ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

An Experimental Study of Crime Investigation using Machine Learning

Sonali Chawla[1], Vijaya Lakshmi Singh[2], Alpana Jijja[3]

Master of Computer Application, Ansal University, Gurgaon, Haryana, India[1]

School of Engineering and Technology, Ansal University, Gurgaon, Haryana, India[2]

School of Engineering and Technology, Ansal University, Gurgaon, Haryana, India[3]

ABSTRACT: Crime Investigation process suffers from various problems. One of the biggest problems is finding crime offense and suspect. Taking into consideration that the knowledge about crime event may not only enough for investigator to deduce the crimes, he may check on the crime cases solved previously that have similar ground than that of current one, finding the generalized information to solve the current case problem. Using this new information would leads to many deductions including the crime offense and even the murderer. This could possibly make the work of the investigator much easier in solving the crime case. This paper provides an overview of the system that deploys a machine learning algorithm which would help in investigating crime scenes. The algorithm would take case evidences, witnesses and forensic reports as inputs, draws some of the inferences on crime offense and give output. The algorithm also provides Bayesian network for more clear understanding of the findings.

KEYWORDS: Crime Investigation; Machine Learning; Confusion Matrix; Bayesian Theorem; Accuracy

I. INTRODUCTION

Criminality is a part of human nature and society which is unavoidable [1]. Hence, there is nothing like completely crime free society. It has been seen that types of criminal behavior prevailing in society makes a pattern according to the social and economic development of that society [16]. Therefore, it is expected that a society with low level of development such as uneven income distribution experienced a high-level of crime rate [10]. The police, moreover, who are the medicine to this injury, are getting overwhelmed by the phenomenon [3]. The criminals are getting advantages of negligence done by police by ignoring the evidence on the crime scene [4]. Computers are used for crime detection in developed countries. Computer-based Criminal Record System was one such system that was deployed for keeping the records of a criminals' history. In traditional system, all the work is done through manual skills using paper work and human brain potential. The traditional approach of criminology is time-consuming which involves a lot of delay and high cost. On the other hand, using computers in criminology includes automation and efficiency of work, less time and accuracy in results. In this paper we mainly focus on using a machine learning approach in solving a type of crime problem. This type of problem needs to be modeled with generic situation that can arise, say diagnosing the crime weapon. It means there are number of possibilities involved in one situation which needs to rank them according to their likelihood to occur. Such ranking can lead to the value of another related situation with another set of possibilities. Hence, these are also known as dynamic investigation problems. Solving each such problem leads the investigation to a new facet of crime. For this, prior knowledge is used to narrow down the domain of interest/investigation. Thus, we need a framework for transforming every problem in quantitative terms and calculate the uncertainties involved with it. The framework can would deal with these uncertainties and help in disclosing new relevant information which can be further used in investigation process.

II. LITERATURE SURVEY

A. *WHAT IS A CRIME?*

A crime can be seen as any activity that violates the norms of a country. The approach use the realities around the crime to understand the changing every aspect of crime and society. These realities are usually arguable, for example, crime rate can fluctuate according to the political environment and cultural difference. Similarly, collection of data can

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

be one issue that can affect public opinion about "crime problem". All these shape the people thinking about crime for which laws should be updated. As per UN Interregional Crime and Justice Research Institute, property crime prevails between 14.8% in New Zealand to 12.7% in Italy, 12.2% in UK, 10% in US, 3.4% in Japan and 31.7% in Nigeria [6]. The reasons for differences between crime rates in these countries are many including distinction in definition of crimes and reporting procedures. It doesn't matter how we look at this, it is still an important subject to be looked to develop the society psychologically and economically. The relationship between society development and crime is a very essential issue. Low-level of crime rate can lead to higher-level of country's development which lead to increase domestic and foreign investment [14] and also affect job vacancy [8]. A crime scene can be a referred to a physical scene that can provide potential evidence to the detective. The evidence includes a person's body, vehicle, objects found on location. Thus, "crime scene examination" includes examination of crime scene with all forensic and scientific techniques to collect all physical evidence of crime.

