Conference Title

Mining Criminal Dataset Using Gradient Boosting Algorithm

Dr Akash Saxena[a], Gaurav Kumar Das[b] and Avinash Sain[a,b]

[1]Department of Compter Science and Engineering, CITM,Jaipur,India
[2]Department of Compter Science and Engineering, CITM,Jaipur,India

ARTICLE INFO

Article history:
Received 00 December 00
Received in revised form 00 January 00
Accepted 00 February 00

Keywords:
Criminology,
Crime Analysis,
Crime Prediction,
Data Mining .

ABSTRACT

Data mining is a process for evaluating & testing huge remaining databases towards produce novel data that can be necessary for association. Using existing datasets is predicted to remove new information. There have been many approaches to data mining analysis and prediction. But there have been only several attempts in the field of crime science. Very few people have tried to compare all of these approaches which are with the information produced by them. Police stations & further related criminal justice agencies maintain big databases of data that may expect or examine criminal activity & criminal activity into community. We may too predict criminals built proceeding crime statistics. On basis of these subjects, we have used the concept of data mining to predict the criminology and the reasons behind the occurrences of the crime. The proposed algorithm is able to predict more significant features with higher accuracy and efficiency.

© 2019 Uttaranchal University, Dehradun. Hosting by Elsevier SSRN (ISN) All rights reserved.

Peer review under responsibility of Uttaranchal University, Dehradun.

1. Introduction

Generally law enforcement & criminal justice experts have right towards solve crime. As computer usage system is used to detect and track off criminals, computer information analysts have started helping law enforcement & investigators towards speed up procedure of explaining crime. Criminology is procedure utilized towards detects crime & criminal features. Criminals & potential offenders may be evaluated by assistance of criminology methods. Culprit helps detective agencies, police department, & crime branch to identify real features of a criminal.

Since 1800s, crime investigation department has been utilized into investigation of crime. Crime is a social nuisance, & it can spend our society into many techniques. Government of India has occupied phases towards improve applications & software aimed at utilize by the state & central police into connection by National Crime Records Bureau (NCRB). Some research that helps into explaining crime quickly will pay aimed at it. Approximately 10% offenders commit 50% offenses [5]. People who study crime science can identify the culprit on the basis of crime, features & crime. In mid-1990s, data mining developed like a powerful device towards remove valuable data after big datasets & to discover relations among data attributes [4]. Data mining formerly derived after statistics & machine learning such an interdisciplinary field, then it has since become one of top 10 leading technologies that will alteration world in 2001 [6]. Conferring towards various researchers, for example Nath [2], explaining crimes is a hard & time-consuming challenge that needs human intelligence & experience, & data mining is a method that helps us find the crime. A data mining paradigm should be developed which implements an interdisciplinary method among computer science & criminal justice for quick resolution of crime.

As previously mentioned, crime science is a process aimed at identifying the characteristics of crime, which very essential area aimed at implementing information is mining. With it, data mining algorithm can produce crime reports and can help identify criminals earlier than any human. Due to this remarkable issue, there is an increasing appeal aimed at data mining into crime science. Crime analysis is a procedure that involves investigation of nature of the crime, identifying the crimes and their relationship with the criminals. The complexity of the relationship between the crime & criminal datasets & complexity of relationship among these kinds of data creates criminology an ideal area aimed at implementing data mining methods. Detecting features of crime is the 1st step into additional analysis. Quality of information analysis be contingent on background information of analyzer. From a criminal terrorist attack to massive attacks such as illegal driving, for example 9/11 attacks, it is difficult towards model best algorithms towards cover them all. [3] Information increased after data mining methods is most useful & may assist & assist police. In specific, we may utilization classification and clustering-based models to help identify types and types of offenders. Extensive range of data mining applications into Criminology creates it a significant research area. Data mining schemes have played an important part into helping humans into this forensic domain & criminology domain. It creates it one of maximum inspiring decision-making atmospheres aimed at research.

Electronic copy available at: https://ssrn.com/abstract=3418733

Aim of this research work is to help young researchers working in areas of criminal analysis & crime forecasts. This paper is planned into a way that provides statistics on the crime analysis process and then applies them to various types of crime analysis activities, which may be applied towards crime analysis into any police station. Investigator Agencies this work will be a useful reference towards those who have completed research work into crime forecasts by crime analysis & data mining methods.

