A Proposed Data Mining Profiler Model to Fight Security Threats in Nigeria

Jackson Akpojaro*
Department of Mathematics & Computer Science
Western Delta University
Oghara, Delta State, Nigeria

Ugochukwu Onwudebelu
Department of Mathematics & Computer Science
Federal University, Ndufu, Alike Ikwo (FUNAI), Abakiliki, Ebonyi State, Nigeria

Princewill Aigbe
Dept of Mathematics & Computer Science
Western Delta University
Oghara, Delta State, Nigeria

ABSTRACT

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 research work attempts to address by proposing an automated data profiler model to detect terrorist activities. The work would be accomplished by using different machine learning methods, in particular, we would combine data mining algorithms with established (or experimental) thresholds to profile users' call data, combined with evidence from the profilers to construct a security detector system that would trigger alarm for prompt arrest and investigation of terrorists by law enforcement agencies. Our model can be generalised to handle other non-information related threats in different domains.

Keywords

Data mining, Security threats, profiler, experimental thresholds, call data, detector system, machine learning.

1. INTRODUCTION

Data mining is the process of finding insights which are statistically reliable, unknown previously, and actionable from a large amount of data [8]. 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 [9]. 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 [7]. 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 individuals' privacy and civil liberties, its benefits to humans and national development are enormous.

The work identifies stakeholders such as 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.

2. RESEARCH MOTIVATIONS

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. The growing erosion of internal insecurity and responses from the populace and the Nigerian state motivated this research work. Because the activities of terrorists, kidnappers and other high level crimes are a challenge to peaceful co-existence and development, our work would address 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.

3. RESEARCH PROBLEM

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 [1]. 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 [10].

False positives are a universal problem as they affect both signature and anomaly-based intrusion-detection systems [11]. A high rate of false alerts [12] 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.

4. OBJECTIVE OF THE STUDY

The objective of this research work is to answer the research questions stated above by studying the behavioural patterns of individual cellular data calls from different telecommunication networks (e.g., MTN, Glo, Airtime, etc). These data calls would be statistically analysed to determine and establish appropriate parameters (or thresholds) to be used in the data mining mathematical framework. The experimenter results from this phase would determine the data mining algorithm(s) that would be employed in the design of the data profiler application. The application will have the capabilities to interface and collect data from different networks and security agencies for the purpose of crime detection/prevention and timely arrest of criminals.

4. RELATED WORK

After the 9/11 terrorist attack on world trade centre, the American government started devising various strategies to monitor and combat (or prevent) the activities of terrorists and other crime perpetrators against the American public. Having seen the devastating effects of the 9/11 and what American government is doing to prevent future occurrence, other nations around the world started investing in technology, and most recently in data mining to detect unusual patterns, terrorist activities and other high level crimes in their countries.

For example, in Nigeria today, many terrorist activities, kidnapping, drug trafficking, and organised crime networks have sprouted in many parts of the country, Boko Haram, Movement for Emancipation of Niger Delta (MEND), etc have been unleashing terror to the Nigerian public. In particular, security agencies and the general public are extremely concerned in how the activities of the Boko Haram can be curtailed, as well as preventing the springing up of other such organised crime networks in the country. The situation has led to constant stream of debates, suggestions and increased research and publications in the literatures on how organised crimes can be curtailed by the security agencies in Nigeria. In [2], the authors reviewed and discussed different data mining algorithms to counter-terrorism without a conceptualized IT-driven tool with Nigerian content. The work concludes that to combat terrorism and other criminal activities requires adequate attention of government and the law enforcement agencies.

Ogedebe et al [3], discussed the role of information technology (IT) in combating security challenges in Nigeria. The work highlighted the applications of IT tools in strengthening Nigeria's National security against potential attacks; but no attempt was made to design a specific tool to assist the security agencies in fighting the menace of terrorist activities and other crimes in Nigeria. Benson et al [4] described the theoretical framework of how some of the most common data mining algorithms could be used to build data mining applications. The work classified data mining techniques into two categories: Classical Techniques: Statistics, Neighbourhoods and Clustering; and Next Generation Techniques: Trees, Networks and Rules.