B. *MACHINE LEARNING*

An expert in machine learning and AI define machine learning as a field of computer science that deals with the technique through which machine can learn to do all the things as human does without explicitly programmed [7]. In other words, machine learning is the way for machine to learn through examples and improve to increase performance results [5]. Machine learning involves directions, concepts and methods which can be used to learn. These three things are distinguished as supervised, unsupervised and reinforcement learning. The choice of these learning is based on the type of data available to solve the problem. Supervised learning is the technique of machine learning in which problems and results are given. The problem data used as training dataset to train the machine and prediction is done for new data [12]. Error rate in prediction and actual results are calculated. Unsupervised learning is the technique of machine learning in which problem data is only given and then it is used as training data for machine. Values are predicted for new data. This is relatively difficult. This is often used for clustering problems [15]. Text classification is also a machine learning problem which involves separating text document in one or more classes or categories [9]. Text categorization can be used for various purposes [2]. Modus Operandi describes the personal character in criminology [1]. Volume crimes impact the community and police ability in solving crimes. Soft forensic evidence includes geographical and temporal features of a crime [13]. Hard forensic evidence includes physical evidence such as DNA and forensic reports [13].

C. *BAYESIAN THEOREM*

In homicide investigation including vehicle, it is difficult to find actual accused. To overcome the problem, Bayesian evaluation is used on forensic evidence. In the proposed method, Bayes' Theorem is used for this. The analysis results can be calculated in the form of true positive and false positive rates on the basis of posterior probability to find the culprit from given evidence [11]. The naïve Bayes classifier is used for machine learning task showing good performance. Naïve Bayes is applied to database application where variable are non-normal. It is found that algorithm works well and predict a class which was derived from the same data. Clear and detailed information is required for performing forensic analysis. Naïve Bayes classifier is applied on reduced dataset [15]. During the prediction made by Bayesian inference, loss occurrence will be higher. The predictions come under model averaging, misspecifications, and non-stationary environment predictions [2]. The Bayesian Theorem can be written as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{eq. (1)}$$

Here, P (A) is the probability of the class which is going to be predicted and P (B) is the probability of the given pattern which predicts the class. For implementing the Bayes Theorem, training set and patterns are required. In the training set, overall probability of the class that needs to be predicted should be found. After finding the probabilities of the required classes, probability for all the given attributes needs to be found.

III. THEORETICAL WORKING OF ALGORITHM

A. *THE MYSTERY OF MURDER*

The aim of this example given in the paper is to solve a criminal mystery. The detective appointed for the case just has discovered the body. He has a vague idea if it's a murder or suicide. The body was of a male and post mortem report shows that he was stabbed in the stomach. The body soon got a name – Mr. Mahesh. Now, the challenge detective face is to deal with the uncertain values. This is what real world problems most suffers from. In this mystery,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

the detective would start with an uncertainty of murder and with the introduction of more and more clues; he would get more certain about the offense. As a result, the detective would need a framework which can represent uncertain quantities and help him to solve the case effectively. The framework that manipulates the uncertain quantities uses probabilities representing degree of uncertainty. There is familiarity with the idea of probability as frequency of occurrence of event. The probability that Mr. Mahesh was murdered ranges from 0% to 100%, where 0% means detective was certain that Mr. Mahesh has committed suicide and 100% means that he has been murdered. From the dataset available and stated above, probability of murder can be found to be 50% or 0.50. To express the belief about the above, detective needs to be precise about what this 50% probability is about. This can be done through representing the offense as a random variable – offense. This variable can take one of the two values: either Murder or Suicide. Given this definition, the 50% belief can be written as:

$P(offense = murder) = 0.5$

$P(offense = suicide) = 0.5$ eq. (2)

Here the notation P() refers to the probability of the uncertain event written inside the brackets. Thus, eq. (2) can be understood as "probability that Mr. Mahesh was murdered is 50%". As it could be only suicide other than murder, the probability that he committed suicide must be 50%. With the fact found that Mr. Mahesh was left handed and a cutter found in his right hand, probability of murder increases to 65% and probability of suicide decreases to 35%. This can be again written as:

$P(offense = murder) = 0.65$; $P(offense = suicide) = 0.35$ eq. (3)

This can also be represented in image form as Fig. 1.

