Survey of Analysis of Crime Detection Techniques Using Data Mining and Machine Learning

IOP Conf. Series: Journal of Physics: Conf. Series **1000** (2018) 012046 doi:10.1088/1742-6596/1000/1/012046

Survey of Analysis of Crime Detection Techniques Using Data Mining and Machine Learning

S Prabakaran and Shilpa Mitra

SRM Institute of Science and Technology, Kattankulathur, 603203, Tamil Nadu, India

E-mail: prabakaran.s@ktr.srmuniv.ac.in and mitrashilpa94@gmail.com

Abstract. Data mining is the field containing procedures for finding designs or patterns in a huge dataset, it includes strategies at the convergence of machine learning and database framework. It can be applied to various fields like future healthcare, market basket analysis, education, manufacturing engineering, crime investigation etc. Among these, crime investigation is an interesting application to process crime characteristics to help the society for a better living. This paper survey various data mining techniques used in this domain. This study may be helpful in designing new strategies for crime prediction and analysis.

1. Introduction

Mining of data is a method of dealing with expansive data indexes to perceive outlines and set up an association to handle issues through information examination. The devices used, allow endeavors to accept future examples. Data mining is a procedure to analyze data from an informational collection to change it into a reasonable structure for additional utilization. It predicts future patterns and also enables the organization to make the learning driven decision. Generally utilized strategies for mining of data are artificial neural networks, decision tree, rule induction, nearest neighbor method and genetic algorithm. They are applied in many fields. One such interesting application is crime investigation. A crime is an unlawful activity for which a man can be penalized by law. Crime against a person is called personal crime like murder, robbery, etc. Property crime means theft of property. Crime analysis is a law implementation task which includes an organized analysis that recognizes and determines the pattern of crime. Crime can be classified into different types but, in this, we focused on four types of crime i.e. Fraud detection, traffic violence, violent crime, web crime and sexual offense. The various techniques used for different crimes have been discussed with an introduction to the concerned crime.

2. Types of Crime

2.1. Fraud detection

A fraud is misdirecting or taking unfair advantage of another. A fraud incorporates any act, exclusion, or concealment, including a breach of legal or equitable obligation or confide in, brings about the damage of other. Different types of frauds include check fraud, internet sale, insurance fraud and credit card fraud etc. Check fraud means issuance of a check when enough money is not present in account; internet sale means selling fake items; insurance fraud means fake insurance claimed for automobile damage, health care expenses and other; credit card fraud

means obtaining credit card information from various means which is used for large amount of purchase without the permission of consumer.

2.2. Violent Crime

A violent crime is a crime in which a guilty party threatens to utilize compel upon a casualty. This entails the two crime of rough act called target, for example, killing or rape. Various sorts of this crime are as follows:

- Murdering of individual by other.
- Murder: Deliberate slaughtering of another individual.
- 1st degree murder: Used to allude to a deliberate slaughtering.
- 2nd degree murder: Used to allude to kill accidently in which the executioner shows, outrageous detachment to life of human.

2.3. Traffic violence

Traffic violations happen when drivers damage laws that manage vehicle operation on roads and highways. The increasing number of cars in cities causes high volume of traffic, and implies that traffic violations become more critical which can cause severe destruction of property and more accidents that may endanger the lives of the people. To solve this problem and prevent such consequences, traffic violation detection systems are needed.

2.4. Sexual assault

Criminal attack is the risk or endeavor to physically strike a man, paying little respect to whether contact is really made, insofar as the casualty knows about the peril included. Level of Sexual assualts include:

- Simple Sexual Assault: It includes constraining a person to participate in any type of sexual action without unequivocal assent.
- Sexual Assault with a Weapon: It incorporates the utilization or danger of the utilization of a weapon or damage to an outsider.
- Aggravated Sexual Assault: It happen when the casualty is truly injured, mangled, fiercely beaten, or in threat of passing on because of a rape.
- Verbal assault: It is a sort of non-physical, oral ambush that outcomes in a passionate, mental, and additionally mental damage to the casualty, instead of physical substantial damage way.

