Infectious Disease-Outbreak-Period Detection Using Discrepancy Scores

Preetika Rani (12535036)
Under Guidance of Dr. Vaskar Raychoudhury
Indian Institute of Technology, Roorkee

1. Introduction

- Since 2012, 2.5 exabytes of data has been generated per day in various areas like finance, railways, banks, academics, meteorology, genomics, biological researches, etc [1].
- With the implementation of data management softwares in hospital, laboratories, medical facilities, etc during recent decades, we have huge dataset for us to learn patterns and methods to perform analysis.
- Consider we have a huge amount of time-series medical database and we want to find the records of outbreak with in a reasonable amount of time. We will try to find the time period which has records related to outbreak.
- Some methods like What's Strange About Recent Event (WSARE), Early Aberration Reporting System (EARS), Shewhart Control Charts, Cumulative Sum (CUSUM), Exponential Weighted Moving Average (EWMA) can be used to determine the outbreak day, hence these are point detection methods.
- However, outbreak can not happen on a single day but it happens over a period of time and there is no common method for finding outbreak period.
- P.Chundhi had used Discrepancy Score to extract hot spots of topic from time-stamped documents [6].
- We have used Discrepancy Score to find the outbreak time period, by changing the definition of parameters in Discrepancy Score formula.

2. Problem Statement

Given a large amount of medical data having the outbreak, we have to identify the time period during which the outbreak has occurred.

3. Related Work

- According to CDC (Centre of Disease Control (US)), the occurrence of more cases of particular disease then normally expected with in a specific place and group of people, over a given period of time is outbreak.
- Outbreak detection is like anomalies pattern detection [3]
- On different basis we can categorizes outbreak detection methods, based on the type of input dataset or number of attributes, type of output and amount of past data require as shown in figure 1.

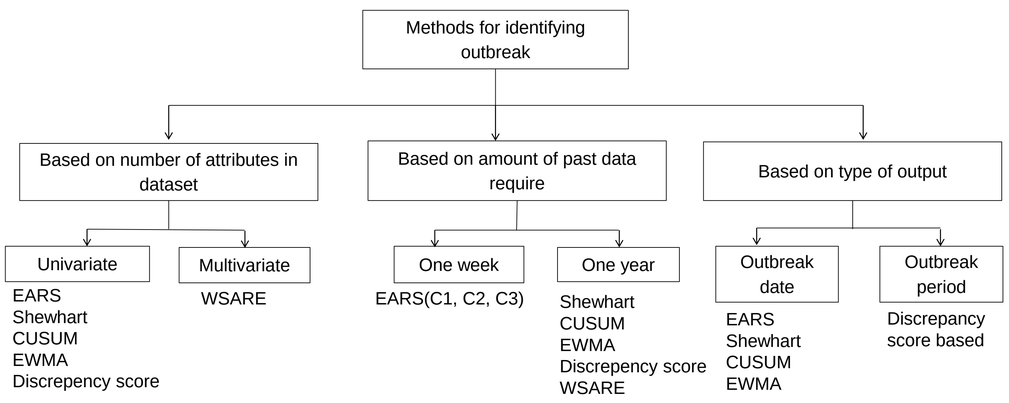

Figure 1: Classification of outbreak detection techniques

What's Strange About Recent Event (WSARE)

- WSARE is a rule based anomalies pattern detection algorithm. It has two phases. In the first phase baseline that determines the "normal pattern" is estimated [3].
- Second phase determines the recent patterns in dataset that are anomalous relative to normal historical patterns. Each pattern is characterized by a rule "*flu=high or action=evisit".*

Early Aberration Reporting System (EARS)

EARS are particularly used in identifying outbreaks by CDC (Centre for Disease Control(US)). EARS contain three methods based on sensitivity C1-MILD, C2-MEDIUM, C3-ULTRA methods [4].

