Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

Crime Detection using Data Mining

Vineet Pande
Student, B. E., Computer Engg.,
D. J. Sanghvi College of Engg.,
Mumbai, Maharashtra, India

Viraj Samant
Student, B. E. Computer Engg.,
D. J. Sanghvi College of Engg.,
Mumbai, Maharashtra, India

Sindhu Nair
Asst. Professor, Computer Dept.,
D. J. Sanghvi College of Engg.,
Mumbai, Maharashtra, India

***Abstract*— As crime rates keep spiralling each day, new challenges are faced by law enforcement agencies. They have to keep their forces on the lookout for any signs of criminal activity. This may only cause more burden on their resources. The law enforcement agencies should therefore be able to predict such increases or decreases or trends in crime, such as the approximate number of murders, rapes, thefts, or any such crimes that may occur in a particular area in a particular month, year, or any timespan, or, the overall number of crimes occurring in a country in a particular year in the future, or any other prediction or projection of future crime statistics. First, our system proposes to extract data from crime record repositories, on which we intend to perform data mining. Data classification and regression algorithms then help in forecasting and predicting this is proposed to be done by first training a set and then applying the learned rules on the test set in order to determine the predicted output. Using this, law enforcement agencies can better understand how the crime pattern across a certain region, or interval of time is, and using this data, such agencies can take proactive action to stem the rise of particular crimes in particular regions, or during particular times. This would save them a lot of time, money and effort. Our system proposes to mine this data and thus run appropriate algorithms on such data. This predicted output could also be presented to the user in the form of clusters using a data visualization algorithm like K-means clustering algorithm. The final end-product could thus be a system where some future predictions would made by training crime data sets, and the output could be visualized in order to be simple to comprehend for the user.**

Keywords*— Data mining, crime, prediction, algorithms, classification, regression.

I. INTRODUCTION

In today’s world criminals are getting more and more technologically savvy. Because of this, law enforcement agencies have to keep up with them as well. Right from large organized terror networks such as Al-Qaeda, to giant drug cartels such as the Medellin Cartel, all criminal elements these days make extensive use of technology in order to stay abreast of the law. Thus, in such an era where there is such a widespread use of technology for many purposes, this project is done in order to show how the policing officials could cross many technological barriers by analyzing the huge amount of crime-related data that is generated in order to make inferences about crime which might be going to happen in a certain area or a certain timespan in the future, by detecting crime patterns.

In recent years, owing to spiralling rates of crime everywhere, it becomes necessary to have a mechanism to understand future crime patterns, so that even if we are not able to prevent particular crimes from happening, we could at least be prepared to deal with them when the time arises. Thus, the problem essentially is to successfully detect and predict crime patterns with high enough accuracy in order to be able to detect and eventually hamper potential future criminal activity.

Traditional policing methodologies to detect and nip criminal elements in the bud include mobilizing the community by encouraging the establishment of neighborhood watches, requesting that citizens ensure informal social control over one another, enforce civil laws in a particular area more tightly if they get the feeling that the security situation in that area is deteriorating, concentrating attention on those people and circumstances that account for a disproportionate share of a problem (e.g., repeat offenders, repeat victims, repeat locations), etc. However, this has not been found to significantly deter or preempt the rise of crime in a region.

This project was undertaken because it can help the law enforcement officials gauge the crime patterns which are going to be prevalent in a particular area or region, or a certain interval of time in the future. Using this data, these officials can then take proactive measures to prevent the seed of crime itself from taking root.

The work of this project is to help ascertain the incidence and the pattern of crime occurrence in the future, which would include the forecasting and predicting the occurrence of it. This is done by first collecting crime data from crime records repositories such as the National Crime Records Bureau (NCRB) of India, US and UK government maintained crime records, or even records maintained by highly feted intelligence agencies such as the FBI or the CIA, then preprocessing this data by cleaning and filtering it, and then finally generating datasets on which to apply various algorithms or models.

