Data Mining: An Active Solution for Crime Investigation

[1]Uddin, Osemengbe O., [2]P. S. O. Uddin
[1]Dept. of Computer Science, Ambrose Alli University, Ekpoma-Edo State, Nigeria
[2]Dept. of Vocational and Technical Education, Ambrose Alli University,
Ekpoma-Edo State, Nigeria

Abstract

The purpose of this paper is to suggest ways by which law enforcement and intelligence agencies can analyze large volume of data using data mining as one of the ways of getting active solutions for crime investigation in Nigeria. The problem is, the general public are extremely concerned on how the activities of the terrorist in Nigeria can be reduced if it cannot be completely eliminated. The results of this paper can be significantly useful to the National Security Advisor's office (NASAO), State Security Service (SSS), Nigerian Police Force (NPF), Nigeria Army, including the Air Force (NAF) and Navy, Immigration, Customs, Economic and Financial Crime Commission (EFFC), Independent Corrupt Practices Commission (ICPC), and the general public. Data mining will help these stakeholders to review various data mining algorithms and then design a framework that would be automated to trigger alarm for timely solution for prevention, arrest and investigation of crimes.

Keywords

Data Mining, Active Solution, Crime, Investigation

I. Introduction

The growing insecurity challenges in Nigeria are of great concern to everyone and every effort must be employed to combat these security threats. Using the proposed data mining profiler model, our work distinguishes between information related threats and non-information related security threats. Information related threats are essentially attacks on computers and networks. That is, they are threats that damage electronic information. Non-information related terrorist threats include terrorist attacks, bombing, shooting and killing someone, vandalism, kidnapping, setting property on fire, etc. The questions asked by all stakeholders are: can the security agencies and their strategies fight the non-information related security threats in Nigeria? Do these agencies have appropriate Information Technology Infrastructure in place for the purpose of information gathering, sharing, dissemination, and decision making? Do they have adequate surveillance systems/ equipment? These are some of the issues this paper attempts to address.

The development and stability of any nation depends on the extent of security of lives and properties of the citizens. A secured atmosphere will encourage individual happiness, investments, bilateral relationships, intellectual minds, which are of great assets to nation building; it will also guarantee an environment for the growth of infrastructural development. Because the activities of terrorists, kidnappers and other high level crimes are a challenge to peaceful co-existence and development, this paper addresses these challenging threats by using appropriate data mining techniques to detect/prevent terrorism and at the same time maintain some level of privacy. The application will assist the security agencies to collect, manage, analyse, and predict certain patterns of an individual behaviour that could lead to timely arrest, prevention, and prosecution of the person.

II. Statement of the Problem

There had been an enormous increase of crime in recent past. The concern about national security has increased significantly since the 26/11 attacks at Mumbi, India and 9/11 terrorist attack on world trade centre in America. In Nigeria today, many terrorist activities like kidnapping, drug trafficking, Boko Haram, Emancipation of Niger Delta (MEND), etc have been unleasing terror to the Nigeria public. In particular, security agencies and the general public are extremely concerned in how these activities can be curtailed. This paper suggests ways by which law enforcement and intelligent agencies can analyze large volume of data using data mining as one of the ways of getting some solutions for crime investigation.

III. Significance of the Study

This paper will be of significance to local law enforcement agencies like National Crime Record Bureau (NCRB), State Crime Record Bureau (SCRB), District Crime Record Bureau (DCRB) and City Crime Record Bureau (CCRB). It will help these agencies to become more alert to criminal activities in their own jurisdictions. Another significance of this paper is that it will hold the promise of making it easy, convenient and practical to explore very large database for these agencies and other organizations that uses data mining.

This paper will also be of significance to the National Security Advisor's office (NASAO), State Security Service (SSS), Nigerian Police Force (NPF), Nigeria Army, including the Air Force (NAF) and Navy, Immigration, Customs, Economic and Financial Crime Commission (EFFC), Independent Corrupt Practices Commission (ICPC), and the general public. Local inputs/experiences from these stakeholders would assist the researchers identify the various terrorist activities, collect and analyse the characteristics/patterns of those activities, review various data mining algorithms, and then design a framework that would be automated to trigger alarm for timely prevention, arrest, and investigation of terrorists.

