Prediction Of Violent-Crime Using Supervised and Unsupervised Learning Algorithms

Duba Sriveni

Research Scholar,VTU
Sriveni.dubas@gmail.com

Dr. Loganathan R

Professor & HOD of CSE,HKBK engineering college
drloganathanr@gmail.com

Abstract

The prediction of violent-crime occurrences such as rape, murder, kidnapping, robbery etc..is very important to avoid the negative consequences like injuries, deaths ,social and economic loss. If this prediction is done accurately many lives could be saved, which can be done primarily by efficient police patrolling.

This paper surveys various prediction methodologies adopted by each researcher in crime prediction using machine learning techniques, also discusses the advantages and disadvantages, various crime prediction methods, put forth the new insights to improve the efficiency .

Introduction

Predictive policing can be effectively forecasted using analytical tools[20]. The following table summarizes various analytics and associated prediction techniques as proposed by www.rand.org[20]

Analytic Category	Predictive Techniques
Hot spot analysis (use crime data)	i. Grid mapping ii. Covering ellipses iii. Kernel density
Regression methods (using a range of data)	i.Linear ii. Stepwise iii. Splines iv. Leading indicators
Data mining (using a range of data)	i.Clustering ii.Classification
Near-repeat (over next few days, using crime data only)	i. Self-exciting point process ii. ProMap
Spatiotemporal analysis (using crime and temporal data)	i.Heat maps ii. Additive model iii.Seasonality
Risk terrain analysis (using geography associated with risk)	i.Geospatial predictive analysis ii. Risk terrain modeling

Table1:Classes of predictive Techniques

Most of the reserchers have used structured data[1][6] on closed areas, such as population, race, income, and education from multiple datasets. Some of them have considered environment context information[21].

As image data has an unstructured data format, conventional strategies, cannot deal with image data. Also, these methods treat multiple datasets equally. These methods result in limitations in predicting crime occurrences because of nonlinear relationships, redundancies, and data dependencies. In other words, to accurately predict crime occurrences and to enhance the accuracy of crime prediction models, it is necessary to use multi-modal data according to deep learning along with environmental context information[21]. To solve this problem, Kang H-W, Kang H-B (2017) employed a deep learning model.

The machine learning algorithms with the help of big data can process structured or unstructured data. A machine learning algorithm can be a supervised[1][6] or unsupervised[19]. The general process followed while predicting using machine learning algorithm is:

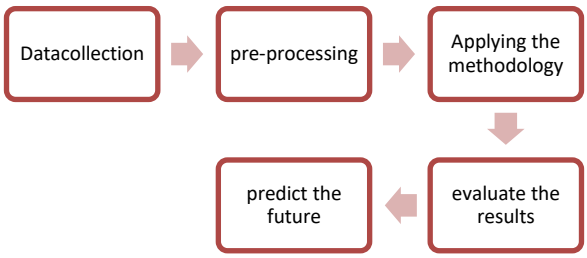

Fig:1 Steps follwed in a machine learning algorithm

Methodologies adopted:

1) Using Linear Regression to Forecast Future Trends in Crime of Bangladesh

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

In this paper[1] Md. Abdul Awal *et.al* have used Bangladesh crime dataset containing 840 instances with 3 predictive features: region ,month , year and one goal feature: predicted value of different types of crime .Dataset is divided into Metropolitan region data and divisional region dataset. Each region is represented using a number.

The model used is Linear Regression Model with a learning parameter of 0.01 and gradient descent was iterated for 400 times. Crimes forecasted for the year 2016 are dacoit,robbery,murder,women and child repression, Kidnapping etc..Results are also divided into Metropolitan region result and divisional region result.The results have shown that most of the crimes are increasing with the increase in population.

This method effectively forecasted the future trends of Bangladesh taking 3 parameters only into account but didn't focus to forecast the location of crime occurance.

2) Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots

In this paper[6] Tahani Almanie et.al have used datasets of Denver in Colorado[8] taken from National Incident Based Reporting System (NIBRS) containing 333068 instances, Los Angeles in California containing 243750 instances and also Denver Neighborhood Demographics Dataset[7] with 78 instances and 127 attributes out of which gender, race, age, family size, housing units, number of occupied and vacant units, and number of rental and owned units are of interest.