2. Literature Review

Sharma [7] He came with the idea of portraying crime in society. In order to detect suspected criminal activity, the company focused proceeding significance of data mining technology & prepared an active application aimed at it. In his dissertation, he suggested a device that appoints a better decision tree algorithm for detecting doubtful e-mails around criminal activity. With an advanced feature choice technique & attribute-importance issue, a better ID3 algo is useful towards create a well & fast decision view built proceeding data entropy, which clearly comes after training information set after some classes. He suggested a novel algo which adds improved feature selection method for improved ID3 classification algorithms and well efficacy of algorithms.

Hamdy et al. [8] Describes a method built proceeding mobile usage, for example call logs for people's conversation with location markers and social networks His work too presented a model to detect doubtful behavior built proceeding social NW feeds, which not only defines a novel way of with people's social interaction then also quickly and accurately judgments of crime analysis Proposes a new mechanism to help in taking. Suspicious offer of unit may be resolute by sequence of assumption instructions.

Bogahawatte and Adikari [9] suggested a method in which they emphasized utilization of data mining methods, classification & clustering aimed at effective exploration of criminal and crimes identification through increasing a structure called Intelligent Crime Investigation System (ICSIS) that might recognize a criminal built up proceeding proof composed after crime location. They utilized clustering towards detect crime designs which are utilized towards obligate crimes perceptive detail that every crime has definite designs.

Agarwal et al. [10] Sharp mining equipment was used to analyze crime rates & to estimate crime rates with various data mining methods. His work is based on crime analysis with K-Means clustering algo. Main purpose of their crime analysis work is towards remove pattern of crime, based on spatial dissemination of remaining information, predict the crime & find out the crime. Their analysis involves detecting domestic crimes after one year towards next.

Kiani et al. [11] a crime analysis work was done built proceeding classification & clustering methods. Based on criminal information available in their work, it is possible to remove the pattern of crime through crime analysis, predict crime based on local dissemination of remaining information & identify crimes. They suggested a model of crime analysis & estimate by optimizing the outlier detection operator parameters implemented by genetic algorithms. In this model, the features are weighted and the selection of the appropriate range is eliminated to reduce the value of the facilities. Clustering of K-media clustering algorithms for classification of crime datasets.

Satyadevan et al. [12] one such work was done in which there was a high probability of crime incidents and visible areas of potential crime. Rather than focusing only on crime, they focus primarily on criminal elements of each day. Every day they use tumor bias, logistic regression and SVM classifier to classify crime factors & crime patterns. Their technique is a design recognition step that can identify the tendency and pattern of crime using empirical algorithms. The prediction decision of crime spots is done by support of the tree algorithm, which identifies areas of crime & their types.

Huang et al. [13] built proceeding location-based social NW interaction, mining focuses on another approach to forecasting forecasts. Using these interventions, they can collect information from people using geographic interventions and information. They prepared a process that categorizes many features after Foursquare & Gowalla utilized into San Francisco Bay Area.

Bruin et al. [14] based on your criminal career, a technique suggested towards define clustering of criminals. Criminal profiles are removed from the database for each crime per year & a profile space is considered. From that, distance matrix into profile is made annually. Distance matrix with frequency value is used towards build a cluster with built-in clustering algo.

Electronic copy available at: https://ssrn.com/abstract=3418733

3. Proposed Methodology

To gain insights about the occurrence of crime & before generate several kinds of crime analysis operations & those which may be applied both aimed at generating an end user product which may be useful towards crime analysis into any detective agencies & police stations, it is important to produce such techniques which can help to reduce criminology acts.

For clean data, removal of missing values was needed to get an appropriate crime data set. Initially, columns that had these missing values or sparse values were deleted as undefined values would have a negative impact proceeding accuracy of model. For dirty data, missing data of a feature were converted into median value of that feature. For predicting feature, 'Per Capita Violent Crimes', a new column called 'High Crime' was created that had a value '1' for 'Per Capita Violent Crime' greater than 0.1 and '0' otherwise.

The threshold of 0.1 was decided upon manual analysis of data by view-through process. All the features had to be predicted using this target feature 'Highcrime'. Clean and dirty data sets were converted into different data frames and the target feature 'High crime' was assigned to a variable 'Target' and the remaining features to 'Features'.

In the previous work, a decision tree named Random forest (ensemble classifier) was used for the implementation of this work. But due to some limitations and issues faced by it, we opted a new machine learning ensemble approach to overcome the deficiencies of the existing classifier.

In our research work, we have used an ensemble approach named as Gradient Boosting. Boosting is an ensemble method in which predictors are not prepared autonomously, then consecutively.

