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 was named COVID-19 and it has killed 6.56 million people. This new outbreak of diseases caused global phenomena and has significantly impacted people's everyday lives. The mortality rate and severity level of the disease have significantly dropped since the help of the vaccine as of 2022. However, it is inevitable that new strains of COVID-19 or a new pandemic will come out in the near future and it is essential to accurately estimate and predict the risk to individuals and communities of COVID-19 to help policymakers and schedule potential interventions.

Topological Data Analysis (TDA) is an analysis using techniques from topology where the study of spatial relations and the shape or size of figures is used to analyze the structure of data. TDA 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 is a large variety of methods to construct a TDA structure and most of the standard methods follow these four steps: 1. Measure the distance or similarity within a finite set of data points, 2. Build continuous shapes with the data, where the shape often is a simplicial complex or a family of simplicial complexes called filtration, 3. Extract topological information from the structures, and 4. Provide analysis of 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 the 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.

Recent papers have since been published about TDA on COVID-19 data using four-dimensional space of latitude, longitude, cumulative confirmed cases, and the day since the first case to generate the topological analysis. Chen and Volic, the initial paper of TDA on COVID-19, and related papers focused on visualization representation and identified the "trunk" and "branch" in the data cloud. The "trunk" represents the connected cloud that contains the majority of the data points and the "branch" represent disconnected or weakly connected components from the trunk. The papers found that the nodes with significant changes lead the branch structures to get detached from the main trunk structure. The later analysis of Chinese data stated that the constant attachment of branches indicated that the pandemic has stabilized. However, this statement is mainly referring the visualization tool and the observations without quantitative measurement of how much the structure data changes.

To capture the time series data and overcome the limitation of visual observation, I suggest standard measurements for evaluating the topological data cloud and using higher dimensional data cumulative daily confirmed cases.