International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 04 | Apr -2017 www.irjet.net p-ISSN: 2395-0072

CRIME ANALYSIS: PROBABILISTIC PREDICTION

P. Monish[1], K. R. Ranjith [2], G. Varun[3], S. Sridhar, B.Tech, M.E.[4]

[1] Student, Dept. of Computer Science Engineering SRM University, TamilNadu, India
[2] Student, Dept. of Computer Science Engineering SRM University, TamilNadu, India
[3] Student, Dept. of Computer Science Engineering SRM University, TamilNadu, India
[4]Assistant Professor, Dept. of Computer Science Engineering, SRM University, TamilNadu, India

---

Abstract – *Events of crime and illegal activities have increased in the past few years. We propose a system which can analyze, detect and predict various crime probability in a given region. To accomplish this, we obtain raw data from police department and pre-process the data as per our needs. In this project we have used real data from san Francisco Police Department's official website. On this pre-processed datasets, by applying Naïve Bayesian and J48 Classifier algorithms we create a predictive model which analyze the data and helps to predict the trends of crimes for a given region in near future. Our project utilizes the hierarchical structure of the given data for prediction. This probabilistic trend is also displayed in form of graphs for easy understanding of the police department.*

Key Words: Analyse, Crime probability, Naïve Bayesian Classifier, J48 Classifier, Predictive model, Hierarchical structure, probabilistic trend.

1. INTRODUCTION

In the recent years, there is a considerable increase in rate of Crimes and Illegal activities. According to Crime Records, in most of the cases a particular type of crime is very limited to specific regions since the criminals associated with those type of crime are also limited and are very aware of the locality. Also with modern technologies, criminals are able to innovate new methods to get away from the hands of justice. The number of crimes starting from day to day petty cases to brutal murder cases has increased drastically over the decade, but the type of crime tends to remain same. Historically solving crimes has been the prerogative of the criminal justice and law enforcement specialists. This issue being a very serious concern for both the innocent victims and the security authorities, the methods used by security authorities are slow and is applicable within a small region or area.

With the use of the computerized systems to track crimes, computer data analysis will help the law enforcement officers and detectives to speed up the process of understanding the crime patterns and trends for a given region, thus helping the police department to take precautionary measures according to the trend of crimes. It is important to understand that Crimes cannot be predicted since it is neither systematic nor random event. Even though the crimes cannot be predicted, the type of crimes are very limited to particular region. This information can be used to predict possible type of crimes associated with a given region.

1.2 Challenges and Our Proposed Solutions

In order to build this predictive model, we face challenges with raw dataset. The real datasets obtained from San Francisco Police Department's repository are very vast and contain many manipulation difficulties such as redundancy of data, unformatted data, irrelevant information and vast number of categories for certain attributes.

To overcome the difficulty with source Datasets, we preprocess the dataset to fit our classification and prediction model. In this preprocessing phase we manipulate the data as required by our system to give better results. Initially, various raw datasets from various sources may contain different names for the attributes. We rename the attributes to our standards. The date and time is formatted to our standards. Unused/ excess information is removed from the dataset. Finally the redundant data is removed to increase the performance of the system. We perform these manipulations with various functions in R Language. The attributes used in our project are: Crime ID, Type Of Crime, Crime Descript., Date, Time, Area, Location Coordinates and Resolution. The pre-processing of the source datasets is very important as a pre-processed dataset gives better and efficient results compared to a raw dataset. Moreover, the raw data set may contain duplicated data or unused data, which reduces the efficiency of the system.

Our other challenge deals with the prediction model and the plotting of the probabilistic values. As we obtain the output in form of probability, it is very important to choose a good algorithm for the work. The data we are using is crime data and data is linear. Thus we are using linear regression for our prediction model, where the models predict the dependency between two variables and further provides us with the results. We also plot these obtained predicted probability in the form of graph to provide ease of use to the police department. Using these graphs we can tell the prior trend and also the change (increase/decrease) in trend of crimes for a region. Hence police department can take precautionary action for those types of crimes in the specified region.

1.3 System Architecture

The Diagram 1.3.1, explains the various steps involved in our project.