2.5. Cyber crime

Cyber-crime is the crime related to computer. It comprises of computer and a network for crime to occur .Offenses that are perpetrated against criminal process to hurt the victims by present day media transmission systems, for example, net and cell. Various types are web extortion, ATM misrepresentation, wire misrepresentation, document sharing and robbery, hacking, and so forth. Cyber-crime analysis is very important responsibility of law enforcement system in any country. It includes breakdown of protection, or harm to the PC framework properties, for example, documents, site pages or programming.

3. Techniques for detection of various crime

3.1. General techniques used in fraud detection

Various techniques are employed in fraud detection. Some prominent techniques are discussed here.

3.1.1. Genetic algorithm Genetic algorithm is a technique for understanding both obliged and unconstrained modifying issues. It is based on a biological choice process that impersonates natural development. It generates points at each iteration. It approaches an optimal solution. It is related to the tremendous class of transformative algorithms which creates solutions to improve issues like inheritance. Application of this algorithm are bioinformatics, computational science, engineering, mathematics, physics and other fields. Advantage of generic algorithm are more efficient, faster, optimize continuous and discrete function.

Five phases in this algorithm are as follows:

- Initial population means beginning of process by individual.
- Fitness function tells about the individual capability to compete with other.
- Selection function select the best individual.
- Crossover is the most significant phase and chosen at random from within the genes and offspring are created by exchanging the genes of parents among themselves until the crossover point is reached, mutation means flipping of some bits.

Steps of Genetics algorithm:

(i) Random population of n chromosomes are generated.
(ii) The fitness *f(x)* of each of the chromosome x is evaluated in the population.
(iii) A new population is created by iterating steps until the new population is finished.
(iv) Newly created sets are used for rest of the algorithm
(v) At last if all the functions are up to the mark then we have to stop and then return to the best solution.
(vi) Return to step 2.

General terms used in Genetic algorithm are as follows:

- Selection: Chromosomes are selected from a group.
- Crossover: With a hybrid likelihood, main set of data form a new set of data. If no hybrid event was performed, new set of data is the copy of the main set of data.
- Mutation: Likelihood of new production at each locus.
- Accepting: New data are placed in the new sets.

3.1.2. Hidden Markov Model(HMM) Hidden Markov Model is a statistical model of process consisting two random variables which change the state sequentially. It has limited arrangement of states in administered set of change probabilities. A Hidden Markov model enables us to discuss about both monitored events and hidden event. It consist of three fundamental technique, which are as follow:

- Likelihood: HMM (λ) = (A, B) and an observation sequence O,used for determining likelihood P (O $|\lambda$).
- Decoding: Observation sequence O and an HMM (λ) = (A, B), finds the good hidden state sequence O.
- Learning: Observation sequence O ,set of states in the HMM, used to learn HMM parameters A and B

It is used in multiple areas like speech recognition, natural language processing, etc. It is efficient and have good statistical foundation.

3.1.3. Naive Bayesian Naive Bayesian is a good classification model. It gives the probability distribution to get optimal result. It is based on probability. This is applied for calculating the posterior from the prior and the likelihood as it is easy to calculate from a probability model. This technique is used when the dimensionality of input is high. The equation used for calculating posterior likelihood *p(c|x)* from *p(c)*, *p(x)* and *p(x|c)* is given below:

$$\text{P(A | B)}=\frac{P(B|A)P(A))}{(P(B))}$$

That is,

$$\text{Posterior}=\frac{((Likelihood)(Propositionpriorprobability))}{(Evidencepriorprobability)}$$

Steps of performing this algorithm are:

(i) Convert the data set into a frequency table
(ii) Create Likelihood table by finding the probabilities
(iii) Naive Bayesian equation is used for calculating the posterior probability for each class.

