Semi-Supervised Deep Geometric Learning for COVID-19 Regional Severity Classification

Prateek Gaddigoudar
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
pgaddigo@andrew.cmu.edu

Andrew Plesniak
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
aplesnia@andrew.cmu.edu

Abstract

In this era, a network system connecting nodes represented by different regions around the world could be seriously jeopardized when these nodes with high connectivity are infected due to a contagious virus that evolves into a pandemic. Timely analysis of virus proliferation among nodes is significant for countermeasure design. High connectivity of nodes and ill - informed graph structure makes it arduous to come up with such an analysis. This paper treats each node as a state of a country and each edge as borderline adjacency between the connecting nodes. It intends to develop an approach for infection analysis using Graph Convolution. The proposed model makes use of GCN, which is a very powerful neural network architecture for machine learning on graphs. Our approach is semi-supervised as it considers infection severity labels of a fixed, small percentage of nodes to predict the severity labels of the remaining nodes. The input ground truth labels are sampled randomly from available ground truth labels' dataset. Extensive simulations are carried out for different percentages of ground truth labels to demonstrate the effectiveness of proposed approach and their results tabulated.

1 Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).[2] It was first identified in December 2019 in Wuhan, China, and has since spread globally, resulting in an ongoing pandemic. As of 9 May 2020, more than 4 million cases have been reported across 187 countries and territories, resulting in more than 277,000 deaths. More than 1.34 million people have recovered.[3]

The catastrophic affect this pandemic has had on economies of different countries, the lives and everyday existence of its citizens is inconceivable. Given such a drastic situation, it is crucial to classify different regions of countries according to their severities. Early prediction and classification of regions into different zones based on their severity incidence would not only help in slowing down the spread of pandemic, but also enable the governments and responsible authorities in taking appropriate steps. These steps might include the redirection of available resources such as medical staff, equipment etc to regions that are more vulnerable to be high risk zones.

We intend to achieve the same through this project, wherein given initially affected states of a country and their severity labels, we make prediction of severity labels of other states of that country by considering the borderline connectivity between states and population densities of each state.

Preprint. Under review.

2 Problem Statement

The goal of this project is to use geometric deep learning to classify the categorical regional severity of the cumulative COVID-19 cases in the United States of America. Each state in the United States will be given a categorical label based on the number of aggregated confirmed cases in that state on May 5th 2020. The categorical labels will be divided follows:

Table 1: Severity Categories

Category	Number of Aggregated Cases
0	more than 100
1	more than 1,000
2	more than 5,000
3	more than 10,000
4	more than 50,000

In order to simulate areas that may not have robust reporting strategies for COVID-19, this project will investigate semi-supervised learning. This means that during training only a subset of the ground truth labels will be available to the classifier. This project will investigate how the accuracy of the predictions of the entire data set changes with the amount of labeled data available.

3 Previous Work

Most of our methods are based of a similar project done for the the states of Mexico [1]. Our project adjusted some of the parameters of the model to account for the the United States having more states than Mexico. In addition, the original project only classified states into three severity (low, medium, high) compared to the five levels in this project.

4 Approach

4.1 Data set

Since there was no readily available data set of COVID-19 cases already in graph form, one was created. The number of aggregated COVID-19 cases in each state up until May 9th 2020 was collected from the University of Washington's COVID-19 map [4]. These raw numbers where then converted to the severity categories as seen in Figure 1. In order to build the graph itself, each state was created as a node. Then, an edge was put in between all states that share a geographic border, or in other words, neighboring states were connected. This was done guided by the assumption that people are more likely to travel from one state to another close by state rather than to a state that is far away. Although this model does not account for things like air travel, it is likely that the general assumption of close travel being more prevalent then distant travel is still founded. Finally, Alaska and Hawaii were excluded as they would be isolated nodes in this model. Thus, the final data set consisted of a graph of the 48 contiguous United States and their respective severity labels.

4.2 Graph Refinement: Population Density

Besides the geographic data in the structure of the graph itself, another feature was used in creating the data nodes: population density. The population per square mile of each state, sourced from [5], was multiplied with the one-hot-vector the respective state giving the refined node data.

