natural factor and the man-made factor that almost impossible to be measure by classic statistic numeration. These factors may contribute towards disease's identification, detection, prediction and

The 2nd International Conference on Applied Science and Technology 2017 (ICAST'17)
AIP Conf. Proc. 1891, 020064-1–020064-6; https://doi.org/10.1063/1.5005397
Published by AIP Publishing. 978-0-7354-1573-7/$30.00

controlling via features classification. However, due to the involvement of integration between the features, the features' clustering method was introduced and had caused the weight adjustment on the input. This features' clustering method may result the less accurate and approximate detection and prediction on the epidemic disease. Hence, the study of finding the optimal result of the integration of clustered features was introduced via the backpropagation method that able to identify and correlate the factors that have impact on the epidemic disease dissemination [2].

This concept paper will discuss on the effectiveness of predictive analysis in determining the epidemic disease infected area by enlightening the process of analysis using backpropagation method of machine learning process that will introduce the cross validation technique. The rest of this concept paper is organized as follows: section 2 consists of a review on the targeted machine learning process; the predictive analysis and the impact on the epidemic disease prediction analysis. The classifiers of the epidemic diseases dissemination are given in the section 3. Conclusion is provided in section 4.

MACHINE LEARNING: PREDICTIVE ANALYSIS

The classical analysis of the epidemic disease had taken place for years before the technologies has had the advance outbreaks in term of compute intensive and data manipulation. The computer scientist had come out with the advance and more urban methods in describing the epidemic pattern, predicting the future outbreaks or the impact size of a population over the infectious outbreaks and controlling the counter measure of the infectious diseases to reduce or overcome the outbreaks via data acquisition [2]. This urban method however needs the machine to learn and train itself for the more dynamic input that cause the needs of cross validation techniques. Hence, this method is known as machine learning process.

Machine learning is the data based analysis process which the analysis only able to be done via gathering data from data mining process and going through the data pipeline process for the knowledge discovery. Machine learning can be used to describe, predict and control the mined data. This learning method will promotes towards the pattern recognition, data classification and features selection in order to produce the accurate and reliable results.

This data learning activity can be done via supervised learning and unsupervised learning. The supervised learning is the directed learning activities which the desire result is known and makes a continuous comparison value to the corrected output in order to find errors. This learning method used the patterns recognition in predicting the correct output. The unsupervised learning is the heterogeneous analysis that will cluster the data functions and implement the data segmentation for features selection. This unsupervised learning will make the predictive analysis become more scalable and high dimensionality that leads towards the inter-correlation of the selected functions' features [2]. However, the unsupervised learning needs high and complex computational process due to the self-organizing and nearest-neighbor mapping of the data. Figure 1 shows the pipeline of the learning process using the epidemic diseases' data.

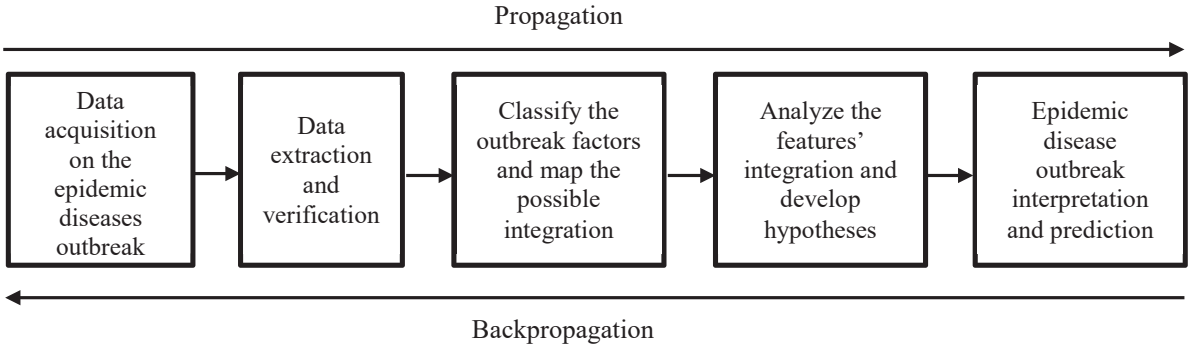

FIGURE 1. The data pipeline of machine learning process on epidemic diseases' data

Based on Figure 1, this computational learning method is the most well know process of finding the classification of the data while inbound the numbers of data features in order to determine data's correlation. The classification of the data then will be analyzed and the hypothesis of the data will be made for the epidemic diseases outbreak prediction. The classification can be made by propagation method or backpropagation method. The propagation method is the normal methods which classification can be done from the root to the tips. However, the backpropagation is the backward process which the classification can be done from the tips to the root.