- C1, C2, C3 methods uses below equations respectively to calculate their respective statistics:

$C_1(t) = \frac{Y(t) - \bar{Y}_1(t)}{S_1(t)}$	$C_2(t) = \frac{Y(t) - \bar{Y}_3(t)}{S_3(t)}$	$C_3(t) = \sum_{i=t}^{t-2} \max[0, C_2(i) - 1]$

Where $C_1(t)$, $C_2(t)$, $C_3(t)$ are C1, C2, C3 methods statistics
Y(t) is total number of patient cases on day t
$\bar{Y}_1(t)$ and $S_1(t)$ is the mean and standard deviation of patient cases from day (t-1) to (t-7)
$\bar{Y}_3$(t) and $S_3(t)$ is the mean and standard deviation of patient cases from day (t-3) to (t-9)

Shewhart Control Chart	$S_t = \left\lvert \frac{\bar{Y} - \hat{\mu}_0}{\hat{\sigma}_{\bar{Y}}} \right\rvert$	where $\hat{\mu}_0$ is mean and $\hat{\sigma}_{\bar{Y}}$ is standard deviation of baseline data of 365 days
Cumulative Sum	$C_t = max[0, C_{t-1} + L_t]$	where L_t is calculated by dividing residuals at time t by standard deviation of all 365 residuals
Exponential Weighted Moving Average	$E_t = \lambda \bar{Y}_t + (1 - \lambda) E_{t-1}$	where λ is weighing factor, $0 \geq \lambda \leq 1$

4. Dataset

- Dataset contains the record of Emergency Department for two years[2].
- First year record does not contain the outbreak, which we consider our baseline.
- Second year records has an outbreak of anthrax on specific date.
- All attributes are categorical attributes for example every person in the city can do any one of the three actions- visit to emergency department, take a sick leave from work or school or purchase medicine.

date_of_visit	*day_of _week*	*weather*	season	location	gender	action	Flu-level	Reported_symptom	drug

5. Proposed Method

Discrepancy Scores based Method

- Discrepancy Scores are used in extracting hot spots of topics from time-stamped documents [6].
- We use this Discrepancy Score to find the time period of outbreak in our case as shown in Figure 2.
- We had modified the parameters definition as in the original formula of Discrepancy Score.

- Discrepancy score is calculated using four variables:

$$\text{Discrepancy score(T)} = \text{m} * \log\left(\frac{m}{b}\right) + (M - m)\log\left(\frac{M - m}{B - b}\right)$$

- **Modified Definitions of used parameters**
  m which is the count of patient cases for that period in second year,
  b is the count of patients cases for same period but in base year,
  B is total count of patients in first year,
  M is total count of patients in second year

Challenges

- Finding the attribute to be used for separating outbreak record from other records.
- To find out the combination of techniques for outbreak detection.
- Finding the method or algorithm which will find the outbreak segment from record.
- What interval size should be used to find discrepancy score.
- What will be the threshold for Discrepancy score.
- Result of which interval should considered.

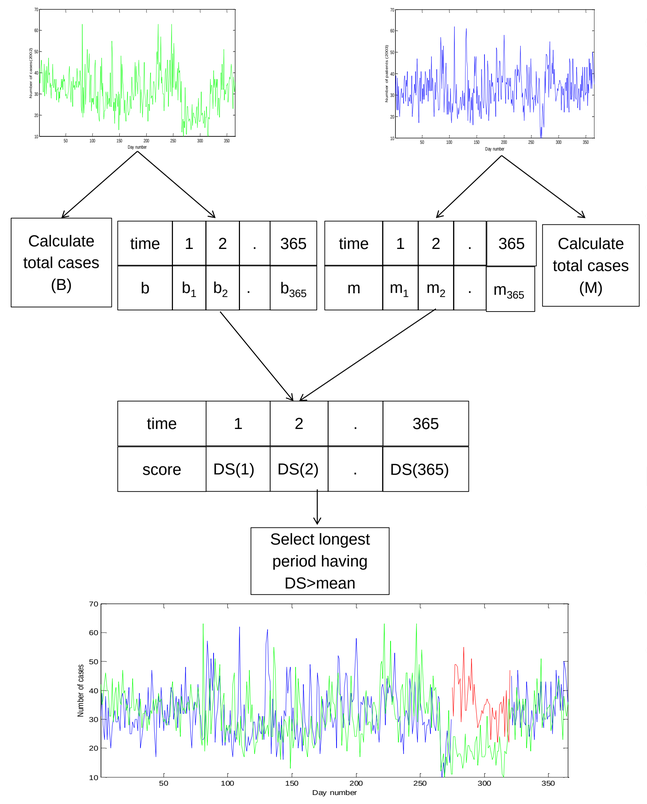

Figure 2: Steps of discrepancy score based method

Advantages

- Very less false signal rate as compared to other outbreak detection methods
- Easy to collect data, data only requires the count of patients do not require to access the other personal attributes information.
- Identify the period of outbreak
- Can be generalised does not require only 365 days data, can be used on much more or less data.
- Like others methods of outbreak detection, this method can be used in other fields like sales, marketing, quality control process, etc where increase or decrease in process output is evaluated.