Fawcett et al [5] studied cellular cloning techniques for perpetuating frauds. Cellular cloning occurs when a customer's Mobile Identification Number (MIN) and Electronic Serial Number (ESN) are programmed unknowingly to the owner (customer) into another cellular telephone with intention to use it to cause criminal activities. When used, the network sees the customer's MIN and ESN and subsequently attributes the calls to the customer and bills the customer accordingly. The study investigated the behaviours of users (e.g., fraudsters) over a specific period of time and designed an automated prototype system using data mining and constructive induction with standard machine learning techniques to detect fraudulent usage of cellular phones based on profiling customer's behaviour. Experimental results indicated that this approach performs nearly as well as the best hand-tuned methods for detecting frauds.

Vialandi et al [6] developed a recommendation system using data mining technique to advise students to take the right decision in relation to their academic itinerary. The study analysed real data corresponding to seven years of students' enrolments at the School of System Engineering, Universidad de Lima. Experimental results showed that the system engine achieved 77.3% of global accuracy.

Other security/fraud detection domains such as e-business and e-commerce on the Internet present a challenging data mining task because it blurs the boundaries between fraud detection system and network intrusion detection systems. However, video-on-demand websites [13], and IP-based telecommunication services [14], and online sellers [15] can be monitored by automated systems. Likewise, fraud detection in government organisations such as tax [16] and customs [17] can be monitored and reported using data mining algorithms.

In the light of the foregoing and to best of our knowledge there has been no research done in this area to build data mining application with local contents to check security threats and other high level crimes in Nigeria. In particular, there have been no systems to check non-information related threats in Nigeria. It is this gap this research is intended to fill and make contributions to the national development and growth of technology application in fighting crimes in Nigeria.

5. ARCHITECTURAL MODEL

The registration of SIM cards initiated by Nigerian Communications Commission (NCC) is in line with the standard best practice. The success of this project is very crucial to this application (a survey of the success of the SIM cards project is being studied in another work). It would enhance the reliability of the profiler system in identifying and detecting terrorists and other high level crime activities. To determine the normal behaviour of each account (user) with respect to certain indicators/thresholds, and to determine that that behaviour has deviated significantly from normal behaviour, three issues are discussed:

- ✓ How to distinguish legitimate calls from fraudulent calls? The application would use combinations of features to distinguish legitimate behaviour from fraudulent behaviour.
- ✓ How should profiles be created? Based on important feature(s) identified in step1, the system would characterise the behaviour of a subscriber with respect to identified feature(s).
- ✓ When should alarm be issued? Given a set of profiling criteria identified in step 2, the system would combine them to detect fraudulent calls and consequently issue

alarm that can lead to intervention or arrest by the security agencies.

Each of the above issue forms component of our design framework. As illustrated in Figure 1, the framework uses data mining to discover indicators of fraudulent behaviour and then builds modules to profile each user's behaviour with respect to those indicators.

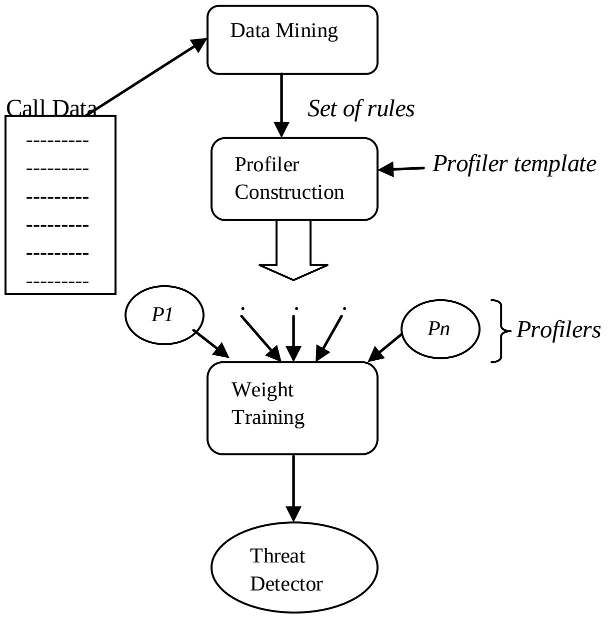


Figure 1: Architectural framework of data mining profiler model

Detailed call data over a period of six months would be collected from two different networks and data on convicted and arrested terrorists would also be collected from the SSS and the Police. The data would be analysed using statistical tools to establish certain parameters/thresholds. The features of the sampled data, combined with the established thresholds would determine the data mining algorithm(s) that would be used in the automated profiler system.