It is upto the end users, what they want as output. In this case, the end users will be none other than law enforcement agencies. These agencies would want to glean out information about the occurrences of crime in the future from the generated datasets, as per their requirements. These requirements could range from wanting to know the frequency of a particular crime in a particular region during a particular time in the future, to simply wanting to know the projected values of all crimes in a general sense, thus, this means that we would have to do forecasting and prediction of crime data. Thus, we plan on running analysis on a host of information from previous years in order to help us perform our predictive analysis.

Through this project, we intend to detect such crime patterns, as well as visualize them on a graph. We intend to gauge crime patterns by applying regression and classification algorithms such as the Bayesian Network algorithm, some decision tree algorithms, forecasting models such as ARIMA (Auto-regressive integrated moving average), or maybe Artificial neural networks (ANN), or maybe use all these algorithms and models and compare their results, thus determining which algorithm leads to the most accurate prediction. Further, we intend to visualize the results of our analysis using algorithms such as K-means clustering. We are intending to use the WEKA tool for our analysis.

II. METHODS AND PROCEDURES

In this project, we are determining patterns of crime as well as intending to forecast, predict and thus detect the occurrence of crime in the future. Various data classification as well as regression algorithms and models are deployed for this purpose. Below discussed are some of the algorithms and models we intend to use:

A. ARIMA (Autoregressive integrated moving average) model

An autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. These models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). Predictive and forecasting techniques such as ARIMA, is used along with artificial neural networks in order to determine a metric called as ‘safety value’ [1], which is deduced using input as crime and census data, and this technology is proposed to be used in smart cities in the future, enabling users to visualize safety levels. [2] ARIMA was also used in order to be able to predict the crime in one particular Chinese city one week ahead, wherein the ARIMA model was derived by using a given property crime data of 50 weeks.

Given a time series of data X_t where t is an integer index and X_t are real numbers, then an ARMA (p', q) model (Autoregressive Moving Average) is given by:

$$\left(1-\sum_{i=1}^{p'}\alpha_i L^i\right) X_t = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$$

Where L is the lag operator, the α_i are the parameters of the autoregressive part of the model, the θ_i are the parameters of the moving average part and the ε_t are error terms. The error terms ε_t are generally assumed to be independent, identically distributed variables sampled from a normal distribution with zero mean.

Assume now that the polynomial $\left(1-\sum_{i=1}^{p'}\alpha_i L^i\right)$ has a unitary root of multiplicity d. Then it can be rewritten as:

$$\left(1-\sum_{i=1}^{p'}\alpha_i L^i\right) = \left(1-\sum_{i=1}^{p'-d}\phi_i L^i\right)(1-L)^d .$$

An ARIMA(p,d,q) process expresses this polynomial factorization property with p=p'−d, and is given by:

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$$

and thus can be thought as a particular case of an ARMA(p+d,q) process having the autoregressive polynomial with d unit roots. (For this reason, every ARIMA model with d>0 is not wide sense stationary.)

The above can be generalized as follows:

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t = \delta + \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$$

This defines an ARIMA (p,d,q) process with drift δ/(1−Σφi)

B. Bayesian Network Algorithm

A Bayesian network, Bayes network, is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. A Bayesian network algorithm can be used to link a certain crime scene [3], with a certain known criminal. This is done by taking all data about the crime scene and comparing it with the data of the respective crime scenes of all criminals in custody; this way, a Bayesian network model is constructed based on this, then finally an inference system is built and if it matches, then the profile of the criminal and the particular crime scene are linked.

X is a Bayesian network with respect to *G* if its joint probability density function (with respect to a product measure) can be written as a product of the individual density functions, conditional on their parent variables

$$p(x) = \prod_{v\in V} p\left(x_v \,\middle|\, x_{\mathrm{pa}(v)}\right)$$

where pa(*v*) is the set of parents of *v* (i.e. those vertices pointing directly to *v* via a single edge).