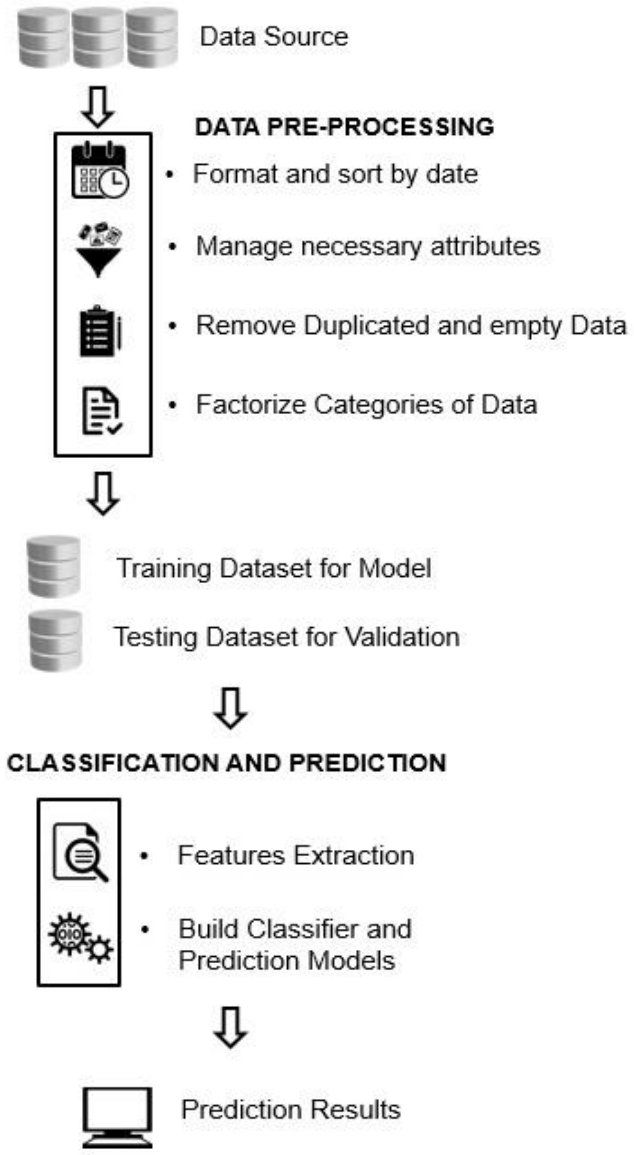

Figure: 1.3.1- System Architecture

2. METHODOLOGY

Steps of methodology followed are:

1. Data gathering: This is the first module of our project. In this step, datasets are gathered from the police department. We have obtained the year wise crime datasets from the San Francisco Police Department's Data repository. The obtained datasets are very huge and contains millions of entries for each year, and also contains many manipulation complexities.

2. Pre-processing: In this step, the obtained dataset is put into place to apply the data mining techniques. To do this, traditional pre-processing techniques such as transformation of variables and data cleansing can be applied. Some other techniques such as the selection of attributes and the re-balancing of data is also applied in order to deal with the problem of high dimensionality and imbalanced data present in these datasets. The pre-processing also makes sure that only the relevant information is kept which will help to produce good accuracy and efficiency in our predicted results.

3. Classification of Data: In this step, classification algorithms are applied to the pre-processed dataset. In our project we have applied J48 classification algorithm and for comparison we have also implemented the same with the Naïve Bayesian algorithm.

4. Predictive Probability: From the predictive classification functions thus obtained data the probability of a crime type to occur in specific area/region is displayed as end result.

A. Data Gathering

The type of crimes associated with a specific region can be affected using various factors. These factors can be location coordinates, date, time, number of resolved crimes etc. Data gathering is the first step in which raw data is collected. We collected datasets of real crime from the data respiratory of San Francisco Police Department. All these datasets are vary raw and contains duplication, missing data, irrelevant attributes etc. Various factors identified from this data are in table 2.1.1

FEATURES	Description
Crime ID	Unique Id for each Case
Type Of Crime	Category of the crime. Ex: Theft, Murder etc
Date and Time	The Date and Time of crime in standard format.
Day of Week	The Day of week of the crime.
Area	The Region/ Area of the Crime.
Location coordinates	The coordinates of the Crime Location.
Descript	Keyword Description of the Crime Incident.
Resolution	Resolved/Unfound.

Table 2.1.1- List of features

The datasets were taken year wise for the past five years (2012-2016) which will be later used for training the model. To predict crime trends for the year of 2017, we use the crime dataset containing data of months: January, February and March of the year 2017. Dataset obtained in .csv format and data preprocessing step is used to make data ready for further classification and prediction process.

B .Pre-Processing

After raw data is collected, data is preprocessed. We do preprocessing to make data suitable for our prediction model. It includes cleaning of data, transformation of variables, selection of features and data balancing. When data is collected it may have some incomplete records, inconsistent values. Pre-processing enable us to get data in a form on which data mining and analysis can be done efficiently. It is important to carry out this step because data quality may affect result obtained. Mining techniques like classification give best results if pre-processed data is used. Data collected and available is processed and represented in a form which is desirable for classification accuracy and good quality results. Initially data gathered was collected in one place in form of excel sheet. Records with incomplete data was removed. Only fully complete rows were considered. Values of few features or attributes were abbreviated for simplicity of representation and understanding. Few Categories were factorized to lesser number of categories for computational simplicity. After data balancing, feature selection is applied. We found out the attributes that affect the output the most. The factors like Descript and Day of week were not of any use so we removed those attributes to avoid the decline in the prediction results. We also converted the scalar data into vector data frames for manipulation simplicity.

C. Classification of Data

We apply various classification algorithms on dataset for knowledge discovery that is useful. There are many Mining techniques that can be applied on data for analysis. These are like classification, association, clustering, etc. We apply classification mining technique on the dataset. The main aim is to obtain the prediction model based on attributes identified. We also use classification technique on various cases to compare their accuracy. We compute the accuracy of various cases and determine which case will give best result. We have also tried Naïve Bayesian algorithm to check the accuracy and found that J48 gives better accuracy for our model. Finally, we apply the J48 Decision Tree algorithm.

