International Journal of u- and e- Service, Science and Technology
Vol.8, No.2 (2015), pp.357-364
http://dx.doi.org/10.14257/ijunesst.2015.8.2.34

Crime Mining: A Comprehensive Survey

Muhammad Arif [1,2], Khubaib Amjad Alam[1] and Mehdi Hussain[1,3]

[1]*Faculty of Computer Science and Information Technology, University of Malaya 50603 Kuala Lumpur, Malaysia*
[2]*Computer Science Department, Comsats Institute of Information and Technology Islamabad Pakistan*
[3]*School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad Pakistan*
arifmuhammad36@siswa.um.edu.my
khubaibalam@siswa.um.edu.my
mehdi141@siswa.um.edu.my

Abstract

Data mining is vast field for mining knowledge in various fields of life. Crime mining is one of the applications focused here. Credit card and web based crime are increasingly as more technologies are rising high. To deal and overcome fraud clustering and classification techniques are implemented. Framework and process models are designed that provide user results, graphs and trees that help user to find criminals without any complex computation.

Keywords: *crime mining, fraud detection, data mining application*

1. Introduction

Many law enforcement agencies are interested in crime detection. Different frameworks have been designed for crime detection. An integrated crime detection systems RECAP[1] allows user to search with respect to location, time and geography. PERPSEARCH has four components one of which is specified for geographical profiling[2].

Credit card crime is the one of the current issues. In [3] two algorithms communal detection and spike detection are introduced which work together for performance. A CLOPE clustering algorithm is implemented for Viet Nam's banking industry[4].

Web based crime or cybercrime is also renowned field. C4.5 algorithm is implemented to find fraudsters using WEKA tool [5]. Trade partners can also find out by keeping in view history of crime. In [6] it is shown that agencies are concerned about blogs having information of crime incidents, events and so on.

Document languages are also converted to extract crime information and then further processing is applied. In [7] and [8] Arabic documents and Chinese web pages are processed. Thus various classification, clustering and outlier detection techniques are implemented. In[9] network security is exploited by creating bots for performing fraud. New classification algorithm can be developed for controlling peer to peer network traffic.

2. Application of Data Mining in Crime Detection

This section analyzes different crime detection frameworks in detail.

The Regional Crime Analysis Program (RECAP): Regional Crime Analysis Program (RECAP) utilizes data fusion and data mining techniques to gain

ISSN: 2005-4246 IJUNESST

knowledge about crime activities[1]. RECAP provides significant information regarding crime activities and it alerts user if activity is unmanageable. It provides interface by utlizing map searching and charts with respect to time, day, week, location etc. kernel density estimation is used for crime determination in a location. K-Means and nearest neighbor algorithm is used for clustering analysis.

Crime Pattern Detection Using Data Mining

This study utilizes K- Means clustering algorithm to find crime patterns[10]. It is implemented on real time data acquired from sheriff's office. Semi-supervised learning is used for forecasting precision. Quality of data and attribute selection are important part of this technique. Every cluster has weight associated with it which is presented to expert for detection of further sub-clusters. On one hand it gives advantage of clustering for large dataset, on the other hand it has disadvantaged that only detective can get help from pattern analysis.

Data Mining in Criminal Career Analysis

This paper introduces a new tool that auto generates the criminal profiles based on four parameters: crime nature, frequency, duration and severity[11]. Distance is measured and clustering is performed on these profiles and similar cases are taken out which are used by police experts. There is lack of data updating in data which can lead to wrong results. After every two years data is updated. The future work can be done in computation process like Progressive Multi Dimension Scaling.

Multivariate Time Series Clustering Approach for Crime Trends Prediction

Traditional clustering approaches do not provide optimal results when data is multi-dimensional. So there is a need of converting multi-dimensional data into a single dimension. For this purpose dynamic wrapping method has been used by [12]. DWT has drawback of managing multivariate data with different values. To overcome this problem parametric Minkowski model is added in finding distance matrix. Determining values Minkowski weights is challenging task. In the present paper it is decided by Indian National Crime Records Bureau.

