Cops Friend: A Crime Pattern Analysis using Machine Learning to Reduce the Crime Rate

Shital B.Jadhav[1] Neelam Rajpal[2] Heena Mulla[3] Nidhi Patel[4] Pooja[5]
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]Bharati Vidyapeeth's College of Engineering for Women, Pune, India

Abstract— Process of Crime analysis involves analysis for identifying and analyzing patterns and trends in crime. Reporting of crime generates huge amount of data every year. The data is collected from different sources providing the information about the crimes occurring. Using this data, we are analyzing and predicting crime and helps in preventing the crime to some extent. The Dataset is loaded in the system which gives information about the crime and it is trained to get the data in the appropriate manner. Clustering technique is used for clustering of data that helps to change in pattern automatically. Using naïve bayes, data is analysed which helps in predicting the crime type and the crime prone areas and it is visualized using bar graph in order to provide better understanding of crime patterns.

Key words: Naïve Baye, K Means Algorithm, Prediction, Clustering

I. INTRODUCTION

Crime around the world has become problem of every nation, around the globe many countries are trying to curb this problem. Crime analysis is a law enforcement function that involves systematic analysis for identifying and analyzing patterns and trends in crime and disorder. Crime analysis also plays a role in devising solutions to crime problems, and formulating crime prevention strategies. Crime analysis is exploring, inter relating and detecting relationships between various crimes and characteristics of crimes .Police department maintains crime data at the record. This data contains huge amount of data set with complex relationships which needs use of data mining techniques in order to be transformed into useful information. The knowledge extracted from the dataset can be a great tool & support to the police department to prevent crimes. The present scope of our project is prediction of crime type that what type of crime may happen in future and crime prone areas i.e. next location where the crime may take place.

Among a large set of crimes that happen every year in a major city, it is challenging, time-consuming, and labor-intensive for crime analysts to determine which ones may have been committed by the same individual(s). If automated, data-driven tools for crime pattern detection are made available to assist analysts, these tools could help police to better understand patterns of crime, leading to more precise attribution of past crimes, and the apprehension of suspects. Our system will predict regions which have high probability for crime occurrence and can visualize crime prone areas. This will be done in five steps. Firstly, collection of data for the training set then classifying it into different categories. Then identification of pattern of crime. Finally prediction of crime prone areas and visualizing it on map.

Before this clustering algorithms have been used for crime analysis. For example, one site it is revealed that suspect is literate and from next site/witness it is revealed that suspect is youth etc. By describing the offender details it gives a complete picture from different crime incidents. Today most of it is manually done with the help of multiple reports that the detectives usually get from the computer data analysts and their own crime logs. The reason for choosing this method is that we have only data about the known crimes we will get the crime pattern for a particular location. The nature of crimes change over time, so we are using clustering technique in order to detect newer and unknown patterns.

II. LITERATURE REVIEW

Shiju Sathyadevan and Surya Gangadharan. S, have used SVM and naïve bayes algorithms for classification. Naïve bayes used is more efficient than SVM as it has high performance, it is easy to implement and gives more accuracy even at less training dataset. It also handles zero frequency problems which SVM cannot[4].

Rasoul Kiani, Siamak Mahdavi and Amin Keshavarzi, used Genetic algorithm that supports best for detection of healthy chromosomes as compared to crime prediction. Even though classification is process of dividing the data to some groups that can act either dependently or independently but this technique relies on existing and known solved crimes only [5].

Mrs. M.Kavitha, Ms.S.T.Tamil Selvi, used Apriori algorithm and frequent pattern growth algorithm for pattern prediction but Apriori algorithm waste more time in producing candidates every time and it also needs more search space. FP growth algorithm performs badly with long pattern dataset[6].

TongWang, Cynthia Rudin, DanielWagner and Rich Sevieri, used Series finder tool which uses series finder algorithm for detecting house breaking crimes only[2].

