Crime Rate Analysis

Amrut Shenoy
aas2903@rit.edu

Ashwini Singh
ars1546@rit.edu

Deep Kataria
drk8921@rit.edu

Siddharth Bidwalkar
ssb9012@rit.edu

***Abstract* — The main aim of 'Crime Rate Analysis' is to predict occurrence of a crime in an area in a particular time interval based on a table having history of crimes occurred in that area. Using classification, we predict offenses based in particular region at a particular time. We use Weka tool to create J48 decision tree which is used for this classification.**

I. Introduction

Gone are the days when a person could walk on the street at any time of the day or night without having the fear of his or her safety. Safety has always been an important aspect for any individual. Its importance is increasing day by day. A crime irrespective of its intensity is a crime that should be prevented in the first place. There is a famous proverb which says "*Charity begins at home*". It wouldn't harm in saying that alertness and safety begins at home too. No doubt the police officials are always doing their duty but, considering the number of civilians that they have to look after, we as civilians have to take up the responsibility for our own safety. This is the sole reason for us to choose our project in this domain. Through our project 'Crime Rate Analysis' we wish to predict the probability of a particular crime occurring in a given place at a given time. This data can definitely act as a harbinger in providing us with a rough idea about any crime that has a high probability of occurring in a given time frame in a given place.

For our project, we needed one or more tables detailing crimes taken place in particular area or areas. We found the dataset as per our requirement on 'denvergov.org' website [1] which has record of all crimes taken place in Denver since year 2013. The dataset contained 18 attributes and 383981 instances. Each instance in the dataset describes a crime incident. The attributes contain details of crime like address where crime has taken place, type of offense, date and time of crime, date and time when crime was reported and neighborhood area where crime has taken place.

During the data cleaning phase, we noticed considerable amount of redundant data. The original dataset had 5 attributes describing location of crime and 5 attributes describing the offense type. To overcome this redundancy, we introduced two new tables. Attributes related to address were placed in address code table. These attributes were Incident address, longitude, latitude, district ID and neighborhood. Attributes related to offense type were placed in Offense Type table. These attributes were Offense ID, Offense Code, Offense Code Extension, Offense Type description and Offense Category. We performed cleaning on the data which is described below -

<u>Cleanup Process</u>

The first step of data cleaning was to remove instances in which data was missing in attributes relevant for classification. Since the aim of project was to predict crimes based on address and time, we removed instances where data of either address of crime, time of crime or type of crime committed was missing. This phase of cleaning removed considerable number of instances and the size of dataset was down to 90000.

After creating Address and Offense type tables, we introduced new attribute in each of them which served as their primary keys. Address ID and Offense ID were primary keys of Address and Offense tables respectively. These two attributes were added to crime table as foreign keys. Due to creation of these new tables, the crime table was reduced to 10 attributes and redundant data was eliminated from it.

Address code table contains Address ID (address_id) as primary key. Each key denotes unique address consisting of detailed address of the location, latitude and longitude of place where crime has taken place, neighborhood ID, description of neighborhood and address bin. For example, if crime has taken place at Crittenden Apartments in Rochester, then detailed address of the location would be exact address of crime location like apartment number, building number, etc. The neighborhood for this example would be Henrietta as Crittenden Apartments is located in Henrietta area. Address bins are formed using neighborhood IDs with all addresses in one neighborhood placed in one address bin. Address ID and Address bins are present in crime table as foreign keys.

Offense codes tables contains Offense ID (offense_id) as primary key. Each key denotes a unique type of offense which has taken place till now. Other important attributes of offense table include offense type description, offense category and offense bin. For example, crime description of 'Criminal Mischief' is included in 'Public Disorder' offense category whereas crime description of 'Heroin Possession' is included in 'Drug & Alcohol' offense category. Offense bins are formed using Offense category with all offenses in one offense category placed in one bin. Offense ID and Offense bins are present in crime table as foreign keys.

The main table which is used for analysis is the crime table. The important attributes of crime table are Incident ID, Address ID, Address Bin, Offense ID, Offense Bin, Crime Date, Crime Time, Reported Date, Reported Time and Time Bin. Incident ID is primary key of crime table which is the ID for each incident recorded. Time Bin attribute can contain only 4 values (1 , 2 , 3 , 4) where time bin 1 represents crime occurred between 12 am to 6 am , 2 represents crime occurred between 6 am to 12 pm , 3 represents crime occurred between 12 pm to 6 pm and 4 represents crime occurred between 6 pm to 12 am.

In original dataset, there were 68000 address codes, 500 offense codes. The bins for these attributes were created to reduce the values in a logical manner, thus making classification process faster. Thus, 68000 address codes were reduced to 50 address bins using their common neighborhood. 500 offense codes were reduced to 8 offense bins using their common crime category.