Machine learning process was introduced in predicting the epidemic disease dissemination by cross validate the impact of the epidemic diseases' training data using backpropagation. The backpropagation method then will identify several main factors that caused the dissemination of epidemic disease. The factors then will be clustered and being analyzed which the correlation of the classifiers had been detected.

PREDICTIVE ANALYSIS: EPIDEMIC DISEASE FACTORS TO PREDICT THE INFECTED AREA

Health predictive analysis is the new outbreak of advance technology that can prevent the contagious epidemic disease [2, 10]. Figure 2 shows the relation of predictive analysis and its role to predict the factors of epidemic disease dissemination.

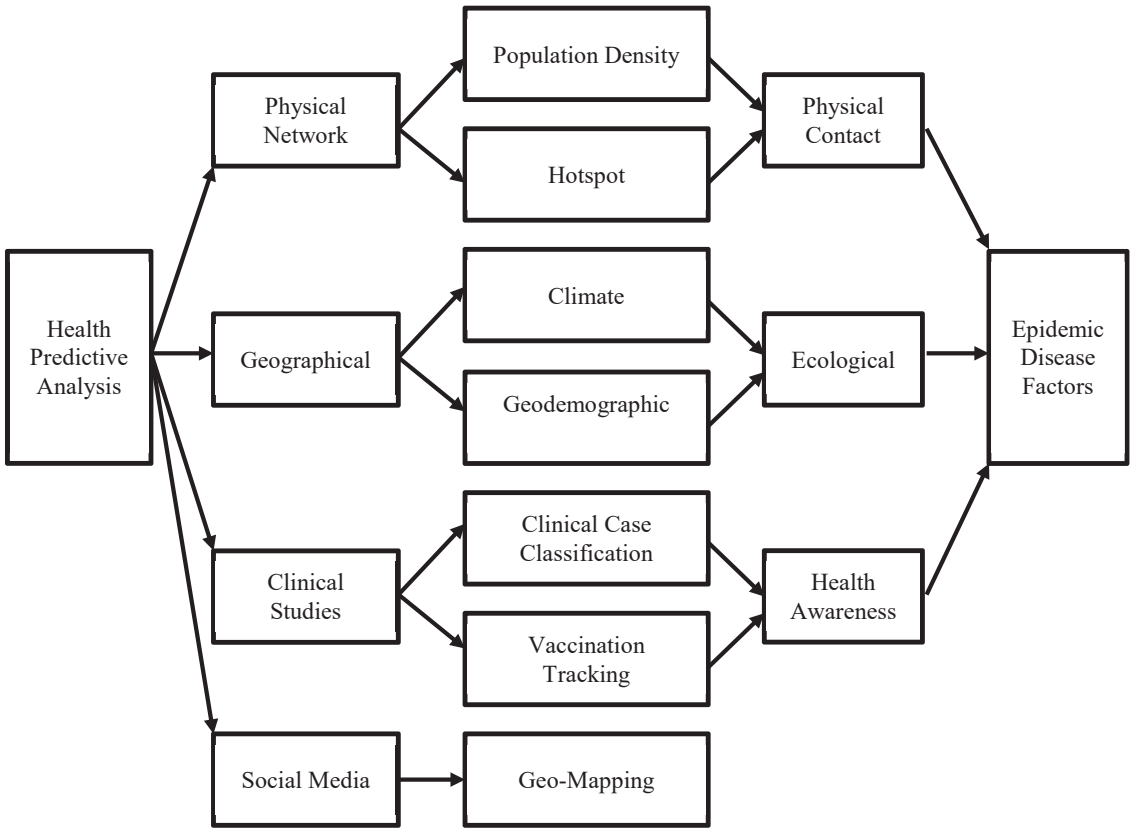

FIGURE 2. The correlation of health predictive analysis and epidemic disease dissemination factors

Based on the Figure 2, the classification of epidemic diseases dissemination can be categorized as physical network, geographical location, clinical studies and social media [2, 5, 7, 8].

The infectious disease dissemination was caused by the harmful virus attack. The viruses have the capability of doing the replication inside the cells of the host and slowly will invade the host's cell ecosystem. However, the viruses have no ability to mobilize themselves from one host to another host without any help from other third party [6]. This mobilization is known as physical network. The physical network is the chain of the virus interchanges between the hosts. There are two factors that cause the physical network that are: 1) population density and 2) hotspot.

