JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 153 - 160
https://publishoa.com
ISSN: 1309-3452

A Survey on Analyzing Crime Patterns Using Data Mining Techniques

S. IMTHIYAS AHAMED
Research Scholar
PG & Research Department of computer science
Khadir Mohideen College, Adirampattinam
(Affiliated to Bharathidasan University, Trichy-24)
Thanjavur, Tamilnadu
India.

Dr. J. CHOCKALINGAM
Research Supervisor
Associate professor of computer science (Retd)
Khadir Mohideen College, Adirampattinam
(Affiliated to Bharathidasan University, Trichy-24)
Thanjavur, Tamilnadu
India

Dr. A. SHAIK ABDUL KHADIR
Research Co-Supervisor
Head & Associate professor of computer science
Khadir Mohideen College, Adirampattinam
(Affiliated to Bharathidasan University, Trichy-24)
Thanjavur, Tamilnadu
India

ABSTRACT

The data mining is data examining strategies that used to investigate crime data recently put away from different sources to find examples and patterns in crimes. In extra, it tends to be applied to build productivity in settling the crimes quicker and furthermore can be applied to naturally inform the crimes. In any case, there are numerous data mining strategies. To expand productivity of crime location, it is important to choose the data mining strategies appropriately. This paper audits the writings on different data mining applications, particularly applications that applied to tackle the crimes. Survey likewise illuminates research holes and difficulties of crime data mining. In extra to that, this paper gives understanding with regards to the data mining for finding the examples and patterns in crime to be utilized suitably and to be an assistance for amateurs in the examination of crime data mining.

I. INTRODUCTION

Crime prevention and detection become an important trend in crime and a very challenging to solve crimes. Several studies have discovered various techniques to solve the crimes that used to many applications. Such studies can help speed up the process of solving crime and help the computerized systems detect the criminals automatically. In addition, the rapidly advancing technologies can help address such issues. However, the crime patterns are always changing and growing [1]. The crime data previously stored from various sources have a tendency to increase steadily. As a consequence, the management and analysis with huge data are very difficult and complex. To solve the problems previously mentioned, data mining techniques employ many learning algorithms to extract hidden knowledge from huge volume of data. Data mining is data analyzing techniques to find patterns and trends in crimes. It can help solve the crimes more speedily and also can help alert the criminal detection automatically.

This paper gives the brief reviews of researches on various implementation of data mining and the guidelines to solve the crimes by using data mining techniques. It also discusses research gaps and challenges in the area of crime data mining. In the next section, the background and the issues of data mining are discussed. Section III elaborately discusses about the uses of data mining techniques to solve the crimes. The research issues and challenges are shown in Section IV. Finally, the study is concluded in Section V.

II. DATA MINING FUNDAMENTALS

Data mining is the analysis process used to analyze the historical data to find trends, patterns and knowledges. To extract the hidden knowledges, there are the initial important factors for analysis as follows: 1) The data used for analysis require the accuracy and sufficiency. 2) Knowledges and ex- periences of specialists.Fig.1 The knowledge results obtained from data mining processes are used to assist in decision makingand to solve the problems. In the data mining, the analyzing techniques are explained in the following sub-sections

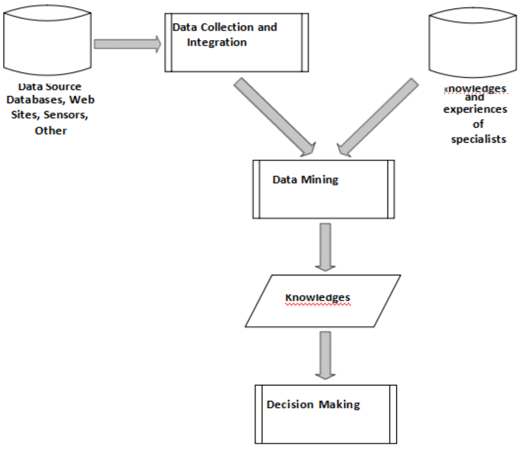

Fig.1. The background of data mining.