3.2. General techniques used in Violent Crime Detection

3.2.1. Fuzzy c-means algorithm Fuzzy c-means algorithm is a technique used for clustering of data. In this dataset are gathered into n number of group with each datum point in the dataset related to every cluster of specific degree. It gives better result than k mean algorithm. Steps of performing fuzzy c mean algorithm. Assume x set of data points i.e. {x1, x2, x3 ..., xn} and v set of center i.e. {v1, v2, v3 ..., vc }.

(i) 'c'cluster centers are selected randomly.
(ii) Fuzzy membership *'μ_{ij}'* is calculated using: $\mu_{ij} = \frac{1}{\sum_c^{i=1}\left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}}$
(iii) Then, fuzzy centers *'vj'* computed using:
$v_j = (\sum_{i=1}^{n}(\mu_{ij})^m x_i)\,(\sum_{i=1}^{n}(\mu_{ij})^m)$
(iv) Step 2 and 3 were repeated until the minimum *'j'* value is achieved or$\left\|\mu^{(k+1)} - \mu^{(k)}\right\| < \beta$
, where
k is the iteration step.
β is the termination criterion between [0, 1],$u = (\mu_{ij})_{n*c}$.
j is the objective function.

3.3. General techniques used in Traffic violence

3.3.1. Cumulative logistics model Cumulative logistics model predicts probability of response in ordinary scale. It has assumption of proportional odds i.e. Coefficients of predictor category which is consistent in all levels of response. Association rule mining is a procedure of finding random patterns, correlations, associations, or causal structures of information sets. Data sets are available in different types of database like relational databases, transactional databases etc.

3.3.2. K-mean Clustering K-means clustering is an unsupervised learning, used for unlabeled data i.e., data are not described as categories or groups. This algorithm find groups in the given data substituted by *k*. It assign each point to one groups on the basis of features. Advantage of k mean algorithm are fast, robust, easy to understand and gives best result. For portioning the cluster there are different ways i.e.

- Dynamically chosen: It means that data are grouped into three different clusters, the initial cluster is first three items of data.
- Randomly chosen: It means initial cluster are randomly chosen values.
- Choosing upper and lower bounds: It depends on the set of data in the datasets, initial cluster means are chosen by highest and lowest data range.

Steps for K-mean Clustering:

- Input dataset, clustering variables and maximum number of clusters needed.
- Initialize the cluster centroid.
- Calculate the Euclidean Distance.
  Euclidean Distance=$\sqrt{(X_H - H_1)^2 + (X_W - W_1)^2}$
  Where,
  X_H Observation value of variable height.
  H_1 : Centroid of Cluster 1 for variable height.
  X_W : Observation Value of variable weight.
  W_1 : Centroid value of cluster 1 for variable weight
- Continue the steps until all observations are assigned and required clusters are found.

3.3.3. K- mode Clustering K-mode clustering algorithm is addition of k-mean clustering algorithm for clustering explicit data. It replaces the Euclidean distance function with the simple matching dissimilarity measure. Mode is used to represent center of cluster. By this algorithm, efficiency of the clustering process is maintained. Steps for performing k mode clustering are listed below:

- K initial modes are selected, each of the cluster is having one mode.
- Data object are allocated to the cluster in which mode is nearer.
- Compute new modes of all clusters.
- Repeat step 2 to 3 until no data object has changed cluster membership.

3.3.4. Neural Network Neural network gives an emerging paradigm for pattern recognition implementation, it involves inter-connected network of relatively simple and typically non-linear unit. This algorithm is non-algorithmic and black box strategy which is trainable. This algorithm is attractive to PR system designer as it requires prior and detail knowledge of internal system operation is minimal. Application of Neural network are character recognition and image processing. Characteristics of Neural network are listed below:

- Network topology or interconnection of neural unit.
- Characteristics of individual unit or neuron.
- Strategy of pattern learning or training.
- High dimensional and complex interaction between problem variable.
- Feature extraction from complex dataset.