Mining Multi-Modal Crime Patterns Using Hierarchical Clustering

Data of any field have different measuring attributes with different values. This paper emphasizes on crimes having different kinds of attribute e.g. type of crime. The main concern is with three modes i.e. weapons, offender demographics and location[13]. Growing self-Organizing Maps have capability of dealing with multi-dimensional dataset. GraMode-CH is also introduced to deal with the matter of multi-modality. Vertical, horizontal and global concept hierarchy has been identified. The future work can be explored on targeted data fusion.

Mining Top-k and Bottom-k Correlative Crime Patterns through Graph Representations

The datasets are represented in the form of graphs in order to overcome the limitations of previous work[14]. Edges are first extracted and then correlation is measured. First normalization is performed to keep original density points. Then kruskal's algorithm is used to find minimum spanning tree. Third step is to find correlation between edges of graph using Pearson's correlation coefficient. Results are limited to top k or bottom k correlated datasets.

Digital Forensic Application in Data Mining

The proposed frame collects data through multiple devices[15]. Presence and authenticated utilization of devices is assured. Data is extracted and preprocessing is

implemented. Before analysis procedure data is check for authenticity. After analysis reports are generated and finally data device is presented in court for evidence. Data in this paper is collected from flash drive and processing is done through ETL. Recuva is used for extraction; Oracle Express Edition is used for warehouse and KMO for loading. First clustering is applied using k-means algorithm and further classification is implemented using C4.5 algorithm. In future other authenticated devices will be used.

Detecting suspicious money laundering cases in an investment bank

The framework proposed consists of three levels for anti-money laundering. Extraction and cleaning of data is done in data preprocessing[5]. Data quality is one of the challenging tasks. Data mining has supervised and unsupervised learning. Classification is done for supervised leaning and clustering is done for unsupervised learning. Results obtained are stored for further analysis. In future work techniques for large datasets will be handled.

Multiple-Phase Modeling to Identify Potential Fraudsters in Online Auctions

This paper emphasizes on identification and forecasting of fraud in online auction stores[16]. Decision tree a classification technique is used. First single phased model is created and then multi phased model is created by attaining attributes from different accounts. Experiment is implemented on yahoo Taiwan data. Ratio of fraudsters is half to that of legitimate users.C4.5 algorithm is implemented using Weka. In future decision tress will be replaced by lazy learners.

Online Auction Fraud Screening Mechanism for Choosing Trading Partners

In real applications of life it is difficult to choose a reliable partner for trade[17]. An instance based approach is implemented. New instances can be frequently added in the model. Experiments have been conducted on yahoo Taiwan's data. K*clustering algorithm is used. Two evaluation parameters are true positive rate and false positive rate are calculated. Results show method is helpful in decision making .New attributes can be analyzed by using new social algorithms.

Using Self Organizing Map to Cluster Arabic Crime Documents

In the paper two techniques information extraction and SOM has been implemented to identify crime related text from crime documents[3]. The framework has five stages. In normalization the words of the same form are translated into a single word. Second step is the information extraction of the grammar of crime related words. Stemming is put into action by removing suffix from words and assigning each word a distinctive number. Last step is the clustering of information extracted documents. Thus system good performance is due to information extraction on syntactic principles.

Event Ontology Construction for cybercrime

In this study Event ontology and support vector machine have been utilized for web crime mining[7]. Text mining includes SVM learning and SVM classification. Event ontology is implemented in case of feature compression and SVM in case of feature expansion. Experiments are implemented by comparing SVM with ontology and simple SVM.

