IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-10, Oct 2019 E-ISSN: 2347-2693

Analysis of Crime Detection using Data Mining Techniques

P. Dineshkumar[1*], B. Subramani[2]

[1]Dept. of Computer Science, Shri Nehru Maha Vidyalaya college of Arts and Science, Tamil Nadu, India & Dr.N.G.P. Arts and Science College, Tamil Nadu, India
[2]Shri Nehru Maha Vidyalaya college of Arts and Science, Tamil Nadu, India

**Corresponding Author: dineshkumarp@drngpasc.ac.in*

DOI: https://doi.org/10.26438/ijcse/v7i10.273279 | Available online at: www.ijcseonline.org

Accepted: 20/Oct/2019, Published: 31/Oct/2019

***Abstract* –** The In recent years the data mining is data analysing techniques that used to analyze crime data previously stored from various sources to find patterns and trends in crimes. Data Mining is the procedure which includes evaluating and examining large pre-existing databases in order to generate new information which may be essential to the organization. The extraction of new information is predicted using the existing datasets. Many approaches for analysis and prediction in data mining had been performed. But, many few efforts has made in the criminology field. In additional, it can be applied to increase efficiency in solving the crimes faster and also can be applied to automatically notify the crimes. However, there are many data mining techniques. In order to increase efficiency of crime detection, it is necessary to select the data mining techniques suitably. This paper reviews the literatures on various data mining applications, especially applications that applied to solve the crimes. Survey also throws light on research gaps and challenges of crime data mining. In additional to that, this paper provides insight about the data mining for finding the patterns and trends in crime to be used appropriately and to be a help for beginners in the research of crime data mining.

***Keywords* –** Data mining; crime analysis, crime detection, criminology; Data analysis

I. INTRODUCTION

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. Crime prevention and detection become an important trend in crime and a very challenging to solve crimes. In addition, the rapidly advancing technologies can help address such issues. However, the crime patterns are always changing and growing. The crime data previously stored from various sources have a tendency to increase steadily. Criminology is process that is used to identify crime and criminal characteristics. The criminals and the crime occurrence possibility can be assessed with the help of criminology techniques. The criminology aids the police department, the detective agencies and crime branches in identifying the true characteristics of a criminal. The criminology department has been used in the proceedings of crime tracking ever since 1800. Crimes are a social nuisance and cost our society in dearly in several ways. 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 implementations 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.

The motivation for proceeding with this survey work is to aid a helping hand to the young researchers who are performing their research in criminal analysis and crime prediction areas. The paper is organized in such a manner to provide insights about the crime analysis procedure and then produce different types of crime analysis operations and those which can be applied together for producing an end user product which can be applied to the crime analysis in any police stations

and detective agencies. This work will be a valuable reference to those who precede their research work in the crime analysis and Crime prediction using data mining techniques.

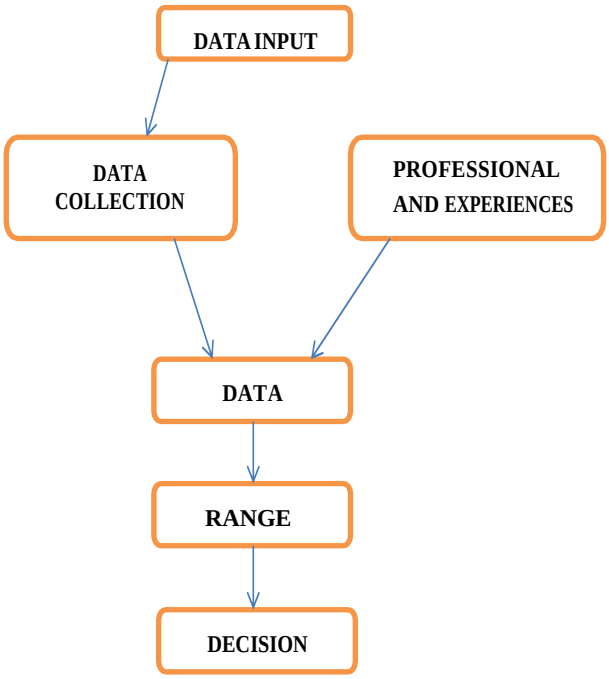

Fig 1. The Background of Data mining.

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 experiences of specialists. The knowledge results obtained from data mining processes are used to assist in decision making and to solve the problems. The data mining diagram is shown in Fig. 1. In the data mining, the analysing techniques are explained in the following sub-sections.

A. Association Rule Mining

This technique is unsupervised learning method that used to find the hidden knowledge in unlabelled data. It is used to solve the issues if the learners get the unlabelled 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 often used to dispose the uninteresting association rules. The confidence measure is used to gauge the reliability of association rules. For a given rule A□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. However, the basic finding of association rule still has the limitations. The finding processes in large item-sets need to use a long period of time. As a consequence, Apriori algorithm is used to help prune the candidates explored during frequent item-set generation to reduce the processing time. Apriori algorithm needs to scan the all item-sets. So, it uses a long period of time as well. Reference proposed the improved Apriori algorithm by using the compressed database algorithm for association rule mining to reduce the amount of time needed to read data from the database. This novel algorithm will delete the transactions that not contained in interesting item-sets. R_Apriori designed and developed by has improved Apriori algorithm to find the effective association rule and to reduce the amount of processing time. Additionally, there are several techniques that have been developed in order to analyze associations between two item sets more effectively such as mutual information concept, association bundle, audio watermarking, etc.