3.4. General techniques used in Sexual Assault

3.4.1. Kernel density estimation Kernel density estimation is a non-parametric method for evaluating the probability of density function in a arbitary variable. It is an information smoothing issue where inferences of the population are made, based on limited information test. It is an effective multi-modal information representation in which consideration of noise for observed data is representation of model/state. The equation of kernel density function is referenced where, n is the sample size, h is the bandwidth, *f(x)* is kernel density.

$f(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i}K\left(\frac{x-x_i}{h_i}\right)$

There are some limitation of kernel density function i.e. the computation is slow and it also stores most of the data in the memory,it also require huge amount of memory, there is an issue in the selection of bandwidth.

3.4.2. Logistic regression Logistic regression is a simple classification algorithm for making decisions. It is a measurable strategy for breaking down a dataset in which there are atleast one free factor that choose the outcome. The objective is to discover the best fitting model for describing the association between dependent variable and independent variables .It generate the coefficients formula for predicting a logit change of the likelihood of presence of interest:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_kX_k$$

Where p = probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$\text{odds}=\frac{p}{1-p} = \frac{probabilityofpresenceofcharacteristics}{probabilityofabsenceofcharacteristics}$$

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

It predicts outcome of variable that is categorical from predictor variables which is continuous. Logistic regression deals these type of situation with logarithmic transformation on the results to model a nonlinear association in a linear way. Advantage of this algorithm are run time of this algorithm is very fast. Logistic regression work well with single decision boundary. Logistic regression is simple, and have low variance that is the reason it is less prone to over-fitting.

3.4.3. Random forest algorithm Random forest algorithm is a supervised classification algorithm and bootstrapping algorithm with decision tree model. It is a learning method of classification and regression. It is operated by constructing decision trees at training time and result comes as individual trees. Advantages of this algorithm are it uses classification and the regression task both, also handle the missing values and it wont over fit when there is more number of trees in the forest and it also classifies categorical values. This algorithm is mostly used in banking, medicine, stock market and e-commerce. Steps for random forest algorithm:

(i) k features are randomly selected from total m features, where k << m
(ii) Node d is calculated.
(iii) Nodes are splited into its branch nodes.
(iv) Steps from 1 to 3 are repeated still number of node become l.
(v) Steps from 1 to 4 are repeated until a forest is created.

3.5. General techniques used in cyber crime

3.5.1. Influenced association rule Influenced association rule mining algorithm is used in pre-processed data set for obtaining the pattern invention. Influenced association classification uses weighted support and confidence. It is stored in the rule base index, if any new data is updated, this will forecast the class label from the rule base. Prediction using j48 occur in two steps

- Formation of tree.
- Validate the built tree over the cybercrime data set.

3.5.2. J48 algorithm The anticipated j48 algorithm classifies the data till the entire categorization and affords utmost accuracy over the training data. It also stabilizes the precision and litheness. The j48 algorithm is the extensive version of decision tree c4.5. The j48 algorithm produces the classifier output in the form of rule sets and decision tree. The rule sets are straightforward to recognize and too easy for employing within the application. It uses pruning method for construction of the tree. This method reduces the size.

4. Related Works

4.1. Survey on fraud detection

Syed Ahsan shabbier *et al.*, [1] described Generic algorithm for preventing credit card frauds. It was used for improving the computing cost with time by creating complex systems. It could analyze a fraudulent transaction in few second. The probability of misrepresentation exchanges could anticipate not long after credit card exchanges and arrangement of hostile to extortion systems could be received to keep banks from incredible misfortunes and minimize dangers.

Naeimeh Laleh *et al.*, [2] discussed supervised methods, semi-supervised methods, unsupervised methods, and real time approaches to detect the type of fraud and compare the different techniques.

Abhinav Srivastava *et al.*, [3] described hidden Markov model. It showed the execution and adequacy of the device. It also demonstrated the needfulness of taking the spending profile. The accuracy of the system was 80 %.

Sammaes *et al.*, [4] proposed Bayesian and Neural networks that provide computational learner which consist of training set having feature and data for detecting fraud so that it can correctly classify the new data as fraud or not. It is concluded that both the technique can be used for detecting fraud.