IV. Research Methodology

Data for this paper were derived from secondary sources of previous researches and analysis of scholars as well as journals, articles that are related to the subject of study

V. Concept of Data Mining

Data mining is defined as the discovery of interesting structure in data, where structure designates patterns, statistical or predictive models of the data, and relationships among parts of the data. Data mining is the framework of crime and intelligence analysis for national security. Data mining is basically used to find out unknown patterns from a large amount of data. There are popular tools of data mining to rub data mining algorithms. There are two approaches for the implementation of data mining, first is to copy data from data warehouse or source and mine it. Other approach is to mine the data within a data warehouse.

Data mining is defined as the identification of interesting structure in data, where structure designates patterns, statistical or predictive models of the data, and relationships among parts of the data

[10]. Data mining is the process of finding insights which are statistically reliable, unknown previously, and actionable from a large amount of data [9]. This data must be available, relevant, adequate, and clean for it to be used. To address a specific problem within a certain domain, the data mining problem must be well-defined, cannot be solved by mere query and reporting tools, and guided by a data mining mathematical framework [15]. Data mining techniques such as pattern recognition, machine learning, artificial intelligence, fuzzy logic, genetic algorithms, neural networks, expert systems, and other technologies have wide use in variety of applications, which include marketing, medicine, multimedia, finance, and recently in counter-terrorism applications [19]. Data mining can be used to detect security threats or fraudulent behaviour of individuals, terrorist activities, money laundering, ATM card cloning, and illegal transfer of money by individuals and corporate organisations. Though the use of data mining could sometimes violate individual's privacy and civil liberties, its benefits to humans and national development are enormous.

Data mining technologies have advanced a great deal. They are now being applied for many applications to discover previously unknown, valid patterns and relationships in large data set [17]. The main question is that, are data mining tools sufficient for detecting and/or preventing terrorist activities? For example, can they be used to completely eliminate false positives and false negatives? False positives could be disastrous for various individuals [4].

False positives are a universal problem as they affect both signature and anomaly-based intrusion-detection systems [2]. A high rate of false alerts [3] is the limiting factor for the performance of an intrusion-detection system. False negatives could increase terrorist activities. The work would address these challenges by identifying key law enforcement agencies and build application tools that can interface with stakeholders‟ systems to gather, analyse, and predict certain behavioural patterns of individuals, which have deviated significantly from normal behaviour and report to appropriate law enforcement agency to prevent, arrest, and investigate terrorist activities.

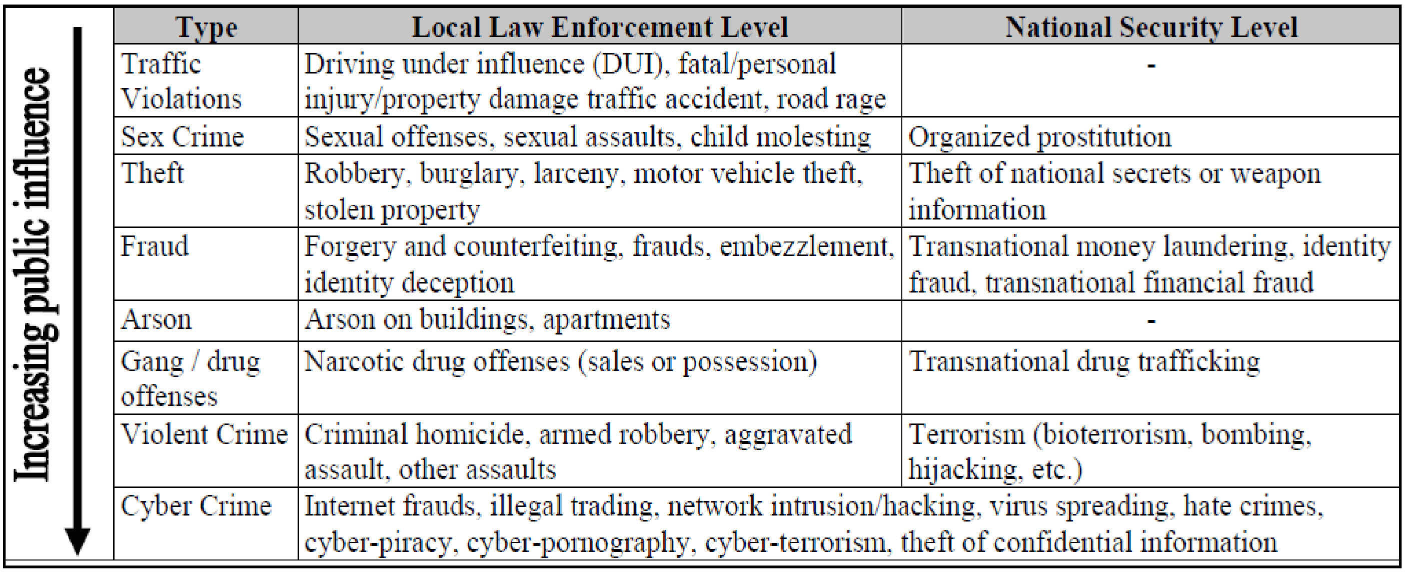
VI. Data Mining as an Active Solution for Crime Investigation