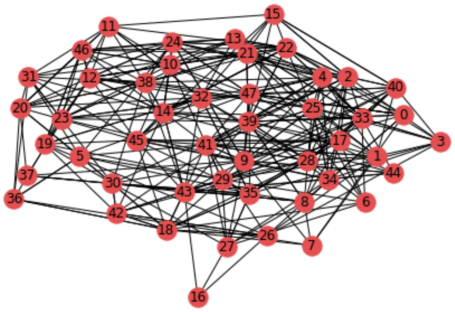

Figure 1: Graph representation of the United States

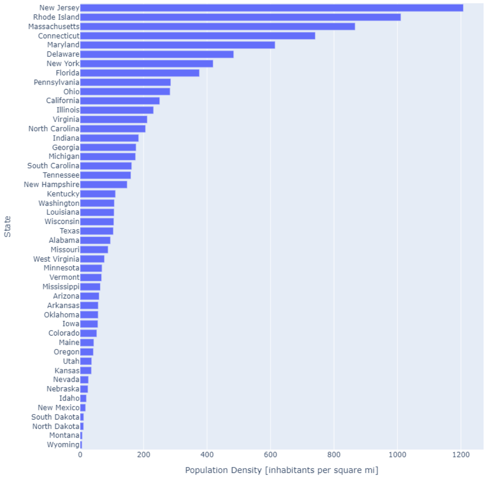

Figure 2: Population density by state

4.3 Network

Since the data set in this problem was very small, a simple graph convolution architecture was used. A Graph Neural Network, also known as a Graph Convolutional Networks (GCN), performs a convolution on a graph, instead of on an image composed of pixels. Just like a CNN aims to extract the most important information from the image to classify the image, a GCN passes a filter over the graph, looking for essential vertices and edges that can help classify nodes within the graph. For this model, two graph convolution layers where used and connected with a hidden size of 10 and a RELU activation function in between. In the first convolution layer, a learn-able feature embedding was used as the inputs to convolve over the graph.

4.4 Training

The network and feature embeddings were trained with an Adam optimizer with a learning rate of 0.01. Semi-supervised trials were run where the percentage of labeled data was incrementally varied from 5% up to 100%. Each network was trained for 500 epochs and the best accuracy from each training was recorded. The code can be found in Github repository. [1]

5 Results

Table 2: Severity Level Classification Accuracy

Percentage of Labeled Data in Training (%)	Prediction Accuracy (%)
5%	29.17%
10%	33.33%
20%	39.58%
30%	54.17%
40%	56.25%
50%	58.33%
60%	72.92%
70%	75.00%
80%	81.25%
90%	91.67%
100%	100.00%