The population density is the percentage of human population over a specific area. The highest of population density in a specific space will leads towards the higher potential of the epidemic disease breakouts will happen. The migration, urbanization and productivity will able to affect the health condition of the population. The high density of population will able to increase the potential of infection's transmission. The hotspot is the point place which has great human attraction that involves the high ingress and egress of human. The place of interest had offer many kind of activities such as tourist visit, amusement park, historical significant, business center, flight connecting point and etc. This point of attraction is the best factors for the speedy spreading of contagious epidemic disease because there are a lot of physical contacts that possible will happened and become one of the best vulnerability factors in infectious disease dissemination.

The physical network that promotes the hosts' interchanges can be stimulated by the direct contact and the indirect contact of an object with the contagious agent. The direct contact of infection will contribute towards the network epidemiology via the vectors such as human or animal that creates a social structure of contagion network that is peer-to-peer contacts such as the human-to-human contacts, human-to-animal contacts and etc. [6, 7]. This direct contact is the physical interchanges of the vector agents and will continuously and diversely distribute the pathogen of the viruses via saliva, blood, body fluids and etc. [6, 7].

A virus almost didn't have a lifeform which only known as a molecule that consist of protein such as DNA or RNA that used to carry their genetic information. Hence, the molecule of viruses also can be spread via indirect contact through the natural distributor such as wind, air, water and etc. This indirect contact usually will cause the viruses to penetrate the respiratory system or any first level filtration system of human body.

Natural distributor was one of the influenced factors in predicting the epidemic disease dissemination. This natural distributor has strong engagement with the geology factors [3, 4]. Based on Figure 2, the predictive analysis based geographical location can be determined by two factors that are: 1) climate and 2) geodemographic. Nowadays, the greenhouse effect had caused the climate changes due to the deforestation and desertification. The global temperature had risen and will cause the warm air that promotes the productivity of insect vectors and other air and water contamination factors [3, 4, 6]. Hence, some of the vector agent and natural-based vectored diseases such as malaria and dengue will highly anticipate to be spread [3, 4, 6].

The greenhouse effect was believed happened due to the man-made disasters. This particular factor had caused by the geodemographic of one's population. Geodemographic is the clustering techniques in classification of a community in a location based on their socio criteria. This classification can be done based on the sanitation level, family structure, neighborhood economic activities, education level and etc. [2, 4, 8]. The poor management of geodemographic features such as improper waste disposal, deforestation, and etc. will cause the vulnerable of the ecology, hence nurture the epidemic disease outbreaks.

The epidemic disease outbreaks based on the physical network and geology can be overcome or reduce by the introduction of advance health awareness. This advance health awareness will able to predict the epidemic disease dissemination based on the clustering method that promotes by the clinical studies. These clinical studies on the epidemic disease can be done by doing the: 1) clinical case classification and 2) vaccination tracking.

The epidemic disease outbreaks had been caused by heterogeneous factors that involved vast area of study and tremendous data collection activities. One of the data collection activities can be done by using the Electronic Health

Record (EHR). EHR is the patient's clinical records and already implemented by almost all of the health organization in the developed country for tracking, data storage and clustering purposes.

This clustering method of machine learning in clinical case classification can help the medical officer to describe, predict and control the epidemic disease outbreaks by collecting the EHRs and classify the data correlation based on the features selection [1, 2]. This clinical study also able to promote the vaccination tracking that was used in finding the classification of the vaccine type and the most used vaccine in an infected area. Hence, the clinical studies had played a great role in health predictive analysis process that promotes the machine learning method which initially requires the training data for the backpropagation method for cross validation.

Nowadays, the predictive analysis of epidemic disease was enhanced by the usage of social media [2, 5, 7]. The social media had become the selected and important medium for information sharing for this 21st century. Due to the technology advancement, the social media is the metadata collection platform that is able to do the geo-tagging, words' classification and personalization [2, 5]. The words' classification can be clustered with the geo-tagging for a personalization of a user. This personalization can be analyzed and the infected area of the epidemic disease can be determined with fine granularity [5].

This new outbreaks of computer services for data segmentation based method will able to help the data scientist in finding the classification of the epidemic diseases factors and will be able to predict the epidemic disease outbreaks faster than the classic clinical studies due to the live feeds of information form the users. This social media medium will able to predict the potential epidemic disease outbreak in a particular area.

