Topological Data Analysis in Computational Epidemiology

Yoonjin Kim

October 19, 2022

Since the speculated pneumonia outbreak in Wuhan, China in December 2019, the new repository coronavirus COVID-19 has killed 6.56 million people and caused global phenomena. With the help of vaccines, the mortality rate and severity level have significantly dropped yet it is inevitable that new strains of COVID-19 will come out in near future and it may cause another outbreak. It is essential to have an accurate estimation and prediction of the risk to individuals and communities of COVID-19 to help policymakers and potential interventions on vaccines and quarantine orders.

Topological Data Analysis is an analysis using techniques from topology. Topology is the study of spatial relations and the shape or size of figures. Topological Data Analysis initially started as focusing on geographical features and homology but as machine learning techniques have developed, it extended into using clustering algorithms to construct a graph with high dimensional data points. There exist a large variety of methods to construct Topological Data Analysis yet most of the standard methods follow four steps: 1. Measure the distance or similarity within a finite set of data points, 2. Build continuous shapes with the data. The shape often is a simplicial complex or a family of simplicial complexes called filtration. Extract topological information from the structures, and 4. Provide analysis on the extracted topological information through typical visualization. Measuring the distance or similarities within the data set point with real numbers can be as simple as adding euclidean distance in high dimensional space. Machine learning algorithms such as autoencoder can also reduce the dimensionality of data set and preprocess the data for the next steps. Building a continuous shape also can be done through clustering algorithms such as DBSCAN and KNN as the popular Topological Data Analysis tool Keppler Mapper suggests. Once the structure is made it is up to the researcher's eyes to make observations and provide analysis.

Recent papers have been published on using Topological Data Analysis on COVID-19 data. The existing two papers used four-dimensional space with latitude, longitude, cumulative confirmed cases, and the day since the first case was used to generate the topological analysis and they were able to identify the "trunk" and "branch" in data cloud and identify the outbreaks where the branch gets disconnected from the main "branch" structures as they go. The latter paper on Topological Data Analysis on Chinese data states that when the new branch is no longer detached from the trunk indicates that the pandemic has stabilized. However, topological data analysis is mainly used as a visualization tool and the observations are limited to visual without quantitative measurement of how the structure data changes.

To able to have a quantitative comparison between graphs and measure the differences, it is important to have a standard measurement and evaluation method on time series patterns in Topological Data Analysis. I plan to develop evaluation methods for higher dimensional data points where the measurement reflects the time series pattern data.