Machine Learning Approaches For Detect Crime Patterns

Research Proposal · September 2017

CITATION
1

READS
1,967

1 author:

Nisal Waduge
University of Moratuwa
8 PUBLICATIONS **3** CITATIONS

SEE PROFILE

Machine Learning Approaches For Detect Crime Patterns

N.D. Waduge (144186R)
Faculty of Information Technology
University of Moratuwa, Sri Lanka
nisalwaduge@gmail.com

Abstract--**Crimes are one of the major threat to the society and also for the civilization. The traditional crime solving techniques unable to live up to the requirement of existing crime scenario. Most challenging area in this crimes is identifying the sets of crimes committed by the same individual or same group. Criminal are humans. Then they tend to do the same work in the same way. Surveys states that 50% of the crimes done by the 10% of criminals. In this scenario Machine Learning can be used to identify the patterns of crimes. The data to feed this Machine Learning approach can be taken from past crime records, social media sentiment analysis, weather data etc. There are five steps in crime prediction by using Machine Learning. Those are data collection, classification of that data, identification of patterns, prediction of events and visualization. The limited resources in law enforcement authorities can be used effectively by using crime prediction methods. Especially data collected through the social media shows that prediction of a crime can be done using Machine Learning and it shows reasonable percentage of hits.**

***Index Terms*—Modus Operandi, Machine Learning, Supervised Learning**

I. Introduction

According to the crime analysis there are four major types of crimes. They are "murder, forcible rape, robbery and aggravated assault . Apart from these major areas there are drug offences, traffic violations, damaging to private and public properties. According to the time based analysis there are two types of crime. Which are accidental crimes and planned crimes.

Studies on these planned crimes observed that there are patterns of happening those kind of crimes on specific geographical areas. Persons who intended to do crimes choose seclude places for committing the crime where the police and other law enforcement patrols are less. Because of this reason there is a higher probability of predicting the crimes which going to happen in the future.

However, earlier days the data about crimes are based on the police reports, newspapers, reports and articles. Those all were in hand written hard copy format, but nowadays police and other law enforcement authorities are maintaining a soft copy of these data along with the hard copy. Those soft data generations in these days are very high. These data are just not a bulk of data but valuable information that can be used to predict upcoming crimes and solve exiting. Criminals are human beings, since they are trying to repeatedly do the same thing. They will use same locales and high crime risk areas repeatedly. This is called "Modus Operandi" (MO). Series of crime have some common attributes which can be used to characterize the modus operandi of the criminal.

Therefore, this paper attempts to identify a machine learning approach to predict the crime series patterns by using past data related to the crimes. Identify the geographical 'hotspots' where a crime can happen frequently.

The rest of the paper organized as follows. Section 2 describes the related work. Section 3 provides a description about methodology while Section 3 gives a conclusion and Section 4 is the acknowledgment.

II. Related work

In Los Angeles police department officers are currently using crime prediction system called predpol[1]. Also the Cambridge police department done a research called "Series Finder". There are significant number of researches done about crime prediction using "Modus Operandi" (Method of Operation) [2]. Some researches show that in the data gathering process, it can be used the social medias [2]. Research done by the University of Alabama used images from social medias to identify most wanted criminals,

in most of the cases the target was to capture terrorists. [3] Also the image processing has been used with CNN (Convolutional Neural Networks). Furthermore, the Surrey Police Department proposed an image processing based crime analysis program [4]. There the researches used the images of criminals' shoe prints to shortlist criminals. The motivation for this research was, accordingly to the statistical data there was above 30% average about criminals left their foot print at the crime scene [5]. Detection of crime patterns by using Artificial Intelligence also have focused into Financial Areas [6]. For financial frauds that shows the fraud documentations has been used in significant number [7]. Some researchers shown that the image processing can be used to analyze the fake signatures, handwritings etc. Research done by Easy Solutions Research shows that it can be used DNS (Domain Name Server) and HTML content analysis for track down the usual phishers of financial institutions. Beihang University also have conducted a research to identify fraud detections which are showing a sequential pattern [8]. The data gathering for this purpose has done using CDR (Call Detail Records). Most of these crime prediction methods showed considerable success rate in crime predictions.

III. Methodology

A. Data Collection Methods

The data collection has been done for crime prediction by using two major data sources. Those are

- Crime data records from law enforcement authorities
- Social media analysis
- IoT devices
- Newspaper articles
- Shoe prints databases
- CRDs (Call Data Records)

The largest data set maintained by a law enforcement authority is "Crime data 2001 to present" [9] by Chicago Police department. There are significant number of Crime Prediction systems tested in Chicago city. [6]
In Sri Lanka these data can be collected from Criminal Records Office (CRO), but those are not in digital format. Because of that reason it is mandatory to convert those written files to digital format before use.

