ECEN 689 - Applied Information Science - Challenge 3

Aditya Lahiri, Kanchan Satpute, Harish Kumar

Abstract

By analyzing the parameters obtained using the Lasso regression analysis method on the World Population dataset, we can visualize the network of all the countries. In this paper we try to visualize how the population of a country can be related to population of other countries using network analysis technique based on graph theory.

I. Problem Description

THE problem provides us a dataset containing the population of 258 countries/geographical aggregations from the year 1960 to 2016. The goal is to identify to what extent the populations of different countries are viable predictors for each other - this is achieved by expressing the population of a chosen country as a linear combination of the populations of several other countries in the same year. To get explainable results from which we can draw insights, the number of predictor countries is restricted to a maximum of 5.

II. Dataset

The dataset consists of population of various countries taken from data.worldbank.org for the years 1960 to 2016. The training data has records from the year $1960 - 1999$, whereas we have used the records from $2000 - 2016$ as the testing data. After cleaning and preprocessing the data, which involves removing the null values from the data, we obtain a matrix of (258×40) for the training and (258×17) for the test data.

III. Mathematical Formulation

We wish to calculate an estimate of the population of country i, $\hat{y_{i,k}}$ in year k as a linear combination of the population of other countries. Mathematically,

$$\hat{y_{i,k}} = \sum_{j \neq i} \lambda_{ij} x_{j,k} \tag{1}$$

Here, $x_{j,k}$ is the population of country j in year k. We have an additional constraint that forces us to limit the number of non-zero coefficients for each country to 5. In other words,

$$||\lambda_i||_0 \leq 5 \forall i \tag{2}$$

$||\lambda_i||_0$ is the L_0 norm of the λ_i vector, and it is simply the number of non-zero values in the vector.

The evaluation metric used is the Mean Sum of Squared Residuals in the estimates $\hat{y_i}$, i.e.

$$E = \frac{1}{258} \sum_{i=1}^{258} \frac{1}{17} \sum_{k=2000}^{2016} (\hat{y_{i,k}} - x_{i,k})^2 \tag{3}$$

IV. Implementation

To predict the population of a country we first initialize a range of α values or penalty parameters and then iterate over these α values and fit the regression model to check if we get the desired number of non-zero coefficients (between 1 and 5) for that country. Once we get a desired α value we measure the accuracy of our model for the corresponding α by computing the mean square training error for that country. If we had more than a single α which gave us the desired number of non-zero coefficients, we chose the α which corresponded to the lowest mean squared error. This process is repeated for each country.

In the cases where we cannot obtain a suitable α for a given country we predict its population using the largest α in the range of α that we initialized at the beginning. Finally, we check if we have less than 5 non-zero coefficient for each and every country. If this condition is not satisfied, we go back and redefine our range of α and repeat the entire process to check if the the number countries that have more than 5 non-zero coefficients has reduced. We repeat this until we bring down the number of countries with more than 5 non-zero coefficients to 0.