6. MINING CALL DATA

The first stage of the data profiler model construction involves circling through large amount of call data, searching for indicators of misbehaviour/fraudulent activity of cell phone users. We use a program to search for rules with certainty factors above a user-defined threshold to establish whether the call is legitimate or not. Each account generates a set of rules. After all accounts have been processed, a rule selection procedure is performed to detect fraudulent or legitimate calls. The set of accounts is circled again and the list of rules generated by each account is sorted by the frequency of occurrence in the entire account set. The highest frequency rule is selected and used in the profiler construction.

7. CONSTRUCTION OF PROFILERS

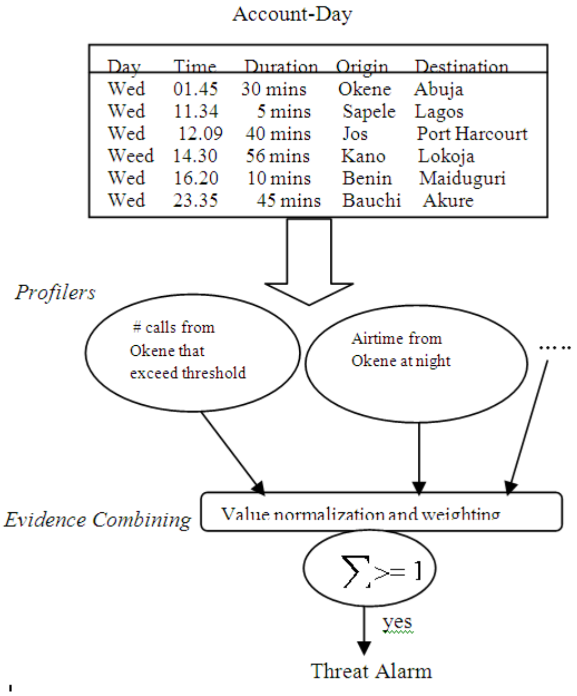
We generate a set of profilers from the discovered fraud rules. The profiler constructor has a set of templates which are instantiated by rule conditions. The profiler constructor is given a set of rules and a set of templates, and generates a profiler from each rule-template pair. Different kinds of profilers are possible. A threshold profiler yields a binary feature corresponding to whether a user's behaviour was above established threshold for a given day. A counting profiler yields a feature corresponding to its count (e.g., the number of calls from a particular 'location' at night or day). A percentage profiler yields a feature whose value is between zero and one hundred, representing the percentage of calls in the account set that satisfy established fraud rules. Each profiler is generated by different types of conditions.

8. COMBINING EVIDENCE FROM THE PROFILERS

The third stage of the model construction is how to combine evidence from the set of profilers generated by the previous stage. For this stage, the outputs of the profilers are used as features to a standard machine learning program. Training (i.e., each profiler has a training step, in which it is trained on typical (non-fraud) account activity; and use step, in which it describes how far from the typical behaviour a current account-day is) is done on account data, and profilers evaluate a complete account-day at a time. In training, the profilers' outputs are presented along with the desired output (the account-day's classification). The evidence combination learns which combinations of profiler outputs indicate fraud with high level of confidence.

8. THREAT DETECTOR/ALARM

The final output of the constructor is a detector that profiles each user's behaviour based on several indicators, and produces an alarm if there is sufficient evidence of threat/fraudulent activity. Figure 2 shows an example of a simple threat detector evaluating an account-day. Before being used on an account, the profilers undergo a profiling period of 30 days during which they measure non-fraudulent usage. In our study, these initial 30 account-days would guarantee free of fraud, but would not otherwise guaranteed to be typical. From this initial profiling period, each profiler measures the characteristic level of users' activities.