For any set of random variables, the probability of any member of a joint distribution can be calculated from conditional probabilities using the chain rule (given a topological ordering of *X*) as follows:

$$\mathrm{P}(X_1=x_1,\ldots,X_n=x_n) = \prod_{v=1}^{n}\mathrm{P}\left(X_v=x_v \mid X_{v+1}=x_{v+1},\ldots,X_n=x_n\right)$$

Compare this with the definition above, which can be written as:

$$\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{v=1}^{n} \mathrm{P}(X_v = x_v \mid X_j = x_j$$

for each X_j which is a parent of X_v)

The difference between the two expressions is the conditional independence of the variables from any of their non-descendants, given the values of their parent variables.

C. *Artificial neural networks*

In machine learning and cognitive science, artificial neural networks (ANNs) are a family of models inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning. For example, a neural network for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function (determined by the network's designer), the activation of these neurons are then passed on to other neurons. This process is repeated until finally, an output neuron is activated. This determines which character was read. Like other machine learning methods - systems that learn from data - neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and speech recognition.

Neural networks can be used for prediction with various levels of success [4]. The advantage of then includes automatic learning of dependencies only from measured data without any need to add further information (such as type of dependency like with the regression) [5]. The neural network is trained from the historical data with the hope that it will discover hidden dependencies and that it will be able to use them for predicting into future. In other words, neural network is not represented by an explicitly given model. It is more a black box that is able to learn something.

III. SYSTEM ANALYSIS

A. *Functional Requirements*

1) *Generating of data sets*

This can be done by getting crime data from institutions like the National Crime Records Bureau. Large volume datasets can be obtained corresponding to what parameter we may want to predict and detect.

2) *Cleaning the Data Set*

This is the pre-processing stage which involves organizing the data collected into a form which is easy to run analysis on and get concrete results. The data should be filtered accordingly.

3) *Analyzing the data*

Finally, data is to be analyzed by using some techniques like those discussed in previous sections. Results of this analysis are to be collected and stored. This is the data that police agencies need in order to act proactively to curb the crime in their jurisdictions. This data can be visualized for better understanding purpose.

B. *Non-Functional Requirements*

1) *Performance Requirements*

The performance of the system completely depends upon how quickly the system will be able to run analysis and prepare crime patterns based on the data and the volume of the data to be extracted. It is necessary to maintain the performance of the system so that the results are accurate.

2) *Safety & Security Requirements*

The system must be safe and should not be susceptible to attacks. Attacks can change the integrity of the data/system and result in loss of confidentiality or data loss which can affect the system hugely.

3) *Data Integrity Requirements*

Data Integrity must be maintained to ensure that the system processes this data efficiently and the output produced is accurate. In data mining and learning algorithms, the result is never accurate but it must be ensured that the data input given is uniform and without errors. It must be free of bugs. No changes must be made to the data or the operated data, once it is in processing. Such an intrusion can cause violation of data integrity. Data received as the result must also maintain integrity because a change in this data can cause changes in the final output, thus reducing the accuracy of the algorithm as well as the system and produce an incorrect output.

4) *Availability Requirements*

The availability of the system is an important attribute since the system should be available to use and implement whenever needed. The system should be up and running whenever it is required for. In cases of exigencies, the system must be up and running since the university and college administrators might require the reports and statistics for office and publication purposes.

5) *Portability Requirements*

A system is said to be portable if it satisfies two conditions. First, the system can be run on any device with the same efficiency, time and accuracy as the original system. Second, it must run on all softwares, platforms and operating systems such as Linux, Mac OS, Windows, etc. Portability is important since it will be used in multiple offices and will be required on multiple occasions. Plus, versatility will be maintained if the system can be outsourced to different organizations and because of portability it can be implemented in computers of different educational organizations all over the world.

6) *Maintainability Requirements*

Just like any commodity needs to be maintained and repaired regularly, any system must also be maintained and repaired for errors and bugs regularly. Maintenance of software is important for smooth functioning of the system on these devices to efficiently produce outputs.