A A. ASSOCIATION RULE MINING

This technique is unsupervised learning method that used to find the hidden knowledges in unlabeled data. It is used to solve the issues if the learners get the unlabeled example data. In additional, association rule can discover the interesting co-occurrences of objects in large data sets. In the basic of association rule, the rule consists of two parts. 1) The antecedent, which is on the left side or called the left hand side (LHS). 2) The consequent, which is on the right side or called the right hand side (RHS). A form of general association rule is LHS → RHS, where LHS and RHS are disjoint item-sets

If the LHS item-set occurs then the RHS item-set will be likely to occur. For the efficient discovery of association rules, the important statistical measurements, the support and confidence measures, should be used together. A value of such measures is in the range of 0-1. If a association rule has very low support, this rule is likely to be uninteresting. As a consequence, the support measure is used to dispose the uninteresting association rules. The confidence measure is used to gauge the reliability of association rules. For a given rule A and B in a transaction set T with higher the confidence, B is likely to be present in T that contain A. To discover co-occurrences between two data sets, support and confidence results should be greater than user-specified thresholds [3].

B. CLUSTERING →

Clustering is a data analyzing technique in unsupervised type. This technique is used to divide the same data into the same group and the different data into the other group. The clustering techniques have a variety of concepts. The use of clustering techniques depends on applied fields [9]. For the simple and effective clustering techniques, there are several algorithms such as K-means, Hierarchical Clustering and Expectation-Minimization that are discussed below.

C. CLASSIFICATION

This technique is supervised learning method that used to assign objects to one of many pre-determined categories. The algorithms of classification have been widely applied to the several problems that include many various applications. For example, it is used to solve the detecting of the suspect vehicles and intruders, the prediction of heart disease, the categorizing the document, etc. The basic concept of classification is described as the following: A collect data, also known as an input data, is used to process in a classification task. Each record consists of the attribute set and a class label. The class label is pre-determined category. A collect data is divided into two sets. 1) Train set is partitioned randomly that is used to create a classification model, also known as a classifier, to predict the class of the new unknown record. 2) Test set is a remaining set that is used to evaluate the performance of the classification model. For building the classification models, there are many systematic approaches such as: decision tree, nearest neighbor, Bayes' Theorem and neural network, etc.

D. CLOUD STORAGE

Cloud storage is the on-demand delivery of compute power, database storage, applications, and other IT resources through a cloud services platform via the internet with pay-as-you-go pricing[1].Cloud storage provides various services in which data storage is the main cloud service. Cloud storage works behind the scene in our day to day activities such as to watch movies, play games, sending mails and listen to music etc, With Cloud storage, we can store, recover and backup data, create new applications, deliver software on demand, host websites and so on. Whenever there is a demand, user can access the services of cloud dynamically via internet[2].

III. THE DATA MINING TECHNIQUES FOR ANALYZING CRIME PATTERNS

Nowadays, the various data mining techniques are used for different objectives such as: criminality, science, finance and banking, email filtering, healthcare and other industries. However, this survey focuses on the following crime types [14].

Traffic Violation and Border Control

Police Eyes [15] is the real-time traffic surveillance system that is developed to enhance the automatic detection capability of traffic violations. To extract the foreground from the back- ground in the scene obtained from IP cameras, they used the gaussian mixture model. Then the foreground extracted is used to analyze the traffic violations by using violation conditions. Cheng et al. [16] used the rough set theory and association rules to find closer relationships with the traffic offense and regular traffic violating data of huge hidden data.

In the field of border control and security, Thongsataporn- watana and Chuenmanus [17] proposed the suspect vehicle detection system using association rule to analyze the vehi- cles with forged license plate crossing the checkpoint that potentially involved the criminal activity. Reference [6] has applied association analysis by using mutual information (MI) and modified the MI formulation with the time heuristic [3] to identify the potential criminal/suspect vehicles at the border. One of the important tools for collecting data is the sensors.The data obtained from the various sensors are analyzed to detect the criminal at the borders. In addition, Geographic Information Systems (GIS) is used to help generate geographic data from the sensors. However, if the system use only the GIS techniques, the geographic data will can not be extract useful hidden knowledge. Hence, Kondaveeti et al. [18] used the GIS techniques with data mining techniques (spatial and associationdata mining techniques) to model the crime patterns and trends.

