Comparative Analysis of Data Mining in Criminal and Fraud Detection

[1]Ayushi Dwivedi, [2]Chintan Singh, [3]*Amarnath Mishra,[4]Ved Prakash Mishra

ABSTRACT-- *This manuscript explains the concept of data mining and its application in cybercrimes. Cybercrimes are becoming very serious day by day due to large data sets are generated by organizations and lack of the awareness of the internet users. The application of data mining in cybercrime and framework of data mining for detection of financial fraud is explained. A comparative analysis on digital forensic tools and techniques are done with their benefits*

Keywords-- *FBI, Data Mining, Computer, KDD, data cleaning, data integration, data selection, data transformation, extraction, pattern, pattern evaluation, interestingness measures, knowledge presentation, database, database warehouse, repository, pattern evaluation, user interface, data warehouse, user interaction.*

I. INTRODUCTION

From the past decade, IT and Computer field is growing enormously and so are their vulnerabilities. Data mining is one of the most recent ways introduced in today's world criminal data mining. National security seems to be at high risk after 09/11 attack. FBI and other agencies are devoting all their time to gather information about the possible upcoming threats in order to prevent it. They started monitoring and analysing the criminal data record and find out any pattern or evidence. The architecture of a basic data mining system has some major components like Databases, Data Warehouse, World Wide Web and other important repository, Data Warehouse server or databases, Knowledge base, Data mining engine, Pattern evaluation, User interface

II. TECHNIQUES USED IN DATA MINING

Data mining is about extrapolating patterns and new knowledge from the big sets of data that were previously collected.Various techniques used are:

A. Tracking patterns

A set of data is taken and observed statistically and thoroughly to find a new and interesting pattern in the given set of data which was not known. It is usually a technique where we find out some aberrations, ebb, and flow of certain variables in a given set of data at a given interval of time.

[1] Amity Institute of Forensic Sciences, Amity University, Noida, India.

[2] Amity Institute of Forensic Sciences, Amity University, Noida, India.

[3] Assistant Professor & Program Leader,Amity Institute of Forensic Sciences, Amity University, Noida, India,*

***Contact detail**: +91-9818978527, amishra5@amity.edu/ drmishraa1@gmail.com*

[4] Department of Engineering & Technology, Amity University, Dubai, U.A.E.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

B. Classification

Classification builds a model to predict different labels according to their category to distinguish between objects of few different classes. These have are predefined, discrete and unordered discrete label [1,2]. Zhang and Zhou [3] state that classification is the way of finding out a set of similar features and models which differentiate data concepts and their classes.

C. Prediction

Old data is recognized and the historical trend is observed that further predict about the future occurring that could take place. Prediction models values functions.

D. Association

Association relates with the data mining function that discovers the probability of the co-occurrence of two similar types of data in a set. Association rules are discarded if it does not satisfy both a minimum support threshold as well as minimum confidence threshold between the correlated attribute value pairs.

E. Outlier detection

Determining anomalies or outlines in data is an important factor to observe and identify a particular type of uncommon event in the previous data set.

F. Clustering

According to Yue et al. [4], p. 5520], “clustering analysis concerns the problem of decomposing or partitioning a data set into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups.”

G. **Regression**

Regression primarily deals with numeric values that help to determine the similarities between the variables.

H. Social Network **Analysis**

This type of data analysis is based on common relation of interest to understand their structure and behaviour.

I. Entity Extraction

An information extraction procedure of identifying and classifying data to pre-defined categories by converting unstructured data into structured one so used for retrieving information.

Keyword used: data, data mining, databases, classification, frequent pattern analysis, association, correlations, prediction, cluster analysis, outlier analysis, evolution analysis, pattern detection, regression, modeling, anomalies, association rules, minimum support threshold, minimum confidence thresh, social network, social network analysis, MMO, email network.

III. METHODOLOGY FOR RESEARCH IN DATA MINING

There are certain phases to divide the framework which are

- Research definition
- Research methodology
- Research analysis

In the first phase that is of research definition, a person finds out the area for research, goals and scope of the particular research area. Here, we take the example of Financial Fraud Detection (FFD) which relies on Data Mining in which the goal is to create a classification framework. Scope here is the literature on the various applications of data mining that are used in FDD between the years 1997 to 2008. In the second phase used for particular research, we define the required portion and the related articles and framework to further elaborate it.s

We took forty-nine articles for classification and were classified according to the given steps [5].

• Classify the selected articles.

• Verify the classification with other author(s) and check it again with another independent co-author.

• Approve the categories assigned to the article if the classification results are consistent, or hold a discussion among the researchers to reach a consensus otherwise.