Data mining as an active solution for crime investigation is discussed under the following headings:

A. Data Mining Techniques for Detecting Crime

Crime is defined as "an act or the commission of an act that is forbidden, or the omission of a duty that is commanded by a public law and that makes the offender liable to punishment by that law" (Webster Dictionary). An act of crime encompasses a wide range of activities, ranging from simple violation of civic duties (e.g., illegal parking) to internationally organized crimes (e.g., the 9/11 attacks). Data mining in the context of crime and intelligence analysis for national security is still a young field. The following describes our applications of different techniques in crime data mining. Entity extraction has been used to automatically identify person, address, vehicle, narcotic drug, and personal properties from police narrative reports [5]. Clustering techniques such as "concept space" have been used to automatically associate different objects (such as persons, organizations, vehicles) in crime records [12]. Deviation detection has been applied in fraud detection, network intrusion detection, and other crime analyses that involve tracing abnormal activities. Classification has been used to detect email spamming and find authors who send out unsolicited emails [8]. String comparator has been used to detect deceptive information in criminal records [21]. Social network analysis has been used to analyze criminals' roles and associations among entities in a criminal network. Table 1 summarizes the different types of crimes in increasing degree of public influence. Note that both local and national law enforcement and security agencies are facing many similar challenges.

Table 1: Crime Type at Different Levels



	Type	Local Law Enforcement Level	National Security Level
Increasing public influence	Traffic Violations	Driving under influence (DUI), fatal/personal injury/property damage traffic accident, road rage	-
	Sex Crime	Sexual offenses, sexual assaults, child molesting	Organized prostitution
	Theft	Robbery, burglary, larceny, motor vehicle theft, stolen property	Theft of national secrets or weapon information
	Fraud	Forgery and counterfeiting, frauds, embezzlement, identity deception	Transnational money laundering, identity fraud, transnational financial fraud
	Arson	Arson on buildings, apartments	-
	Gang / drug offenses	Narcotic drug offenses (sales or possession)	Transnational drug trafficking
	Violent Crime	Criminal homicide, armed robbery, aggravated assault, other assaults	Terrorism (bioterrorism, bombing, hijacking, etc.)
	Cyber Crime	Internet frauds, illegal trading, network intrusion/hacking, virus spreading, hate crimes, cyber-piracy, cyber-pornography, cyber-terrorism, theft of confidential information	

B. Data Mining for Predicting Problems

Data mining (DM) is an attempt to answer the long-standing question "what does all this data mean?". Such investigations are inherently an attempt to automate and "predict problem" in database security. Predicting is basically the process of establishing relationships between data sets, the same objective as data mining. That is, given that certain attributes apply to a set of data, we "know" that certain other attributes also apply to that set of data. This is equivalent to stating that one set "implies" the other.