7) *Software Quality Assurance Attributes*

a) Reliability: High reliability must be sustained in order to maintain that the system works well in each and every circumstance. The system works in tough environments where it has to make decisions which are entirely corruptible. Thus, the system can be relied upon on any acute circumstance.

b) Robustness: Robustness is the ability that the system works in unusual situations. Situations where large data sets are to be operated on, with different parameters must be handled effectively.

c) Efficiency: Efficiency is attained when the system processes the required information and data in less amount of time, the processing is hassle free and is user friendly. It requires that the system does not lag and will run smoothly with the required base hardware. Also, removal of errors and bugs will alleviate the efficiency and processing of the system.

d) Compatibility: Compatibility is the ability of the software to work with other systems. The system complies strictly with industry standards and there are no violations in rules and regulations that. It also implies that the system is compatible on any platform mentioned in the documentation and can implemented easily. Compatibility also plays a huge role in portability. If portability of the system can be maintained, it will make the respective system also compatible.

e) Modularity: It maintains that the system is divided into modules for better functioning.Modularity is the measure of the extent to which software is composed of separate, interchangeable components, each of which accomplishes one function and contains everything necessary to accomplish this.Modularity increases cohesion and reduces coupling and makes it easier to extend the functionality and maintain the code. Thus, the system inherits a larger database of functionalities, reduced coupling also reduces system overhead.

C. *Use-Case Diagram and Description*

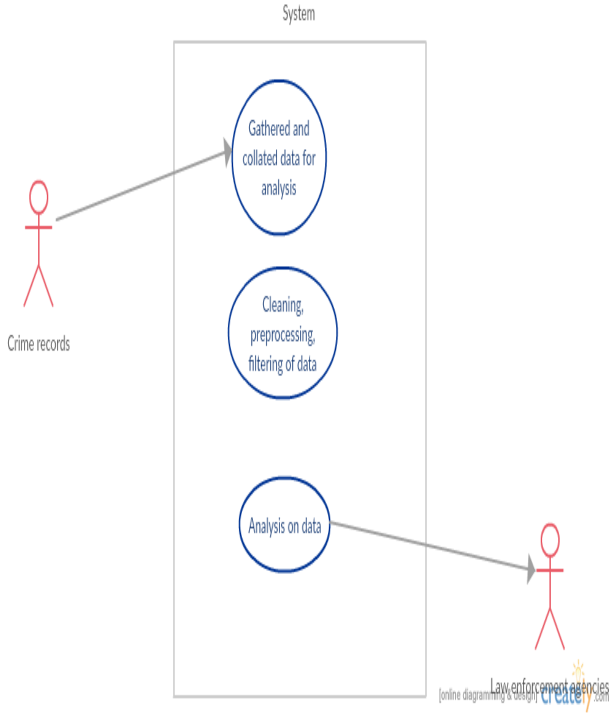
System
Gathered and collated data for analysis
Crime records
Cleaning, preprocessing, filtering of data
Analysis on data
Law enforcement agencies
(online diagramming & design) creately.com

Fig. 1: Use case diagram

In the above diagram, the user, i.e., the law enforcement agencies are presented data after analysis has been done on the data supplied as input data sets in the form of crime records.

First, the data is collected and organized, such that the process of determining the target variables from the predictor variables becomes much simplified later. Then, this data is pre-processed in order to eliminate any redundant fields which might be used in order to predict our target variables. Pre-processing also means that any empty fields are removed, any wrongly supplied data is rectified, etc., thus the data is cleaned and filtered, so that it becomes easy to be analyzed by data mining algorithms. Finally, after the pre-processing of the data has been done, data mining algorithms are applied to it so that the predicted or scaled value of our target variables are determined.