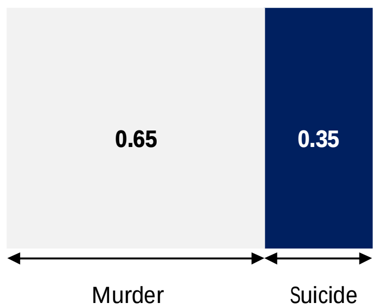

Fig. 1: Probability distribution of random variable 'offense'

Fig. 1 shows the unit square which has been separated into two proportions representing two probabilities of random variable offense. The grey area shows the probability that Mr. Mahesh was murdered and red area shows that he committed suicide. The total of both the areas comes out to be 1.0. This can be written as eq. (4).

$P(offense = murder) + P(offense = suicide) = 1.0$ eq. (4)

We can rewrite the above as:

$\sum_{offense} P(offense) = 1.0$ eq. (5)

Here subscript 'offense' is the sum of all the values this random variable could take i.e. suicide and murder. This is known as normalizing the probability distribution.

B. *CONSIDERING EVIDENCE*

Till now, we have just seen one random variable in the mystery: offense. With more intense observation performed of the crime scene by the detective he found a knife. This introduces a new random variable: weapon. This random variable weapon can now take two values: cutter and knife. The task now is to use this new random variable to evaluate the random variable offense. From the dataset collected of previous cases, it has been seen that the probability of usage of knife in murder is 80%. Therefore, if Mr. Mahesh was murdered and the probability of the murderer uses knife is 80%. This proves to be true when blood on the knife was found related to Mr. Mahesh on the basis of forensic report. On the other hand, if he has committed suicide then the knife could not be found so far from him, which in turn makes the probability of cutter being the murder weapon 20%. This states that the probability distribution over the random variable weapon depends upon whether the offense was murder of suicide. This is called conditional probability distribution. Now, for example, knife was the murder weapon. The conditional probability of offense being murder is

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

0.8 and that of suicide is 0.2. Therefore, the conditional probability of knife being the murder weapon when it was a murder can be expressed as:

$$P(weapon = knife|offense = murder) = 0.8 \quad \text{eq. (6)}$$

This can be understood as "the conditional probability of knife being the murder weapon given it was a murder is 0.8". The other possibility of the knife being the murder weapon when it was a suicide is 20%, this can be written as:

$$P(weapon = knife|offense = suicide) = 0.2 \quad \text{eq. (7)}$$

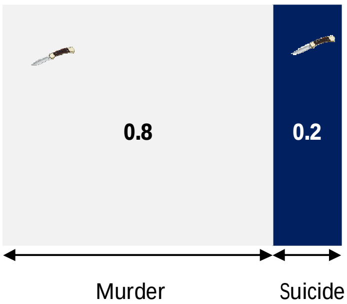

Fig. 2: Probability distribution of random variable 'weapon' to be knife depending upon variable 'offense'

Fig. 2 indicates that the total area 1.0 is divided into two areas where one represent the probability of knife being the murder weapon when it was a murder (80%) and other represent the probability of knife being the murder weapon when it was a suicide (20%). The above observations can be combined in the following form:

$$P(weapon = knife|offense) = \begin{cases} 0.8, if\ offense = murder \\ 0.2, if\ offense = suicide \end{cases} \quad \text{eq. (8)}$$