AK-Modes: A Weighted Clustering Algorithm for Finding Similar Case Subsets

An AK-Modes algorithm consisting of two parts is implemented to find out similar case subsets[18]. Firstly, attributes are selected and weights are calculated using information domain ratio. Secondly, the results of the first step are used in the

second as input. Clustering is performed to find similar case subsets. Real crime datasets show considerable results. Selecting a threshold in algorithm is a difficult task. Semantic distance will be a good direction for future work.

General crime matching framework

The proposed framework[6] has three major components. First one is Entity Extraction as a branch of text mining which uses lexical lookup approach. Crime data clustering is divided into two phase. In first phase a self-organizing neural network was used to mine the attribute map and then k-means algorithm group the output. Third component is Neural Network as an engine for crime matching process. Multi-Layer perceptron is used for crime matching process. This framework can be implemented as integrated enterprise software in future.

Automatic Online Monitoring and Data Mining Internet Forums

With the advancement in technology news is available more frequently on internet rather other media communication systems. To obtain issues and news related to crime law enforcement agencies are working. A framework[19] consisting of five components has proposed in this regard. Crawler continuously download posting and information is stored in database. Statistical engine forms meta-summary of a specified time. Then clustering technique is applied on obtained information. User interface is also available to show hot topics of crime.

Money Laundering Detection

Money laundering detection has been implemented using CLOPE clustering algorithm for money transfer in banking[4]. The architecture consists of four steps. Data conversion takes place by converting the transaction data of accounts into dataset. In data fragmentation data is converted into different parts by converting number data type to nominal. Then data will be clustered by CLOPE algorithm. The system is implemented for Viet Nam's banking industry.

Resilient Identity Crime Detection

Credit fraud application falls under the category of identity crimes. This paper presents a multi layered detection system with two important detection layers: communal detection and spike detection[8].

Communal detection searches actual social affiliation which reduces the degree of fraud. Spike detection searches for the fraudsters thus increases the degree of fraud. Both algorithms perform well for large datasets of sliding window even are complements of each other. It has some restrictions of changeable and time constraints.

Data Mining and Predictive Analytics in Public Safety and Security

This process model consists of six phases[20] where initial phase is to specify the question which data mining could answer. Modus Operandi which is an important part of data is also collected with it. In data preprocessing phase data is cleaned, attributes are selected and data quality is checked in recoding. Data analyzing has two technique categories: supervised and unsupervised. Public safety and security evaluation answers to the question developed in the first state. The output is brought to a proper usable format.

PerpSearch: An Integrated Crime Detection System

PerpSearch is based on LETS. It has four main components as geographic profiling, social network analysis, crime patterns and physical matching[2]. Geographic profiling defines the location of the crime, while other components

answer for 'who'. It has social networking analysis of COPLINK system which is called concept space. It provides GUI for input physical description data and crime type. After entering output, officers can mark the area they are interested in. Serial crimes link can be added in future to automate link between same crimes.

Correlation analysis for Money Laundering Crimes

In this paper a new technique for Link Discovery based on Correlation Analysis (LDCA) for money laundering has been introduced[22]. Automatic data community items are developed for MLC. LDCA components are Link hypothesis, Link generation and Link identification. In link hypothesis data is collected from individuals. Correlation is found out between individuals transaction record. Monetary vectors are determined by k-means clustering. Histogram is generated and local and global correlation is discovered in Link generation phase. A new segmentation threshold is defined in Link identification.

Data Mining for Security Applications:

The aim of project is to develop a peer to peer botnet for network traffic[9]. A bot is an agent that performs certain actions on behalf of intruders. This paper utilizes mining techniques including clustering, classification and outlier detection over concept-drifting data streams to detect peer-to-peer Botnet traffic.