4.2. Survey on violent crime

Chao Yangt *et al.*, [5] discussed about rough-fuzzy c-means algorithm for analysis of violent crime, rough set and information entropy. It was combined to upgrade the capacity so that it could deal with the uncertainty, vagueness, and incompleteness. This algorithm was used for resolving overlapping data.

Chao Yang *et al.*, [6] proposed swarm rough algorithm to investigate the mix components of brutal crime and break down three sorts of mix factors, i.e. Genetic, natural and psychological factors and assessed the execution and the fuzzy swarm optimization technique by getting numerous diminishments for the mix factor datasets. It works better in a mixed dataset group.

Jorge E *et al.*, [15] discussed about outdoor physical actions and violent crime among internal city youth. Multiple regression analysis was performed using outdoor physical actions. This survey was performed for demonstrating connections between adolescents outdoor physical action and for measuring violent crime densities along other natural key variables.

4.3. Survey on Traffic Violence

Sachin Kumar *et al.*,[13] discussed k-mode clustering and association rule mining algorithm which were used to examine various design or pattern of accidents occurred in the road. After applying the algorithm EDS was made basis of month and hour to monitor the accidents occurred. Aaron Christian *et al.*, [7] proposed genetic algorithm. The system provided detection for both violation but detected swerving violations faster than blocks the pedestrian lane violation and process one data at a time but runtime of the system is slow but can be improved.

Jieling jin *et al.*,[8] described about cumulative logistics model, neural network model and bayesian network model and used for analyzing the traffic violation and compared different model. Accuracy of Bayesian networks was about 70%, the cumulative logistic model was about

47%, and the neural network model was about 51%. Bayesian networks model better predicted the level of traffic violations.

Sachin Kumar *et al.*,[12] proposed k-means clustering and association rule mining algorithm It was used for indicating the rate of accident prone areas i.e. high, low and moderate. Association rule mining was used for finding the association between various attributes that frequently happened together when an accident takes place. Both the algorithm could be used for recognizing factors related with road accidents.

4.4. Survey on Sexual Assault

Elise Clougherty *et al.*, [9] discussed kernel density estimation, logistic regression and random forest modeling was used to conduct spatial and temporal analysis of sexual assault.Kernel density estimation was used to compare the probability density functions of sexual assaults over daily, weekly, and monthly time periods. They constructed time series using logistic regression, and random forest models to assess correlation between point-locations of sex crimes, weather conditions. These results indicate that sexual assault is more likely to occur near the homes of registered sex offenders.

4.5. Survey on Cyber Crime

Anshu sharma, *et al.*, [10] proposed k means clustering algorithm which was used for constructing patterns of data. Data were collected and distributed, two third of true data and misrepresentation history information were utilized for preparing and remaining information were utilized for forecast and web crime discovery. The precision of the proposed work was 94.75 % and it productively recognized the false rate of 5.28%.

K. K. Sindhu *et al.*, [11] explained scientific investigation ventures in the capacity media and hidden data investigation in the record framework, network forensic and cyber-crime mining. Device was proposed by combining digital forensic investigation and mining of crime data intended for discovering motive and pattern of attacks and hecks of assaults sorts occurred in that time period.

K. Chitra lekha, *et al.*, [14] discussed about k-means algorithm, influenced association classifier and j48 prediction tree for detecting web crime data sets and for solving the problem. It also recognize patterns in crime for predicting participating criminal activity so that it can be controlled. They developed a crime tool for recognizing aim of crime instantly and to detect future cybercrime pattern.

5. Resources available for Crime Data Analysis

To develop unique classification forecasting model, bench mark datasets are necessary. In case of crime analysis various dataset resources are available which are listed in the following table.

