Rule-Based Anomaly Detection for Outbreak Detection

This presentation is going over two papers to compare how anomaly detection (AD) has changed over time. Our first paper Wong et al. (2002), an almost two-decade old paper that goes over the use of creating rules from the data to preform AD. This paper is relatively short and shows a naïve approach to AD that has a lot of room to improve. The main idea we're showing with this paper is how useful historical data is for AD systems. Even when the paper is vague in techniques it shows that using historical data, we can build what's normal for a hospital to see through the year. If we see data that's anomalous, we can raise a 'flag' to signify unusual behavior. These systems aren't meant to predict when an outbreak will hit but rather sound an alarm as soon as the possibility of one is being show in the data.

A few things are noted in Wong et al. (2002) additional is the importance of looking at more than a singular anomaly. For example, it's easy to detect an outbreak that causes a loss of taste, if we're only looking for an abnormal amount of that symptom. There are possibilities of certain age groups, genders, or locations being affected in higher amounts than the average population getting a sickness. This would be much like, how COVID-19 changed through spread and started to effect younger children with Kawasaki disease like symptoms[1].

Though Wong et al. (2002) has interesting points inside it, there are problems as well. The paper uses a simple statistical approach to create these rules, and as seen in the generated rules, the point of their paper is forgotten by their system. The point was to find multiple attributes, at least two, to find unusual behavior of a symptom and a population characteristic. However, due to their loose approach many times, the time of day was the second attribute being found anomalous. This was noted in the paper as human input error at the hospital, but for their system to be highly susceptible to this problem raises questions on how proper their approach is.

1. https://www.cdc.gov/kawasaki/index.html

Turning over to the second paper, Karadayi et al. (2020), we didn't want anyone to deep dive into this paper unless they were interested. Rather the point here was to see that some of the ideas presented in Wong et al. are being used here as well.

Most importantly that, historical data is incredible useful for building what's normal in a system. However, without as much historical data, their approach allowed them to start prior to COVID-19 spread in some states to build what's normal for other states. Using states like Washington or New York which were hit sooner, to build what a spread would look like to 'flag' what begins to look like anomalous.

The main technology used in the second paper is an interesting technology, so we'll include a definition from the web as well as a link to read more about.

> **Autoencoder** is an unsupervised artificial neural network that learns how to efficiently compress and encode data then learns how to reconstruct the data back from the reduced encoded representation to a representation that is as close to the original input as possible.[2]

2. https://towardsdatascience.com/auto-encoder-what-is-it-and-what-is-it-used-for-part-1-3e5c6f017726