IV. ANALYSIS MODELING AND ARCHITECTURE DESIGN

A. Data Modeling

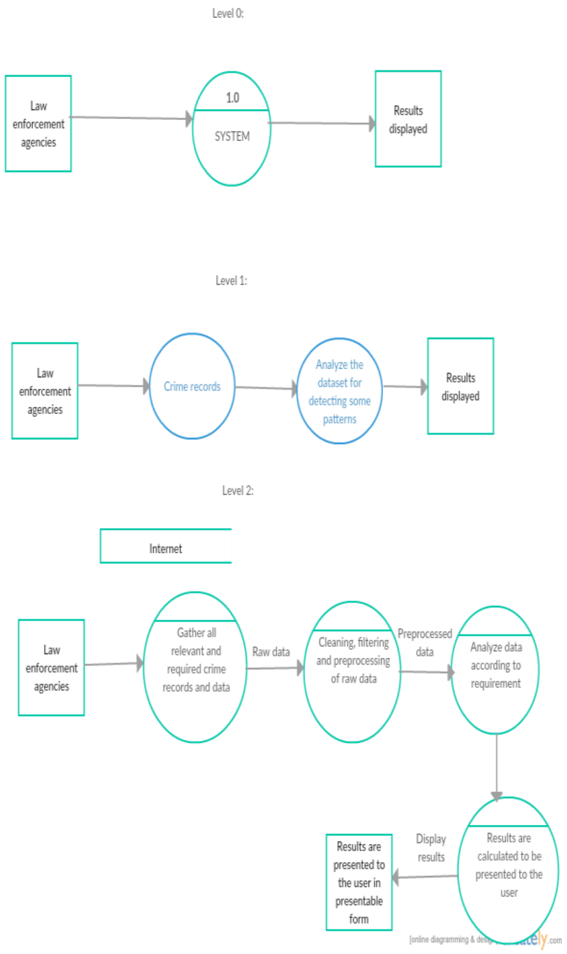
Level 0:
Law enforcement agencies
1.0
SYSTEM
Results displayed
Level 1:
Law enforcement agencies
Crime records
Analyze the dataset for detecting some patterns
Results displayed
Level 2:
Internet
Law enforcement agencies
Gather all relevant and required crime records and data
Raw data
Cleaning, filtering and preprocessing of raw data
Preprocessed data
Analyze data according to requirement
Results are calculated to be presented to the user
Display results
Results are presented to the user in presentable form
[online diagramming & design] creately.com

Fig. 2: Data modeling

1) Level 0: This level shows the crime data being taken in from law enforcement agencies. This raw data is fed into our system, after which, the predicted values of the specified fields are displayed as results.

2) Level 1: This level shows us the general scheme of working of our system. First, data is going to be fed into our system in the form of crime records. Then, these crime records are going to be analyzed using some algorithms to search for patterns, after which, the predicted values of the specified target variables will be displayed.

3) Level 2: This level goes into the most detail. Here, the data which is going to be gathered from law enforcement agencies is going to be in the form of crime records. Now, this data is going to be pre-processed by cleaning and filtering it. After having pre-processed the data, data mining algorithms are going to be applied on the dataset and predicted values of specified target variables are going to be displayed in a presentable format to the end-user or viewer.