Now, in a multi-level secure (MLS) database, we do not want Low-classified data to infer High-classified data. Data mining processes cannot be used to compromise such rules, of course.

This is because each DM process must operate at a specified level (i.e. Low) and must have access to the High data in order to "discover" the rule. However, such Low-to-High rules may be "common knowledge" but unknown to the database designer. Data mining could then be used to combine Low information until the tail of the common-knowledge rule is derived. This is the process of predicting. Data are put together "in a surprising way" until some common-knowledge rule, relating Low and High data, can be applied.

Fortunately, data mining can be used effectively to enforce security. The most straightforward way is to search for rules relating Low and High data. We need not be concerned with chains of predictions, merely what conjunction of attributes for a High set may be implied by Lowclassified attributes for that set. The security officer doing this analysis has some advantages over an attacker, since he/she has access to both the High and Low data. In most systems, there is relatively little High data, so the number of rules relating High data to Low data is much fewer than the total number of possible rules.

C. Data Mining for Predicting Crime Trends Using Clustering in Weka

The first task is the prediction of the size of the population of a city [6]. The calculation of per capita crime statistics helps to put crime statistics into proportion. However, some of the records were missing one or more values. Worse yet, half the time, the missing value was the "city population size", which means there was no per capita statistics for the entire record. Over some of the cities did not report any population data for any of their records. To improve the calculation of "yearly average per capita crime rates", and to ensure the detection of all "per capita outliers", it was necessary to fill in the missing values. The basic approach to do this was to cluster population sizes, create classes from the clusters, and then classify records with unknown population sizes [3]. Why use clustering to create classes? Classes from clusters are more likely to represent the actual population size of the cities. The only value needed to cluster population sizes was the population size of each record. These values were clustered using "weka. clusterers. EM -I 100 -N 10 -M 1.0E-6 -S 100"

The next task is the prediction of future crime trends. This meant we tracked crime rate changes from one year to the next and used data mining to project those changes into the future. The basic method here is to cluster the cities having the same crime trend, and then using "next year" cluster information to classify records [11]. This is combined with the state poverty data to create a classifier that will predict future crime trends. Few "delta" attributes were applied to city crime clustering: Murder for gain, Dacoity, Prep.&Assembly For Dacoity, Robbery, Burglary, Theft, Murder, Attempt to commit murder, C.H.Not Amounting to murder, Hurt/ Grievous Hurt, Riots, Rape, Dowry Death, Molestation, Sexual Harassment, Kidnapping & Abduction of others, Criminal Breach of Trust, Arson, Cheating, Counterfeiting, and Others IPC crimes. These attributes were clustered using 'Weka 3.5.8's, Simple EM (expectation maximization)' with parameters of "EM -I 100 -N 4 -M 1.oE-6 -S 100" [4]. EM is a deviation of K-Means clustering. Four clusters were chosen because it produced a good distribution with a relatively easy to interpret set of clusters [5]. Usually, the high level interpretation of clusters from an unsupervised algorithm is not easily defined.

However, in this case, the four clusters produced had the following attributes: Note: The clusters are ordered from best to worst.

1. C0: Crime is steady or dropping. The Sexual Harassment rate is the primary crime in flux. There are lower incidences of: Murder for gain, Dacoity, Preparation for Dacoity, rape, Dowry Death and Culpable Homicide.
2. C1: Crime is rising or in flux. Riots, cheating, Counterfeit, and Cruelty by husband and relatives are the primary crime rates changing. There are lower incidences of: murder and kidnapping and abduction of others.
3. C2: Crime is generally increasing. Thefts are the primary crime on the rise with some increase in arson. There are lower incidences of the property crimes: burglary and theft.
4. C3: Few crimes are in flux. Murder, rape, and arson are in flux. There is less change in the property crimes: burglary, and theft. To demonstrate at least some characteristics of the clusters.