Now, let's take cutter was the murder weapon. The conditional probability of offense being murder is 0.15 and that of suicide is 0.85. Therefore, the conditional probability of cutter being the murder weapon when it was a murder can be expressed as:

$$P(weapon = cutter|offense = murder) = 0.15 \quad \text{eq. (9)}$$

This can be understood as "the conditional probability of cutter being the murder weapon given it was a murder is 0.15". The other possibility of the cutter being the murder weapon when it was a suicide is 85%, this can be written as:

$$P(weapon = cutter|offense = suicide) = 0.85 \quad \text{eq. (10)}$$

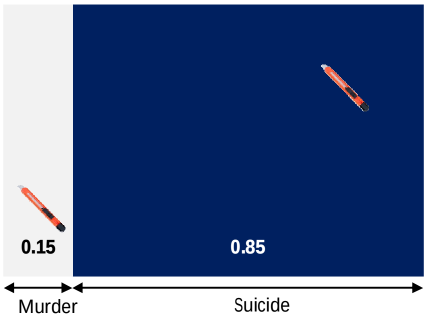

Fig. 3: Probability distribution of random variable 'weapon' to be cutter depending upon variable 'offense'

Fig. 3 indicates that the total area 1.0 is divided into two areas where one represent the probability of cutter being the murder weapon when it was a murder (15%) and other represent the probability of cutter being the murder weapon when it was a suicide (85%). The above observations can be combined in the following form:

$$P(weapon = cutter|offense) = \begin{cases} 0.15, if\ offense = murder \\ 0.85, if\ offense = suicide \end{cases} \quad \text{eq. (11)}$$

This can be written as *P(weapon|offense) in the following terms:*

$$\sum_{weapon} P(weapon|offense) = 1 \quad \text{eq. (12)}$$

Here the sum of the values that weapon can take based on the variable offense is shown. These conditional probabilities can be shown in the form of Table 1.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

Table 1: The Conditional Probability for P(Weapon|Offense)

Weapon	Murder	Suicide
Knife	0.80	0.20
Cutter	0.15	0.85

This indicates that the probability of weapon depends on the value of offense. Therefore, these two variables are dependent variables.

C. *THE CRIME MODEL INTRODUCTION*

So it is been confirmed that the murder weapon was the knife, not the cutter. This fact strongly indicates that Mr. Mahesh was murdered. An effective way is to think of all the probabilities we have encountered so far as a description of the crime. So, initially we tried to find out the crime offense with the help of Fig. 1, which shows that there was 65% possibility for Mr. Mahesh to be murdered and 35% possibility that he has committed suicide. Let's take that it was a murder. Using Fig. 2 and 3, murder weapon can be determined. There is 80% possibility that it would be a knife and 15% possibility that it would be a cutter. In case, cutter is the weapon, the joint probability of choosing murder and cutter would be 65% x 15% = 10%. Repeating this for the other combinations of offense and weapon would give joint probability distribution over two random variables (Fig. 4).

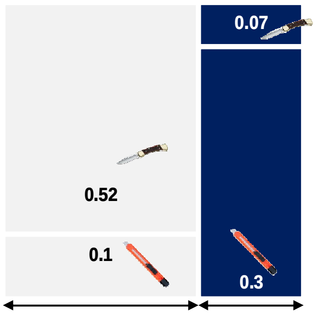

Fig. 4: Joint Probability of the two random variables 'offense' and 'weapon'

The joint probability *P(weapon,offense)* can be read as "the probability of weapon and offense". Now, we have the two required things: A set of random variables and a joint probability distribution over these variables. With these, we can build the probabilistic model to predict answers of any possible problem in terms of random variables included in the model. Probabilistic model can be understood as the set of assumptions involving uncertainty using probabilities made to solve the problems. The joint probability of offense and weapon in Fig. 4 can be again written as:

$$P(weapon, offense) = P(offense)P(weapon|offense) \quad \text{eq. (13)}$$

We have the encountered now that the weapon was a knife. Correspondingly, it increases the chances for Mr. Mahesh being murdered but to confirm this fact we need to use updated probability. This process of revising the probabilities after observing the random variable values is called inference. Inference can be used to reason out the model, learning from the dataset, making predictions using model and achieving machine learning tasks. After knowing that it was a knife that was used in crime, we can therefore rule out the region from Fig. 4 which uses cutter (Fig. 5).