Violent Crime

Reference [1] proposed the use of naive bayes algorithm with the concept of named entity recognition (NER), also known as entity or element extraction, to classify the news articles into the crime type and to create a crime model. In addition to that, Apriori algorithm is used to find and create frequent patterns in crime by training crime data from the different web sites. For prediction in crime, they used the decision tree concept. As a tested results, their system can classify and predict the crimes more than 90% accuracy. For a crime predicting model implemented in collaboration with the police department of a United States city in the Northeast crime, the hotspots are the best method for crime forecasting proposed by [19]. To improve the accuracy of clustering technique, the segmented multiple metric similarity measure (SMMSM) is proposed by [20] that used to find the crime suspects.

The Narcotics

In the narcotics networks, the main component consists of nodes or actors and connections or relationships among them. the narcotics network is characterized which changes over time that might be from the removal and increment of the nodes and relationships. As a consequence, Kaza et al. [21] developed the predicting criminal relationship algorithms that used to predict automatically the vehicles that are a co-offender to prevent the future attacks. They used the dynamic social network analysis (SNA) methods and multivariate survival analysis by using the hazard ratios of Cox regression analysis. Reference [22] proposed the use of evolved neural networks and evolved rule-based classifiers. Both methods are useful to distinguish between toxic via narcotic and reactive mechanisms of action (MOAs) of small molecules. The CRISP-TDMn approach with support for temporal data mining, proposed by [23], is used to distinguish correlating the heart rate variability (HRV) with the respiratory rate variability (RRV) to identify the patients receiving narcotics or other drugs and the patients with imminent sepsis. They used creating momentary abstractions of hourly briefs to analyze relationships between HRV and RRV. Chau et al. [24] has focused on data collection and text extraction which these data processing is a important challenge. Therefore, they proposed a neural network-based entity extractor by using named-entity extraction techniques such as lexical lookup, machine learning, and minimal hand-crafted rules.

Cyber Crime

For the detection and prevention on cyber crime for Chi- nese web pages, Reference [25] has presented comparing the performance of the event ontology method as the priori knowl- edge and the method based on Support Vector Machine (SVM) to analyze the attributes and relations in web pages. Also these methods are used to reconstruct the scenario for crime mining. A web based crime analysis system is proposed by [26]. This system can extract the news article entities from news website, blog, etc. Then the newspaper article entities are classified as crime and non-crime articles. It has a duplicate detector used to identify exact or near duplications of newspaper articles and remove them from the database. For the crime analysis processes, the system used hot spot detection to identify the crimes and the crime frequencies. Sharma [27] proposed an improved ID3 algorithm, an enhanced feature selection method and an attribute-importance factor to classify e-mails as either maybe-suspicious or non-suspicious e-mails. Also they used a tool that is named as zero crime to help the system detect e-mails in relation to criminal activities. Framework of Market- ing or Newsletter Sender Reputation System (FMNSRS) [28] is developed from applying classification method called as sender reputation algorithm with the centralized user feedback database.

Performed Researches	**Techniques**	**Tasks**	**Research Gaps**	**Research Challenges**