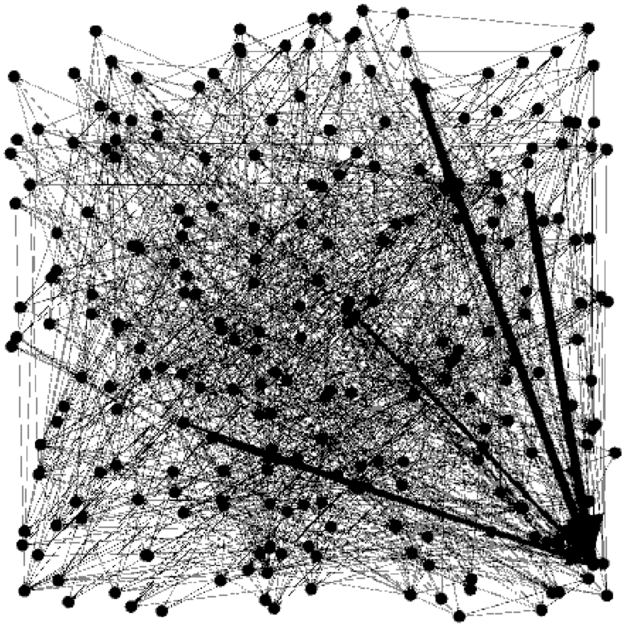

Fig. 1. The graph structure from the coefficient matrix

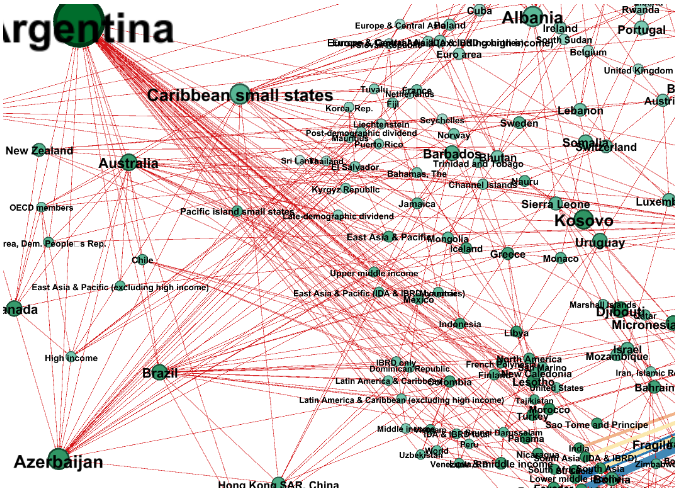

Fig. 2. Graph after preprocessing

V. Graphical Analysis

In this section, our objective is to learn the relationships that may exist among different countries using the coefficient matrix that we have generated while using regression to predict the population of the countries. We make use of the software Gephi and the coefficient matrix to build a directed network. The coefficient matrix essentially serves as a weighted adjacency matrix for building the graph structure. In graph structure every node (circle) represents a country whereas an edge represents a causal relationship between the countries. The thickness of the edges is determined by the weight of the edges that is obtained from the coefficient matrix. This means that larger the coefficient the thicker is the edge. Hence thicker edges have a larger flow of influence through them. We first load the coefficient matrix in Gephi to build the overall graph structure. Figure 1 below depicts the entire graph structure.

The graph in **Fig. 1** is very dense and is not very insightful. To make our graph more insightful we first apply the Yifan Hu layout scheme, followed by filtering using degree criteria. We also determine some statistics such as the betweenness centrality to rank our nodes. Hence, larger the node in the graph larger is the betweenness centrality statistic for that node, which corresponds to a large degree for that node. Following these preprocessing steps, we obtain the graph in **Fig. 2**.

The objective is to infer useful information from the graph. We need to select a node of interest. Here we will select Canada and Brazil as examples. We do so by selecting the Ego Network filter and entering the node of interest Canada. By

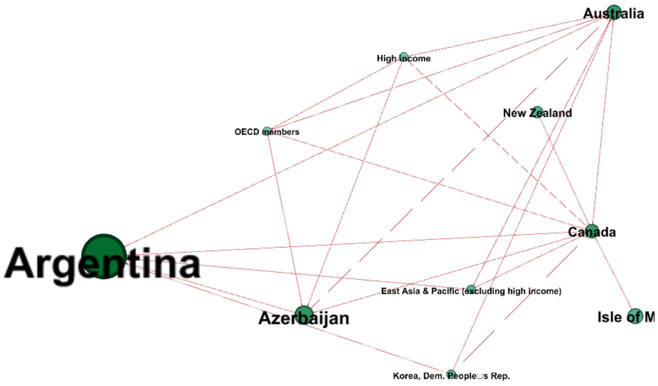

Fig. 3. Isolate Canada and its neighboring nodes from the entire graph.

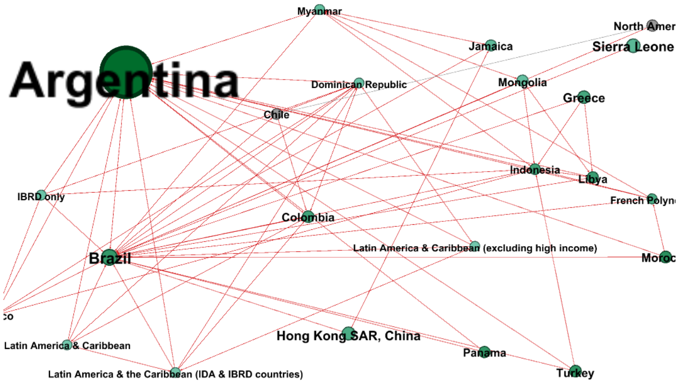

Fig. 4. Isolate Brazil and its neighboring nodes from the entire graph

filtering based on Canada we get the subgraph shown in **Fig. 3**. We can see that Canada and Australia are connected to OECD members Node which is consistent with the current list of OECD countries. While certain connections are insightful some are not, for instance, we see that New Zealand is not connected to OECD while being a current OECD member and Azerbaijan is connected to OECD while not being a current OECD member.

Similarly, we filtered our graph to get insights on Brazil. Refer to **Fig. 4** for this filtered graph. We see that Brazil is connected to Argentina, Colombia and many other nodes involving Latin America. Therefore, we can perhaps infer that it is reasonable to expect countries within the same continent in this case Latin America to be closely connected or clustered together.

Since our graph structure involves multiple nodes and are very densely connected, we used the filtering feature to draw inferences regarding our specific country of interest. While we got some interesting information from our graph structure we must verify it with the ground truth to determine the accuracy of our representation.

VI. Conclusion

We used L1 regularized multiple linear regression to predict the population of the countries from 2000 to 2016. We had to regularize our model since we had more independent variables than data samples which can lead to overfitting. We designed our regressors for each country in such a way so as to limit the number of non-zero coefficients to maximum of five. This design requirement emphasized sparsity hence we used Lasso over Ridge. Once we trained and predicted the population we used the coefficient matrix as a weighted adjacency matrix to construct the graph structure in Gephi. This graph had to be preprocessed and filtered to perform inference. While some of the inference were insightful and consistent with the ground truth others were not. Therefore the quality of information obtained from the graph structure is entirely dependent on the design of the regressors.

References

[1] http://www.briansarnacki.com/gephi-tutorial