Crime Rate Inference with Big Data Summary

Carlos Petricioli (cpa253)

October 31, 2017

Abstract

Crime is one of the most important social problems in the country, affecting public safety, children development, and adult socioeconomic status. Understanding what factors cause higher crime is critical for policy makers in their efforts to reduce crime and increase citizens life quality. We tackle a fundamental problem in our paper: crime rate inference at the neighborhood level. Traditional approaches have used demographics and geographical influences to estimate crime rates in a region. With the fast development of positioning technology and prevalence of mobile devices, a large amount of modern urban data have been collected and such big data can provide new perspectives for understanding crime. In this pa- per, we used large-scale Point-Of-Interest data and taxi flow data in the city of Chicago, IL in the USA. We observed significantly improved performance in crime rate inference compared to using traditional features. Such an improvement is consistent over multiple years. We also show that these new features are significant in the feature importance analysis.

Link

- Direct link to pdf:
  `http://www.kdd.org/kdd2016/papers/files/adp1044-wangA.pdf`
- Link to DOI:
  `https://dl.acm.org/citation.cfm?doid=2939672.2939736`

Authors

- Hongjian Wang, hxw186@ist.psu.edu, College of Information Sciences and Technology Pennsylvania State University, University Park, PA, USA

- Daniel Kifer, dkifer@cse.psu.edu, Department of Computer Science & Engineering Pennsylvania State University, University Park, PA, USA
- Corina Graif, corina.graif@psu.edu, Department of Sociology and Criminology Pennsylvania State University, University Park, PA, USA
- Zhenhui Li jessieli@ist.psu.edu, College of Information Sciences and Technology Pennsylvania State University, University Park, PA, USA

Summary

In the paper [1] the authors propose to study two newer types of urban data: POI and taxi flow, data reflect how people commute in the city. They hypothesize that taxi flows may be considered as hyperlinks in the city that connect the locations and use such data to estimate crime rates. Taxi flows may be a proxy for broader patterns of population routine activity and mobility, commuting flows, and other forms of social and economic exchanges between two communities over space.

They experiments including a systematic comparison between linear regression and negative binomial models, tests of different combinations of features, detailed discussions of how to construct features, analysis of the relative importance of features, and theoretical interpretations of the results from a social scientist (a co-author in the paper). In summary, the contribution of this paper are:

1. They study crime inference problem by utilizing new urban data: POIs and taxi flows.
2. Utilizing these data improves the crime rate inference.
3. Experiments are used to compare different results and feature combinations.

All the related work can be categorized in the following categories:

Time-centric paradigm. This line of work focuses on the temporal dimension of crime incidents.

Place-centric paradigm. Most existing work adopt a place-centric paradigm, where the research question is to predict the location of crime incidents. The predicted crime location is usually referred by the term hotspot, which has various geographical size.

Population-centric paradigm. In the last paradigm, research focuses on the criminal profiling at individual and community levels.

They use POI to enhance the demographics information and use taxi flow as hyperlinks to enhance the geographical proximity correlation. They do not consider the temporal dimension of crime in depth.

Their problem is population-centric because they try to profile the crime rate for Chicago community areas, where, community areas are well-defined and stable geographical regions. The proposed POI features and taxi links provide new perspectives in profiling the crime rate across community areas.

The crime data collected in Chicago has detailed information about the time and location of crime and the types of crime. The term crime count refers to number of crime incidents in a region in a year. The community area is used as geographical unit of study, since it is well-defined, historically recognized and stable over time. Crime rate is the crime count normalized by the population in a region. We use vector to denote the crime rates in regions. The crime rate inference problem is to estimate the crime rate in one region using the crime rate of other regions in the same year by considering the features of regions and correlations between regions.

They study the crime rate inference problem, estimate the crime rate of some regions given the information of all the other regions. Without loss of generality, assume there is one community area t with crime rate $y_t$missing, and use the crime rate of all the other regions $\{y_i\} - y_t$ to infer this missing value. The problem is

$\hat{y_t} = f(\{y_i\} - y_t, X)$,

where X refers to observed extra information of all those community areas. They consider two types of features X:

Nodal feature. Describe the characteristics of the focal region, demographic information and POI.

Edge feature. (1) Geographical influence, crime rate of the nearby locations. (2) Hyperlink by taxi flow. Locations are connected through the frequent trips made by humans, which can be considered as the hyperlinks in space. Their hypothesis is that two regions that are more strongly connected through social flow will influence each others crime rate.

References

[1] H. Wang, D. Kifer, C. Graif, and Z. Li, “Crime rate inference with big data,” in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, (New York, NY, USA), pp. 635–644, ACM, 2016.