This technique uses the logic of successive predictions, learning from the mistakes of previous predictions. Thus, there is an uneven possibility of appearing in subsequent models in observations, in which the highest error is most visible. (So the observations are selected on the basis of error, not on the basis of the bootstrap process). Predictors can be selected from models like decision-making, registrars and classifier. As new predictions learn from the mistakes made by the previous predictors, sometime / repetition seems to get close to the actual forecast. But we need to carefully select those criteria that we are determining or it may be over-fitting of training data. Slowly an example of boosting the boosting algorithm.

Dealing with categorical features efficiently is one of major tasks into machine learning.

Algorithm: Catboost

Step 1.	Collect dataset Communities and Crime from UCI Repository
Step 2.	Convert text file into csv format
Step 3.	Start
Step 4.	Import dataset
Step 5.	Clean the dataset by removing missing values
Step 6.	Replace "Per Capita Violent Crimes" with a new column "High Crime"
Step 7.	If
	Per Capita Violent Crimes >0.1
	High Crime =1
	0, Otherwise
Step 8.	Assign High Crime column to a variable "Target"
Step 9.	Remove features "Per Capita Violent Crimes" and "Target"
Step 10.	Now train dataset with 10 fold cross validation
Step 11.	Apply Gradient Boosting
Step 12.	Run the algorithm
Step 13.	Obtain the best 10 features for this model
Step 14.	Obtain the final predicted result
Step 15.	End

Electronic copy available at: https://ssrn.com/abstract=3418733

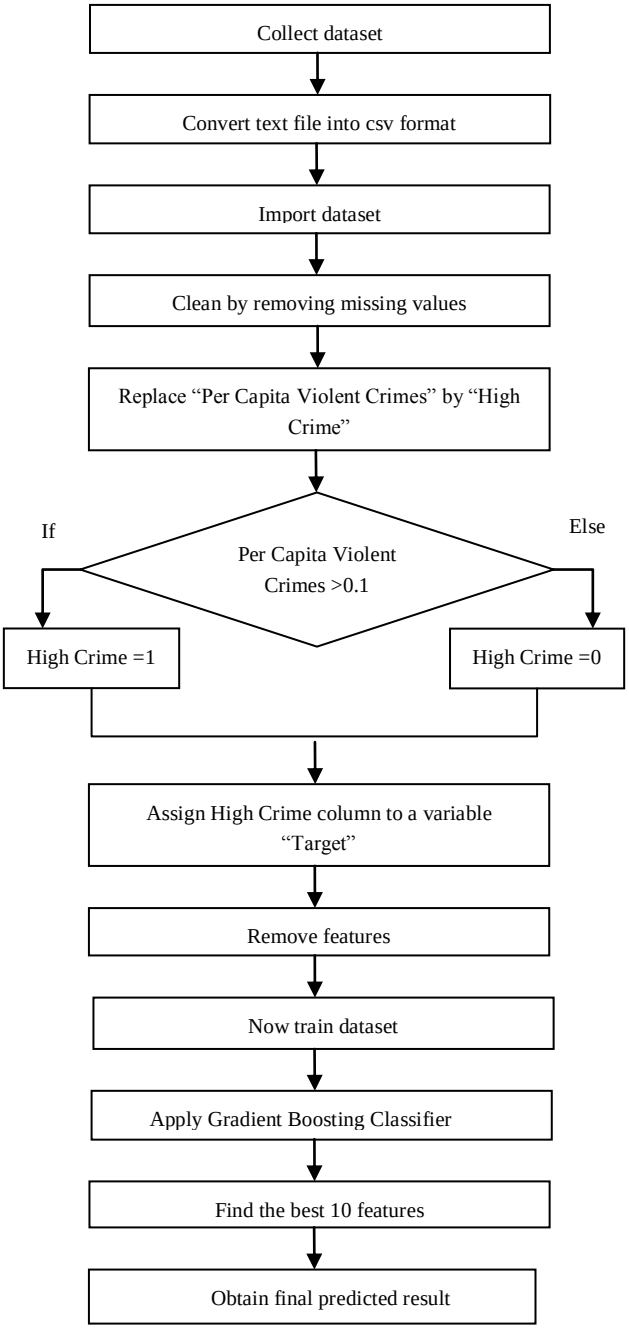


Fig. 1 - Data Flow diagram of the proposed model

Electronic copy available at: https://ssrn.com/abstract=3418733

4. Result Analysis and Discussions

The data set as mentioned above is taken from a UCI Repository: Communities and Crime dataset. It has total 1994 instances and 128 attributes like population, race, and age. The attributes are real and of multivariate characteristics. This data was first converted into CSV file using JSON file from the website using Python. For naming convention, original data was assumed as 'dirty data' and the data with no missing values as 'cleaned data'.