Table 1.Summary Of Researches In Crime

[1]	Apriori algorithm	Extracting elements from data sources and analyzing the crime patterns	False detection	Improve precise detection
[6], [17]	Association analysis concept	Analyzing the crime patterns	No concerning with solving pro- cessing time and visualization	Improve performance of process- ing time and visualization
[18]	Spatial and association rule techniques	Analyzing the crime patterns andtrends in geographic data	No concerning with performance	Improve performance
[20]	Segmented Multiple Metric Similarity Measure (SMMSM)	Classifying attributes into similar and related groups for detecting the crime suspects	No data collection and visualiza- tion	Collect data and improve visual- ization
[21]	Cox regression	Analyzing the co-offending relation- ships using dynamic social network analysis (SNA) method and multivari- ate survival analysis	No data collection and visualiza- tion	Collect data and improve visual- ization
[24]	Neural network-based entity ex- tractor	Collecting and extracting the data ob- tained from police reports	No crime model and visualization	Model the crime future attacks and improve visualization
[25]	Event ontology and SVM algo- rithm	Analyzing the crime patterns	No flexibility of crime model and visualization	Model the crime future attacks and improve visualization
[26]	Crawler, document classifier, entity extractor, duplicate detec- tor, data base handler, analyzer and graphical user interface	Extracting element data and analyzing the crime and frequency	No crime prediction and crime model creation	Improve performance and model the crime future attacks
[27]	An enhanced Decision Tree Algorithm and an improved ID3 Algorithm with enhanced feature selection method and attribute- importance factor	Classifying e-mails in relation tocrime activities	No data collection, crime predic- tion and crime model creation	Collect data and model the crime future attacks
[28]	Sender reputation algorithm based on the centralized user feedback database	Classifying unwanted e-mails sent from attackers or spammers	No crime model creation	Model the crime future attacks

IV. ISSUES AND CHALLENGES ON CRIME

The summaries of research gaps and challenges in crime are shown in Table I.

Data Collection and Integration

In the crime analysis processes, input data is very important to used in training process and testing process. The training process is used to conduct the crime model and the testing process is used to validate the algorithm. Input data can be obtained from various sources such as news, social medias, dif- ferent sensors, criminal records obtained from the government agencies, etc. As a consequence, the collected data is large volumes of data. In additional, these data are in many formats that may be unstructured data. The collected data is stored into different databases. The issues of data collection lead to the challenge of preparation, transformation and integration of data. The many researches are concerning with solving these issues. However, one challenge is the difficulty and complication in analyzing and extracting hidden knowledge from large volumes of data. The methods may be useful to collect and integrate data such as entity extraction [1] or grouping and filtering method [29].

Crime Pattern

The issues of crime pattern are concerning with finding and predicting the hidden crime. Nowadays, the crime rate is increase continuously and the crime patterns are alway changing. As a consequence, the behaviours in crime are difficult to be explained and predicted. The research interests on crime prevention and detection are concerning with finding and conducting the crime model to detect crimes. The chal- lenge is modeling the crime attack behaviours that support crime detection although the crime patterns are changing. The predictive and statistic methods may be useful to find and conduct the crime model.

Performance

The issues on performance are concerning with precision, reliability and processing time. The uncertainty in crime patterns effects the precision of crime detection. Besides that, the algorithms used properly and the transformed data also effects the processing time. Many researches attempt to develop algorithms to detect crimes efficiently. Most of them used a combination approach. However, the challenge onperformance is developing the detecting algorithms to increase the crime detection accuracy although crime patterns are alway changing or the crime data increases continuously.

Visualization

The main responsibility of the data visualization is to create images, diagrams, or animations to provide data summarization. It can help the text data and mining results provide more interesting and more easily understood. The current issue is that the amount of data is growing rapidly, which leads to the difficulty and complication to display the hidden knowledges. One of the greatest challenges is finding out how to display the data summaries of important crime patterns and trends from huge data. To visual the low-dimensional data, there are many visualization methods used for visualization such as chart, maps, scatter diagram, coxcomb plot, etc. Additionally, the visualization for multi-dimensional data needs to use the visu- alization methods such as geometric projection, image-based visualization technology, pixel-oriented visualization methods,distortion techniques, etc. [30].

V.CONCLUSION

Crime are characterized which change over time and in- crease continuously. The changing and increasing of crime lead to the issues of understanding the crime behaviour, crime predicting, precise detection, and managing large volumes of data obtained from various sources. Research interests have tried to solve these issues. However, these researches are still gaps in the crime detection accuracy. This leads to the challenges in the field of crime detection. The challenges include modeling of crimes for finding suitable algorithms to detect the crime, precise detection, data preparation and transformation, and processing time.