SNO	Type of data	Description	Source
1	Communities and crime un- normalized dataset	Actual crime statistical data for the state of Mississippi in which 4 non-predictive features,125 predictive features and 18 potential goal features are present.	www.neighborhoodscout.com, University of California-Irvine repository

SNO	Type of data	Description	Source
2	Cybercrimes by motives	Data regarding various cybercrime like revenge, greed, extortion, harassment and others.	Https://data.gov.in /catalog/cases-registered-under-cyber-crimes-motives
3	Communities and crime dataset	Multivariate characteristics of data with real attribute characteristic associated with regression, which consists of 1994 instances and 128 attributes.	Http://archive.ics.uci.edu/ml /datasets
4	Persons arrested for crime against women	The various crimes covered for a given time period, like rape kidnapping and abduction, insult to the modesty of women, immoral traffic act, dowry deaths, commission of sati prevention act and assault on women	Https://data.gov.in/catalog/persons-arrested-under-crime-against-women
5	Victims of murder	A total number of cases reported under murder and number of victims of murder by gender on the basis of age group.	Https://data.gov.in/catalog/age-group-wise-victims-murder
6	Criminal victimization data	Data consists of burglary, robbery, theft, fraud, assault, murder, thuggery and rape cases.	In-home, face-to-face personal interview using a stratified multistage random selection procedure
7	Real-word crimes in two cities of the US	Dataset information is based on the national Incident-based reporting system for five years. It consists of 19 attributes with 333068 instances, also gives the exact occurrence time and location of a district where a crime occurred	http://data.denver gov.org/dataset /city-and-county-of-denver crime. Http://us-city.census.okfn.org/dataset/crime-stats.
8	Crime data of Tamilnadu	Six cities (Chennai, Coimbatore, Salem, Madurai, Thirunelvelli, Thiruchirapalli) crimes are listed for the year 2000-2014 which consist of 1760 instances and 9 attributes.	National crime records bureau (NCRB)

References

[1] Syed Ahsan Shabbir and Kanna Dasan R 2013 An Effective Fraud Detection System Using Mining Technique *International Journal of Scientific And Research Publications* **3(5)**

[2] Abhinav Srivastava, Amlan Kundu, Shamik Sural and Arun K. Majumdar 2008 Credit Card Fraud Detection Using Hidden Markov Model *IEEE Transactions On Dependable And Secure Computing* **5**

[3] Sam Maes, Karl Tuyls, Bram Vanschoenwinkel, Bernard Manderick Credit Card Fraud Detection Using Bayesian And Neural Network *Https://Www.Researchgate.Net/Publication/254198382*

[4] Chao Yangt, Shiyuan Chet, Xueting Cao, Yeqing Sun, Ajith Abraham 2013 A Rough-Fuzzy C-Means Using Information Entropy For Discretized Violent Crimes Dat *13th International Conference On Hybrid Intelligent Systems* .

[5] Chao Yang, Hongbo Liu, Yeqing Sun, Ajith Abraham 2012 Multi-Knowledge Extraction From Violent Crime Datasets Using Swarm Rough Algorithm *12th International Conference On Hybrid Intelligent Systems (His)*

[6] Jieling Jin, Yuanchang Deng 2017 A Comparative Study On Traffic Violation Level Prediction Using Different Models *4th International Conference On Transportation Information And Safety (Ictis)*

[7] Anshu Sharma, Shilpa Sharma 2012 An Intelligent Analysis Of Web Crime Data Using Data Mining, International Journal Of Engineering And Innovative Technology *(Ijeit)* **2**(3)

[8] K K Sindhu and B B Meshram 2012 Digital Forensics And Cybercrime Data Mining, Journal Of Information Security *3, 196-201 Http://Dx.Doi.Org/10.4236/Jis.2012.33024 Published Online July 2012 (Http://Www.Scirp.Org/Journal/Jis)*

[9] Sachin Kumar and Durga Toshniwal 2016 A Data Mining Approach To Characterize Road Accident Locations *Journal Of Modern Transportation* **24**(1) pp.6272

[10] Sachin Kumar and Durga Toshniwal 2015 A Data Mining Framework To Analyze Road Accident Data *Journal Of Big Data*,