Other major way of collection of data to crime prediction are social medias. Most of the researches has been used Twitter [10] [11] [2] because of the flexibleness of the APIs. They have used users tweets for a selected geographical area for predict crimes. In a research from South Africa, they have used Twitter (tweets) to collect data of crimes in South Africa. The tweets' data about crimes in specific geographical areas being recorded. According to those records they have determined that which specific geographical areas are famous for what specific crime pattern.

There was a one research conducted about how weather can be a reason and how weather analysis can be used to forecast crimes. [11] In this research they sorted out that in hotter days there is an increment of occurring crimes. Following set of constraints has been checked to the relevancy for crimes.

- Minimum, Maximum and Mean temperatures in Fahrenheit
- Humidity
- Sea level pressure
- Visibility per Miles
- Wind speed in Mph
- Cloud Cover
- Events (Sunny, Rain, Snow, Fog)

Main data source of this research was tweets from Chicago region and those tweets are embedded with GPS coordinates. In addition, they have used crime records of Chicago, which were maintained by Chicago Police Station. These data have been combined with weather data to determine whether there is a relationship weather conditions and crimes and the results shown that there is a significant effect from the weather conditions to occur a crime.

In a research conducted from University of Colombo, Sri Lanka about this crime data prediction, they have mentioned that "Modus Operandi" in other words Method of Operation can be used to predict crimes [12]. Criminals are creatures of habit, repeatedly using the same locales for committing crimes, or are attracted to certain high crime risk areas [13]. In this research they have defined MO as, is the set of habits that the offender follows, and is a type of motif used to characterize the pattern. As more crimes are added to the set, the M.O. becomes more well-defined. Following 21 characteristics have been extracted as MO table to their research.

MO defining attributes	MO supportive attributes	Identification attributes
• Characteristics of crime • Did use force • Method of entry • Method of exit • Property attacked • Place attacked • Date and Time (From-To) • Weapons used	• Age of criminal • Partners involved • Location • State of the property • Entry Orientation • Exit orientation • Traveling Method • Characteristics of theft • Remarks of the criminal	• Name of the police station • Crime Number • Criminal Number

Figure 1: Modus Operandi Table[12]

When a crime happens most likely that will be published on a newspaper. The data retrival part from the text and newspapers was a complicated problem. As a slotion for this matter

reseachers have been used Web Crawlers as Cralwler4j[14]. This a modiafiable open source application. Currently all most all newspapers are also published in Digital Format. Then there is an ability to combine these digital files and crawlers to extract data from newspapers to determine crimes. Among lots of text classification algorithms supervised learning approach called SVM(Support Vector Machine) [15][16] has choosen as the optimal alogorithm to extract data. The extracted data then be used to do crime analysis while saving to an external database for further use. The following component digram describes a high level view how newspaper data converted to crime analysis data.

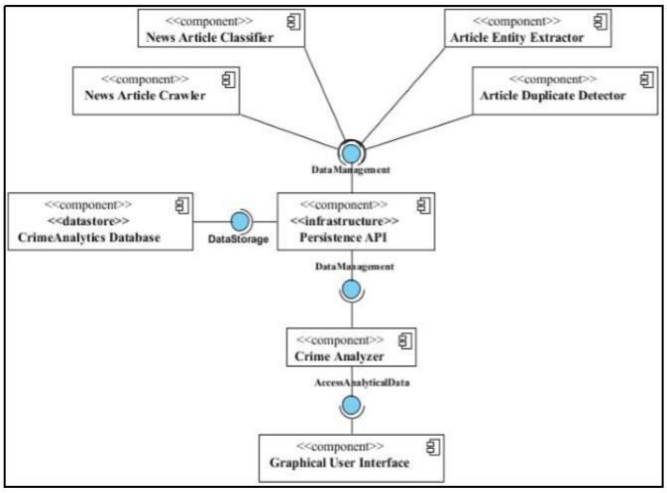

Figure 2: Component Diagram for Newspaper Analysis[17] Using web baed data is one other huge data source used by the researchers. In a research conducted about financial institution based phishing [6], have used HTML structures and DNS (Domain Name Server) RRSets (Registration Records). Researchers were able to understand who are the phishers target the financial institution and what is their method of operation (Modus Oprenadi) by feeding above data in to machine learning pragram. The following diagram depicts the framework for above scenario.

