Visualizing Brazilians in United States

Junhe Chen / Haoyu Zhang / Hui Li
Department of Computer Science/Boston University
College of Art & Science

Overview

▪As more and more Brazilians choose to live in United States, data of Brazilians in United States also also grows rapidly. With demographic state population of Brazilians, which is convenient for people to better understand the population distribution and its structure by visualizing the data on map.

Problem Definition

▪Cluster states based on location and population.
▪Build classification model to predict an individual's location based on demographic information.
▪Visualize Brazilians, and the clustering result mentioned above.

Data Processing

Remove Redundancy → Compute Probabilities → Output Clean Data

▪The data actually is arranged in a readable way, but not friendly to programs. Thus, we take the following steps to preprocess the data.

▪Remove redundant rows and columns. Delete any useless cell.

▪Convert absolute values into proportions. This is to prepare for Naive Bayes Model, which requires probabilities instead of original values.

▪Re-decode the data using regular expression in order to prepare it for visualizing. The result behind is the data are too convoluted to be used to display with JavaScript. It is necessary to prepare a clean and organized version of data for visualization only. Surely, we also output the data that used to perform clustering and classification.

▪The data preprocessing step is significant for this project due to the data are arranged in a human friendly way but not program friendly.

Visualization

Here is the screenshot of our visualizing. We take the attributes into groups and we can have charts of specific group of attributes on the side when the cursor of mouse is moved on the state. Here is the example of visualization of the group of attribute 'Age distribution'.

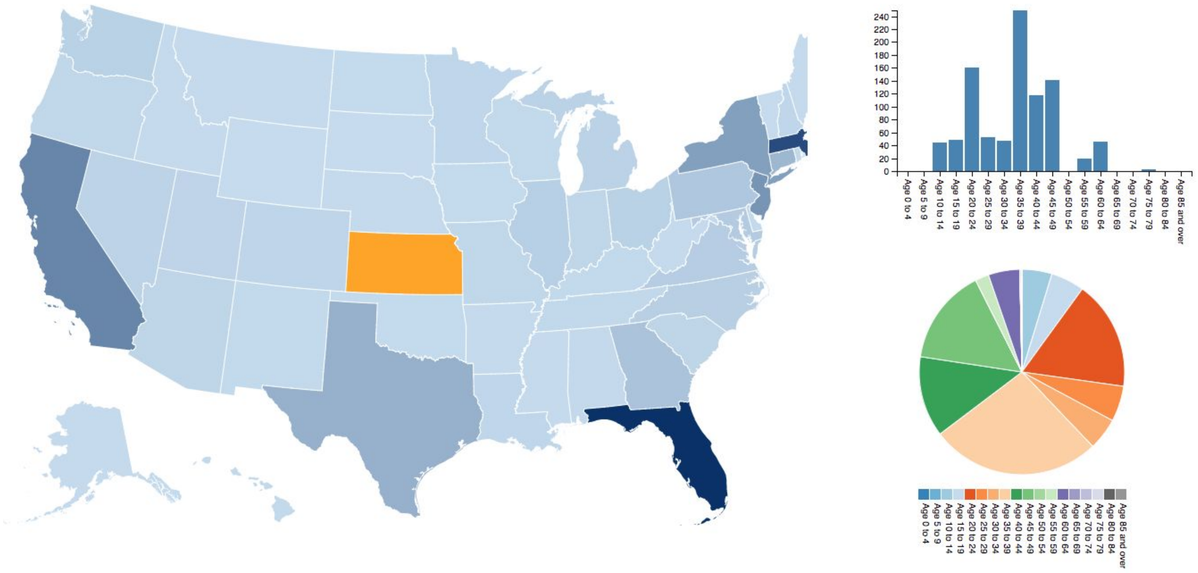

Figure 3. Example of Visualizing Brazilians in a state on age atrributes

Classification

▪Here we are using the data processed which has probabilities as statics instead of exact numbers.
▪We had an input reading file written in python, which stores every input into a list.
▪Then according to the inputs, we can allocate those attributes in the possibility version data, hence we can calculate the overall possibilities that the individual might appears in each state and out put them as a result list.
▪ Finally, we can get the max value from the list and return the name of the state according to the index of the max value in the result list.

Conclusion

For clustering, the areas consists of similar states are successfully detected. It is obvious to conclude that main states the most Brazilians live in are California, Massachusetts, and Texas. The conclusion is reasonable considering these states are more economics developed than neighbor areas. For prediction, due to the lack of individual data, we are not able to evaluate the prediction accuracy. Theoretically, it is able to perform prediction roughly. For data visualization, the visualized data shows the distribution of Brazilians in United States straightforward. Visualization successfully helps the display of data, and it helps to convey more information than original data.

Clustering

- **Preprocessing:** We use altitude, longitude and population as features which are scaled for clustering quality.
- **Algorithm:** We apply hierarchical clustering to preprocessed data. Our hierarchical clustering L2-norm as distance to cluster adjacent states with similar population. To find the best clustering number, we try from 5 to 35, and finally find 18 is the most suitable number by evaluating Silhouette Score.

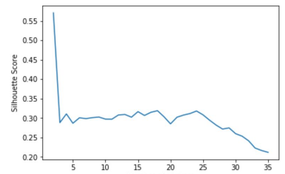

Figure 1. Silhouette Score for Different Number of Clusters

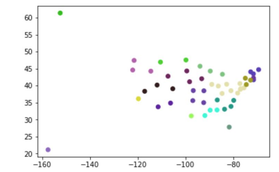

Figure 2. Visualized Clustering Result

- Massachusetts
- Florida
- California
- Michigan, Wisconsin, Minnesota
- Wyoming, Iowa, South Dakota, Nebraska
- Montana, North Dakota
- Oregon, Washington, Idaho
- Alaska
- New Mexico, Arizona
- Utah, Colorado, Nevada
- Missouri, Oklahoma, Arkansas, Kansas
- Louisiana, Mississippi, Alabama
- North Carolina, Georgia, Tennessee, South Carolina
- Texas
- Connecticut, New Jersey, New York
- Rhode Island, New Hampshire, Vermont, Maine
- Delaware, Virginia, Kentucky, Pennsylvania, Indiana, Maryland, West Virginia, District of Columbia, Illinois, Ohio
- Hawaii

- **Results:** The result basically satisfies our clustering goal: divide states into groups according to their similarity in location and population. There are some individual states consists of a group on its own. It means these states are not like their neighbors in population. For example, California itself is a group, and it is to say that California has a population that is distinguished from its neighbors. Oppositely, states in the middle area form groups with their neighbors since the population are very similar for these state, and it may be a result of the fact that Brazilians are less willing to live in these area compared to more economic developed area such as Massachusetts and California.

Future Work

In the future, the clustering would be performed with more features, such as weather, economic status and etc..The clustering would contain more informations. Besides, if we obtain individual features, we are able to evaluate our model.