In this section, we have performed number of experiments. These experiments have been simulated through Python programming language. The attribute or feature towards be expected is 'Per Capita Violent Crimes' which was pre-calculated into information by population & sum of crime variables deliberated violent crimes into United States: robbery, murder, rape, & attack. Experiments have done for two type of data that are cleaned data and other one is dirty data.

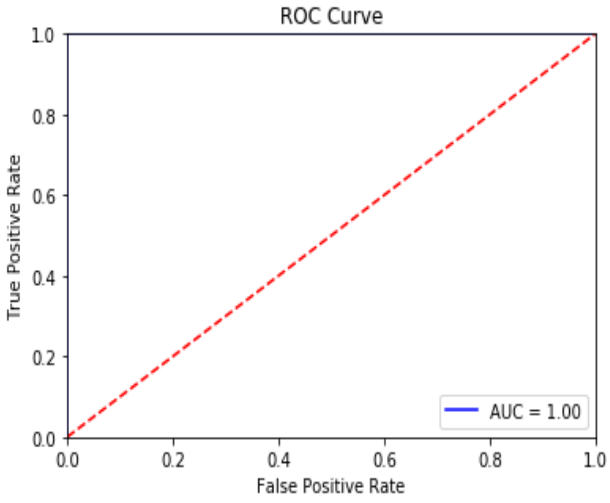
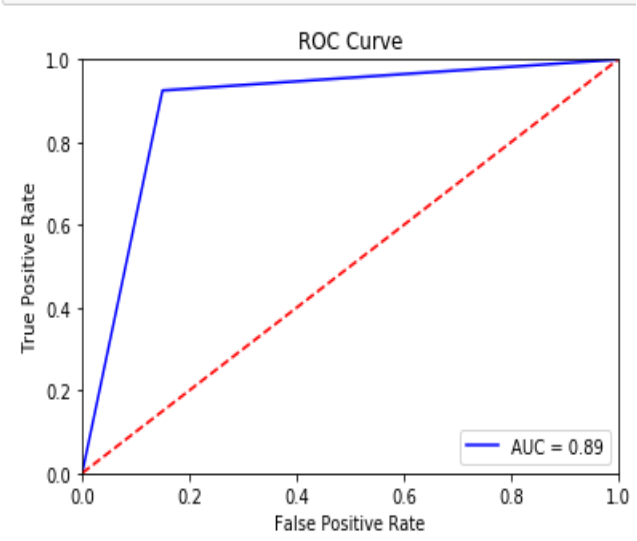


Fig. 2 - ROC Curve for Cleaned Data training through random Forest.

In above fig 2 plots the ROC Curve aimed at training of cleaned data. The ROC has false positive rate at X-axis and true positive rate at Y-axis.



Electronic copy available at: https://ssrn.com/abstract=3418733

Fig. 3 - ROC curve plot with respect of TP and FP rate.

In figure 3 plots the ROC Curve for training of cleaned data. Here, AUC =0.89.

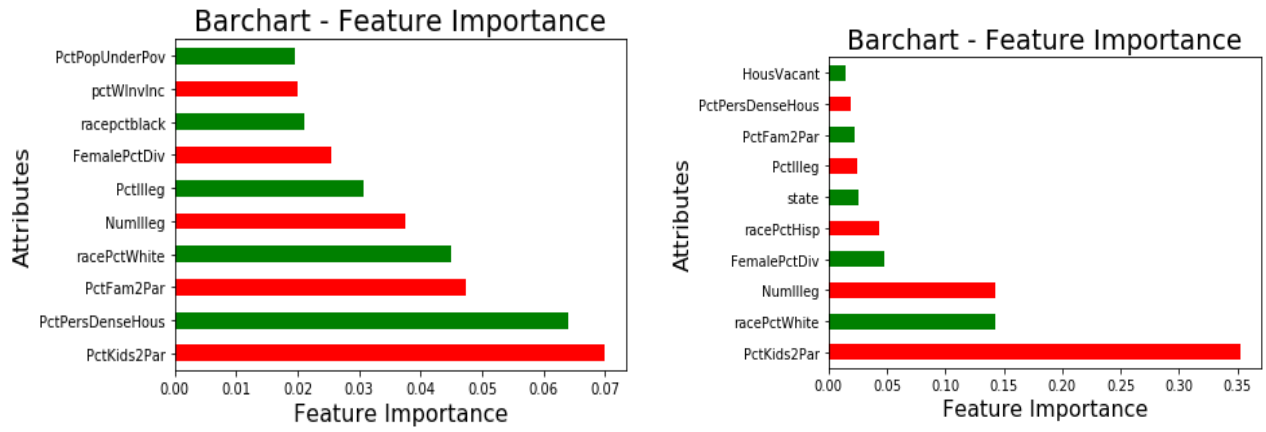


Fig. 4 – (a) Bar Graph of Feature Importance for cleaned data through random forest (b) Bar Graph of Feature Importance of clean data using Gradient Boosting a.

