Heinz 95-845: Project Proposal

Kaung M. Khin KKHIN/KKHIN@ANDREW.CMU.EDU
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

Shawn Leahy SLEAHY/SLEAHY@ANDREW.CMU.EDU
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

1. Introduction

Influenza, or simply the flu, is a seasonal contagious respiratory disease that affects millions of people in the United States yearly with outcomes ranging from just mild symptoms to even death. (CDC, 2017a) It is caused by the influenza virus with the most infamous version in recent time being the 1918 flu pandemic nicknamed the Spanish Flu. The economic cost of the flu on the United States is estimated to be in the billions (Rolfes et al., 2017) so the Center for Disease Control (CDC) tracks the seasonal flu using Flu Surveillance which is a collaborative effort between the CDC and many healthcare providers.
The use of big data to track flu trends is not a novel idea (Ginsberg et al., 2009), in fact many studies have attempted to learn from the mistakes of Google Flu Trends. The challenges of using big data to attempt this type of prediction is well-documented. (Lazer et al., 2014) However, there has been some success in the use of social media data and search data combined with health surveillance data to track and predict the number of cases for the Zika virus outbreak in Latin America. (Mcgough et al., 2017) We will be using a variety of data sources to build a machine learning model to predict the number of actual flu cases in each state.

2. Analysis and Outcomes

- **Y:** Number of actual cases reported in a particular state/region
- **U:** There is no applicable treatment for our study
- **V:** We will use a number of data sources for our covariates but we will focus on the social media data, the health expenditure data and Google search data
- **W:** People who have filled out the MEPs survey for a given year

3. Importance and Contribution to Existing Literature

Demographic, hospital condition, and influenza surveillance data has been used to predict occurrences and investigate trends for influenza infections on U.S. military personnel. (Buczak et al., 2016) Machine learning methodologies have been used on medical data and

have been found to be successful in extrapolating from incomplete data (Chen et al., 2017; Santillana et al., 2016). This paper seeks to merge these analyses together to predict influenza outbreaks on a more diverse dataset utilizing machine learning methodologies that may have to deal with incomplete information and using social media data to track changes in infection rates.

4. Methods

4.1 Data

The Medical Expenditure Panel Survey is a set of large scale surveys of individuals, families, providers, and employers across the United States. MEPS collects data health services, how frequently they are used, the cost of these services, and how they are paid for, as well as data on the cost, scope, and breadth of health insurance. MEPS has been collected continually since 1996 (AHRQ). We will be utilizing the full year data files, medical conditions files, and prescribed medicines files. We will also be utilizing the Centers for Disease Control FluView Portal data in order to validate our results. This portal accesses state, regional and national influenza statistics as well as mortality information (CDC, 2017b). We are also attempting to include social media data to better track how fast and which geographic locations are experiencing influenza infections.

4.2 Modeling Approach

Our approach involves utilizing demographic, location, medical condition, and prescription drug information to predict susceptibility and frequency of influenza outbreaks in regions of the United States. We will create regional models to classify if we predict an individual to contract influenza or not. Variables for regional infection frequencies, infection rates, and hospital visits due to influenza will be created. Modeling approaches under consideration include logistic regression, SVMs, decision trees (and random forests), as well as neural networks.

4.3 Evaluation Measures

We will build a Naive Bayes model to predict influenza and this will be our baseline model to compare against due to its assumption of conditional independence. We will measure the accuracy, sensitivity, specificity, and F1-score for our models to try and gain insights into the results. We need an array of measures due to the fact that false positives and false negatives should be given a different weight, but we also need to measure an overall model accuracy of some sort.

5. Limitations and Possible Avenues of Continuing Work

This is not a real-time analysis tool that could be used ahead of time. However, this could be used to iteratively improve preparedness year after year as well as to track changes in disease spread over time. Further analysis into how disease spread changes year by year could be performed. Also, our data comes from a survey methodology and not EHRs or medical claims data. This introduces some uncertainty into every level of our analysis.

References

AHRQ. Medical expenditure panel survey. *Medical Expenditure Panel Survey*. URL `https://meps.ahrq.gov/mepsweb/`.

Anna L Buczak, Benjamin Baugher, Erhan Guven, Linda Moniz, Steven M. Babin, and Jean-Paul Chretien. Prediction of peaks of seasonal influenza in military health-care data. *Biomedical Engineering and Computational Biology*, 7s2, Apr 2016. doi: 10.4137/becb.s36277.

CDC. Influenza (flu). Oct 2017a. URL `https://www.cdc.gov/flu/about/index.html`.

CDC. Influenza (flu) interactive data. *Centers for Disease Control and Prevention*, Nov 2017b. URL `http://www.cdc.gov/flu/weekly/fluviewinteractive.htm`.

Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, 5:88698879, Apr 2017. doi: 10.1109/access.2017.2694446.

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):10121014, Feb 2009. doi: 10.1038/nature07634.

D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of google flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, Mar 2014. doi: 10.1126/science.1248506.

Sarah F. Mcgough, John S. Brownstein, Jared B. Hawkins, and Mauricio Santillana. Forecasting zika incidence in the 2016 latin america outbreak combining traditional disease surveillance with search, social media, and news report data. *PLOS Neglected Tropical Diseases*, 11(1), 2017. doi: 10.1371/journal.pntd.0005295.

Melissa A Rolfes, Ivo M Foppa, Shikha Garg, Brendan Flannery, Lynnette Brammer, James A Singleton, Erin Burns, Daniel Jernigan, Carrie Reed, Sonja J. Olsen, and et al. Estimated influenza illnesses, medical visits, hospitalizations, and deaths averted by vaccination in the united states. *Centers for Disease Control and Prevention*, Apr 2017. URL `https://www.cdc.gov/flu/about/disease/2015-16.htm`.

M. Santillana, A. T. Nguyen, T. Louie, A. Zink, J. Gray, I. Sung, and J. S. Brownstein. Cloud-based electronic health records for real-time, region-specific influenza surveillance. *Scientific Reports*, 6(1), May 2016. doi: 10.1038/srep25732.