GEORGE MASON UNIVERSITY Department of Computer Science

CS 584: Data Mining

Crime Analysis (Chicago) Final Report

TEAM- ALL THAT DATA

G: G01398498 Pramod Kumar Catari (pcatari@gmu.edu)

G: G01387625 Stephen Simon Dias (sdias3@gmu.edu)

G: G01329585 Srikar Pratap Susarla (ssusarl@gmu.edu)

1 Introduction

Data provide by: Chicago Police Department.
Important Labels in Data set : Primary Type(type of crime), Location Description , District , FBI code , etc..
Type of crime : This will help us classify the data in a given group.
FBI code : The law which was violated by the convict.
District : This will help us understand which district ranks high in crime.
Location Description: This will help us to group the crimes into different location clusters where the crime has occurred.
Sub labels : longitude, latitude.
Other Labels – Crime Id, Date, Year, Arrested, Weapons, Community. etc.

1.1 Project motivation

To understand and gain perspective over why and where crimes occur in terms of the location of the crimes, weapons used, laws broken and hopefully contribute to creating more awareness and predictability of these occurrences.

1.2 Our project goal is to study

What are the major types of crimes and where do they generally occur by classifying the crimes based on the type of crime and incident place (private house, apartment, parking lot. etc).

The location of the crimes and find if certain locations are prone to increased frequency of certain crimes and help with response time.

Be able to predict the type of crime based on certain factors such as the time of the crime, description, weapons used, incident place, location in the city and other factors.

1.3 Contributions

- Pramod Kumar Catari :

  My part in the project was to first explore the data-set and perform pre-processing steps and analyse the approaches to the data-set. I have taken ownership for the K-Modes implementation on categorical attributes to find similarities and gain insights on crimes based on the attributes such as Crime Type, Description and Location(premises). Also, contributed on PART-3 of the project to determine how to compare level of safety between districts where we have done it based on the crime frequency of the top occurring crimes.

- Stephen Simon Dias :

  My part in the project was to understood the data through cleaning raw data, finding patterns ,then with the help of longitude and latitudes given in the data set plotting the data of longitude and latitudes on the co-ordinate scale of Chicago, looking at the plot we discussed what method we should use and and which will give us the best result.
  After which my next plan was to implemented the K-Means Clustering part with the help of elbow method and and testing those models and found the best centroid so as give the location where the team should look for contact office (emergency rescue team) so that they cover all the areas and also be present over there in minimum time.

- Srikar Pratap Susarla :

  I have contributed in the initial planning and discussions on the approaches to be taken with the team. Later, I helped in PART-3 of the project to cluster the districts based on the level of safety. Then, the clusters were plotted over the map of Chicago City with different color codes representing the level of safety.

2 Methodology

In the work that we present we have used three different algorithms that we have studied in our present and previous bachelors degree class.Which are K-Means, K-Modes,We can implement Random Forest Classifier and the Decision Tree. All of which are implemented For better result which we decided to use after looking at the data-set of crime of Chicago city by the police of Chicago.

2.1 K-Means

To process the learning data, the K-means method in data mining begins with a first set of randomly selected centroids, which serve as the starting points for each cluster, and then performs iterative (repetitive) computations to optimize the placements of the centroids. In data science, K-Means Clustering is a straightforward but effective technique. K-Means Clustering has numerous real-world applications (a few of which we will cover here). This is introduction to clustering and K-Means Clustering, along with a Python implementation on a real-world data set.

The primary principle underlying partitioning methods like k-means clustering is to construct clusters in such a way that the total intra-cluster variation [or total within-cluster sum of squares (WSS)] is minimized. We want the overall WSS to be as small as feasible because it measures the compactness of the clustering. The Elbow technique examines the relationship between the total WSS and the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.

Non-numerical (categorical) data cannot be handled using K-Means clustering. However, category value can be mapped to 1/0. However, with high-dimensional data, this mapping is unable to produce reliable clusters. The K-Modes approach, also known as k-modes clustering, is then requested by those who wish to replace the clusters' means with their modes.

2.2 K-modes

A large portion of data in the actual world is categorical, such as gender and occupation, and categorical data is discontinuous and unordered in contrast to numeric data. As a result, categorical data cannot be clustered using the same procedures utilized for numeric data. We can instead turn to K-Modes since K-Means cannot handle categorical data because converting the categorical values to 1/0 cannot produce good clusters for large dimensional data.

By substituting the Euclidean distance characteristic with the truthful matching dissimilarity measure, the use of modes to represent cluster facilities, and updating modes with the maximum everyday categorical values

in each generation of the clustering technique, the okay-Modes technique modifies the traditional okay-method manner for clustering specific information. With those adjustments, the clustering procedure is positive to reach a local minimal outcome. seeing that clusters serve as centroids, the variety of modes may be equal to the wide variety of clusters wanted. The Hamming distance from information idea is the dissimilarity degree applied to k-Modes, as visible in Fig. here, the values of attribute j in items X and Y are represented by means of x and y.
The extra category fee mismatches there are among X and Y, the extra exceptional the two items are from one another. inside the specific dataset, an characteristic's mode is either "1" or "0," depending on which value is greater conventional within the cluster. The sum of the distances between each item inside the cluster and the cluster middle is minimized by way of the mode vector.

