Data Mining for Crime Analysis

Julia Holland Chen
Applied Computer Science
University of Colorado Boulder
Boulder, CO USA
julia.chen-1@colorado.edu

ABSTRACT

Although violent crime has generally decreased over the last 20 years[1], the crime rate remains an important issue for Americans[2]. Available crime data continues to increase, as does our need to process and interpret the large amounts of raw data. Data mining techniques such as association rule mining, decision trees, and k-nearest neighbors will be used in this study in order to better understand and predict crime incidents in the City of Chicago.

KEYWORDS

Association Rule Mining, Apriori Algorithm, Crime Analysis, Data Mining, Decision Tree, K-Nearest Neighbor

1 Motivation

Determining when and where crime takes place in our cities is an important topic and relevant to everyone. In this study, we will examine data collected from the City of Chicago. In 2019, Chicago was found to be the 31st most dangerous city in America[3] which makes it a useful location for crime analysis. Using the dataset, this study will attempt to find patterns in location and time that may help police target their resources where they are most needed and lead to a better understanding of crime patterns in the City of Chicago. The dataset spans 22 years, 2001 until September 2023, and features include the date of incident, primary type of crime and description, and the ward and location where the incident occurred.

This study will focus on the following three questions:

- For each ward, which type of crime is the most likely to occur?
- Are there any significant frequent patterns of crime for time of year, type of crime, and type of location of the incident?
- Given incident features, is an arrest likely?

2 Related Work

Khatun et al[4] studied how data mining may be used by crime investigation agencies to discover relevant precautionary measures from prediction rates. The authors used decision trees, k-nearest neighbors (KNN), and random forest algorithms in their investigation. Forecasting was performed for the most frequently occurring crimes like robbery, assault, and theft. While their paper focused on crime forecasting, this study will also test frequent patterns to see if additional data may be extracted along with decision trees and KNN for crime prediction.

Jantan et al[5] explored the relationship between the category of location and area of crime incidents by applying the Apriori algorithm for association rule mining. The authors focused on type of crime in specific locations with the goal to produce rules for pattern visualization in crime analysis. This study will also look at location and types of crime but include other predictive techniques for analyzing the dataset.

Yerpude et al[6] studied prediction of crimes when looking at data from the City of Chicago. Some of the techniques they used in their study included decision trees, random forest classification, Naïve Bayes classification, and linear regression. They focused on determining features of high crime

areas and comparing the results for each of the prediction methods. This study will examine multiple areas of crime and include both association analysis and prediction.

3 Proposed Work

3.1 Data Collection

Data was collected from Data.gov and downloaded to personal computer.

3.2 Preprocessing

Dataset cleaning will be performed by first changing each feature to correct type. For NaN feature values of interest, remove rows if possible. Most of the features of interest have few NaN values. As one of the items of investigation is to determine crime during times of year, a new feature will be derived where it will contain one of four values (0,1,2,3) which signify winter, spring, summer, and fall and will be named "season."

3.3 Association Rule Mining

Using the Apriori algorithm, this study will first find all frequent itemsets from the features: season, primary type, and location description. This will investigate relationships between the time of year, type of crime, and what type of location the incidents occur. The selection of the minimum support will be determined after test runs on smaller portions of the dataset. The goal will be to have a minimum support that gives a good number of frequent itemsets that are not so large the list would be unmanageable. Once the frequent itemsets are determined, strong association rules will be calculated. Not all strong association rules are interesting so a pattern evaluation method will be added which will help to determine if the rule is truly interesting. Jantan et al[5] also studied the Apriori algorithm for association rule mining, with a focus towards producing a visualization system for police department uses. This study will not be visualization-based for exploring the results and will be focused on discrete rules, which meet the evaluation goals.

3.4 Classification

This study will use a decision tree to predict the type of crime given a ward. There are 50 wards in Chicago and 405 unique ICUR values used in our dataset. An additional topic which may be explored is to generate the top two or 3 types of crime in each ward. The classifier performance will be evaluated by computing a confusion matrix and reviewing the values for accuracy, recall, specificity, and precision.

The k-nearest neighbors algorithm will be used to predict whether or not there was an arrest made for the incident. This was also investigated in Khatun, et al[4] on the Chicago dataset so this study will attempt to replicate, noting that our dataset includes more recent samples and differing features for input data. The features for input will include ICUR, location description, domestic, season, and ward. As in the decision tree task, the performance evaluation will be determined by a confusion matrix and values for accuracy, recall, specificity, and precision.

4 Data Set

The dataset was published by the City of Chicago[7]. The dataset includes approximately 7.91M rows and 22 features.