Predictive features:type of crime,the occurance time and the crime location,month,day.

Under preprocessing

- **data cleaning** for missing values
- **dimensionality reduction** using attribute subset selection
- **data integration** using uniform key attribute name ,considering military time system hour part was separated
- **data transformation and Discretization**- mapping attributes to 6 new groups

Algorithms used:

Apriori – mining frequent crime patterns using an open source tool provided by github[9], with minimum support value of 0.0012 and 0.0018 for Denver ,Los angeles datasets respectively. Also used constraint-based mining with three specific itemsets-Location,Day,Time

Naive Bayesian classifier – A supervised learning algorithm for classification. They have used Multinomial Naïve Bayes to support categorical data .80% of the data was taken as training data and 20% as test data. Same model was trained to work with both datasets.

Decision tree classifier- A supervised learning algorithm for classification. Entropy function is used for information gain with time attribute as the tree node.

Both classifiers are implemented using Scikit-Learn[10] an open source tool for python. A 5-fold cross validation strategy is used to compare the prediction accuracy of each city.

As a result this paper using Apriori algorithm ,spatial and Temporal hotspots were founded successfully. Bayes classifier was selected as appropriate classifier .Decision tree classifier ran successfully but was complex.

More classification models can be applied to increase crime prediction accuracy and to improve the overall performance. Also the income information for neighborhoods can be considered ,to check the relationships between neighborhoods income level and their crime rate, analysis of Los Angeles demographics information can be considered to find its crime pattern.

3) Using Machine Learning Algorithms To Analyze Crime

In this paper[11] Lawrence McClendon *et.al.* have used dataset from FBI UCR containing 2215 instances to focus on violent crime , the features that were analyzed are the murders, murdPerPop, rapes, rapesPerPop, robberies, robbbPerPop, assaults, assaultPerPop, and ViolentCrimesPerPop. WEKA an open source software[17] is used to conduct a comparative study between the violent crime patterns from the Communities and Crime Unnormalized Dataset provided by the University of California-Irvine repository and actual crime statistical data for the state of Mississippi that has been provided by neighborhoodscout.com.

Algorithms used:

Linear Regression- The algorithm uses linear regression for prediction

Additive Regression- To enhance the performance of a regression base classifier. The output of one

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

iteration is given as input to the next iteration and finally all outputs are summed up.

Decision Stump algorithms - The decision stump is basically a decision tree, however, with a single layer .stump stops after the first split. They are typically used in population segmentation for large data and in smaller datasets to aid in making decisions in simple yes/no models .

WEKA outputs five metrics :Correlation coefficient, Mean absolute error, Root mean squared error, Relative absolute error, and the Root relative squared error which evaluate the effectiveness and efficiency of the algorithms.

The linear regression algorithm performed the best among the three selected algorithms.

The Decision Stump Algorithm can be replaced with a random forest for better performance.

4) Prediction of crime occurrence from multimodal data Using Deep Learning

In this paper[19] Hyeon-Woo Kang *et.al.* . proposed a feature-level data fusion method with environmental context based on a deep neural network (DNN). Their dataset consists of data from seven domains: crime occurrence reports, demographic, housing, economic, education (http://factfinder.census.gov) , weather , and image data by using the Weather Underground API (https:// www.wunderground.com/) and the Google Street View Image API (https://developers. google.com/maps/documentation/streetview/) respectively . Finally they trained their DNN, which consists of the following four kinds of layers: spatial, temporal, environmental context, and joint feature representation layers. We trained the DNN by using the deep learning framework Caffe [21]. They set the batch size, initial learning rate, and dropout rate to 256, 0.01, and 0.5, respectively.

Algorithms Used:

regression analysis,kernel density estimation (KDE) , support vector machine (SVM)

statistical analysis software package SPSS 18.0 is used to conduct the Pearson correlation coefficient analysis with a correlation coefficient in the range from -0.2 to 0.2 and with a p-value greater than 0.05

To capture the environmental context information **Alex net** is used.