Table 1. Techniques to Handle Security Threats

Reference	Proposed Method	Mining Technique	Mining technique	Other Techniques
[1], [2]	The Regional Crime Analysis Program (RECAP):a Framework for Mining Data to catch criminals	clustering	K means and nearest neighbor clustering algorithm	Kernel density estimation
[3]	Using Self Organizing Map to Cluster Arabic Crime Documents	clustering	Self-Organizing Map	rule-based approach
[4]	Applying Data Mining in Money Laundering Detection for the Vietnamese Banking Industry	Clustering	CLOPE algorithm	
[5]	A data mining-based solution for detecting suspicious money laundering cases in an investment bank	clustering	centre-based clustering algorithm	neural network
[6]	Detecting and investigating crime by means of data mining: a general crime matching framework	clustering	SOM, k-means	MLP Neural Network
[7]	An Event Ontology Construction Approach To Web Crime Mining	Classification	Support Vector Machine	Event Ontology
[8]	Resilient Identity Crime Detection			communal detection (CD) and spike detection (SD).
[10]	Crime Pattern Detection Using Data Mining	clustering	k-means clustering technique	Semi supervised learning
[11]	Data Mining Approaches to Criminal Career Analysis	clustering	Multi-dimensional clustering	
[12]	A Multivariate Time Series Clustering Approach for Crime Trends Prediction	clustering	Nearest neighbor clustering	single linkage method and dynamic time wrapping and parametric M inkowski model
[13]	Mining Multi-Modal Crime Patterns At Different Levels of Granularity Using Hierarchical Clusterings	Clustering	Hierarchical clustering-Growing Self Organizing Maps	
[14]	Mining Top-*k* and Bottom-*k* Correlative Crime Patterns through Graph Representations	clustering	Density based clustering	Kruskal algorithm
[15]	A Novel Data Generation Approach for Digital Forensic Application in Data Mining	clustering, classification	**k-means algorithm, C4.5 decision tree model**	Recuva, Oracle Express Edition, Bartlett's test of sphericity and Kaiser-Meyer-Olkin

[16]	A Multiple-Phased Modeling Method to Identify Potential Fraudsters in Online Auctions	classification	decision trees	
[17]	An Online Auction Fraud Screening Mechanism for Choosing Trading Partners	Clustering	K* algorithm	
[18]	AK-Modes: A Weighted Clustering Algorithm for Finding Similar Case Subsets	clustering	AK-Modes algorithm	
[19]	Automatic OnLine Monitoring and Data Mining Internet Forums	Clustering		Text mining

Table 2. Techniques to Handle Security Threats

Sr. No	Quality measuring attribute	Strength	weakness	comment
[1], [2]	Automating spatial analysis	Map-Oriented Searches, Control and Time Charting, Hot Spot analysis using Kernel Density Estimation		GIS expert requirement removal is good.
[3]	Arabia documents crime	ability to extract keywords based on syntactic principles		New application gives new idea for Arabic documents
[4]	Bank transactions	Detecting money laundering	system can't run standalone absolutely without analyst	
[5]	running time performance	simple and efficient data mining-based solution		
[6]	Automatic insertion of data into database	Crime data analysis		SOM is widely used for clustering
[7]	cybercrime on Chinese websites	Feasible and effective in Web crime mining.	web pages in Chinese	Event ontology increases performance rate
[8]	resilience (multilayer defence), adaptivity (accounts for changing fraud and legal behavior), and quality data (real-time removal of data errors)	successful credit application fraud patterns	extreme imbalanced class, and time constraints	Only good for credit card application
[10]	Increase productivity of detectives	Requires Time ,crime type and area Selection	crime pattern analysis can only help the detective, not Replace them.	Data quality and attributes are important factors.
[11]	Crime Nature, Frequency, Seriousness and Duration	Digital profile creation	Lack of information of crime	behavior of criminal can be predicted
[12]	Reducing data dimensions	Suitable for sequences of differed lengths	Difficult to find appropriate weights for M inkowski model	Other techniques can be used for reducing dimensions.
[13]	Pattern recognition at different levels of granularity	necessity and significance of mining patterns		Weapon example clears the idea of multi-modality
[14]	Correlation analysis among datasets	Efficient calculation		Real crime datasets are provided for approach.
[15]	generating, storing and analyzing data, retrieved from digital devices	economic preprocessing	cannot restore files if Windows OS is being used	open source tools but complex processing.
[16]	legitimacy of users before a fraud occurs	rules are generated by the legitimate accounts extracted features	decision trees detect the minority of fraudsters, of which results are not as accurate as the majority of instances	User can be altered before fraud
[17]	Instance based learning	Helpful in decision making	Lesser capability of inducing general rules	
[18]	finding the similar case	improve the efficiency	reasonable threshold	Still needs more