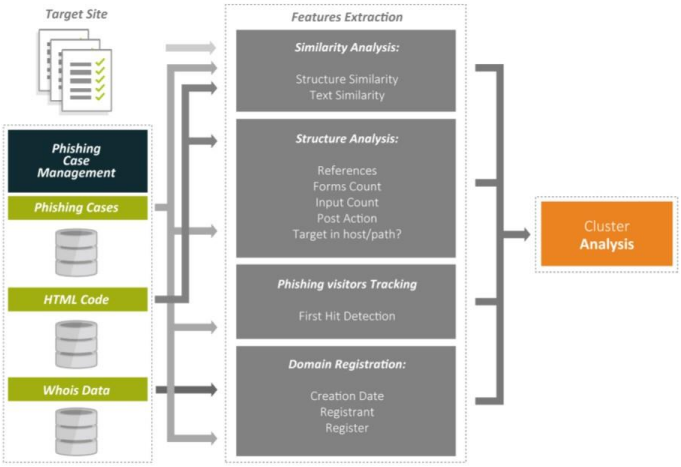

Figure 3: Newspaper Data Extracion Framework[17]

B. Classification and Pattern Recognition

In the above researches Clustering Algorithms has been used to check whether it is possible to short list crimes using MO. Most of the researches including the above one implemented with 'K-Means' clustering [18]. K-Means has used in most of crime the crime pattern detecting researches and is one of the best clustering approaches[19]. Some of the researches used supervised learning methods, while some used un-supervised learning methods. "Apriori Algorithm", Decision Trees has been used in a research done by "Amrita Center for Cyber Security" [20] by supervised learning method. Apriori can be used to determine association rules which highlight general trends in the database. [21]
Following are some of the clustering algorithms which can be used in predicting the patterns of criminals.

1) Hierarchical Clustering Algorithm

In this algorithms it implements a hierarchical based clustering solution. There are two main approaches which are Agglomerative and Divisive. This Agglomerative is a 'top-down' approach while the Divisive is a 'bottom-up'.

2) Squared Error Based Algorithm

Statistical function which uses average of squares of deviations or errors. The K-Means Algorithm is a subset of this.

3) Neural Network Based

According to the researches up to now, it depicts that this Artificial Neural Network (ANN) non-algorithmic approach will give much sophisticated answer for crime predictions. [22] The main advantage of using ANN is, the training to crime regarding data capability and also self-organization of data. After the ANN is fed, trained with a proper model of old crimes it will be able to identify the patterns of the crimes and predict the upcoming ones. ANN was the first implication about Artificial Intelligence which is proposed in 1943 and even today it is widely used to generate information in vast areas of industries.

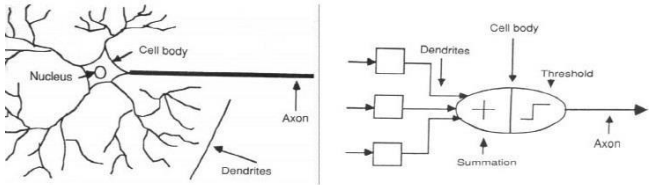

Figure 4: Neuron Vs Model of Nueron[23]

4) Fuzzy Algorithm

This algorithm helps to determine the best choice when there are multiple options available. Can be used to filter the outputs given by a crime prediction solution to identify the most accurate answer.

5) K-Means Clustering Algorithm

According to the researches in this crime prediction criteria is shows that most of them have used this 'K-Means' algorithm to cluster the data [24] . This paper will discuss about 'K-Means' algorithm to some extent. In 'K-Means' algorithm 'K' stands for number of clusters initialized to clustering process. These clusters' data points have to in numerical format. Categorical data cannot be clustered by K-Means Algorithm.

Following list depicts the information about K-Means Algorithm

J =Objective function

K =Number of c

N =Number of cases

Cj =centroid of cluster 'j'

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

distance $= |x_i - c_j|$ (1a)

This distance is the gap between a data point and a centroid. For each and every data point (x_i) a nearest cluster has to be found. Then assign that point to that cluster.

For each cluster, the centroid has to be re-initialized according to an attribute value of 'x_i'. This attribute value has to be a numerical value. The following function which generates the new centroid gets the average of attributes of data points inside that a specific cluster.

C(S) =New centroid

$$\vec{C}(S) = \sum_{i=1}^{n} \vec{X}_i / n \quad (2)$$

If the original attribute is categorical, those has to be converted to numerical using suitable method.