2.3 Random Forest Classifier

Random forest classifier is a top-tier classification algorithm known for its flexibility and practicality. It is an algorithm based on ensemble-based learning. Decision trees are the vital units of the Random forest classifier. These simple decision trees are produced in the training stage and then processed into the classification stage to collect the best vote. The tendency of decision trees to overfit their training set is corrected by random decision forests.Although their accuracy is not the best compared to a few algorithms, random forests typically outperform the high-performing decision trees. However, data variables might influence their performance.
Random forests apply the general bagging strategy to each individual tree in the ensemble during the training phase. Bagging continually chooses a random sample from the training set with replacement and then fits trees to these samples. No pruning is done as the trees grow. A free parameter that is quickly learned automatically utilizing the out-of-bag error is the ensemble's number of trees. The random forest is made up of a set of bootstrap samples that are produced from the original data set and a collection of decision trees. The entropy of a chosen subset of the characteristics is used to divide the nodes. The subsets that are generated using bootstrapping from the original data set are the same size as the original data set.

2.4 The Decision Tree

The Decision Tree is a Supervised learning approach that may be used to solve both classification and regression issues, however it is most commonly employed to solve classification problems. It is a tree-structured classifier, where internal nodes stand in for a dataset's characteristics,

branches for the decision-making process, and each leaf node for the classification result.
The Decision Node and Leaf Node are the two nodes of a decision tree. While Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches.The provided dataset's characteristics are used to execute the test or make the judgments.
A graphical representation to get all possible solutions to a problemdecision based on given conditions.It is called a decision tree because, like a tree, it starts from a root node and expands to other branches to build a tree-like structure.To create the tree, we use the CART algorithm, which stands for Classification and Regression Tree Algorithm.A decision tree simply asks a question and subdivides the tree into subtrees based on the answer (yes/no).

3 Experiment And Conclusion

PART 1: Location based clustering

Here, we have applied the K-Means clustering algorithm on the location data of the crimes .ie. the longitude and latitude attributes for each crime and divided the city into different clusters based on the nearest points.

Steps:
1. Created a data frame with the Longitude and Latitude data for each crime.
2. Normalized the data using the Standard Scaler.
3. Plotted the elbow-visualizer to find the optimal number of clusters.
4. Clustered the given location into 4 different clusters using K-Means algorithm.
5. Visualized the clusters by overlaying them over the map of Chicago city

Conclusion:
We can see the different clusters represented =by different colors in the generated image. The red points represent the centroids of theses clusters. Police stations and so on

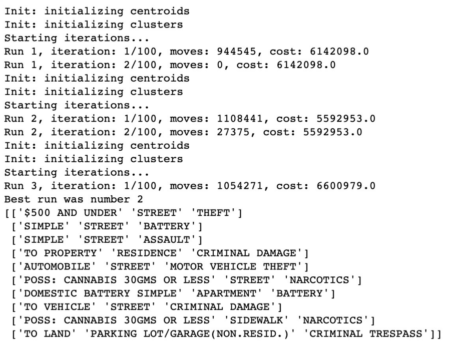
PART 2: Crime similarity using K-Modes Clustering

We will use the K-Modes algorithm, to understand similarities between crimes and gain insights on the commonly occurring crime types.
K-Modes defines clusters based on the similarities of categories between data points. Here, we want to identify specific attribute combinations on

crime type, location and description.

Steps:
1.Created a new data frame with the attributes ‘Description’, ‘Primary. Type’ and ‘Location Description’.
2. Run the K-Modes algorithm with 3 clusters on the data frame.
3. The centroids of these clusters would represent the similarities in attributes between the crimes in that cluster.

Image



```
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 1, iteration: 1/100, moves: 944545, cost: 6142098.0
Run 1, iteration: 2/100, moves: 0, cost: 6142098.0
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 2, iteration: 1/100, moves: 1108441, cost: 5592953.0
Run 2, iteration: 2/100, moves: 27375, cost: 5592953.0
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 3, iteration: 1/100, moves: 1054271, cost: 6600979.0
Best run was number 2
[['$500 AND UNDER' 'STREET' 'THEFT']
 ['SIMPLE' 'STREET' 'BATTERY']
 ['SIMPLE' 'STREET' 'ASSAULT']
 ['TO PROPERTY' 'RESIDENCE' 'CRIMINAL DAMAGE']
 ['AUTOMOBILE' 'STREET' 'MOTOR VEHICLE THEFT']
 ['POSS: CANNABIS 30GMS OR LESS' 'STREET' 'NARCOTICS']
 ['DOMESTIC BATTERY SIMPLE' 'APARTMENT' 'BATTERY']
 ['TO VEHICLE' 'STREET' 'CRIMINAL DAMAGE']
 ['POSS: CANNABIS 30GMS OR LESS' 'SIDEWALK' 'NARCOTICS']
 ['TO LAND' 'PARKING LOT/GARAGE(NON.RESID.)' 'CRIMINAL TRESPASS']]
```