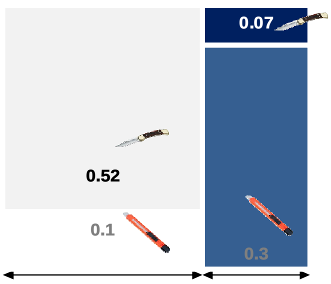

Fig. 5: Joint probability distributions with ruled out region containing cutter

Now, the probability of Mr. Mahesh being murdered is the fraction of remaining areas:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

$P(offense = murder \mid weapon = knife) = \frac{0.52}{0.52+0.07} \cong 0.88$ eq. (14)

This is higher than the 85% probability we had before knowing that the weapon was the knife. This makes the deduction of Mr. Mahesh being murdered correct. The probability that was assigned to the offense being murder before seeing the evidence knife is called prior probability, while the updated probability after seeing the new evidence is called posterior probability. The probability that offense could be a suicide is given by:

$P(offense = suicide \mid weapon = knife) = \frac{0.07}{0.52+0.07} \cong 0.12$ eq. (15)

Because the offense could either be murder or suicide, these two probabilities summed up to be 1. This can be shown through Fig. 6.

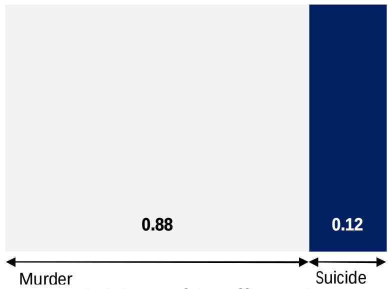

Fig. 6: Posterior probabilities of the offense when weapon is knife

As the crime offense has been solved as murder now we need to find the murderer. The eq. (13) can be written as:

$P(weapon, offense) = P(weapon)P(offense|weapon)$ eq. (16)

Equating the right side and re-arrangement of eq. (13) and eq. (16) gives:

$P(offense|weapon) = \frac{P(offense)P(weapon|offense)}{P(weapon)}$ eq. (17)

This is known as Bayes' Theorem/Rule. Now as we know that the murder weapon is knife, we can calculate the posterior probability that offense was murder.

$P(offense = murder|weapon = knife) = \frac{0.52 \ X \ 0.65}{0.52 \ X \ 0.65+0.07 \ X \ 0.35} \cong 0.88$ eq. (18)

Although we have arrived at the same result by a different route, Bayes Theorem is preferable as it didn't incorporate computing the joint distribution.

IV. EXPERIMENTAL PROCEDURE

A. *DATA COLLECTION*

The dataset available in this paper has been collected from the Sony TV's Crime Patrol Series. This dataset contains twelve types of tables that can be joined together to form a collective dataset. Table 2 shows the 3-row sample dataset.

Table 2: Sample Dataset used from Crime Patrol Series

a) Case Data

Case Id	Reg Yr	Reg Mon	Reg Date	Solved Yr	Country	State	City	District	Offense
1	2012	1	2	2012	India	Maharashtra	Mumbai	Naigaon	Murder
2	2012	4	27	2012	India	Gujarat	Mansa	Ajol Village	Kidnapping
3	2011	11	4	2011	India	Maharashtra	Thane	Ulhas Nagar	Murder

b) Victim Data

Victim Id	Case Id	Name	Sex	Spouse	Country	State	City	District	Occupation
1	1	Shrinath Mhapsekar	Male	Kumud Mhapsekar	India	Maharashtra	Mumbai	Virar	Lender
2	1	Sunita	Female	Ganesh	India	Maharashtra	Mumbai	Naigaon	Housewife
3	1	Ganesh	Male	Sunita	India	Maharashtra	Mumbai	Naigaon	Worker