Disadvantages

- Amount of baseline dataset required is equal to the testing dataset.
- Threshold value is not fixed, may vary.
- Can not find outbreak period when the number of patients can not increase sufficiently.
- Can not be used for early outbreak detection.

6. Initial Experimental Results

- C1, C2, C3, Shewhart, Cumulative sum, EWMA these methods detect the starting date of outbreak.
- Discrepancy based method give the time period of outbreak.
- We know day number 277 have outbreak of anthrax.
- C1 is less sensitive, while C3 is more sensitive on baseline data
- EARS three methods require very less baseline, while our method require a large baseline data.
- EARS gives false signals of outbreak very frequently which is not the case in our method.
- Shewhart control charts also have high false signal rate as compared to our method. It gives signal during outbreak, but not continuously as our method.
- EWMA detects outbreak later as compared to all other methods. This method gives freedom to assign different weightage to current day or baseline which is not required in our method.
- CUSUM gives continuous signal during outbreak but frequency of false signal is still more as compared to our method.
- Our method will gives the time period in which discrepancy score is greater than mean of discrepancy score.
- Figure 3 shows a comparison of above mentioned methods

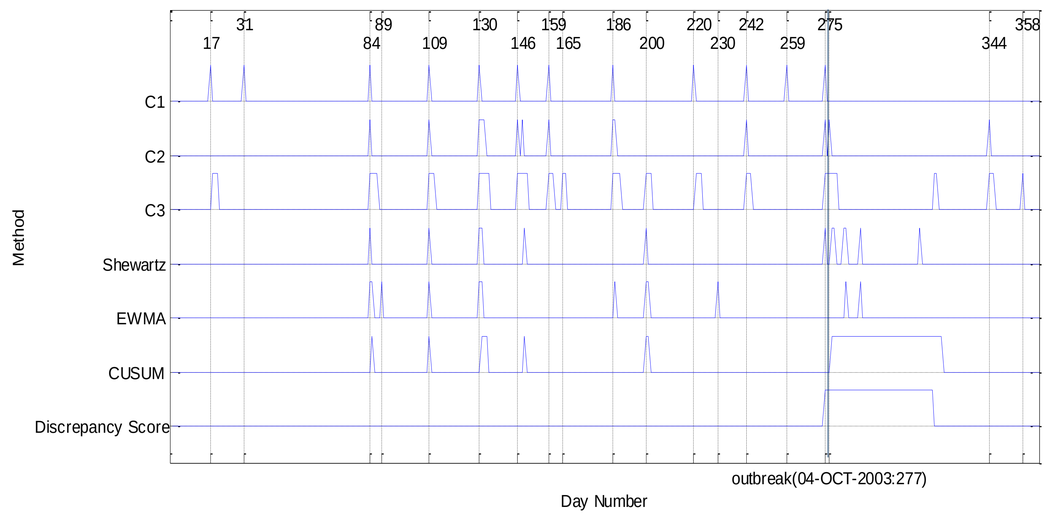

Figure 3: Comparison of our method with other outbreak detection techniques

Key References

[1] "IBM What is big data?-Bringing big data to the enterprise", source: www.ibm.com.
[2] "WSARE Dataset", source: www.autonlab.com.
[3] Wong W.K., Moore A., Cooper G., Wagner M.,"What's Strange About Recent Events (WSARE): An Algorithm for the Early Detection of Disease Outbreaks", in the proceeding of Journal of Machine Learning Research 6, page no. 1961–1998, 2005.
[4] Fricker Jr, R. D.,Hegler B. L, & Dunfee D. A., "Assessing the Performance of the Early Aberration Reporting System (EARS) Syndromic Surveillance Algorithms", Statistics in Medicine, 2007.
[5] Fricker,Jr., "Univariate Temporal Methods", in Title of his published book Introduction to Statistical Methods for Biosurveillance in year 2012.
[6] Chundi P. and Chen W., "Trends Analysis of Topics Based on Temporal Segmentation", in the proceedings of Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg, page no. 402-414, 2009.