Note that features which include an exact location are shifted slightly from the true location of the incident to protect privacy, but will still fall on the same block.

4.1 Feature Information

- ID – unique identifier, nominal
- Case Number – police department records division number, nominal
- Date – date of incident, ordinal
- Block – partially redacted address of incident, nominal
- IUCR – Illinois uniform crime reporting code, nominal
- Primary Type – primary description of the IUCR code, nominal
- Description – secondary description of the IUCR code, nominal
- Location Description – description of the incident location, nominal
- Arrest – if arrest was made, Boolean

- Domestic – if incident was domestic-related, Boolean
- Beat – beat where the incident occurred, nominal
- District – police district of incident, nominal
- Ward – City Council district of incident, nominal
- Community Area – community area of incident, nominal
- FBI Code – crime classification as outlined in the FBI National Incident-Based Reporting System (NIBRS), nominal
- X Coordinate – x coordinate of incident location, continuous
- Y Coordinate – y coordinate of incident location, continuous
- Year – year of incident, ordinal
- Updated On – date and time record was last updated, ordinal
- Latitude – latitude of incident location, continuous
- Longitude – longitude of incident location, continuous
- Location – latitude and longitude of incident location formatted for maps, nominal

5 Evaluation Methods

The evaluation methods will vary depending on the technique used. For association rule mining the results of this study will validate a strong association with a significant correlation between the itemsets. For the classification tasks, a confusion matrix will be calculated so accuracy, recall, specificity, and precision may be reviewed.

The results will also be evaluated against the following criteria: easily understood, valid on new or test data, potentially useful, and give new insights into crime patterns in the City of Chicago.

6 Tools

This study will use the Python language and Jupyter Notebook for analysis. Libraries used will include:

- Pandas, a powerful data analysis and manipulation tool[8] will be used to load and process the dataset.
- Numpy, a library for scientific computing[9] will be used to convert dataset into n-dimensional arrays for numerical computations.
- Scikit-learn is a library for predictive data analysis[10]. It will be used for decision tree and k-nearest neighbor algorithms.
- Mlxtend is a library which contains machine learning extensions[11]. It will be used for the Apriori algorithm.
- Matplotlib is a visualization library[12] and will be used for simple visualizations of the dataset.

7 Milestones

1. Data cleaning and transformation, 1 week, complete 11/6.
2. Initial algorithm tests with subset of data, 1 week, complete 11/13.
3. Review of initial tests and modifications if needed for larger run of algorithms, 1-2 weeks, complete 11/20-11/27.
4. Review of algorithm results on entire dataset, 1 week, complete 12/4.
5. Write up and create visualizations for results, 1 week, complete 12/11.

REFERENCES

[1] Crime in the United States. https://en.wikipedia.org/wiki/Crime_in_the_United_States

[2] Domenico Montanaro 2023. Poll: Dangers for both parties on the economy, crime and transgender rights. https://www.npr.org/2023/03/29/1166486046/poll-economy-inflation-transgender-rights-republicans-democrats-biden

[3] Elisha Fieldstadt 2020. The most dangerous cities in America, ranked. https://www.cbsnews.com/pictures/the-most-dangerous-cities-in-america

[4] Most. RokeyaKhatun, SafialIslam Ayon, Md. RahatHossain, Md. JaberAlam, Data mining technique to analyseand predict crime using crime categories and arrest records. Indonesian Journal of Electrical Engineering and Computer Science Vol.22, No.2, May 2021, DOI: 10.11591/ijeecs.v22.i2.pp1052-1060.

[5] Hamidah Jantan, Aina Zalikha Mohd Jamil, Association Rule Mining Based Crime Analysis using Apriori Algorithm. International Journal of Advanced Trends in Computer Science and Engineering Vol.8, No.1.5, 2019. DOI: 10.30534/ijatcse/2019/0581.52019.

[6] Yerpude, Prajakta, Predictive Modelling of Crime Data Set Using Data Mining (July 21, 2020). International Journal of Data Mining & Knowledge Management Process (IJDKP)

Vol.7, No.4, July 2017, Available at SSRN: https://ssrn.com/abstract=3656953
[7] Crimes - 2001 to Present. https://catalog.data.gov/dataset/crimes-2001-to-present
[8] Pandas Library. https://pandas.pydata.org
[9] NumPy Library. https://numpy.org
[10] Scikit-learn Library. https://scikit-learn.org/stable/
[11] Mlxtend Library. https://rasbt.github.io/mlxtend/
[12] Matplotlib Library. https://matplotlib.org/