Algorithms to predict a future crime using Data Mining

Aarthi Vishwanathan
Department of Computer Science and Engineering
PES University
Bangalore, India

Bhargava Bodas
Department of Computer Science and Engineering
PES University
Bangalore , India

Chandini Velilani
Department of Computer Science and Engineering
PES University
Bangalore, India

Harshit Pandey
Department of Computer Science and Engineering
PES University
Bangalore, India

***Abstract*—** In this project, given time and location, we predicted the category of crimes that occurred from 1/1/2003 to 5/13/2015 in San Francisco's neighborhoods.

We chose to perform classification using naive bayes after analysing the most appropriate among investigated classification models including Na¨ıve Bayes, k-NN, decision trees and SVM and analyze their pros and cons on this prediction task.

The results obtained were fairly accurate predicting the type of crime which could occur next

***Keywords*—**Crime-patterns,clustering,datamining, law-enforcement

INTRODUCTION

In almost every period of western civilization, the inexorable increase in crime has been lamented in the corridors of power, the media, and the public.
Haunted by recollections of a previous golden age, people have used crime statistics, research, and the almost daily barrage of media stories as a basis to conjecture about the changing nature and scope of crime, including dire predictions for the future.Security is considered to be one of the major concerns and the issue is continuing to grow in intensity and complexity. Security is an aspect that is given top priority by all political and government organizations worldwide and measures are being taken to reduce crime incidence.
Criminology is an area that focuses on the scientific study of crime and criminal behavior and law enforcement and is a process that aims to identify crime characteristics. It is one of the most important fields where the application of data mining techniques can produce important results. Crime analysis, a part of criminology, is a task that includes exploring and detecting crimes and their relationships with criminals.The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Identifying crime characteristics is the first step for developing further analysis.

SUMMARY OF LITERATURE REVIEW

Below are summaries of the most relevant references in the literature review.

- The paper 'An enhanced algorithm to predict a future crime using data mining proposes the use of a mix of of data mining techniques for developing such a crime analysis tool. Use of classification algorithms in order to predict future crime behaviour based on previous crime trends was the main highlight.The C4.5 decision tree algorithm was used to predict the crime trends for the subsequent year. Experimentally,

results have proven that the technique used for prediction is speedy with high accuracy.

- The second reference is a framework for crime trends was introduced by comparing all individuals using a new distance measure and then clustering them as needed.This procedure provided identification of various criminal types and provides a visual clustering.

PROBLEM STATEMENT

Crime rates are increasing considerably day by day. Crime cannot be predicted easily since it is neither systematic nor random. Also the modern technologies and hi-tech methods help criminals in achieving their misdeeds, According to Crime Records Bureau , crimes like burglary, arson etc have decreased while crimes like murder, sex abuse, gang rape, etc. have increased. Although we cannot predict who all may be the victims of crime , we can try and predict the place that has greatest probability for its occurrence.

The negative impacts of crime to society involve the following:

- Depopulation, particularly in urban areas.
- High levels of crime may damage community spirit and result in less neighbour interactions..
- High crime levels can contribute to environmental poverty
- Once a region with a high level of crime is labelled as a potential area, it might become a ghetto

From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz.

Today, the city is known more for its tech scenes rather than its criminal past. But, with rising wealth inequality, housing shortages, and a proliferation of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay.

The knowledge that is gained from data mining approaches is a very useful tool which can help and support police forces.The idea here is to try to capture years of human experience into computer models via data mining. In the present scenario, crimes are increasingly influenced by the role of technology. Therefore, police needs such a crime analysis tool to catch criminals and to remain ahead in the eternal race between the criminals and the law enforcement. The police should use the current technologies to give themselves the much-needed edge.

Availability of relevant and timely information is of utmost necessity in conduction of daily business and activities by the police, particularly in crime investigation and analysis of criminals. Police organizations everywhere have been handling a large amount of such information and huge volume of records.

An ideal crime analysis tool should be able to identify crime patterns quickly and in an efficient manner for future crime pattern detection and action. However, in the present scenario, the following major challenges are encountered.