This process has to be continued until a cluster don't include a x_i from another cluster, because of that multiple iteration will be executed. If that argument is valid, the clustering process has been completed. The following figure show a data set clustering to two clusters. Step 1 depicts two clusters using boundary while Step 2 shows the re-arrangement of centroid, and other steps show how the clustering process get completed.

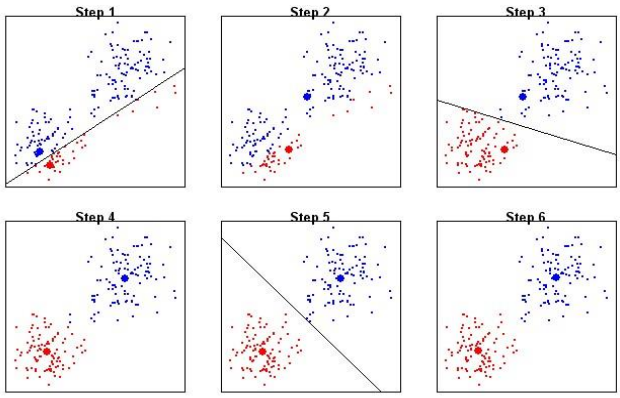

Figure 5: Steps of Clustering by K-Means[25]

Step 5 it can be identified that no blue color data point is on red color cluster. This means the inserted data has been successfully clustered around two centroids.

Also some researches conducted in supervised learning methods by using "Naïve Bayes" classifier and "Back Propagation Algorithm". Back Propagation Algorithm is a model of reasoning based on the human brain [26].

IV. DISCUSSION

This paper discuses about crime prediction systems which are using machine learning approaches and those systems' methods of execution the crime prediction process. Algorithms used to predict events in machine learning such as 'K-Means', Neural Network based, Squared Error Based, Hierarchical Clustering, Fuzzy etc. From these used data analysis algorithms "K-Means" has widely used to cluster the crime data. This happened because of the reliability and relativity to the criminal data. But K-Means won't give productive result while dealing with noisy data, also when the number of clusters are higher the algorithm won't show a better average. Because of that reason using of K-Means is more suitable when the data is not noisy and the number of clusters are less. The following table shows the average of accuracy and the clusters used.{Citation}

Number of Clusters	K Means
20	62.439544
40	53.969942
60	53.643471
80	49.760920
100	47.628894

Figure 5: K-Means Avg. vs. Number of Clusters [12]

The basic data collection methods are social media analysis, weather analysis, IoT devices' data analysis, newspaper articles

about crimes etc. The discussed area has got widen because even the image processing has been involved with the crime pattern detection systems.

V. Conclusion

Crimes has to be stopped or limited for a better civilization in a region if it is possible. Apart from limiting old crime data to criminal records, those has to be used to get better idea about the crimes and criminals. Also the day today activities of people may link to crime incidents, because of that reason analyze of these data has to be done, but because of the excessive number of data are there to process a human being may be unable to clearly analyze them. That is where the Machine Learning, Artificial Intelligence and Data Mining Techniques has to be used for an efficient data gathering and analyzing process. Chicago city has done number of above kind of researches to predict crimes because of the excessive number of crimes happening in there and Chicago city is successful in implementing these systems. Not only the Chicago city according to the results all around the world of researches done regarding to crime prediction using Machine Learning shows higher success rate often over than 50%. Because of these evidence it can be seen that Machine Learning can be used to predict crimes. Since the systems already using these crime prediction systems shows significant results, which gives a motivation to enhance these crime predictions not by sticking to one specific algorithm or data collection method but a better combination of those.

VI. Acknowledgement

The author of this paper would like acknowledge with gratitude to the supervisor Dr. Lochandaka Ranathunga, Head of the Department of Information Technology for his guidance and valuable advices to make this independent study a success.

References

[1] "Predict Crime | Predictive Policing Software," *PredPol*. [Online]. Available: http://www.predpol.com/. [Accessed: 22-Jun-2017].

[2] S. Aghababaei and M. Makrehchi, "Mining Social Media Content for Crime Prediction," in *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*, 2016, pp. 526–531.

[3] P. Chitrakar, C. Zhang, G. Warner, and X. Liao, "Social Media Image Retrieval Using Distilled Convolutional Neural Network for Suspicious e-Crime and Terrorist Account Detection," in *Multimedia (ISM), 2016 IEEE International Symposium on*, 2016, pp. 493–498.

[4] N. E. Sawyer and C. W. Monckton, "'Shoe-fit'-a computerised shoe print database," 1995.