The last phase of this methodology is used to analyze the final research to lead us to the upcoming research and give us a proper conclusion.

Keywords used: World scientific net, journals, scientific data, methodology, framework, Research definition, Research methodology, Research analysis, Financial Fraud Detection (FFD), computer science, management, marketing, engineering, social work, information science, Transactions, medical research, Springer-Link Journals, protocols, World Scientific Net.

IV. APPLICATION OF DATA MINING TECHNIQUES IN CRIME DETECTION

Data mining techniques used to resolve various types of crime including financial frauds which are better analysed and crime pattern were recorded successfully by data mining technique.

Following are the techniques used for detection of various crime patterns-

V. ENTITY EXTRACTION

These techniques were used in frauds discussing in [6, 7, 8, 9, 10]

Key Terms-Named Entity Extraction (lexical look up, rule-based, SPSSLexiQuest, Natural Language Processing

Objective- Location, Time, Vehicle, Nationality,

Phone, Gender and Race

B. Cluster analysis- Application of this technique described widely in- [11, 12, 13, 14, 15, 16]

Key terms-GIS, Self-Organizing Map, Partitioning Clustering Technique, Hierarchical Clustering Technique

Objective-Detect crime hotspots; automatically identify relation from available crime data and weigh relationships to find out all associate possible chances of crime with other densities.

VI. ASSOCIATION RULE

association rule application effectively used in [17, 18, 19, 20, 21]

Key terms- Distributed Association Rule Mining and Apriori Algorithm.

Objective- to connect crime incidents, narrow down all possible suspects, gives informative association among criminal entities or items.

VII. SOCIAL NETWORK ANALYSIS

pattern analyzed and described in [22, 23, 24, 25]

Key terms- K-core, Core/periphery Ratio; Measure of Centrality, Closeness and Between.

Objective- It Provide analyses of functions, related structures and the combination measurements, which in crime domain

VIII. ADVANTAGES AND DISADVANTAGES OF DATA MINING TECHNIQUES FOR FRAUD DETECTION

Data mining techniques day by day proving themselves a key factor for data extraction and analysis. Varieties of tools and brands competing each other to prove themselves best in this field. Below is the table which points out some of the mostly used techniques for data extraction along with their advantages and disadvantages- (Table 1)

Table 1: Advantages and disadvantages comparison of widely used techniques

S. No.	Techniques used	Advantages	Disadvantages
1.	Entity Extraction	Clear, Simple Gives lower estimate of evaluated system Discriminative method Readability and maintainability	Too Precise Data Sparseness
2.	Cluster Analysis	Efficient Simple Not complex Arbitrary shape Works in presence of noise	Maybe possible that the number of cluster is given in advance Sometimes can't work well in the

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

		Good visualization capability	presence of outlier
3.	Classification	Easy to understand Easy to implement Fast training and noisy data robust Compatible with multimodal classes Easy interpretation Handles uncertainty Stochastic relationships are identify	Limited memory Slow process Sometimes Complex Duplication could be present Solution depends on direction of the decision
4.	Association Rule	Subset of a frequent item set will also be frequent (Apriori) Number of transaction in database will be more than number of entries (AprioriTid)	Unnecessarily generating and counting so much information which is actually small (AIS) Same entities as its support value(SETM)
5.	Social Network Analysis	It is a type of predictive analysis Non expensive and revenue improves It creates awareness Enforce government regulations	Affected privacy Serious security threats Quite expensive initially
6.	Regression	Relative influence of the predictor variables to criterion value is determined It can identify anomalies or outliers so works well even in its presence	Use of incomplete data False result that correlation is caused

IX. THE CONCEPTUAL FRAMEWORK FOR APPLICATION OF DATA MINING IN FINANCIAL FRAUD DETECTION

This classification framework is based on the knowledge on the research of data mining and fraud detection research. The research in the field of application of data mining algorithms and techniques useful for financial accounting fraud detection is a well-studied area.

Below is the main areas of the framework-

A. Prediction- For prediction the value should be continued value rather than discrete or unordered value. This attribute as termed as predicted attribute

B. Clustering- It is analysed that data objects in individual cluster should possess high intra-cluster similarity among the similar cluster but also have low inter-cluster similarity to those as in different clusters.

C. Regression- This technique used in fraud related to detection of credit card, automobile insurance, crop and corporate sector.

D. Outlier detection- Data which have the different properties than the rest of the remaining ones is termed as outliers.

E. Classification- neural networks, the Naïve Bayes technique, decision trees are some basic classification techniques.

F. Visualization- It gives the answers on easy way for difficult
complex action.

X. COMPARATIVE ANALYSIS OF THE TOOLS USED FOR DIGITAL FORENSIC ANALYSIS

As described about the techniques in previous sections this section compare the tools used for data extraction and analysis. Table 2 describe such tools-

Table 2: Comparison between different tools use worldwide for data mining and fraud detection.