The J48 Decision tree classifier follows simple algorithm. To classify a new object, J48 algorithm initially creates a decision tree based the attribute values available in the training dataset. Now taking that as the root node it further starts to identify the other variables that discriminates the various instances. This explains about the data instances so that we can classify the data with highest information gain. If in the process it finds a data instance same as the target value, the branch terminates and assigns the value to the target value. In other cases, we keep looking for another attribute which gives us the highest information gain. We continue the process till we reach the target value or the attributes gets over. In case all the attributes are utilized, we assign the value of majority of objects in that branch to the target value. Now using the decision tree obtained we can predict and assign the target value

In our project, for J48 Algorithm, we can consider an attribute say, the coordinates of the Location, which will be the root node and thus the tree starts to structure with various other independent variables to find the target (dependent) variable.

BASE ALGORITHM	Accuracy
Naïve Bayesian	92.65 %
Zero R	38 %
One R	95.38 %
J48	**97.59 %**

Table 2.1.2 – Accuracy with different Base algorithms

D. Predictive Probability

Finally, we use the linear regression function on the J48 model to predict the trend of various types of crimes over different regions. Using the predict function, we obtained the desired results, with good accuracy. We also obtain the confusion matrix to cross reference the obtained results. Using the result obtained we further perform probabilistic analysis and display the results in probabilistic values.

We also plot the obtained probabilistic results in form of a graph, which shows us the various type of crime Vs rate of crime graph for a particular region, where we can see the prior predicted rate of crimes and the predicted rate of crimes for the given region.

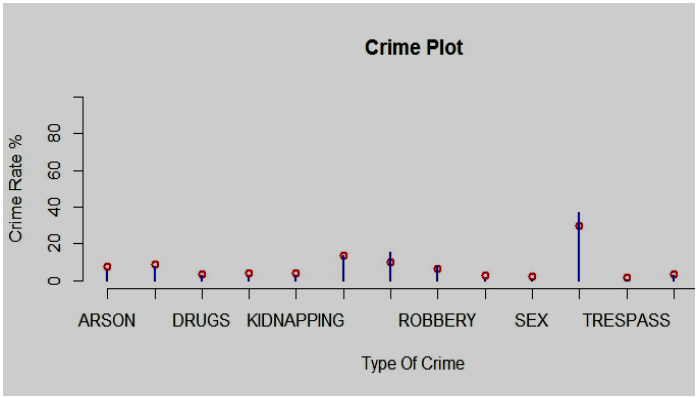

Chart -1: Crime plot for a given region

The chart displays the prior predicted probability and the predicted probability of various types of crimes in a specific region.

Crime Incidents (Jan 1st , 2012 ~ Dec 31st , 2016)

Type	Incidents
Arson	39487
Assault	63822
Drugs	33385
Fraud	18392
Kidnapping	24486
Warrants	33058
Robbery	54128
Sex	7242
Theft	225269
Trespass	6765
Weapon Laws	7881
Non-Criminal	94776
Secondary Codes	9617
Other Offenses	131631

Table 2.1.3- Crime Data Sheet

3. FORECAST TO THE FUTURE

We intend to keep on working and find a way to make the system even better and improve accuracy.
To include suspect specific predictions, where we will be able to predict the possible suspects associated with a crime using the crime patterns and Modus Operando of various criminals.

4. CONCLUSIONS

With the increase in crimes, there needs to be an advancement in the technology, to help the police department to take precautionary actions and stop the crimes from happening than taking actions on crimes later. This is a step towards that, where we predict the trend of crimes in specific region, giving Police Department a view on the possible types of crimes which can take place in a specific region. Thus, Police Department can concentrate to avoid those specific types of crimes.

REFERENCES

[1] M. S. Chen, J. Han, P. S, Yu, "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, December 1996.

[2] W. Yathongchai, C. Yathongchai, K. Kerdprasop, N. Kerdprasop, "Factor Analysis with Data Mining Technique in Higher Educational Student Drop Out", Latest Advances in Educational Technologies, 2003.

[3] Hina Gulati, "Predictive Analytics Using data Mining Techniques", 2nd International Conference on Computing for Sustainable Global Development, IEEE Conference, 2015.

[4] H.Chen, W.Chung, J.J.Xu,G.Wang ,Y.Qin, and M.Chau," Crime data mining: a general framework and some examples, "Computer, vol. 37, no. 4, pp. 50–56, 2004.

[5] J. H. Ratcliffe, "Aoristic signatures and the spatio-temporal analysis of high volume crime patterns," Journal of Quantitative Criminology, vol. 18, no. 1, pp. 23–43, 2002.

[6] Data Set - San Francisco Police Department Incidents Database- https://data.sfgov.org/Public-Safety/SFPD-Incidents-from-1-January-2003/tmnf-yvry