To analyze the difference in the number of crime incidents according to environmental context information, we conducted a **Kruskal-Wallis H test** (also known as a "**one-way ANOVA on ranks**"), Dividing the environmental context information into ten groups and using k-means clustering to conduct the Kruskal-Wallis H test and Dunn's test with Bonferroni-type adjustment of p-values for a post hoc test after the Kruskal-Wallis H test is prformed. The pairwise multiple comparisons of mean rank sums (PMCMR) package in the R software package were adopted.

All the results obtained using this method are higher than the corresponding values produced by the traditional methods.

The limitation of this study is that this method cant run with insufficient data. Furthermore, this method is unable to provide information such as specific crime type at a given time slot.

Conclusion:

After analyzing all the supervised and unsupervised machine learning techniques it is observed that deep learning outperformed all the supervised learning techniques as it can work with both structured and unstructured data. Furthermore various auto encoders like denoising auto encoders,Variational auto encoders can be adopted for data cleaning ,predicting and forecasting ,Dimensionality reduction etc.

References

[1] Md. Abdul Awal, Jakaria Rabbi, *et.al*. Using Linear Regression to Forecast Future
Trends in Crime of Bangladesh, 5th International Conference on Informatics, Electronics and Vision (ICIEV).2016.

[2]http://www.dmp.gov.bd/application/index/page/crimedata

[3]http://www.police.gov.bd/Crime-Statistics-yearly.php?id=337

[4] A. Malathi, S. S. Baboo, "Enhanced Algorithms to Identify Change in Crime Patterns", International Journal of Combinatorial Optimization Problems and Informatics, Aztec Dragon Academic Publishing, vol. 2, no.3, pp. 32-38, 2011.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

[5] J. R. Zipkin, M. B. Short, and A. L. Bertozzi, "Cops on the dots in a mathematical model of urban crime and police response", Disc Cont Dyn Syst, B 19 (2014).

[6] Tahani Almanie, Rsha Mirza and Elizabeth Lor, "CRIME PREDICTION BASED ON CRIME TYPES AND USING SPATIAL AND TEMPORAL CRIMINAL HOTSPOTS" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5,No.4, July 2015

[7] Imgh.us, 2015. [Online]. Available: http://imgh.us/neighborhood_map.jpg. [Accessed: 20- May2015].

[8] Data.denvergov.org, 'Denver Open Data Catalog: Census Neighborhood Demographics (2010)', 2015. [Online]. Available: http://data.denvergov.org/dataset/city-and-county-of-denver-censusneighborhood-demographics-2010. [Accessed: 20- May- 2015].

[9] GitHub, 'asaini/Apriori', 2015. [Online]. Available: https://github.com/asaini/Apriori. [Accessed: 20- May- 2015].

[10] Scikit-learn.org, '3.3. Model evaluation: quantifying the quality of predictions — scikit-learn 0.17.dev0 documentation', 2015. [Online]. Available: http://scikit-learn.org/dev/modules/model_evaluation.html. [Accessed: 20- May- 2015].

[11] Lawrence McClendon and Natarajan Meghanathan, "USING MACHINE LEARNING ALGORITHMS TO ANALYZE CRIME DATA" Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.1, March 2015

[12] Violent Crime. http://www2.fbi.gov/ucr/cius2009/offenses/violent_crime/.

[13]Murder. http://www2.fbi.gov/ucr/cius2009/offenses/violent_crime/murder_homicide.html.

[14]Forcible Rape. http://www2.fbi.gov/ucr/cius2009/offenses/violent_crime/forcible_rape.html.

[15]Robbery. http://www2.fbi.gov/ucr/cius2009/offenses/violent_crime/robbery.html

[16]Assault. http://www2.fbi.gov/ucr/cius2009/offenses/violent_crime/aggravated_assault.html.

[17] Mississippi Crime Rates and Statistics - NeighborhoodScout. Mississippi Crime Rates and Statistics - NeighborhoodScout. Accessed February 17, 2015. http://www.neighborhoodscout.com/ms/crime/.

[18] Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/. Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.1, March 201

[19] Hyeon-Woo Kang, Hang-Bong Kang "Prediction of crime occurrence from multimodal data using deep learning", PLoS ONE 12(4): e0176244. 2017.

[20] Elizabeth R. Groff, Nancy G. La Vigne, Forecasting the future of predictive crime mapping, Crime Prevention Studies , volume 13, pp.29-57.2013.

[21] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia. ACM; 2014. p. 675–678.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731