J. Agarwal, R. Nagpal and R. Sehgal in have analyzed crime and considered homicide crime taking into account the corresponding year and that the trend is descending from 1990 to 2011. They have used the k-means clustering technique for extracting useful information from the crime dataset using RapidMiner tool because it is solid and complete package with flexible support options.

Priyanka Gera and Dr. Rajan Vohra in have used a linear regression for prediction the occurrence of crimes in Delhi (India). They review a dataset of the last 59 years to predict occurrence of some crimes including murder, burglary, robbery and etc. Their work will be helpful for the local police stations in decision making and crime supervision.

Brown (1998) constructed a software framework called ReCAP (Regional Crime Analysis Program) for mining data in order to catch professional criminals using data mining and data fusion techniques. Data fusion was used to manage, fuse and interprets information from multiple sources. The main purpose was to overcome confusion from conflicting reports and cluttered or noisy backgrounds. Data

All rights reserved by www.ijsrd.com

mining was used to automatically discover patterns and relationships in large databases.

De Bruin et.al. (2006) introduced a framework for crime trends using a new distance measure for comparing all individuals based on their profiles and then clustering them accordingly. This method also provided a visual clustering of criminal careers and identification of classes of criminals.

Malathi.A and Dr. S. Santhosh Baboo, have used MV algorithm that handles only missing values of crime attributes but does not support for prediction of crime[1].

III. PROPOSED SYSTEM

- In the proposed system, we are introducing the application which will predict the crime that criminals can do in the future and the crime prone areas.
- As our system is client based it provides us these different features which the existing standalone system don't.
  1) Centralization: Servers help in administering the whole set-up. Access rights and resource allocation is done by Servers.
  2) Proper Management : All the files are stored at the same place. In this way, management of files becomes easy. Also it becomes easier to find files.
  3) Back-up and Recovery possible: As all the data is stored on server, its easy to make a back-up of it. Also, in case of some break-down if data is lost, it can be recovered easily and efficiently. While in peer computing we have to take back-up at every workstation.
  4) Upgradation and Scalability : Changes can be made easily by just upgrading the server. Also new resources and systems can be added by making necessary changes in server.
  5) Accessibility: From various platforms in the network, server can be accessed remotely.
  6) As new information is uploaded in database, each workstation need not have its own storage capacities increased (as may be the case in peer-to-peer systems). All the changes are made only in central computer on which server database exists.
  7) Security: Rules defining security and access rights can be defined at the time of set-up of server.
- We will perform mining on all the previously stored criminal records and will calculate the possibility and predict the next type of crime and next area that may get affected by crime.
- This prediction will be done on the basis of attributes like criminal record, education, occupation, family background and other various attributes.
- In the proposed system, we are introducing the application which will predict the crime that criminals can do in the future and the crime prone areas.
- As our system is client based it provides us these different features which the existing standalone system don't.

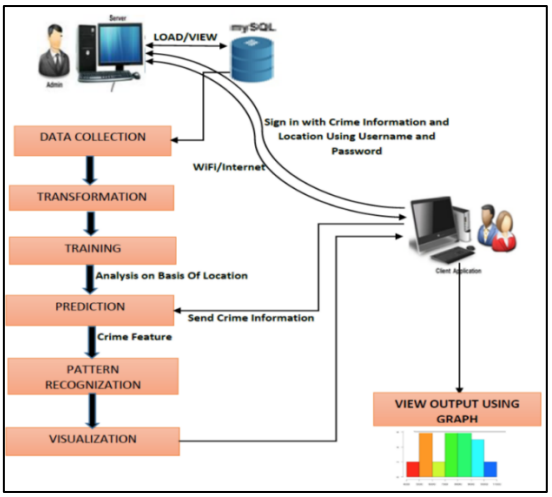

Fig. 1: Proposed architecture

Following are steps of our proposed system in doing Crime Analysis:

A. *Load Dataset*

For Training Purpose of Data Set First System Needs to introduced with some data set. Data set contains the parameters which can be used to detect the crime pattern and further analysing the crime and detecting the crime prone areas.