VI. REFERENCES

1. S. Sathyadevan, M. Devan, and S. Surya Gangadharan, “Crime anal- ysis and prediction using data mining,” in *Networks Soft storage (ICNSC), 2019 First International Conference on*, Aug 2019, pp. 406– 412.
2. T. Pang-Ning, S. Michael, and K. Vipin, *Introduction to Data Mining*, 1st ed. Pearson, 5 2005.
3. S. Kaza, Y. Wang, and H. Chen, “Suspect vehicle identification for border safety with modified mutual information,” in *Proceedings of the 4th IEEE International Conference on Intelligence and Security Informatics*, ser. ISI’06. Berlin, Heidelberg: Springer-Verlag, 2018, pp. 308–318.
4. V. Vaithiyanathan, K. Rajeswari, R. Phalnikar, and S. Tonge, “Im- proved apriori algorithm based on selection criterion,” in *Computational Intelligence storage Research (ICCIC), 2012 IEEE International Conference on*, Dec 2012, pp. 1–4.
5. C. Chu-xiang, S. Jian-jing, C. Bing, S. Chang-xing, and W. Yun-cheng, “An improvement apriori arithmetic based on rough set theory,” in *Circuits, Communications and System (PACCS), 2017 Third Pacific- Asia Conference on*, July 2017, pp. 1–3.
6. S. Kaza, T. Wang, H. Gowda, and H. Chen, “Target vehicle identification for border safety using mutual information,” in *Intelligent Transporta- tion Systems, 2005. Proceedings. 2005 IEEE,* Sept 2005, pp. 1141–1146.
7. W. Huang, M. Krneta, L. Lin, and J. Wu, “Association bundle - a new pattern for association analysis,” in *Data Mining Workshops, 2018. ICDM Workshops 2018. Sixth IEEE International Conference on,* Dec 2018, pp. 601–605.
8. N. Sasaki, R. Nishimura, and Y. Suzuki, “Audiowatermarking based on association analysis,” in *Signal Processing, 2018 8th International Conference on,* vol. 4, Nov 2018.
9. A. Ben Ayed, M. Ben Halima, and A. Alimi, “Survey on clustering methods: Towards fuzzy clustering for big data,” in *Soft storage and Pattern Recognition (SoCPaR), 2019 6th International Conference of,* Aug 2019, pp. 331–

336.

10. A. Thammano and P. Kesisung, “Enhancing k-means algorithm for solv- ing classification problems,” in *Mechatronics and Automation (ICMA), 2018 IEEE International Conference on,* Aug 2018, pp. 1652–1656.

11. Y. Zhao and G. Karypis, “Evaluation of hierarchical clustering algorithms for document datasets,” in *Proceedings of the Eleventh International Conference on Information and Knowledge Management,* ser. CIKM ’02. New York, NY, USA: ACM, 2016, pp. 515–524. [Online]. Available:http://doi.acm.org/10.1145/584792.584877

12. C.-N. Hsu, H.-S. Huang, and B.-H. Yang, “Global and componentwise extrapolation for accelerating data mining from large incomplete data sets with the em algorithm,” in *Data Mining, 2018. ICDM ’06. Sixth International Conference on*, Dec 2018, pp. 265–274.

13. X.-M. Zhao, Y. ming Cheung, and D.-S. Huang, “Microarray data analysis using rival penalized em algorithm in normal mixture models,” in *VLSI Design and Video Technology, 2005. Proceedings of 2005 IEEE International Workshop on*, May 2005, pp. 129–132.

14. H. Chen, W. Chung, Y. Qin, M. Chau, J. J. Xu, G. Wang, R. Zheng, and H. Atabakhsh, “Crime data mining: An overview and case studies,” in *Proceedings of the 2016 Annual National Conference on Digital Government Research*, ser. dg.o ’03. Digital Government Society of North America, 2016, pp. 1–5. [Online]. Available: http://dl.acm.org/citation.cfm?id=1123196.1123231

15. R. Marikhu, J. Moonrinta, M. Ekpanyapong, M. Dailey, and S. Sid- dhichai, “Police eyes: Real world automated detection of traffic viola- tions,” in *Electrical Engineering/Electronics, Computer, Telecommuni- cations and Information Technology (ECTI-CON), 2018 10th Interna- tional Conference on,* May 2018, pp. 1–6..