- Increase in the size of crime information that has to be stored and analyzed.
- Developing appropriate techniques that can accurately and efficiently analyze this growing volume of crime data
- The data available is inconsistent and are incomplete thus making the task of formal analysis a far more difficult.
- Investigation of the crime takes longer duration due to complexity of issues

All the above challenges motivated us to focus on providing solutions that can enhance the process of crime analysis for reducing crime rates. The main focus is to develop a crime analysis tool that assists the police in

- Detecting crime patterns and perform crime analysis
- Identify and analyze common crime patterns to predict and reduce further occurrences of similar incidences.
-

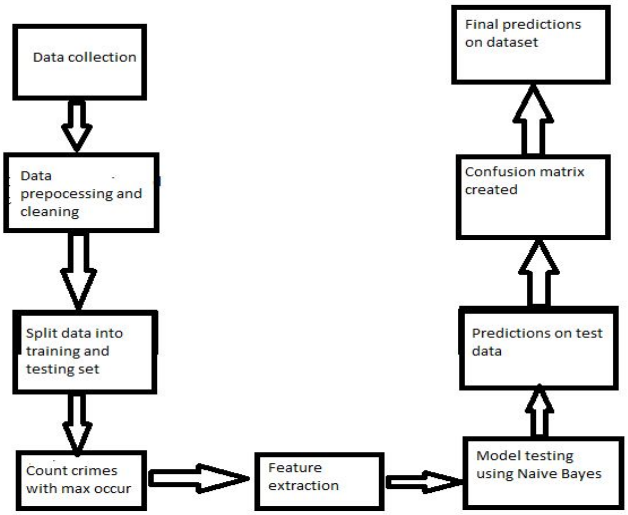
From Sunset to SOMA, and Marina to Excelsior, our chosen dataset provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods. Given time and location, we predict the category of crime that occurred.

The dataset has the following attributes :

- **Dates** - The date and time of crime occurrence.
- **Category**- The category of the crime eg: Robbery,Larceny/Theft,Vandalism etc.
- **Descript** - Brief detail of crime event.
- **DayOfWeek** - Which day the crime occurred
- **Pd District**- The police district holding the crime.
- **Resolution**- How was the crime resolved
- **Address** - The detailed address of crime occurrence
- **X** - x-coordinate of event place.

- **Y** - y-coordinate of event place.

Proposed System



Explanation of steps involved

Step1 : Data collection

The San Francisco crime dataset was collected using online sources. Websites were searched to find an interesting crime dataset that could be used for prediction. Using a sample dataset , we performed initial testing and prediction,however they proved inaccurate and thus had to be discarded.

The final dataset that we deemed fit was the San Francisco crime dataset that can be found here :

dataset - https://www.kaggle.com/c/sf-crime/data

Step2 : Data preprocessing and cleaning

The amount of null values were extremely less, so we decided

to remove them from our dataset. Attributes that were not required for prediction were removed to make the task simpler.

Step3 : Split Data into training and testing set

The dataset was split using a split factor of 0.8 into training and testing set. The training set was used to come up with predictions while the test set was used to evaluate the accuracy of our model.

Step4 : Count crimes with maximum occurrence

In order to predict the next type of crime that could occur in a district, the occurrence of prevalent crimes were counted.

Step5 : Feature selection and extraction

We adopted proper representation of Dates, Hour as our feature set

Dates: We extracted year, month, and hour from this field and used their trivial representation as our features. Day wasn't included, for it could be reflected by day of week.

Hour: We represented Hour from 0-24 and gave names to them.

Step 6 : Model testing using Naive Bayes

The Naive Bayes model was applied to find out probabilities of occurrence of the crime categories in various districts.

Step7 : Predictions on test data

Predictions were made on test data to verify the probabilities.

Step8 : Confusion Matrix created

The confusion matrix was created to check for accuracy of the results.

Step9: Final predictions on given dataset

The final prediction was made on the given dataset

Experiments and Results

We collected data for the San Francisco dataset which has crime data for 12 years from 2015 onwards.

In this dataset, there are 39 types of crimes in total, among which the top 5 most frequent ones are theft, other offenses, non-criminal, assault, and drug/narcotic.
These made up for more than half of the total crimes.

We also explored which police department handled the most number of crimes so that it may give an indication of the frequency of crimes in that place.The top five turned out to be Southern, Mission, Northern, Bayview and Central.

To verify the above findings , we used a map depicting crime concentration in San Francisco from Google Images as follows and found around 60% similarities as the maps weren't exactly available for the required timeframes but gave an idea of most concentrated crime areas.