D. Detecting Criminal Identity Deceptions: An Algorithmic Approach

Criminals often provide police officers with deceptive identities to mislead police investigations, for example, using aliases, fabricated birth dates or addresses, etc. The large amount of data also prevents officers from examining inexact matches manually. Based on a case study on deceptive criminal identities recorded in the TPD, Hsinchun et al (2002) have built a taxonomy of criminal identity deceptions that consisted of name deceptions, address deceptions, date-of-birth deceptions, and identity number deceptions. They found criminals usually made minor changes to their real identity information. For example, one may give a name similarly spelled or, change the sequence of digits in his social security number. Based on the taxonomy, they developed an algorithmic approach to detect deceptive criminal identities automatically [21]. Their approach utilized four identity fields: name, address, date-of-birth, and socialsecurity-number and compared each corresponding field for a pair of criminal identity records. An overall disagreement value between the two records was computed by calculating the Euclidean Distance of disagreement measures over all attribute fields. A deception in this record pair will be noticed when the overall disagreement value exceeds a pre-determined threshold value, which is acquired during training processes. They conducted an experiment using a sample set of real criminal identity records from the TPD. The results showed that our algorithm could accurately detect 94% of criminal identity deceptions.

Authorship Analysis in Cybercrime

The large amount of cyber space activities and their anonymous nature make cybercrime investigation extremely difficult. Conventional ways to deal with this problem rely on a manual effort, which is largely limited by the sheer amount of messages and constantly changing author IDs. Hsinchun et al (2002) proposed an authorship analysis framework to automatically trace identities of cyber criminals through messages they post on the Internet. Under this framework, three types of message features, including style markers, structural features, and content-specific features, are extracted and inductive learning algorithms are used to build feature-based models to identify authorship of illegal messages. To evaluate the effectiveness of this framework, they conducted an experimental study on data sets of English and Chinese email and online newsgroup messages produced by a small number of authors. They tested three inductive learning algorithms: decision trees, backpropagation neural networks, and Support Vector Machines. Their experiments demonstrated that with a set of carefully selected features and an effective learning algorithm, they were able to identify the authors of Internet newsgroup and email messages with a reasonably high accuracy. They achieved average prediction accuracies of 80% - 90% for email messages, 90% - 97% for the newsgroup messages, and 70% - 85% for Chinese Bulletin Board System (BBS) messages. Significant performance improvement was observed when structural features were added on top of style markers. SVM outperformed the other two classifiers on all occasions. The experimental results indicated a promising future of using their framework to address the identity-tracing problem.

E. Data Mining and Knowledge Discovery

To x-ray the definition of data mining and Knowledge discovery database according to [10], they defined data mining as a process in the Knowledge Discovery Database (KDD) which is a nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Their views are diagrammatized in fig. 1 and show data mining as a continuous process; from a large dataset, valid data are selected, processed and transformed into a more useful dataset before data mining techniques are applied for valid patterns. Dissecting further the definition and concept of data mining according to [10], the following are evident:

Datasets: Data are set of facts (database) and pattern describes a subset of the dataset.

Model: Designates extracting and fitting a model to the data

Process: The fact that KDD and data mining comprise many processes

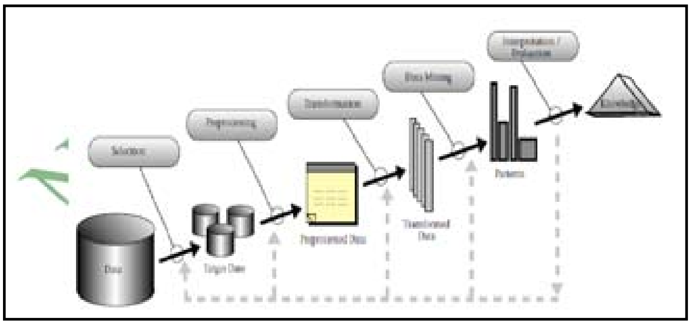


Fig. 1: The Process of Data Mining in the Knowledge Discovery Database

VII. Conclusion

Data mining applied in the context of law enforcement and intelligence analysis holds the promise of alleviating crime related problems. In this paper, data mining techniques has been discussed as a means of detecting crimes like sex crime, theft, fraud, arson, gang drug offences, violent crime and cyber crime. Data mining for predicting problems and crime trends using clustering in weka was extremely discussed. Other areas the paper covers center on authorship analysis in cybercrime and data mining in relation to knowledge discovery. It is hoped that the encouraging results from these discussions on data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis and give active solutions for crime investigation in Nigeria.

