Chicago Crime: From Analysis to Prediction

Noemi Derzsy

11.02.2017
Analysis and Modeling of Criminal Activities

Centuries of studies:
• Criminology: branch of Sociology, studies causes and impacts of crime
• Social Sciences: social behavior aspect

Recently:
• Machine Learning: e.g. Kaggle competition: https://www.kaggle.com/c/sf-crime
• Physics: e.g. agent-based modeling, diffusion model
• Biology: e.g. epidemic spread

Data:
• public crime records available for major cities (i.e. San Francisco, New York, Chicago, etc.)
  -> continuously growing collection of data (smart cities)
Datasets:
- Police crime data (2002-2015)
- IUCR (Illinois Uniform Crime Reporting) code
- Census data
- Shapefile: geographic information of community areas
Violent Crimes in Chicago as Categorized by IUCR (Illinois Uniform Crime Reporting Codes)

- Robbery: 204978
- Theft: 1136878
- Motor vehicle theft: 255398
- Homicide: 6790
- Assault: 586633
- Offense involving children: 36251
- Burglary: 324542
- Arson: 8861
- Ritualism: 15
- Battery: 995615
- Criminal sexual assault: 20605
Criminal hotspots: clustered communities

High criminal neighborhoods exhibit dynamics over time

Spatio-Temporal Dynamics

Chicago Communities Crime Rate
Year 2002

Number of crimes
- 0 - 25
- 25 - 75
- 75 - 100
- 100 - 150
- 150 - 200
- 200 - 250
- 250 - 300
- 300 - 350
- 350 - 400
- 400 - 500
- 500 - 900
Criminal Activity Patterns
Criminal Activities by Day and Hour

Assault highest in weekend!

Battery highest in weekend!

Arson highest in weekend!

Robbery highest Friday, Saturday!
Criminal Activities by Day and Hour
Correlation:
Which demographic features are in relationship with criminal activities?
No correlation between high school degree and crime
Strong negative correlation between higher education and crime
Negative correlation between owned property and crime
Correlation Analysis Census Information and Crime

![Graph showing the correlation between poverty rate and crime rate. The Pearson correlation coefficient is 0.58.](image-url)
Correlation Analysis

Same correlation is visible for each year
The size of a circle is proportional to the number of crimes in that community area.
Clustering:
Do community areas cluster together based on some criteria?
K-Means Clustering

• popular clustering method

• given a set of observations \((x_1, x_2, \ldots, x_n)\), where each observation is a \(d\)-dimensional real vector, \(k\)-means clustering aims to partition the \(n\) observations into \(k\) (\(\leq n\)) sets \(S = \{S_1, S_2, \ldots, S_k\}\) so as to minimize the within-cluster sum of squares

• calculate Euclidean distance between data points -> data must be normalized!

• performance evaluation: silhouette score
Community Areas Clustering

Clearly delimited clusters of communities
Community Areas Clustering

Identical clustering seen for each year (2002-2015)
Community Areas Clustering

- **crime**: no/very low
- **housing** (property ownership): high
- **education**: high rate of high school
- **race**: mixed

- **crime**: no/very low
- **housing** (property ownership): very low
- **education**: no/very low
- **race**: mostly Hispanic/Latino and Asian

- **crime**: no/very low
- **housing** (property ownership): no/very low
- **education**: high rate of high school and university degree
- **race**: mostly Asian, White and Others

- **crime**: high
- **housing** (property ownership): no/very low
- **education**: very low
- **race**: mostly Black
Discover Feature Importance with Decision Trees

- Features: race, education, poverty rate
- Gini coefficient: class purity
- Classes: quantiles or split crime rate in equal sized classes
- Performance evaluation: pure classes in leaves
Discover Feature Importance with Decision Trees

- **Features**: race, education, poverty rate, property ownership
- **Gini coefficient**: class purity
- **Classes**: quantiles
Regression Analysis:
Can we forecast the future number of crime incidents from historical data?
Predicting the Number of Crime Incidents with Linear Regression

**Linear regression:** \[ y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_n y_{t-n} + b \]

Our approach:

- \( a_{i,t} \) - the number of crime incidents at location \( i \) at time \( t \)
- \( y_t = \{a_{1,t}, a_{2,t}, a_{3,t}, \ldots, a_{77,t}\} \)
- \( X_t \in \mathbb{R}^{L \times \tau \times n} \)

- \( L \) - the number of samples (total number of community areas)
- \( n \) - number of features
- \( \tau \) - timestamp

\[ y_t = wX_t \rightarrow w = \arg\min_w \|y_t - wX_t\|_2^2 \]
Predicting the Number of Crime Incidents with Linear Regression

With census data the prediction accuracy is improved!
the accuracy of short time window (less than 8 months) is improved a lot
- time window size = 12 months is a critical point
- predicting the crime number of next year is easier than that of next month
- when window size is too large, the accuracy decreases for the lack of training data
Classification:
Can we label accurately the type of crime incidents?
Classifying Crime Incidents by Category

Issues to solve:
- imbalanced data: classes are not represented equally
  -> balance it: bootstrapping or use same sample size

Features:
- prior (historical) number of crimes by category
- location (community area)

Naïve Bayes classifier:
- based on Bayes’ theorem
- naïve: assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature
Classifying Crime Incidents by Category

Classifiers:
- Decision Tree
- SVM
- kNN
- Naïve Bayes

Features:
- prior (historical) number of crimes by category
- location (community area)
- year
- month
- day of week
- time of day
- weekday/weekend

Performance evaluation:
- ROC curve (true positive/false positive)
Other Approach: Crime as Infectious Disease Epidemic

**Network Properties**

Nodes: 77
Edges: 204
\[ \langle k \rangle = 5.3; \quad k_{\text{min}} = 1; \quad k_{\text{max}} = 9 \]

**Hypothesis:** crime spreads in neighboring community areas as an infectious disease, can be modeled by epidemic models (i.e. SIR, SIS)

Will social contagion improve predictive capabilities of criminal activities?

Chicago Community Areas Network
Total violent crimes committed in 2002-2015
Conclusions

✓ revealed dynamics of criminal hotspots

✓ analyzed correlation between neighborhood demographic features and violent crimes

✓ discovered different patterns in regions of low/high criminal activities

✓ developed prediction algorithm to forecast the number of crimes in community areas

✓ showed that incorporating census information in the prediction algorithm improves the prediction capability

✓ we can classify crime incidents with increased accuracy with feature engineering

✓ machine learning tools are effective in analyzing, modeling and predicting criminal activities

✓ public crime record data continuously growing -> great time to study crime modeling capabilities