[1] https://github.com/andrewplesniak/COVIDGeometricDeepLearning.git

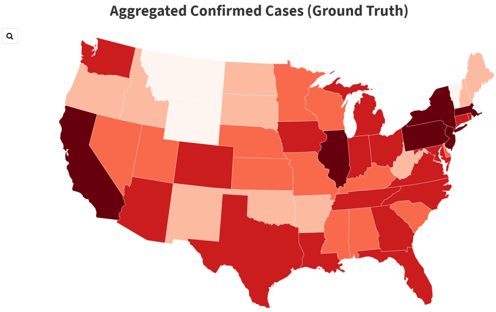

Figure 3: Ground Truth

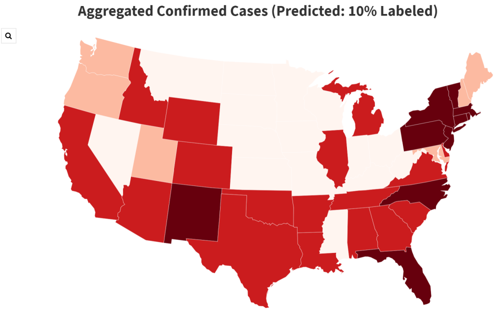

Figure 4: 10% Labeled Training Data

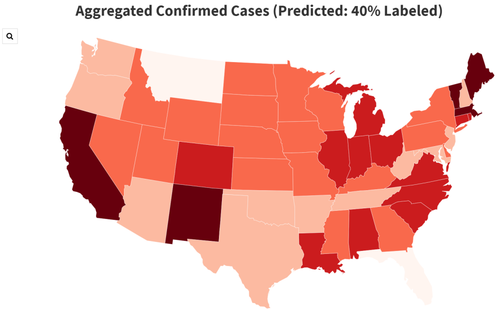

Figure 5: 40% Labeled Training Data

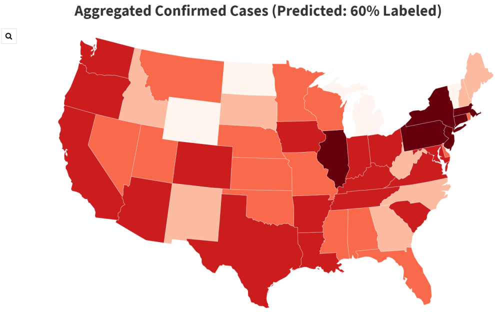

Figure 6: 60% Labeled Training Data

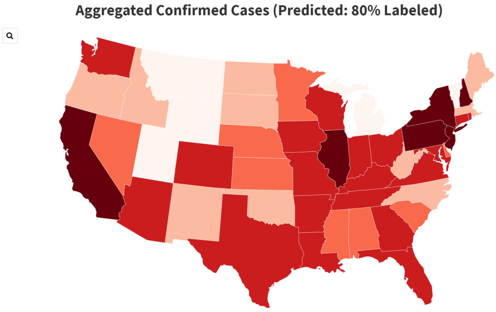

Figure 7: 80% Labeled Training Data

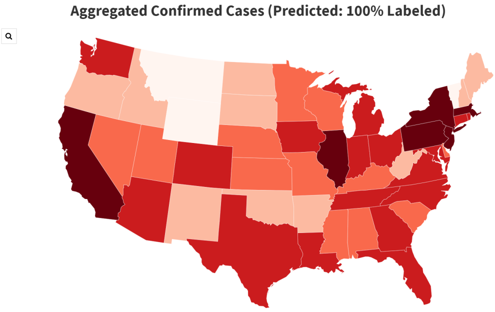

Figure 8: 100% Labeled Training Data

6 Discussion

Looking at the objective results for our model, we found that it could provide a decently accurate prediction of severity labels of all the nodes given just a small percent of ground truth labels (around 40 percent). This accuracy could be further improved by handpicking the input labels for training purpose from each of the five levels. This is partly due to the fact that the data set is somewhat imbalanced, only having a few examples of category 0 and category 4 states. However since handpicking is not always possible when it comes to reality, we decided to randomly sample a given percentage of inputs from the ground truth label dataset[4].

As witnessed from Figure 1 and Figure 2, higher the centrality of a node and higher its population density, higher is its predicted severity label. This can be seen explicitly in case of states like New Jersey (node 27), New York (node 29) and California (node 3). However, states like Wyoming (node 47), Montana (node 23)and South Dakota (node 38) are least affected owing to their small - scale population density although they are well connected. This result seems to be consistent across all the states and thereby indicates that population density has a crucial role to play when it comes to spread of COVID-19.

7 Conclusion

Given a borderline connectivity between United States as input feature, along with the population densities of these states, COVID-19 severity labels have been predicted by making use of semi-supervised geometric learning. Graph Convolution Network (GCN) represents a node as a sum of its neighbors feature representations before it is transformed by applying the weights and activation function. The model makes an assumption that borderline adjacency establishes communication between nodes.

As the paper indicated that higher the population density of a state and higher its connectivity, higher are the chances of it being severely affected. Clearly, isolation seems to be the primary solution to the increase in pandemic as it reduces the connectivity between the nodes and also maintains a hold on the population density in areas of public interest. In addition, there could be other factors beyond the scope of this paper such as availability of vaccines and medicines that might play a significant role. To further evaluate the model, other modes of communication between the nodes such as connecting flights could be considered. Furthermore, other features of nodes such as climatic temperature, humidity and availability of medical resources could be taken into account to improve the model efficiency.

References

[1] https://towardsdatascience.com/graph-convolutional-nets-for-classifying-covid-19-incidence-on-states-3a8c20ebac2b

[2] https://en.wikipedia.org/wiki/Coronavirus$disease$2019

[3] https://www.cdc.gov/coronavirus/2019-ncov/faq.html

[4] https://hgis.uw.edu/virus/

[5] https://state.1keydata.com/state-population-density.php