B. *View Dataset*

Dataset can be viewed so that further training of the dataset is done considering all the parameters collected in the dataset.

C. *Train Dataset*

Dataset which contains the parameters of the crime done in the past history is then trained so that the data can be obtained in more precise form which can be useful to analyse the crime for detection. Clustering technique is applied to train the dataset which forms the appropriate clusters.

D. *Analysis of data on the basis of location*

The data provided as an input are analyzed on the basis of location and the crimes are predicted by comparing the data from the trained dataset. Naïve Bayes classifier is used for the predicting the crime.

E. *Probability Estimation*

Initial probability and Individual probability calculation comes under the training phase of system.
And final probability calculated in detection phase of system.

F. *Prediction*

Final probability calculation has been done under this block.
Final probability = initial probability * individual probability

G. *Output*

System gives the detection of crime and visualize the crime prone areas using graph.

H. *Algorithms Used*

- Naive Bayesian Classification Algorithm

1) Bayes Rule

A conditional probability is the likelihood of some conclusion, *C*, given some evidence/observation, *E*, where a dependence relationship exists between *C* and *E*.

This probability is denoted as P*(C |E)* where

All rights reserved by www.ijsrd.com

$$P(C|E)=\frac{P(E|C)P(C)}{P(E)}$$

The naive Bayesian classifier, or simple Bayesian classifier, works as follows:

- Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X=(x1, x2,…, xn), depicting n measurements made on the tuple from n attributes, respectively, A1, A2,.., An.
- Suppose that there are m classes, C1, C2,…, Cm. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple x belongs to the class Ci if and only if

$$P(Ci|X)>P(Cj|X) \quad \text{for } 1\le j\le m, j\ne i$$

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(Ci|X)=\frac{P(X|Ci)P(Ci)}{P(X)}$$

- As P(X) is constant for all classes, only P (X|Ci) P (Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1)=P(C2)=…=P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci)=|Ci,D|/|D|, where |Ci,D| is the number of training tuples of class Ci in D.
- Given data sets with many attributes, it would be extremely computationally expensive to compute P(X|Ci).In order to reduce computation in evaluating P(X|Ci), the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$P(X|Ci)=\prod_{k=1}^{n}P(xk|Ci)$$

$$=P(x1|C_i)x\ P(x2|C_i)x\ldots\ P(xm|C_i).$$

We can easily estimate the probabilities P(x1|Ci), P(x2|Ci),… , P(xm|Ci) from the training tuples. Recall that here xk refers to the value of attribute Ak for tuple X. For each attribute, we look at whether the attribute is categorical or continuous-valued. For instance, to compute P(X|Ci), we consider the following:

1) If Ak is categorical, the P(Xk|Ci) is the number of tuples of class Ci in D having the value xk for Ak,divided by |Ci,D|, the number of tuples of class Ci in D.
2) If Ak is continuous valued, then we need to do a bit more work, but the calculation is pretty straightforward. A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation σ, defined by

$$g(x,\mu,\sigma)=\frac{1}{\sqrt{2\Pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

So that

$$P(xk|C_i)=g(xk,\ \mu ci,\ \sigma ci)$$

We need to compute μci and σci, which are the mean and standard deviation, of the values of attribute Ak for training tuples of class Ci. We then plug these two quantities into the above equation.

- In order to predict the class label of X, P(X|Ci)P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of tuple X is the class Ci if and only if

$$P(X|Ci)P(Ci)>P(X|Cj)P(Cj)$$

for $1\le j\le m, j\ne i$

In other words, the predicted class label is the class Ci for which P(X|Ci)P(Ci) is the maximum.

2) K Means Algorithm

K-means Algorithm is the algorithm which is used for clustering.

a) Steps in K_MEANS algorithm

Let $X = \{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.

1) Defined value of k.
2) Read length values.