The percentage of the infected people per area can be obtained by setting the area as a grid and setting the targeted area as the cells of the grid. The percentage of infection disease outbreak in a day can be determined by:

$$Infection\ \%\ in\ an\ Area\ = \frac{Negative\ Update\ Post\ in\ an\ i\ Area}{Total\ Numbers\ of\ Live\ Feed\ in\ i\ Area}\ x\ 100 = \frac{N_T}{L_T}\ x\ 100$$

Where T is the number of day and i is the selected area to be analyze.

The epidemic disease can be considered as contagious and heavily spread if the symptoms are remain constant or increase after 7 days of infection [5]. Based on the percentage of infection disease outbreak in a day, this percentage measurements by using the geo-tagging and mapping can be monitored for at least the first 7 days. Hence, the prediction of the spreading of the contagious disease can be determined by:

$$Prediction\ of\ Contagious\ Disease\ = \frac{N_1}{L_1}\ \%, \qquad \frac{N_2}{L_2}\ \%, \ldots\ldots\ldots, \frac{N_7}{L_7}\ \%$$

From this prediction of contagious disease, the contagious percentage of the disease can be predicted if the percentage value of the infection disease outbreak are constant or increase for over 7 days monitoring process. If there are some continuations on the infections, the selected area can be quarantine and the counter measure can be implement based on this early prediction.

CONCLUSION

The epidemic disease had become the most dangerous disease for this 21st Century if the infectious disease had still gone through the outbreaks despite of the modern medical treatment. This classical model of modern medical treatment has had turned far side due to the epidemic disease dissemination factors such as the increasing of population density and the speedy outbreaks of new infectious diseases. Hence, due to the nowadays' outbreaks of the computer technology that promotes the high compute intensive, the epidemic disease factors had become the selected features over the unsupervised learning on the data collected from EHR, social media record, and etc. for the predictive analysis purposes. The output determines will be used as the training data for the future output expectation determination. Based on the correlation of the health predictive analysis and epidemic disease dissemination factors, this backpropagation method had shown that the predictive analysis is able to predict the infected area of epidemic diseases outbreak.

ACKNOWLEDGMENTS

Research reported here is pursued under the Short Term Grant Scheme by Universiti Sains Malaysia for "Pedestrian Simulator and Heuristic Search Methods for Spatial Layout Design" [304/PKOMP/6313169].

REFERENCES

1. 1. Boivin, G., et al., *Predicting Influenza Infections during Epidemics with Use of a Clinical Case Definition.* Clinical Infectious Diseases, 2000. **31**(5): p. 1166-1169.
2. 2. Ravì, D., et al., *Deep Learning for Health Informatics.* IEEE Journal of Biomedical and Health Informatics, 2017. **21**(1): p. 4-21.
3. 3. Shope, R., *Global Climate Change and Infectious Diseases.* Environmental Health Perspectives, 1991. **96**: p. 171-174.
4. 4. Shuman , E.K., *Global Climate Change and Infectious Diseases.* New England Journal of Medicine, 2010. **362**(12): p. 1061-1063.
5. 5. Sadilek, A., H. Kautz, and V. Silenzio, *Predicting Disease Transmission from Geo-Tagged Micro-Blog Data*, in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012, AAAI Press: Toronto, Ontario, Canada. p. 136-142.
6. 6. Andrick, B., et al. *Infectious Disease and Climate Change: Detecting Contributing Factors and Predicting Future Outbreaks*. in *Geoscience and Remote Sensing, 1997. IGARSS '97. Remote Sensing - A Scientific Vision for Sustainable Development., 1997 IEEE International*. 1997.
7. 7. Masuda, N. and P. Holme, *Predicting and Controlling Infectious Disease Epidemics Using Temporal Networks.* F1000Prime Reports, 2013. **5**: p. 6.
8. 8. Kimura, Y., et al., *Geodemographics Profiling of Influenza A and B Virus Infections in Community Neighborhoods in Japan.* BMC Infectious Diseases, 2011. **11**(1): p. 36.
9. 9. Abas, A., *Five diphtheria deaths in Malaysia so far*, in *New Straits Times*. 2016, New Straits Times Press (M) Berhad (4485-H).
10. 10. Yu, W.D., et al. *Big Data Approach in Healthcare Used for Intelligent Design*. in *2016 IEEE International Conference on Big Data (Big Data)*. 2016.