S. No.	**Digital Forensic Tools Used**	**Advantages**	**Daadvantages**
1.	SANS .investigation Forensic Toolkit	-Analysis of Expert witness format (E01) -Advanced Forensics Format (AFF)	-Poor user documentation -Need to resort to the Command -Line for any serious forensics work.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

		-update and customize DFIR package automatically.	
2.	Sleuthkit + Autopsy	-analysis of timeline and time zone -filter hash values -analysis of file system and keyword search in advance manner -GUI (Graphical user interface) can display system events in form pf pictorial representation.	-not be able to manually carve out data -tool would freeze -disoriented at times -hard to navigate to places that a search provided
3.	FTK imager	-Data can be preview as well as files and folders. -analysis of data from different sources -also image mounting can be done.	-No progress report -unable to perform multi-tasking operation - scripting can't run in this tool -no MAC support -file limit is 2 million -PSD and AVI not supported
4.	EnCase	- complex file can be breakdown easily for examination, such as the registry files, dbx & pst files, thumbs db etc. - time line present - full scripting abilities and can automatically	- EnCase Index needs work - No Outlook 2003 PST/OST support - No Internal Mail Viewer - Rough looking Report - No full Indexing of the Drive

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

		decryption and carving of report.	
5.	Linux dd command	- can done backup and restoration of master boots records - Easy modification of data. - dd can duplicate data across files, devices, partitions and volumes - copy the entire disk	-can wipe a disk completely so caution is required. -dd uses the kernel to read or write to device files instead of accessing hardware
6.	CAINE (Computer Aided Investigative Environment)	-full of tools and utilities to aid every stage of a digital investigation -very helpful scripts that are mated to the Caja file manager -host device is mounted with a read-only software write blocker	-It lacks documentation.
7.	ExifTool	-verbose and HTML-based hex dump output format can also be analysed. -can duplicate meta-data information between files. -can back up the original image	-Information stored in different places within a single format. -The writing logic for ExifTool is the reverse of the reading logic -Can't edit or create most of the formats.

		without command or permission.	
8.	Xplico	-Port Independent Protocol Identification (PIPI) for each application protocol; -Multithreading; -Output data and information in SQLite database or Mysql database and/or files; -At each data reassembled by Xplico is associated a XML file that uniquely identifies the flows and the pcap containing the data reassembled; -Realtime elaboration	-no instructions or support for installing the software on Windows via Red Hats Cywin or other similar tools -online manuals are out dated and refer to very old versions of the tool -The command line interface of Xplico does not have a manual -all third party documentation are either outdated
9.	Last Activity View	-Records many user actions -it will create a timeline of events on launching. -To monitor kids or family persons activity	-It can only log files opened and saved in the standard Windows Open/ Save dialogs -System cleaning tools may wipe information
10.	DSi USB Write Blocker	-protect the USB for modification by making it only readable. So that no data can be changed or overwrited.	-For newer versions of window require changes. -Receiver also should have same tool to unlock USB.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

		-application status can be seen in taskbar simultaneously. -Mostly runs on windows	-May damage USB if try to unlock from other source.
11.	FireEye RedLine	-this tool helps to identify a compromised file that was introduced and how such file survives in the system -data can be filter out by using whitelist indicators. -collects information from live running sytems. -It is free to use in any sized environment - As for general rootkit protection, Redline uses raw disk access by default where possible to avoid being subverted by rootkits	-Redline only works on memory images and live hosts -not support remote collection -Always have to analyse by GUI cannot work on command line method.
12.	Helix3	-Data-folds are used to tag different memory sections -Have a RAM editor	-Pro version is paid. -Live system capture in windows not there.

XI. LIMITATIONS OF DATA MINING METHOD

Beyond all these profits data mining technique have limitation also which are as follows-

- Validity must be made by the user

 - The system violates the privacy of the user

 - High chance of critical data must be hacked.

- The safety and security measure

- Become it less prone to misuse

XII. CONCLUSION

This paper is a review study which begins with brief introduction of fraud and data mining and its techniques for crime detection, preventions and analysis. By correct application of data mining savings, security, and extraction could be attained. Data mining techniques generally aimed to uncover the hidden relationship under large data base. By following a thorough research, we wrote list of application and framework of data mining techniques. This is presently most accepted technique for fraud analysis.