While true

- Calculate distance between each centroid.
- Calculate mean.
- Get new cluster.
- If (previous cluster and current cluster is not equal)

then process again.
Else
Stop
End if
End while.

b) Pseudocode

```
doubleminDist, currDist;
arraytotalPointsInCluster, sumClusterPoints;
dataPoints = new int[datapoints][2];
for (int i = 0; i <datapoints length; i++)
{
dataPoints[i][0] =value
dataPoints[i][1] = -1;// -1 means invalid cluster...
}
//Apply K-Means to find High And Low Cluster
if (datapoints>= 2) {
k = 2; //Consider Total cluseter value 2
clusters = new int[k];
} else
{   return;
}
// calculate minimum and maximum data
intmaxLimit = dataPoints[0][0];
intminLimit = dataPoints[0][0];
for (int i = 0; i <parent.dbPacket.list.size(); i++)
{
if (parent.dbPacket.list.elementAt(i).len != 0)
{
minLimit = min of all
maxLimit = max of all
}  }
inttotalRange = maxLimit - minLimit;
// initialise cluster
// either use min & max Limit...
System.out.println("\n K  :" + k);
for (int i = 0; i < k; i++)
{
```

All rights reserved by *www.ijsrd.com*

```
clusters[i] = (((totalRange / k) * i) + ((totalRange /
k) / 2));
System.out.println(" Centroid Is  " + clusters[i]);
}
totalPointsInCluster = new int[k];
sumClusterPoints = new int[k];
boolean change = true;
while (change) {
change = false;
// update data points clustering...
for (int i = 0; i < data points lengh; i++) { // data
points
minDist = 999999999;
for (int j = 0; j < k; j++)
{
 // clusters
currDist=minDistance(dataPoints[i][0], clusters[j]);
if (currDist<minDist)
 {
dataPoints[i][1] = j; //update clustering
minDist = currDist;
                    }
                  }
               }
  // update centroids
for (int i = 0; i < k; i++) {
sumClusterPoints[i] = 0;
totalPointsInCluster[i] = 0;
     }
for (int i = 0;i <parent.dbPacket.list.size(); i++)
 {
sumClusterPoints[dataPoints[i][1]]+=
dataPoints[i][0];
totalPointsInCluster[(int) dataPoints[i][1]]++;
               }
for (int i = 0; i < k; i++) {
if (totalPointsInCluster[i] == 0) {
continue;  }
intcurrCentroid= sumClusterPoints[i]
                    / totalPointsInCluster[i];
if (clusters[i] != currCentroid) {
clusters[i] = currCentroid;
change = true;
   }  }
```

IV. METHODOLOGY

A. *Data Collection*

In data collection step data from different web sites like news sites, blogs, social media etc. are collected and is stored into database for further process. The database used is MySQL as it is open source relational database. We will process the data and will find the frequently occurring crimes. Then we will relate this crime type to location.

B. *Transformation*

In this phase we will perform clustering. The reason for choosing this method is that if we have data of known crimes we can get a particular place's crime pattern. But the criminal data in the database keeps on changing so in order to handle new and unknown patterns clustering is a good technique to work with.

C. *Training*

In this phase we will apply naïve bayes algorithm on the transformed data set we get from transformation step. We have choose Naive Bayes because it is the basis for many machine-learning and data mining methods. The algorithm is used to create models with predictive capabilities. It provides new ways of exploring and understanding data.

1) It can handle high amount of data with ease.
2) It gives more efficient output as compared to other methods.
3) Prediction

In the prediction step the transformed current value and training data set will be analysed using frequent count method and at the end of this phase frequently occurring crime type will be predicted and stored into database.

4) Pattern Recognition

In pattern recognition phase we will again apply naïve bays algorithm to training dataset to find the frequently occurring value of location attribute and crime type attribute on the basis of output of prediction phase and pattern recognition phase we will get the final output.

5) Visualization

The crime prone areas will be then represented using maps or bar graphs.

V. RESULTS

The results shown at the end contains Loading and viewing dataset stored in the system in .csv file. Generate element and Cluster i.e. Calculate the clusters of age give in the dataset to classify into different age groups.