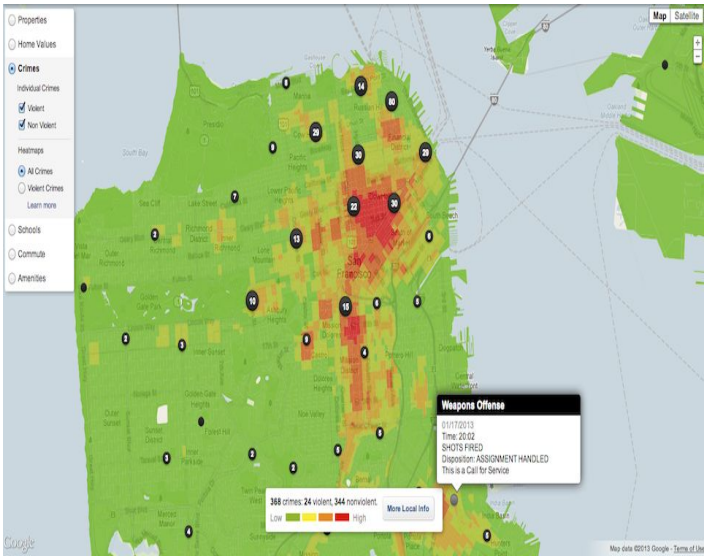


Figure 1

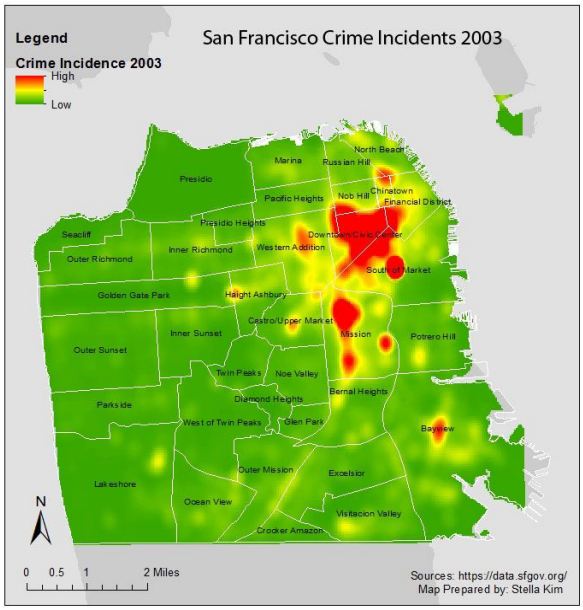


Figure 2

We plotted graphs for analysing Crime Category vs Occurrence as well as Crime Category vs District and obtained graphs as below.We observed that Friday had the most number of crime incidents and Saturdays had the least.

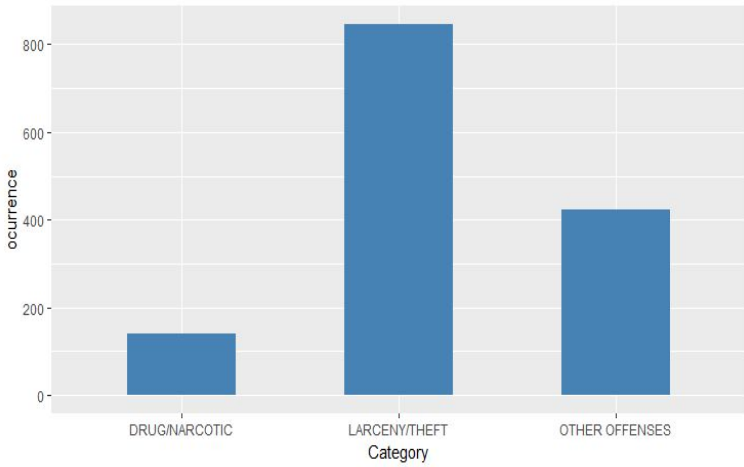


Figure 3

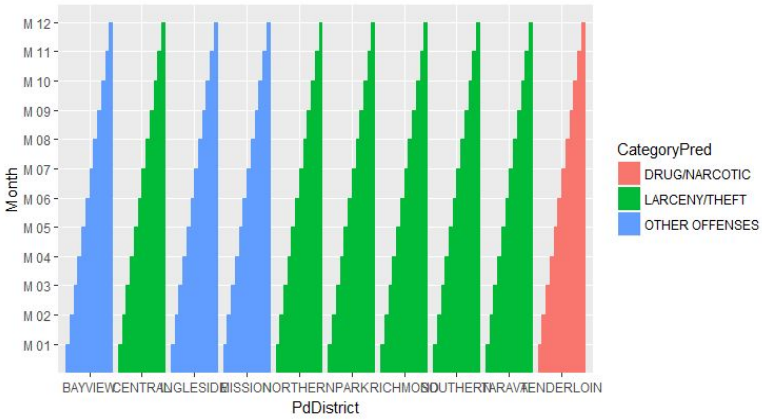


Figure 4

The following steps describe our approach for the prediction process:

- The few NAs observed in the dataset were removed as they seemed insignificant.
- No noticeable outliers were present.
- We divided the dataset into training and testing sets.
- We added new columns into the training set to separate out the day,month,year and hour.
- Then we used a function to classify the hour into Night,Pre-Job,Morning,Afternoon,Evening.
- Next the training set was used to train the naive bayes model which we implemented in R.
- Test set was then input to the model for prediction.

The results obtained were visualised by drawing the confusion matrix. Around 65% of accuracy in predictions was obtained consistently. Most of the records were predicted to be larceny/theft and other offences which hold similar to our initial findings above.So the result seems reasonable as the most occuring crimes still remain a majority in the prediction.

Conclusion

The project results overall seemed to agree with the existing data. However we notice that the crime trends are susceptible to change..So to ensure robustness of the prediction model we need to collect more data and also dynamically keep adding to dataset as and when new events occur to ensure reasonable accuracy over a continuous time range.

Naive Bayes performed reasonably well but we found that it was relatively time consuming. Thus, we need to look for faster ways of implementation.

References

Reference 1:

The paper 'An enhanced algorithm to predict a future crime using data mining' authored by Malathi. A (Assistant Professor Post Graduate and Research in the Dept. of Comp Sc. , Government Arts College,Coimbatore, India). and Dr. S. Santhosh Baboo (Reader,Post Graduate and Research,Dept. of Comp Sc. ,DG Vaishnav College Chennai, India) proposes the use of a mix of of data mining techniques for developing such a crime analysis tool. For this purpose, the following specific approach were formulated:

- To create a cleaning algorithm that cleans the given dataset by deleting unwanted data and fills in missing values with various techniques.

- Use of clustering algorithms to search for crime patterns in historical data.

- Use of classification algorithms in order to predict future crime behaviour based on previous crime trends.

The C4.5 decision tree algorithm was used to predict the crime trends for the subsequent year. Experimentally, results have proven that the technique used for prediction is fast with high accuracy.

Result1: Crime is either steady or dropping. The rate of sexual harassment is the main issue , along with slight incidences of murder , dowry death , dacoity , homicide.

Result2: Crime is either rising or is unstable due to Rioting, cheating, counterfeiting, cruelty by family members.

Result3: Crime is generally increasing. Thefts are the primary crimes on the rise with some increase in fire-raising. Lesser cases of the property crimes: burglary and theft are observed.

Result4: Few crimes are unstable, the main ones being murder,rape and arson which are also unstable. Property crimes such as burglary and theft have less change. Issue is to demonstrate at least some characteristics of these clusters. Experimental results prove that the tool is effective in terms of analysis speed, identifying common crime patterns and future prediction.No limitations were reported as such for the process.

Overall we perceived this as a focused approach however noticed lacuna in the predicting ability of the system to possible future crimes.

Reference 2:

A framework for crime trends was introduced by comparing all individuals using a new distance measure and then clustering them as needed.This procedure provided identification of various criminal types and provides a visual clustering.

CONTRIBUTIONS

01FB14ECS057 - Chandini Velilani - Exploration of dataset, field study, decision on problem statement, literature review, data preprocessing,code implementation and testing of code, analysing results , final report :-

1.Title, authors, affiliations

2. Abstract

3. Introduction to the context

4. A very brief summary of the literature survey report

5. Problem statement

6. Experiments and results with a discussion on the results

7. Conclusions

01FB14ECS004 - Aarthi Vishwanathan - Exploration of dataset, field study, decision on problem statement, literature review,code implementation and documentation,testing of code,debugging , analysing results,video creation, final report:-

1. Title, authors, affiliations

2. Abstract - summary of what you have done and the results you have obtained

3. Introduction to the context

4. A very brief summary of the literature survey report

5. Proposed system - a block diagram summarizing approach

6. Detailed explanation of each component of the system

7. References

01FB14ECS054 - Bhargava Bodas

Data preprocessing and cleaning,code implementation and documentation, Final report (Audio).

01FB14ECS080 - Harshit

Choosing the dataset,editing lit survey, code testing and training, final report(Audio) .