REFERENCES

1. Han, M. Kamber, Data Mining: Concepts and Techniques, Second ed, Morgan Kaufmann Publishers, 2006, pp. 285–464.
2. P. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, First ed.AddisonWesley Longman Publishing Co., Inc, 2005
3. D. Zhang, L. Zhou, Discovering golden nuggets: data mining in financial application, IEEE Transactions on Systems, Man and Cybernetics 34 (4) (2004) Nov.
4. Yue, X. Wu, Y. Wang, Y. Li, C. Chu, A review of data mining-based financial fraud detection research, international conference on wireless communications Sep, Networking and Mobile Computing (2007) 5519–5522
5. E.W.T. Ngai, L. Xiu, D.C.K. Chau, Application of data mining techniques in customer relationship management: a literature review and classification, Expert Systems with Applications 36 (2) (2009) 2592–2602
6. Chau, M., Xu,J.J., & Chen,H. (2002,May) .Extracting meaningful entities from police narrative reports. In Proceedings of the 2002 annual national conference on Digital government research (pp. 1-5). Digital Government Society of North America.
7. Chen, H., Chung, W.,Xu. J., Wang,G., Qin,Y.,& Chau,M.(2004). Crime data mining: a general frame work and some examples. computer, (4),50-56.
8. Cocx, T. K., &Kosters, W.A. (2006, July). A distance measure for determining. Similarity between criminal investigations. An Industrial Conference on Data Mining. (pp.511-525). Springer, Berlin, Heidelberg
9. Ku,C. H., Iriberri,A.,and Leroy.G.(2008).Crime information extraction from police and witness narrative reports. In: Proceedings of the IEEE Conference on Technologies for Homeland Security,12-13May, Waltham, MA:193–198.

10. Pinheiro, V., Furtado, V.,Pequeno, T., and Nogueira,D.(2010). Naturallan-guage processing based on semantic inferentialism for extracting crime information from text. In: Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI), 23-26May, Vancouver, BC, Canada:19–24.
11. Nath, S. V. (2006, December). Crime pattern detection using data mining. In2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops (pp. 41-44).
12. Pang-Ning,T., Steinbach,M.,and Kumar,V.(2014). Introduction to Data Mining. First Edition,Pearson:UK.
13. Dietterich,T. G.,Becker, S.,& Ghahramani,Z.(Eds.). (2002).AdvancesinNeural Information Processing Systems 14:Proceedings ofthe2001Conference (Vol.2). MIT Press.
14. Anderberg, M.R. (1973). Cluster analysis for applications (No.OAS-TR-73-9). Office of the assistant for study support kirtlandAFBnmex.
15. Ng,R. T., &Han,J.(1994, September).Efficientand Effective clustering methods for spatial data mining.In Proceedings of VLDB (pp. 144-155).
16. Schroeder, J., Xu,J., Chen,H.,and Chau, M. (2007). Automated criminal link analysis based on domain knowledge. Journal of the American Society for Information Science and Technology,58(6):842–855.
17. Brown,D.E., and Hagen,S.(2003). Data association methods with applications to law enforcement. DecisionSupportSystems,34(4):369–378.
18. Lin, S., & Brown D. E. (2006). An outlier-based data association method for linking criminalincidents. DecisionSupportSystems,41(3),604-615.
19. Appavu, S., Pandian, M., andRajaram,R.(2007).AssociationRuleMiningfor Suspicious Email Detection: A Data Mining Approach In:Proceedings ofthe IEEE IntelligenceandSecurityInformatics,23-24May,NewBrunswick,NJ:316–323.
20. Ng,V., Chan,S., Lau,D.,&Ying,C.M. (2007,March).Incremental mining for temporalassociationrulesforcrimepatterndiscoveries.InProceedings of the eighteenth conference on Australasiandatabase-Volume63(pp.123-132). Australian Computer Society,Inc..
21. Buczak,A.L., and Gifford, C.M. (2010). Fuzzy Association Rule mining for community crime pattern discovery. In: Proceedings of the ACM SIGKDD Work-shop on Intelligence andSecurityInformatics,25-28July, WashingtonDC.
22. Sparrow, M.K. (1991). The application of network analysis to criminal intelligence: An assessment of the prospects. Social networks, 13(3), 251-274.
23. Wang,J.C., and Chiu,C.Q.(2005). Detecting online auction inflated-reputation behaviors using social network analysis.In: Proceedings of the Annual Conference of the NorthAmerican Association for Computational Socialand Organizational Science,Notre Dame:26–28
24. Qin, J.,Xu,J. J., Hu,D., Sageman,M.,and Chen,H. (2005). Analyzing terrorist networks:In Intelligence and Security informatics,3495:287–304.
25. McNally, D., &Alston,J. (2005). The use of social network analysis (SNA)in the examination of an outlaw motor cycle gang. Journal of Gang Research,13(3),1.