16. W. Cheng, X. Ji, C. Han, and J. Xi, “The mining method of the road traffic illegal data based on rough sets and association rules,” in *Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on*, vol. 3, May 2010, pp. 856–859.

17. U. Thongsatapornwatana and C. Chuenmanus, “Suspect vehicle de- tection using vehicle reputation with association analysis concept,” in *Tourism Informatics*, ser. Intelligent Systems Reference Library,T. Matsuo, K. Hashimoto, and H. Iwamoto, Eds., vol. 90. Springer Berlin Heidelberg, 2019, pp. 151–164.

18. A. Kondaveeti, G. Runger, H. Liu, and J. Rowe, “Extracting geographic knowledge from sensor intervention data using spatial association rules,” in *Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2017 IEEE International Conference on,* June 2017, pp. 127–130.

19. C.-H. Yu, M. Ward, M. Morabito, and W. Ding, “Crime forecasting using data mining techniques,” in *Data Mining Workshops (ICDMW), 2017 IEEE 11th International Conference on,* Dec 2017, pp. 779–786.

20. G. Yu, S. Shao, and B. Luo, “Mining crime data by using new similarity measure,” in *Genetic and Evolutionary storage, 2008. WGEC ’08. Second International Conference on*, Sept 2008, pp. 389–392.

21. S. Kaza, D. Hu, H. Atabakhsh, and H. Chen, “Predicting criminal relationships using multivariate survival analysis,” in *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, ser. dg.o ’07. Digital Government Society of North America, 2007, pp. 290–291. [Online].Available:http://dl.acm.org/citation.cfm?id=1248460.1248524.

22. G. Fogel and M. Cheung, “Derivation of quantitative structure-toxicity relationships for ecotoxicological effects of organic chemicals: evolving neural networks and evolving rules,” in *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, vol. 1, Sept 2005, pp. 274–281 Vol.1.

23. C. McGregor, C. Catley, and A. James, “Variability analysis with ana- lytics applied to physiological data streams from the neonatal intensive care unit,” in *Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on*, June 2012, pp. 1–5.

24. M. Chau, J. J. Xu, and H. Chen, “Extracting meaningful entities from police narrative reports,” in *Proceedings of the 2016 Annual National Conference on Digital Government Research*, ser. dg.o ’02. Digital Government Society of North America, 2016, pp. 1–5. [Online].Available:http://dl.acm.org/citation.cfm?id=1123098.1123138

25. L. Cunhua, H. Yun, and Z. Zhaoman, “An event ontology construction approach to web crime mining,” in *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, vol. 5, Aug 2010, pp. 2441–2445.

26. I. Jayaweera, C. Sajeewa, S. Liyanage, T. Wijewardane, I. Perera, and A. Wijayasiri, “Crime analytics: Analysis of crimes through newspaper articles,” in *Moratuwa Engineering Research Conference (MERCon), 2019,* April 2019, pp. 277–282.

27. M. Sharma, “Z - crime: A data mining tool for the detection of suspicious criminal activities based on decision tree,” in *Data Mining and Intelligent storage (ICDMIC), 2019 International Conference on,* Sept 2019, pp. 1–6.

28. A. Kawbunjun, U. Thongsatapornwatana, and W. Lilakiatsakun, “Framework of marketing or newsletter sender reputation system (fmn- srs),” in *Advanced Information Networking and Applications (AINA), 2019 IEEE 29th International Conference on,* March 2019, pp. 420– 427.

29. L. Alfantoukh and A. Durresi, “Techniques for collecting data in social networks,” in *Network-Based Information Systems (NBiS), 2019 17th International Conference on,* Sept 2019, pp. 336–341.

30. H. Jin and H. Liu, “Research on visualization techniques in data min- ing,” in *Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on*, Dec 2009, pp. 1–3.