Correlation matrix analysis on spatiotemporal data

Heejong Kim (hk2451), Jun Yuan (jy50), Jiin Nam (jn1664)

Big data 2018 Spring

Objectives

We want to understand a spatial-temporal data in real world. The collected data will be processed in a format that we can use for the correlation analysis. We want to learn to analyze correlations across data attributes in different ways.

Related Work

Nowadays, we can get access to a lot of urban data and many of them are either high dimensional or high volume. To ensure successful data analysis, data preprocessing is necessary. The set of techniques used prior to the application of a data mining method is named as data preprocessing for data mining[1]. It includes a wide range of disciplines, such as data transformation, integration, cleaning and normalization for data preparation; and feature selection, instance selection or discretization to reduce the data complexity.

In our project, we try to preserve the relationship among the attributes and maintain the high-quality of data. So we will mainly use data cleaning techniques. In the paper from Rahm et al.[2], they introduced related problems and approaches to data cleaning. As for the correlation, there are several correlation coefficients measuring the degree of correlation. The most common of these is Pearson correlation coefficient.

Dataset

Since one of our objectives is dealing with a spatial-temporal data, we will use dataset including spatial and temporal information. "NYPD Complaint Data Historic[3]" will be an example because it includes all valid felony, misdemeanors, and violation crimes from 2006 to 2016 in the New York City. To analyze more correlation from the different dataset, we can use The City Record Online (CROL) database from NYC open data[4].

For data cleaning, We will fill in missing values. Along with that, we will identify outliers and smooth out noisy data using the techniques such as binning, clustering, and regression. We will also correct inconsistent data.

Correlation analysis

Pearson correlation coefficient will mainly be used in this project. Other measurements like Intra Class Correlation (ICC) and Distance Correlation will also be covered. We will construct a correlation matrix using the classification data over time between spatial bins (e.g. zip code) that we created in preprocessing step. This correlation matrix will represent the relationship between the areas over time. If we consider it as a graph matrix, the areas will be nodes and the correlation coefficient will be edge weight. We can construct matrices with the different time range. The correlation matrices constructed from same time range from different data can be used to form another correlation for further experiments. In addition to that, we can extract clusters from the correlation matrix.

Evaluation

We will perform the evaluation of three aspects. Firstly, we want to check the correctness of our tool. Because there is not too much ground truth based on the NYC urban data, we will use our tool to process the data, and check the correlation result using our prior knowledge. Not only will we test the correlation between the columns among the same dataset, we will also test the correlation between different datasets. In addition, we will visualize the results for better understanding. Secondly, we want to check the efficiency of our tool. So we will run several experiments executing different queries and test the time it takes. In the end, we will conduct a detailed study on multiple NYC related data and find interesting correlations. This step should be based on the first task we proposed above. Because there could be some unexpected correlations which should be tested and checked.

References

[1] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Analytics*, vol. 1, no. 1, p. 9, 2016.

[2] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.

[3] "NYPD Complaint Data Historic (August 2017 updated); Police Department (NYPD); available from: https://data.cityofnewyork.us/public-safety/nypd-complaint-data-historic/qgea-i56i,"

[4] "City Record Online (Oct 2017 Created; Department of Citywide Administrative Services (DCAS); available from:https://data.cityofnewyork.us/city-government/city-record-online/dg92-zbpx),"