Parameter Prediction will show the details about the parameter classified on the basis of history data and entered parameters.

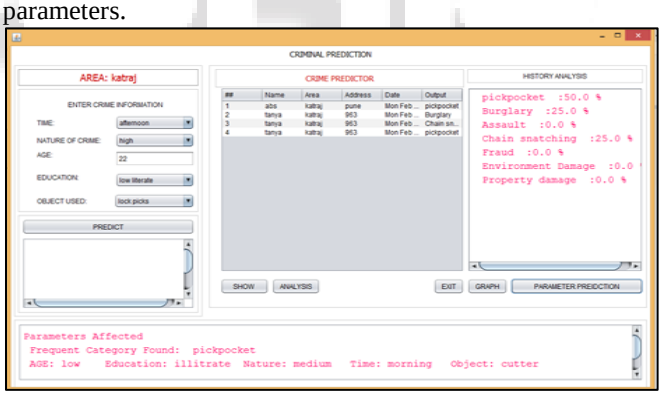

Fig. 2: Parameter Prediction

Results will be shown using the bar graph for better understanding.

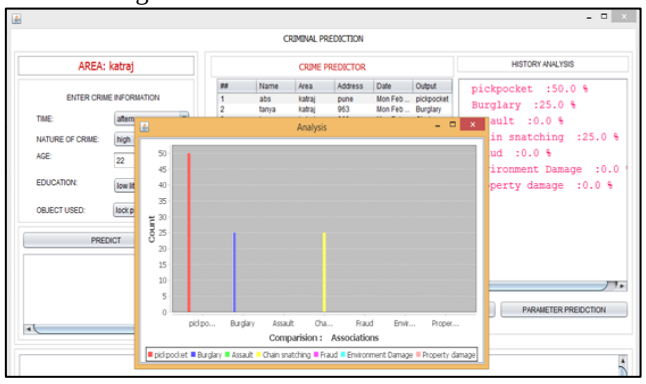

Fig. 3: Visualization using bar graph

All rights reserved by *www.ijsrd.com*

VI. CONCLUSION AND FUTURE WORKS

The present scope of our project is prediction of crime type that what type of crime may happen in future and crime prone areas i.e. next location where the crime may take place. We can also predict the estimated time for which the crime to take place as a future scope of our project. Along with this, one can try to predict identity of an individual criminal who may perform the crime. We will do clustering on different data sets. This will help system to learn change in pattern of crime and location automatically. Our software will predict the crime prone area i.e. location where chances of happening of next crime is maximum. We will use naïve bays algorithm for this purpose, this will find the crime prone areas. Instead of this web based application, an android application can be developed for this as a future so that it can easily be accessible anywhere.

REFERENCES

[1] Malathi. A and Dr. S. Santhosh Baboo.,"an enhanced algorithm to predict a future crime using data mining. "International Journal of Computer Applications,6, May 2011. Published by Foundation of Computer Science.
[2] Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. "Detecting patterns of crime with series finder. "In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2013.
[3] Clifton Phua,Kate Smith-Miles,Vincent Cheng-Siong Lee, and Ross Gayler. "Resilient Identity Crime Detection", IEEE transactions on knowledge and data engineering, vol. 24, no. 3, March 2012.
[4] Shiju Sathyadevan, Devan M.S,Surya gangadharan .S."crime analysis and prediction using data mining", IEEE 2014 international conference on network and soft computing.
[5] Rasoul Kiani, Siamak Mahdavi and Amin Keshavarzi "Analysis and Prediction of Crimes by Clustering and Classification" (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.8, 2015
[6] Mrs. M.Kavitha [1], Ms.S.T.Tamil Selv."Comparative Study on Apriori Algorithm and Fp Growth Algorithm with Pros and Cons" International Journal of Computer Science Trends and Technology (I JCS T) – Volume 4 Issue 4, Jul - Aug 2016.

All rights reserved by *www.ijsrd.com*