c) Suspect Data

Suspect Id	Case Id	Name	Gender	Country	State	City	District	Occupation
1	1	Madhav Gorpade	Male	India	Maharashtra	Mumbai	Naigaon	Owner
2	1	Gurudev	Male	India	Maharashtra	Mumbai	Naigaon	Worker
3	1	Ganesh	Male	India	Maharashtra	Mumbai	Naigaon	Worker

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

d) Evidence Data

Evidence Id	Case Id	Type	Whose	Additional	Place Found	Basis	Report Id
1	1	Body	Shrinath Mhapsekar	Nothing	Behind Madhav House	Post Mortem Report	1
2	1	Body	Sunita	Nothing	Behind Madhav House	Post Mortem Report	2
3	1	Body	Ganesh	Nothing	Behind Madhav House	Post Mortem Report	3

e) Report Data

Report Id	Case No	Date	Month	Year	Type	Evidence/Victim Id	Results	Additional
1	1	1	2	2012	Post Mortem	1	Can't Say	None
2	2	28	4	2012	Call Record	4	Match	Wasim Shekh
3	2	28	4	2012	Automobile Report	5	Match	Sanjay Parmar

f) Weapon Data

Weapon Id	Case Id	Type	License No	Additional	Main Weapon
1	1	Steel Rod	Nothing	None	Yes
2	3	Ax	Nothing	Blood Found	Yes
3	5	Rope	Nothing	None	Yes

g) Criminal Data

Criminal Id	Case Id	Suspect Id	Victim Id	Motive Id	Status
1	1	1	1	1	Arrested
2	1	2	3	5	Surrender
3	1	3	2	6	Killed

h) Witness Data

Witness Id	Case Id	Name	Gender	Year	Seen	Eye Witness
1	1	Sheela Gorpade	Female	2012	Murder	Yes
2	3	Kadhir	Male	2012	Blood on Shirt of Sharad	No
3	6	Viraj	Male	2012	Sanjay and Sakshi with Vijay	No

i) Crime Motive Data

Motive Id	Motive
1	Loan from Friend
2	Extra Marital Affair
3	Rebel

B. *DATA PREPROCESSING*

The experiment is conducted using the R workbench. To impose the final algorithm on the dataset, it is necessary to make data in the standard and normalize form. Data preprocessing is often neglected but it is a very vital part for normalizing any kind of data. Data Preprocessing consists of data cleaning, data integrity, data transformation and data reduction. Data transformation consists of transforming the data into appropriate form. Table 3 shows the sample table name with its corresponding attributes and their data types.

Table 3: Attributes and their Desired Data Types

Table Name	Column Name	Data Type
Case	District	Character
Complaint	Age	Numerical
Evidence	Whose	Character
Missing	Name	Character

Data integrity refers to combining the data from different sources. Here, all the twelve tables are merged using 'case id' as common key between them. Data reduction means reducing the volume of data in a way that it would not affect the data integrity. Due to the merge, some extra column appears, hence, are removed. Data cleaning refers to the elimination of noisy incomplete and inconsistent data from the given dataset. Here, missing value of numeric variables is replaced by the mean value; that of categorical variables is replaced by their mode and that of character variables is replaced by the string 'Not Known'. Table 4 shows few column names along with their missing values' replacements.