Account-Day

Day	Time	Duration	Origin	Destination
Wed	01.45	30 mins	Okene	Abuja
Wed	11.34	5 mins	Sapele	Lagos
Wed	12.09	40 mins	Jos	Port Harcourt
Weed	14.30	56 mins	Kano	Lokoja
Wed	16.20	10 mins	Benin	Maiduguri
Wed	23.35	45 mins	Bauchi	Akure

Profilers
# calls from Okene that exceed threshold
Airtime from Okene at night
.....
Evidence Combining
Value normalization and weighting
$\sum >= 1$
yes
Threat Alarm

Figure 2: Threat activity detector process for a single account-day.

9. APPLICATION OF THE PROFILER MODEL IN OTHER DOMAINS

The profiler model is useful in other domains in which typical behaviour of individuals is to be distinguished from unusual behaviour. In this respect, our profiler model can be customised to handle other problems relating to:

- Money Laundering. Money laundering has been a serious issue that needs concerted efforts from both the government and people of Nigeria. The Central Bank of Nigeria (CBN), Economic and Financial Crime Commission (EFCC), and other agencies of government can employ a customised version of the profiler model to monitor and report fraudulent transfer of money by individuals and financial institutions. Work on this area is being studied in other research work.

- Credit-Card Fraud: Financial institutions can employ the application to detect debit-card frauds. Data mining can identify locations that arise as new hot-beds of fraud and proffer solution. The constructor can incorporate profilers that notice if a customer begins to charge more than usual from that location.

- Cellular Phone Cloning Fraud: Every cellular phone periodically transmits two unique identification numbers: its Mobile Identification Number (MIN) and its Electronic Serial Number (ESN). These two numbers are broadcast unencrypted over the airwaves, and can be received, decoded and stored using special equipment that is relatively inexpensive. Cloning occurs when a customer's MIN and ESN are programmed into a cellular telephone not belonging to the customer. When this telephone is used, the network sees the customer's MIN and ESN and subsequently bills the usage to the customer. This costs telecommunications industry hundreds of millions every year.

- Toll-Fraud Detection: The toll-fraud detection identifies and protects the telephone companies from costly fraudulent use of the toll network. It applies thresholds of usage and number of calls for both hourly and daily time periods. Domestic, international and hot spot threshold sets can be defined for users. When any of the thresholds are exceeded, the system alerts the user so that toll calling from the offending number can be disabled.

10. RESEARCH CHALLENGES

The research challenges include, but not limited to the followings:

- Identify Stakeholders: This is the process of identifying all organisations (e.g., NCC, GSM Network Operators, EFCC, the Police, ICPC, etc) and people impacted by the project. This is critical for the success of the project and identifying and analysing their levels of interest, expectations, importance, and influence early enough would enable the researcher to develop the strategy to approach each stakeholder and determine the level and timing of stakeholder's involvement to maximise positive influences and mitigate potential negative impacts.
- Plan Communications. This is the process of determining the project stakeholder information needs and defining a communication approach to handle them. The project would appoint a communication manager to handle the plan communications process and respond to the information and communications needs of the stakeholders; for example, who needs what information, when they will need it, how it will be given to them, and by whom. Identifying the information needs of the stakeholders and determining a suitable means of meeting those needs pose a serious challenge to the success of the project.

- Managing Stakeholder Expectations: Managing stakeholders' expectations is the process of communicating and working with stakeholders to meet their needs and addressing issues as they occur. It involves communication activities directed toward project stakeholders to influence their expectations, address concerns, and resolve issues bothering on increasing the likelihood of project acceptance by negotiating and influencing their desires to achieve and maintain the project goals. It is also concerned with addressing issues which have not yet become problems but are related to the anticipation of future problems to the success of the project. These concerns need to be uncovered and discussed, and the risks need to be assessed, clarified and resolved. The resolution may result in a change request that is capable of improving the outcome of the researched problem.

- Plan Risk Management: This is the process of defining how to conduct risk management activities for the success of the project. Planning risk management process is important to ensure that the degree, type, and visibility of risk management are commensurate with both the risks and the importance of the project to the stakeholders and the general public. Planning is also important to provide sufficient resources and time for risk management activities, and to establish an agreed-upon basis for evaluating risks.