References

[1] Akpojaro J.; Onwudebelu U., Aigbe P.,"A Proposed Data Mining Profiler Model to Fight Security Threats in Nigeria", International Journal of Computer Applications & Information Technology Vol. 2, Issue 1, January, 2013.

[2] Axelsson S.,"Intrusion Detection Systems: A Survey and Taxonomy", Technical Report Chalmers University, pp. 99-15, March, 2000.

[3] Axelsson S."The base-rate fallacy and the difficulty of intrusion detection", ACM Trans. Inf. System Security (TISSEC), Vol. 3 No. 3, pp. 186–205, 2000b.

[4] Bolzoni D., Etalle S.,"APHRODITE: an Anomaly- based Architecture for False Positive Reduction", University of Twente, The Netherlands, 2005.

[5] Chau, M., Xu, J., Chen, H.,"Extracting meaningful entities from police narrative reports", In: Proceedings of the National Conference for Digital Government Research (dg.o 2002), Los Angeles, California, USA, 2002.

[6] Chen H., Chung W; Qin Y; Chau M; Xu J. J; Wang G.; Zheng R., Atabakhsh H.,"Crime Data Mining: An Overview and Case Studies", [Online] Available: http://www.ai.bpa.arizona.edu/

[7] "Classification via Decision Trees in weak", Depaul University, [Online] Available: http://maya.cs.depaul.edu/~classes/ect584/WEKA/classify.html

[8] De Vel, O., Anderson, A., Corney, M., Mohay, G.,"Mining E-mail Content for Author Identification Forensics", SIGMOD Record, 30(4), 55-64, 2001.

[9] Elkan C.,"Magical Thinking in Data Mining", Lessons from CoIL Challenge, Proceeding of SIGKDD01, pp. 426- 431, 2001.

[10] Fayyad, U.M., Uthurusamy, R.,"Evolving data mining into solutions for insights", Communications of the ACM, 45(8), 28-31, 2002.

[11] Frank Dellaert,"The expectation Maximization Algorithm, Technical Report", Georgia Institute of Technology, 2002.

[12] Hauck, R.V., Atabakhsh, H., Ongvasith, P., Gupta, H., Chen, H.,"Using Coplink to analyze criminal-justice data", IEEE Computer, 35(3), pp. 30-37, 2002.

[13] Malathi. A. S.; Santhosh Baboo, Anbarasi A.,"An intelligent Analysis of a City Crime Data Using Data Mining", International Conference on Information and Electronics Engineering IPCSIT, Vol. 6, IACSIT Press, Singapore, 2011.

[14] Manish Gupta,"Crime Data Mining for Indian Police Information System", Proceeding of the Computer Society of India, 2008.

[15] Motoda L. N., Fawcett H., Holte T., Langley R., Adriaans P.,"Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving", Machine Learning 57(1-2), pp. 13-34, 2004.

[16] Okonkwo R. O., Enem F. O.,"Combating Crime and Terrorism Using Data Mining Techniques", Nigeria Computer Society (NCS) 10th International Conference, July, 2010.

[17] Seifert J. W."Data Mining: An Overview", Congressional Research Service, CRS Report for Congress, Order Code RL31798, December, 2004.

[18] Stephen Schneider,"Predicting crime: A review of the research", Department of Justice Canada, 1-2, 2002.

[19] Thuraisingham B.,"Data Mining for Counter-Terrorism", The MITRE Corporation, Burlington Road, Bedford, MA, pp. 191-218, 2004.

[20] University Crime Data Mining, [Online] Available: http://www.cse.msu.edu/~kingstua/Team3/

[21] Wang, G., Chen, H., Atabakhsh, H.,"Automatically detecting deceptive criminal identities", Communications of the ACM (Accepted for publication, forthcoming), 2002.

[22] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. 7 Dec. 2008, [Online] Available: http://www.cs.waikato.ac.nz/ml/weka/