Table 4: Columns with their Missing Value Treatment

Column Name	Value (replacing missing values)
Registered Year	2013
Registered Month	7
Country	"India"

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

C. *DETERMINING VARIABLE RELATIONSHIP*

For applying Bayes' Theorem, the relationship between the variables of the given dataset needs to be found. For this, we need to find the number of observation for every kind of offense. Table 5 shows the relationship of variables.

a) Offense and Registered Year

Offense	Registered Year			
	2011	*2012*	*2013*	*2014*
Kidnapping	0	6	0	0
Murder	16	28	10	0
Robbery	0	0	0	128
Sexual Assault	0	40	0	0
Suicide	2	1	0	0

Table 5: Relationships between Variables

b) Offense and Case State

Offense	Case State			
	Gujarat	*Haryana*	*Maharashtra*	*West Bengal*
Kidnapping	6	0	0	0
Murder	0	0	54	0
Robbery	0	128	0	0
Sexual Assault	0	0	0	40
Suicide	0	0	3	0

c) Offense and Victim Gender

Offense	Victim Gender		
	Female	*Male*	*None*
Kidnapping	6	0	0
Murder	31	23	0
Robbery	0	0	128
Sexual Assault	40	0	0
Suicide	3	0	0

d) Offense and Evidence Type

Offense	Evidence Type				
	Ax	*Bicycle*	*Blood*	*Body*	*Car*
Kidnapping	0	3	0	0	0
Murder	2	0	4	30	2
Robbery	0	0	0	0	0
Sexual Assault	0	0	0	0	5
Suicide	0	0	0	3	0

e) Offense and Report Type

Offense	Report Type			
	Automobile Report	*Forensic*	*Medical*	*Post Mortem*
Kidnapping	3	0	0	0
Murder	0	4	0	39
Robbery	0	0	0	32
Sexual Assault	5	0	5	30
Suicide	0	0	0	3

f) Offense and Weapon Type

Offense	Weapon Type			
	Ax	*Knife*	*Rope*	*Steel Rod*
Kidnapping	6	0	0	0
Murder	41	4	0	9
Robbery	128	0	0	0
Sexual Assault	40	0	0	0
Suicide	0	0	3	0

Using the confusion matrix in Bayes' Theorem, it is possible to identify the necessary information about the crime investigation, say Crime Offense.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

D. *USING BAYES' THEOREM*

Sample Size of 230 observations is taken and Bayes' Theorem is applied. With code implemented on workbench predicted the crime offense. Table 6 shows the 3 observations with actual and predicted values.

Table 6: Predicted and Actual Values

Sample#	Predicted	Actual
1	Kidnapping	Murder
2	Murder	Murder
3	Murder	Murder

E. *DETERMINING ACCURACY*

Accuracy refers to the quality or state of being correct or precise. For this purpose, we will use confusion matrix approach showing true-positive, false-negative, false-positive and true-negative count of the predicted observations (Table 7).

Table 7: Confusion Matrix of Predicted and Actual Values

		Predicted	
		Positive	Negative
Target	Positive	A (True Positive)	B (False Negative)
	Negative	C (False Positive)	D (True Negative)

$$Accuracy = \frac{Number\ of\ correctely\ predicted\ observations}{Total\ number\ of\ predictions}$$

$$Accuracy = \frac{A+D}{B+C} \quad \text{eq. (19)}$$

There are 207 correctly predicted observations on 230 total observations. Hence, accuracy is 0.9. Fig. 7 shows the accuracy of the predicting model in graphical form.

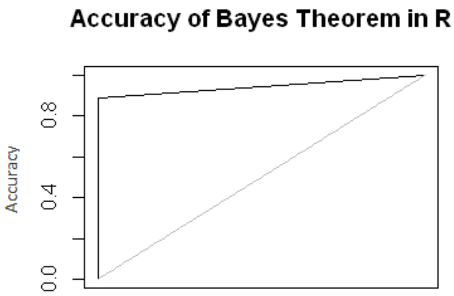

Fig. 7: Graphical representation of Accuracy

V. CONCLUSION

This paper is based on the deploying the Bayesian Theorem in Machine Learning to predict the crime offense using the crime files available. A theoretical framework with its experimental procedure is given. The dataset was collected from Crime Patrol Series. The dataset is then preprocessed including data cleaning, transforming, integrating and reducing. Using the given dataset of over 230 observations, we have investigated the crime scene to predict the crime offense and crime criminal. The results are predicted and accuracy of the algorithm is calculated using true and false positive. The algorithm works with 90% of accuracy. The accuracy is also plotted in form a graph.