	subsets	compared with the traditional approaches, Assists in the decision-making process.		improvement.
[19]	Finding news witness	helps investigators to control internet forums		No specific clustering technique

3. Conclusion and Future Work

With different searching methods of crimes it can be concluded that Clustering is more widely used than classification. Data quality can be one the challenging issue. Cyber mining is the one of main concern as agent are developed to perform fraud thus exploiting network security. Criminal network analysis is one important for police to track criminal via location, vehicle etc. Credit card fraudster is also revolutionary to be solved by different mining methods.

For money laundering techniques for large datasets will be handled. Semantic distance will be a good direction for future work in ontology. Credit application fraud detection can be extended

References

[1] D. E. Brown, "The Regional Crime Analysis Program (ReCAP): a framework for mining data to catch criminals," SMC'98 Conf. Proceedings. 1998 IEEE Int. Conf. Syst. Man, Cybern. (Cat. No.98CH36218), vol. 3, pp. 2848–2853, 1998.
[2] L. Ding, D. Steil, M. Hudnall, B. Dixon, R. Smith, D. Brown, and A. Parrish, "PerpSearch: An integrated crime detection system," 2009 IEEE Int. Conf. Intell. Secur. Informatics, pp. 161–163, 2009.
[3] M. Alruily, A. Ayesh, and A. Al-Marghilani, "Using Self Organizing Map to cluster Arabic crime documents," Proc. Int. Multiconference Comput. Sci. Inf. Technol., pp. 357–363, Oct. 2010.
[4] D. K. Cao and P. Do, "Applying Data Mining in Money Laundering Detection," pp. 207–216, 2012.
[5] N. A. Le Khac, S. Markos, and M.-T. Kechadi, "A Data Mining-Based Solution for Detecting Suspicious Money Laundering Cases in an Investment Bank," 2010 Second Int. Conf. Adv. Databases, Knowledge, Data Appl., pp. 235–240, 2010.
[6] M. R. Keyvanpour, M. Javideh, and M. R. Ebrahimi, "Detecting and investigating crime by means of data mining: a general crime matching framework," Procedia Comput. Sci., vol. 3, pp. 872–880, 2011.
[7] C. Li, Y. Hu, and Z. Zhong, "An event ontology construction approach to web crime mining," 2010 Seventh Int. Conf. Fuzzy Syst. Knowl. Discov., no. Fskd, pp. 2441–2445, Aug. 2010.
[8] C. Phua, K. Smith-miles, S. Member, V. C. Lee, and R. Gayler, "Resilient Identity Crime Detection," vol. 24, no. 3, pp. 533–546, 2012.
[9] B. Thuraisingham, "IEEE ISI 2008 Invited Talk (I) Data Mining for Security Applications : Mining Concept-Drifting Data Streams to Detect Peer to Peer Botnet Traffic," no. I, 2008.
[10] S. V. Nath, "Crime Pattern Detection Using Data Mining Florida Atlantic University / Oracle Corporation," vol. 1, no. 954, pp. 1–4, 2006.
[11] J. De Bruin, T. Cocx, W. Kosters, J. J. Laros, and J. Kok, "Data Mining Approaches to Criminal Career Analysis," Sixth Int. Conf. Data Min., pp. 171–177, Dec. 2006.
[12] B. Chandra, M. Gupta, and M. P. Gupta, "A multivariate time series clustering approach for crime trends prediction," 2008 IEEE Int. Conf. Syst. Man Cybern., pp. 892–896, Oct. 2008.
[13] Y. L. Boo and D. Alahakoon, "Mining Multi-modal Crime Patterns at Different Levels of Granularity Using Hierarchical Clustering," 2008 Int. Conf. Comput. Intell. Model. Control Autom., pp. 1268–1273, 2008.
[14] P. Phillips, E. Peterphillipsjcueduau, and I. Lee, "Mining Top- k and Bottom- k Correlative Crime Patterns through Graph Representations," pp. 25–30, 2009.
[15] V. H. Bhat, P. G. Rao, A. R.V., P. D. Shenoy, V. K.R., and L. M. Patnaik, "A Novel Data Generation Approach for Digital Forensic Application in Data Mining," 2010 Second Int. Conf. Mach. Learn. Comput., pp. 86–90, 2010.
[16] W.-H. Chang and J.-S. Chang, "A Multiple-Phased Modeling Method to Identify Potential Fraudsters in Online Auctions," 2010 Second Int. Conf. Comput. Res. Dev., pp. 186–190, 2010.
[17] I. Conference and E. Technology, "2010 2nd International Conference on Education Technology and Computer (ICETC)," pp. 56–60, 2010.
[18] L. Ma, Y. Chen, and H. Huang, "AK-Modes: A weighted clustering algorithm for finding similar case subsets," 2010 IEEE Int. Conf. Intell. Syst. Knowl. Eng., pp. 218–223, Nov. 2010.
[19] Y. M. Lai, X. Zheng, K. P. Chow, L. C. K. Hui, and S. M. Yiu, "Automatic Online Monitoring and Data-Mining Internet Forums," 2011 Seventh Int. Conf. Intell. Inf. Hiding Multimed. Signal Process., pp. 384–387, Oct. 2011.
[20] C. Mccue, "and Security," no. August, 2006.