An image of part 2 output

Conclusion:
Analyzing the clusters, some of the below insights were gathered:
Based on the clusters, we can gain the following insights such as:
1. ’Simple battery’ and ’Simple assult’ are the most common types of crimes occuring on the streets.
2. Thefts of ”$500 and under” type which occur on the street are the majority followed by Motor Vehicle Thefts of automobiles
3. Most of ’Criminal Damage’ cases takes place on residences where property is damaged and streets where vehicles are damaged.
4. Typically, domestic violance cases are of type ’Battery Simple’ and are seen mostly at apartment units.
5. Generally, Criminal Tresspass concerns arise in parking lots/garages of non-residential units. ’POSS: CANNABIS 30GMS OR LESS’ are the majority of the narcotic cases usually occuring on streets and sidewalks.

PART 3: District Clustering used on safety

We will again use the K-Means algorithm to classify the districts in Chicago City into different safety levels based on the crime density.

Steps:
1. Find the top 100 crimes based on value counts over the dataset.
2. Find the frequency of these crimes in each district. Create a pivot table to represent the same with districts as rows and Crime types as columns
3. Run K-Means to cluster the districts into different levels of safety.
4. The average values of centroids for each clusters will represent the level of safety for that cluster (lower means better) as it represents the no of crimes taking place.
5. Plot the districts color coded with the cluster they belong to.

Conclusion:
Here, we see that the districts color coded according to the level of safety. Red represents dangerous and Green represents safety.
Near West(12), Shakespeare(14), Austin(15), Morgan Park(22) and Rogers Park(24) are comparatively safer police districts
Grand Crossing(3), South Chicago(4), Gresham(6), Englewood(7) are comparatively more dangerous
The image on the right shows the districts marked on basis of levels of safety

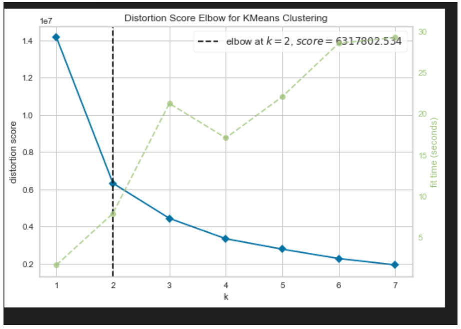
4 Future Work

Future Work: We tried to perform Crime Prediction on the attribute 'Primary Type' using the classifier algorithms such as Decision Tree Classifier, Random Forest Tree, Extremely Randomized Trees .etc . The resulting accuracy of 71% can be imporved using techniques such as better feature selection and our focus will be on the same going forward.

To access code for the project: `https://github.com/pcatari/584`

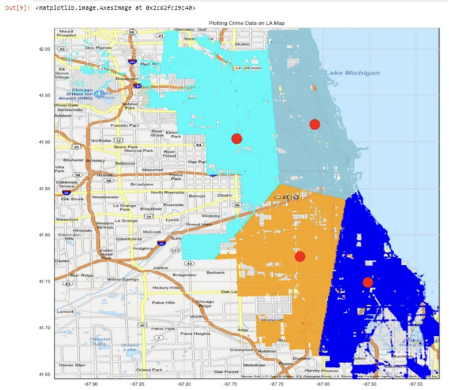
5 Images

Image 1



An image of First elbow methord produced

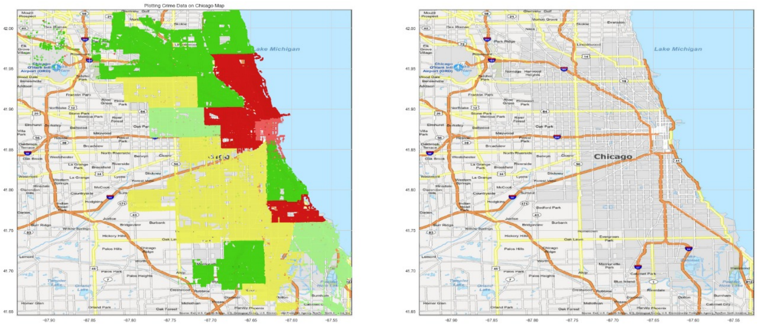
Image 2



```
Out[9]: <matplotlib.image.AxesImage at 0x2c62fc29c40>
```

An image of centroide maped on the Map of Chicago City

Image 3



An image of Safe and Unsafe districts in the city of Chicago