- Identify Risks: This is the process of determining which risks may affect the project and documenting their categorisations and characteristics. This is an iterative process because new risks may evolve or become known as the project progresses through its life cycle. The format of the risk statements should be consistent to ensure the ability to compare the relative effect of one risk event against others on the project. The researcher would take this into consideration while developing and maintaining a sense of ownership and responsibility for the risks and associated risk responses actions throughout the project.

- Plan Risk Responses: This is the process of developing options and actions to enhance opportunities and to reduce threats to project objectives. It includes the identification and assignment of one person, which is known as the risk owner, to take responsibility for each agreed-to and funded risk response. Plan risk responses addresses the risks associated with the research work by their priority, inserting resources and activities into the budget, schedule research project management plan as needed. Planned risk responses must be appropriate to the significance of the risk, cost effective in meeting the challenge, realistic within the research work, agreed upon by all parties involved, and owned by a responsible person within the project.

- Monitor and Control Risks: This is the process of implementing risk response plans, tracking identified risks, monitoring residual risks, identifying new risks, and evaluating risk process effectiveness throughout the research work. The researcher would employ such techniques as variance and trend analysis to determine if the research assumptions are still valid and if performance information generated during research execution meet the research objectives.
- Non-availability of skilled data mining personnel and system analysts: The project would employ and train personnel in areas of data mining tools and system analysis to ensure the project is successfully implemented to meet stakeholders' expectation.

11. CONCLUSIONS AND FUTURE RESEARCH WORK

We have presented the architectural framework of our data mining profiler model to fight terrorist threats and other high profile crimes in Nigeria. As criminals learn and change their strategies, our model would be deigned generically to adapt, profile, and characterise individuals' data calls based on new features exhibited by individual users. The project would identify stakeholders that would impact on the project positively or negatively and address their expectations accordingly. The research framework would be presented to some law enforcement agencies for evaluation in a view to securing funds to carry out the project execution. The outcomes of this work would motivate and lead to consistent and constant stream of research and publications in this field. The work would advocate for the establishment of national data mining centre in Nigeria.

While trying to secure funding, our next research direction is to collect secondary data (users' call data and detailed of convicted criminals) and study them using statistical tools to draw some parameters and thresholds to enable us develop a mathematical framework that fits into our local content. The framework would take into account data mining algorithm(s). The threat detector alarm system would be developed using modern information and communications technology (ICT) tools. The experimental results of this work would be published elsewhere.

12. REFERENCES

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

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

[3] P. M. Ogedebe and B. P. Jacob, "The Role of Information Technology in Combating Security Challenges in Nigeria," Academic Research International, Vol. 2, No. 1, pp. 124 – 130, January 2012.

[4] A. Benson, S. Smith and K. Thearling, "An Overview of Data Mining Techniques," 2010. retrieved 11-07-2012, http://www.thearling.com/text/dmtechniques/dmtechniques.htm,

[5] T. Fawcett and F. Provost, "Combining Data Mining and Machine Learning for Effective User profiling," KDD-96 proceedings, pp. 8-13, 1996.

[6] C. Vialandi, J. Bravo, L. Shafti, and A. Ortigosa, "Recommendation in Higher Education Using Data Mining Techniques," Educational Data Mining, pp190 – 1999, 2009.

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

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

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

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

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

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

[13] E. Barse, H, Kvarnstrom, and E. Jonsson, "Synthesizing Test Data for Fraud Detection Systems," In Proceedings of the 19th Annual Computer Security Applications Conference, pp. 384-395, 2003.

[14] J. McGibney and S. Hearne, "An Approach to Rules-based Fraud Management in Emerging Converged Networks," In Proceedings of IEI/IEEE ITSRS, 2003.

[15] B. Bhargava, Y. Zhong, and Y. Lu, "Fraud Formalisation and Detection," In Proceedings of DaWaK2003, 330-339, 2003.

[16] F. Bonchi, F. Giannotti, G. Mainetto, and D. Pedreschi, "A Classification-based Methodology for Planning Auditing Strategies in Fraud Detection," In Proceedings of SIGKDD99, pp. 175-184, 1999.

[17] H. Shao, H. Zhao, and G. Chang, "Applying Data Mining to Detect Fraud Behaviour in Customs Declaration," In Proceedings of 1st International Conference on Machine Learning and Cybernetics, pp. 1241-1244, 2002.