[5] A. Alexander, A. Bouridane, and D. Crookes, "Automatic classification and recognition of shoeprints," 1999. [6] R. Kumar, N. R. Pal, B. Chanda, and J. D. Sharma, "Detection of fraudulent alterations in ball-point pen strokes using support vector machines," in *India Conference (INDICON), 2009 Annual IEEE*, 2009, pp. 1–4.

[7] J. Vargas, A. C. Bahnsen, S. Villegas, and D. Ingevaldson, "Knowing your enemies: leveraging data analysis to expose phishing patterns against a major US financial institution," in *Electronic Crime Research (eCrime), 2016 APWG Symposium on*, 2016, pp. 1–10.

[8] Y. Yu, X. Wan, G. Liu, H. Li, P. Li, and H. Lin, "A combinatorial clustering method for sequential fraud detection," in *Service Systems and Service Management (ICSSSM), 2017 International Conference on*, 2017, pp. 1–6.

[9] "Crimes - 2001 to present - Data.gov." [Online]. Available: https://catalog.data.gov/dataset/crimes-2001-topresent-398a4. [Accessed: 24-Jun-2017].

[10] M. Wang and M. S. Gerber, "Using Twitter for Next-Place Prediction, with an Application to Crime Prediction," 2015, pp. 941–948.

[11] X. Chen, Y. Cho, and S. Y. Jang, "Crime prediction using Twitter sentiment and weather," in *Systems and Information Engineering Design Symposium (SIEDS), 2015*, 2015, pp. 63–68.

[12] M. Munasinghe, H. Perera, S. Udeshini, and R. Weerasinghe, "Machine Learning based criminal short listing using Modus Operandi features," 2015, pp. 69–76.

[13] W. Gorr, A. Olligschlaeger, and Y. Thompson, "Shortterm forecasting of crime," *Int. J. Forecast.*, vol. 19, no. 4, pp. 579–594, 2003.

[14] "crawler4j/README.md at master · yasserg/crawler4j · GitHub." [Online]. Available: https://github.com/yasserg/crawler4j/blob/master/README.md. [Accessed: 22-Aug-2017].

[15] Z. Weifa, "A SVM Text Classification Approch Based on Binary Tree," 2009, pp. 455–458.

[16] L. Youwen, X. Shixiong, and Z. Yong, "A Supervised Local Linear Embedding Based SVM Text Classification Algorithm," 2009, pp. 21–26.

[17] I. Jayaweera, C. Sajeewa, S. Liyanage, T. Wijewardane, I. Perera, and A. Wijayasiri, “Crime analytics: Analysis of crimes through newspaper articles,” in *Moratuwa Engineering Research Conference (MERCon), 2015*, 2015, pp. 277–282.

[18] T. Aljrees, D. Shi, D. Windridge, and W. Wong, “Criminal pattern identification based on modified K-means clustering,” in *Machine Learning and Cybernetics (ICMLC), 2016 International Conference on*, 2016, vol. 2, pp. 799–806.

[19] X. Zheng, Y. Cao, and Z. Ma, “A mathematical modeling approach for geographical profiling and crime prediction,” in *Software Engineering and Service Science (ICSESS), 2011 IEEE 2nd International Conference on*, 2011, pp. 500–503.

[20] S. Sathyadevan, S. Gangadharan, and others, “Crime analysis and prediction using data mining,” in *Networks & Soft Computing (ICNSC), 2014 First International Conference on*, 2014, pp. 406–412.

[21] T. Chauhan and R. Aluvalu, “Using Big Data Analytics for developing Crime Predictive Model.”

[22] J. Azeez and D. J. Aravindhar, “Hybrid approach to crime prediction using deep learning,” in *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, 2015, pp. 1701–1710.

[23] A. Ghazvini, M. Z. B. A. Nazri, S. N. H. S. Abdullah, M. N. Junoh, and Z. A. bin Kasim, “Biography commercial serial crime analysis using enhanced dynamic neural network,” in *Soft Computing and Pattern Recognition (SoCPaR), 2015 7th International Conference of*, 2015, pp. 334–339.

[24] W. Wu and M. Peng, “A Data Mining Approach Combining K -Means Clustering With Bagging Neural Network for Short-Term Wind Power Forecasting,” *IEEE Internet Things J.*, vol. 4, no. 4, pp. 979–986, Aug. 2017.

[25] S. McElwee, “Active learning intrusion detection using kmeans clustering selection,” in *SoutheastCon, 2017*, 2017, pp. 1–7.

[26] A. Babakura, M. N. Sulaiman, and M. A. Yusuf, “Improved method of classification algorithms for crime prediction,” in *Biometrics and Security Technologies (ISBAST), 2014 International Symposium on*, 2014, pp. 250–255.