[21] S. Wang, "A Comprehensive Survey of Data Mining-Based Accounting-Fraud Detection Research," 2010 Int. Conf. Intell. Comput. Technol. Autom., pp. 50–53, May 2010.
[22] Z. M. Zhang, J. J. Salerno, and P. S. Yu, "Applying Data Mining in Investigating Money Laundering Crimes," no. Mlc, pp. 747–752, 2003.

Authors

Muhammad Arif is a PhD student at Faculty of CS and IT, University of Malaya. Currently he is working on Medical image Processing. His research interests include image processing, E learning, Artificial intelligence and data mining. He joined UM as a Bright Spark Scholar in September 2013 for the period of 3 years. Before this he completed masters and bachelor degrees in Pakistan. He received his BS degree in Computer Science from University of Sargodha, Pakistan in 2011. He obtained his MS degree in Computer Science from COMSATS Islamabad 2013 Pakistan.

Khubaib Amjad Alam is a Ph.D. candidate at Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, under the prestigious Bright Spark Fellowship. He received Master of Computer Science degree in 2013 from Comsats Institute of Information Technology, Pakistan. His current research interests are Business Process Management (BPM); Service oriented architecture (SOA), Web services, distributed systems, Change management, Maintenance and Evolution issues of Enterprise level software systems and engineering applications of soft methodologies.

Mehdi Hussain received his BS degree in Computer Science from The Islamia University Bahawalpur, Pakistan in 2005. He obtained his MS degree in Computer Science from SZABIST Islamabad 2011 Pakistan. He has 8 years working experience in a renewed Software House (Streaming Networks (private)). He recently selected as funded scholar at National University of Science and Technology (NUST) under faculty development program 2014. Presently he is research scholar at University of Malaya, Malaysia. Research interests are multimedia security, steganography, data mining. He can be reached at *mehdi141@hotmail.com*.