REFERENCES

1. Alemika, E. 2004. Crime and Policing in Nigeria.
2. Alessio Sancetta – "Universality of Bayesian predictions", International Society for Bayesian Analysis, Volume 7, Issue 1, October 2012, Pages 1-36.
3. Ash, H.S. 1971 Police Authority and the Rights of Individual.
4. Awake Magazine. 2002. "The Police: Why do we need them". Watch Tower and Tract society of Pennsylvania. July 8.
5. Daniele Soria, Jonathan M. Garibaldi, Federico Ambrogi, Elia M. Biganzoli, Ian O Ellis - "A non-parametric version of the naïve Bayes classifier" , Knowledge-Based Systems, Volume 24, Issue 6, August 2011, Pages 775-784.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

6. Imrohoroglu, A., A. Merlo, and P. Rupert. 2006. “Understanding the Determinants of Crime”. Journal of Economics and Finance. 30(2):270 – 283.
7. J. McCarthy, "Arthur Samuel: Pioneer in Machine Learning," AI Magazine, vol. 11, no. 3, pp. 10-11, 1990.
8. Krkoska, L. and K. Robeck. 2006. “The Impact of Crime on the Enterprise Sector: Transition Versus Non-Transition Countries”. EBRD Working Paper No.97.
9. Lawrence M.Rudener & Tahung Liang - “Automated essay scoring using bayes theorem” JTLA - The Journal of Technology, Volume 1, Number 2 , June 2002, Pages 1966-1994, University of Maryland.
10. Madden, P. and W.H. Chiu. 1998. “Burglary and Income Inequality”. Journal of Public Economics. 69:123-141.
11. Michael D. Freeman, Annette M. Rossignol, Michael L. Hand - "Applied forensic epidemiology: The Bayesian evaluation of forensic evidence in vehicular homicide investigation", Journal of Forensic and Legal Medicine, Volume 16, Issue 2, February 2009, Pages 83- 92.
12. Muhammad Naeem Ahmed Khan - "Performance analysis of Bayesian networks and neural networks in classification of file system activities”, Computers & Security, Volume 31, Issue 4, June 2012, Pages 391-401.
13. R. Bermudez, B. Gerardo, J. Manalang and B. Tanguilig, III. "Predicting Faculty Performance Using Regression Model in Data Mining," Proceedings of the 9th International Conference on Software Engineering Research, Management and Applications, pp. 68-72, 2011
14. Rivera-Batiz, F. and L. Rivera-Batiz. 2002. “Democracy, Participation, and Economic Development: An Introduction”. Review of Development Economics. 6(2)”135-150.
15. Saurabh Mukherjee, Neelam Sharma - "Intrusion Detection using Naive Bayes Classifier with Feature Reduction”, Procedia Technology, Volume 4, 2012, Pages 119-128.
16. Sohnen, E. 2012. Paying for Crime: A Review of the Relationships between Insecurity and Development in Mexico and Central America.

BIOGRAPHY

Sonali Chawla is the student of Master of Computer Applications in Ansal University, Gurgaon, Haryana, India. Her area of interest includes big data analysis and machine learning.

Vijaya Lakshmi Singh is the Assistant Professor in School of Engineering and Technology, Ansal University, Gurgaon, Haryana, India. She has 6 years of teaching experience. Her areas of interest includes ad hoc network and cloud computing.

Alpana Jijja is the Assistant Professor in School of Engineering and Technology, Ansal University, Gurgaon, Haryana, India. She has over 15 years of extensive experience in IT Industry and Academics. Her interdisciplinary research interests span over Neural Networking and Machine learning.