

Utilizing news article data to predict infectious disease outbreak and spread

Juhyeon Kim(jhk@kisti.re.kr), Insung Ahn

Korea Institute of Science and Technology Information

Abstract

Background

Many different types of infectious diseases are emerging and threatening global human health conditions. For this reason, the detection of emerging infectious disease pattern is critical. However, as the epidemic spread of infectious diseases occur sporadically and rapidly, it is not easy to predict if an infectious disease will emerge or not. Furthermore, accumulating some specific infectious disease related data is not easy. For these reasons, identifying useful data and building a prediction model with these data are necessary steps toward predicting if some infectious disease will emerge at some countries or the number of patients will increase or decrease.

Methods

The Internet press releases numerous articles every day that rapidly reflect currently pending issues. Thus, we accumulated Internet articles related to infectious disease from several sources such as the Centre for Health Protection (CHP) which is published by Hong Kong Department of Health, medisys, or ProMED Mail to see if news text data could be used for predicting disease emergence or spread.

Results

We evaluated the predictive ability of article text data through this research. The support vector machine (SVM) model was used for prediction in order to examine the use of information embedded in the web articles and detect the pattern of influenza spread variance and infectious disease emerging pattern. The prediction result using news text data with SVM delivered a mean accuracy of 86.7 % on predicting if weekly ILI patient ratio increased or decreased.

Conclusions

In order to remedy the problems of conventional data, using news articles can be a suitable choice, because they can help estimate if ILI patient ratio will increase or decrease as well as how many patients will be affected or if some disease will emerge at some countries, as shown in the result of research. Thus, advancements in research on using news articles for influenza prediction should continue to be pursued, as the result showed acceptable performance.

Keywords: *Machine Learning, Article Data, Infectious Disease, Influenza*

References

- [1] Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, et al. Influenza Forecasting with Google Flu Trends. PLOS ONE 2013;8(2): e56176. doi: 10.1371/journal.pone.0056176
- [2] eC.E.B. Carslake, Harvesting Real Time and Historical Disease Outbreak Data from the ProMED-Mail Database: Pitfalls and Proposed Solutions. International Journal of Infectious Disease. 2017;53: 111
- [3] Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lipsitch. Real-time influenza forecasts during the 2012-2013 season. Nature Communications 2013;4: 1-10. doi: 10.1038/ncomms3837
- [4] M. Biggerstaff, D. Alper, M. Dredze, S. Fox, I.C.-H. Fung, K.S. Hickman, B. Lewis, R. Rosenfeld, J. Shaman, M.-H. Tsou, et al. Results from the centers for disease control and prevention's predict the 2013-2014 Influenza Season Challenge. BMC Infectious Disease 2016;16(1): doi: 10.1186/s12879-016-1669-x
- [5] John S Brownstein, Clark C Freifeld, Ben Y Reis, Kenneth D Mandl. Surveillance Sans Frontieres: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLOS ONE 2008;5(7): e151. Doi: 10.1371/journal.pmed.0050151
- [6] K. Park, T. Hou, H. Shin. Oil price forecasting based on machine learning techniques. J. Korean Inst. Ind.

Eng. 2011;37: 64-73

- [7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. ICLR Workshop, 2013. arXiv1301.3781
- [8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In NIPS, 2013
- [9] Schölkopf B. and Smola A.J. Learning with Kernels. MIT Press; 2002

Biography

Biography

Researcher (11/01/2015 ~ present, Data scientist)

Dept. of data-centric problem solving research,
Korea Institute of Science and Technology Information,
Daejeon, Republic of Korea

Affiliated Researcher (08/01/2016 ~ present, Data Scientist)

Center of Convergent Research of Emerging Virus Infection,
Korea Research Institute of Chemical Technology,
Daejeon, Republic of Korea