A bar graph of feature importance for cleaned data has been displayed in figures 4 (a) and (b). These are done on various included attributes of crime.

5. Conclusion

Enthusiasm aimed at happening by this research work is towards assistance an assisting hand towards young researchers who are execution their research into crime prediction and criminal analysis zones. paper concludes that Gradient Boosting algorithm (GBA) is able to predict more balanced results with respect to accuracy, precision, recall & F1 score out for prediction of 'Per Capita Violent Crimes' feature compared to that of Random Forest. Some common features having high importance scores that proved to be highly predictive of 'High crime' features are 'PctFam2Par': 'PctPersDenseHous': 'racepctblack': 'PctIlleg': 'NumIlleg': 'PctKids2par': using Random Forest Classifier model and Gradient boosting ensemble model.

These predicted features will be useful for the Police Department to utilize their resources efficiently and take appropriate actions to reduce criminal activities in the society. This can be used to enhance security and protection of criminal data by a desktop or a mobile application to track the crime rate and take any safety measures based on the relevant features. By maintaining dynamic databases with the criminal records across various countries, this technique can be implemented widely all over the world. The present dataset consists of all types of crimes; this type of analysis can be narrowed down to a single category of crime.

REFERENCES

Manish Gupta, B. Chandra and M.P. Gupta, "Crime Data Mining for Indian Police Information System", Journal of Crime, Vol. 2, No. 6, pp. 43-54, 2006.

P. Thongtae and S. Srisuk, "An Analysis of Data Mining Applications in Crime Domain", Proceedings of IEEE 8th International Conference on Computer and Information Technology Workshops, pp. 122-126, 2008.

Shyam Varan Nath, "Crime Data Mining", Proceedings of Advances and Innovations in Systems, Computing Sciences and Software Engineering, pp. 405-409, 2007.

Tong Wang et al., "Learning to Detect Patterns of Crime", Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 515-530, 2013.

Kamal Taha and Paul D. Yoo, "SIIMCO: A Forensic Investigation Tool for Identifying the Influential Members of a Criminal Organization", IEEE Transactions on Information Forensics and Security, Vol. 11, No. 4, pp. 811- 822, 2016.

Electronic copy available at: https://ssrn.com/abstract=3418733

Kevin Sheehy et al., “Evidence-based Analysis of Mentally 111 Individuals in the Criminal Justice System”, Proceedings of IEEE Systems and Information Engineering Design Symposium, pp. 250-254, 2016.

Mugdha Sharma, “Z-Crime: A Data Mining Tool for the Detection of Suspicious Criminal Activities based on the Decision Tree”, International Conference on Data Mining and Intelligent Computing, pp. 1-6, 2014.

Ehab Hamdy, Ammar Adl, Aboul Ella Hassanien, Osman Hegazy and Tai-Hoon Kim, “Criminal Act Detection and Identification Model”, Proceedings of 7 th International Conference on Advanced Communication and Networking, pp. 79-83, 2015.

Kaumalee Bogahawatte and Shalinda Adikari, “Intelligent Criminal Identification System”, Proceedings of 8th IEEE International Conference on Computer Science and Education, pp. 633-638, 2013.

Jyoti Agarwal, Renuka Nagpal and Rajni Sehgal, “Crime Analysis using K-Means Clustering”, International Journal of Computer Applications, Vol. 83, No. 4, pp. 1-4, 2013.

Rasoul Kiani, Siamak Mahdavi and Amin Keshavarzi, “Analysis and Prediction of Crimes by Clustering and Classification”, International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No. 8, pp. 11-17, 2015.

Shiju Sathyadevan, M.S. Devan and S. Surya Gangadharan, “Crime Analysis and Prediction using Data Mining”, Proceedings of IEEE 1st International Conference on Networks and Soft Computing, pp. 406-412, 2014.

Yu-Yueh Huang, Cheng-Te Li and Shyh-Kang Jeng, “Mining Location-based Social Networks for Criminal Activity Prediction”, Proceedings of 24th IEEE International Conference on Wireless and Optical Communication, pp. 185-190, 2015.

Electronic copy available at: https://ssrn.com/abstract=3418733