B. Activity diagram

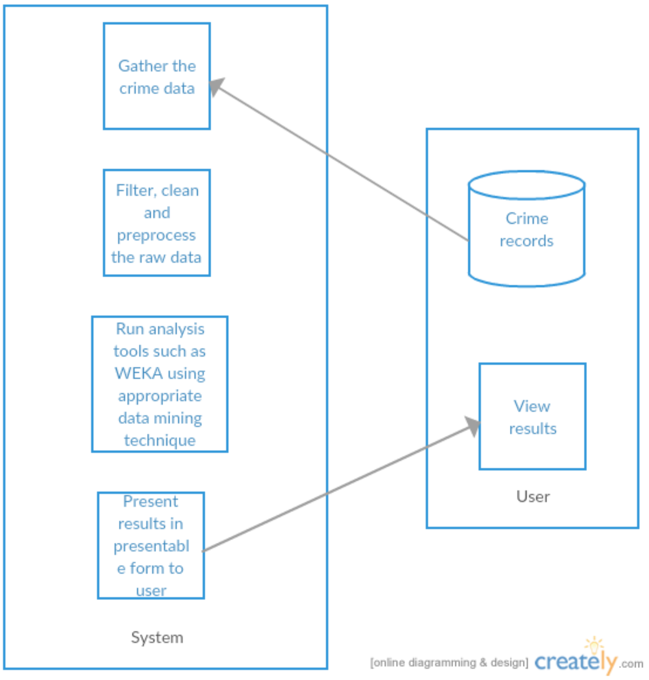
Gather the crime data
Filter, clean and preprocess the raw data
Run analysis tools such as WEKA using appropriate data mining technique
Present results in presentabl e form to user
System
Crime records
View results
User
[online diagramming & design] creately.com

Fig. 3: Architecture

The architecture diagram gives an overview of the software used in the system, and thus, tells us in brief about the functionalities of the system. The system's architecture consists of the functionalities of gathering and aggregating the raw crime data, preprocessing it, using tools such as WEKA, and presenting to the user, while at the used end, the system is also supplied with data which it mines for patterns; also, the system displays the results at the user end.

As shown, data from law enforcement agencies is going to be given as input into our system in the form of crime records. Then, a succession of steps are going to take place: the gathering and integration of all the crime data, the pre-processing of the raw data to make it more easily analyzed and worked upon, application of data mining algorithms; which would find patterns, generate learning rules, and predict future values of the specified fields of the crime dataset, and finally presentation of these results.

V. CONCLUSION AND FUTURE SCOPE

In today's world, where the average quantity of data that a person handles has been increasing by leaps and bounds over the past few years, the utilization of data mining techniques in order to extract useful information from the huge amounts of raw data becomes important.

This project mines the huge amounts of raw data by first generating it in the form of a dataset and then preprocessing it. The various data mining techniques, algorithms and models mentioned when applied on such datasets produces results which could be of great potential use to law enforcement agencies especially. In conclusion thus, we hope that this project performs its functions well, and works with the highest efficiency possible, and that it surely proves to be a boon to law enforcement agencies.

The functionalities of this project can be scaled up in the future. These functionalities could be:

•Real-time data analysis of crime data: This could help us obtain crime patterns and forecasts of the future instantly using real-time datasets.
•Data mining of social media to generate datasets, and then preprocess and analyse them to spot trends of the current crime situation in a particular place or region.
•Compare and display the results of all available and applicable forecasting, predicting and classification models side by side, such that the user can select any of those methods.

ACKNOWLEDGMENT

We would like to sincerely thank our professor, Mrs. Sindhu Nair, for being a part of our effort and guiding us in the right direction. We would also like to thank the entire Computer Engineering department of D. J. Sanghvi College of Engineering for supporting us in all ways.

REFERENCES

[1] Ballesteros, J.; Rahman, M.; Carbunar, B.; Rishe, N., "Safe cities. A participatory sensing approach," in Local Computer Networks (LCN), 2012 IEEE 37th Conference on , vol., no., pp.626-634, 22-25

[2] Peng Chen; Hongyong Yuan; Xueming Shu, "Forecasting Crime Using the ARIMA Model," in Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on , vol.5, no., pp.627-630, 18-20 Oct. 2008

[3] Baumgartner, K.C.; Ferrari, S.; Salfati, C.G., "Bayesian Network Modeling of Offender Behavior for Criminal Profiling," in Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC '05. 44th IEEE Conference on , vol., no., pp.2702-2709, 12-15 Dec. 2005

[4] Deep belief networks at Scholarpedia.

[5] Hinton, G. E.; Osindero, S.; Teh, Y. (2006). "A fast learning algorithm for deep belief nets" (PDF). Neural Computation 18 (7): 1527–1554.