After cleaning, we had 90000 instances and 12 attributes. Out of these, the important ones were Address bin, Time bin, Precinct bin and Offense bin which were used for classification.

II. Database Design

Database contains three tables. They are crime, Address codes and Offense codes. Primary key of Address codes table is Address ID and that of Offense codes table is Offense ID. From Address codes table, values of Address ID and Address bin are included in the crime table. From Offense codes table, values of Offense ID and Offense bin are included in the crime table. Both tables share 1:1 relationship with crime table.

III. Analysis

For our project, we decided to analyze the data and predict the occurrence of a crime at a given place in a given time frame. In order to do so, we considered the Address ID, Time bin and Precinct bin. Using these values, we tried to predict Offense bin. Here, Offense bin represents the nature of crime that can occur at a location obtained from Address ID and Precinct bin in a time-span specified by the Time bin. Hence, we decided to use of these four attributes from Crime table in order to predict the type of crime that might occur.

Firstly we divided the instances in the data into training and testing sets. This was required as we needed to train the model first. On the completion of training sets, we wanted to test whether the model developed was able to accurately predict the outcome. Approximately 60 percent of the data was used as a training set whereas the remaining instances constituted the testing set. The instances in the training set were chosen randomly. It was made sure that the training set contained all the possible values of every attribute present in the table. This ensured that when testing set was provided, the model did not encounter any unseen values.

We started off by making use of ZeroR classifier in Weka. Although this classifier built a model whose accuracy was acceptable, the confusion matrix generated was highly skewed towards one particular Offense bin. We realized that ZeroR classifier simply predicts the majority category as the predicted outcome. Suppose we have five categories out of which category 1 had the highest value, then the outcome of the remaining categories were also directed towards category 1. This was of no use to us as we wanted to predict the outcome for every type of crime that could be committed at different locations. Hence, we decided to use a different classifier to solve this classification problem.

We used the J48 decision tree classifier for this purpose. The model we used was able to predict the type of crime at a given location and at a given time. The accuracy of the model was acceptable. Also, on inspecting the confusion matrix of this model we found that the model was able to predict all the categories with a good amount of accuracy. We also tried using Naive Bayes classifier to solve the classification problem. Although this model was also able to predict the type of crime correctly, J48 model was able to classify and predict the outcomes slightly more accurately. Hence, we decided to stick to using J48 decision tree classifier in order to solve the classification problem and be able to predict the type of crime that could occur at a given location in a given time-span.

IV. Results

The model generated using J48 decision tree classifier had an accuracy of 82% approximately for the training set. On testing the model with the testing set we found that the accuracy of the model was around 72%. The tree developed by the model has Precinct bin as the root node of the tree. This tree splits the data on the basis of the Address bin. Next it splits the data based on the time in Time bin attribute. Finally, at the leaves of the tree we find the Offense bin. Offense bin contains the type of crimes.

For every Precinct bin, we are able to predict the type of crime that can occur in that Precinct at any given point of time in different addresses covered under that Precinct. For example, if for a Precinct 159(present in the Precinct bin as 100-199), if we want to predict what type of crime occurs at 12:30 am(present in the Time bin as Night) at address ID 1298(which corresponds to Address bin 1000-1999), then we do that by predicting the value of Offense bin(it contains a range). The range is common for a particular type of crime. So assume if we find that "Burglary with forced entry into residence" was the kind of crime committed and its Offense Code is 2202. This falls under the bin 2000-2999(in Offense bin) which contains all Offense Codes related to "Burglary". This is what we try to predict. The Offense bin helps us identify the nature of crime that can occur.

V. LESSONS LEARNED

One of the most important thing we learned is that data cleaning is an important aspect of any data project. Although it was time consuming and tedious task, it helped us achieve desired classification results more accurately. Proper planning and careful analysis of the data is required so that we don't end up modifying or removing crucial data as a part of the cleaning process. Another important lesson we learnt is that deciding on the attributes that we are going to use for classification is an important part of the project. It requires careful evaluation of the data at hand such that the attributes we select lead us to our ultimate project goal.

After data cleaning process, the data in hand with us was still large. Then we realized the importance of binning the data of specific attributes. Although this process made the value range of certain attributes small, it kept essence of data intact. Also, when we tried out different decision trees for classification, we realized not to be satisfied with just one approach of solving a problem even if it gave us desired results. There is always another approach which can give more accurate results.

REFERENCES

[1] The dataset can be found at- http://data.denvergov.org/dataset/city-and-county-of-denver-crime