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. For the simple and effective clustering techniques, there are several algorithms such as K-means, Hierarchical Clustering and Expectation-Minimization that are discussed below.

1) K-means Algorithm: First, the user specifies the k centroids number. The K is the number of the wanted clusters. Each cluster must have a centroid that is a mean of a cluster. Then each data record is assigned to the nearest centroid. When all input data records have been assigned, the centroid changed of each cluster is updated by calculating the mean cluster. These processes will be repeated the assignment and

improvement the centroids until the latest centroids do not change.

2)Hierarchical Clustering: As with K-means technique, the hierarchical clustering is used to segment the similar data into the similar group by using the following methods to measure the similarity and dissimilarity such as Hamming Distance, Euclidean Distance, Minkowski Metric, etc. The viewed form is similar as the tree or sequence. Each node or cluster in the tree consists of its related member or child node which it is viewed the different layers based on data different. For generating the hierarchical clustering, there are two algorithms as follows. 1) Agglomerative algorithm has the main idea that is each object as a single cluster to be merged with the nearest pair of clusters repeatedly until a single. 2) Partitional algorithm is the idea to split a cluster repeatedly.

3) Expectation-Maximization (EM):
This is the most popular algorithm or a recursive method in statistics used to segment the incomplete data into the cluster based on probability. EM algorithm performs best when the models involve the missing value or to unknown parameters. The iterative process to find the best parameter is divided into two steps. 1) Estimation process that is used to generate the first parameter. 2) Improvement process that is a process of parameter adjustment. Then the parameter has been improved, it will be sent back to step 1 to calculate again. These processes will be repeated the estimation and improvement the parameter until the latest parameter and the parameter obtained from step 1 are very similar. In the situations where there is the huge missing data sets and many parameters, EM algorithm will converge slowly. Therefore, Hsu et al has proposed the triple jump extrapolation method to solve the performance of EM by reducing the number of iterative convergence. In additional to that, EM algorithm is low effective when there is not the class number pre-assigned appropriately. So, Rival Penalized Expectation- Maximization (RPEM) algorithm proposed by is used to solve this issue. The class number will be determined automatically.

A. 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 neighbour, Bayes' Theorem and neural network, etc.

1) Decision Tree: The main component consists of root node, internal nodes and leaf or terminal nodes. The root node is the top of tree which is chosen from important attribute in data set and used to separate records. The internal nodes contain the attributes that is test condition. Each internal node has dissimilar characteristics. The leaf or terminal nodes represent the class label or the result of prediction. In decision process, the root node is first considered by comparing the test condition. The test outcome determines the next appropriate branch to consider the next internal nodes. The internal node is considered steadily until a leaf node is reached. The class label or the result of prediction is assigned to the record.

2)Nearest Neighbour: This approach is used to find the similarity between a new test record and a train record. When a train record closest to a new test record is discovered, the class label of a new test record is defined as the same class label of a train record. These processes can classify a new test record into the same group. However, nearest neighbour approach still has the limitation. If the number of records of train set is too less, train set does not cover all the possibilities of the attributes. To improve the performance of nearest-neighbour classification, the distance measurement may be useful to solve this problem such as euclidean distance, minkowski distance and mahalanobis distance. In addition, the issue that the number of train records is more than one record closest to a new test record. K- Nearest Neighbour (KNN) method is used to solve the problem. This method will use the majority vote to find the class label.

3) Neural Network: The main objective of neural network is to generate the algorithm that has the ability to learn and recognize patterns and to deduct knowledges. The component of neural network consists of an input, a hidden and an output layer. The neuron represents the connections between an input, a hidden and an output layer and each neuron is assigned the connective weights. Each input is assigned user-specified value. The processes are introduced as

follows. When an input reaches the network, the input is computed by multiplying the connective weights. The sum of all of the connective weights is compared with the threshold. If the sum is greater than the threshold, the sum will be considered as an output and sent to output layer. The connective weights and threshold have the uncertainty. To determine the appropriate value to the connective weight and threshold, neural network can adjust these values automatically by using back propagation algorithm.

III. DATA MINING TECHNIQUES FOR DETECTING CRIME AND ITS ANALYSIS CRIME ANALYSIS PROCEDURE

Usually, the crime analysis tasks can be a tedious process for the police or the Investigation team to work with. The criminals when leaving the crime scene does leave some traces which can be used as a clue to identify the criminals. The crime sequence and the patterns which several criminals follow when committing a crime make it easy for analysing the crime. This process includes several procedures to be followed in order to identify the criminals and getting more information based only on the clues or information given by the local people. The criminal can be analysed based on the information from the crime scene which is tested against the previous crime patterns and judging by the method which is implied to test and proceed with the information that can affect the prediction results. The prediction can be further made useful for detecting the crimes in advance or by adding more cops to the sensitive areas which are identified by the system .The police stations can put up special force when there are chances for crime ahead of time. This type of the system will ensure there are peace and prosperity among the citizens.

The crime analysis can be performed procedure which is similar to figure Fig.2 which specifies each module which is used for machine learning to predict the crime or form group of clusters of criminals according to crime records. The criminals can hold certain properties and their crime characteristics and crime careers may vary from one criminal to another. Such a type of information can be taken as the input dataset. The input dataset is given to a pre-processor which performs the pre-processing based on the requirements. Once the pre-processing is completed the features or attributes from those information are extracted which may be in the form of text content from emails, the crime factors for a day, criminal characteristics, geo-location of the criminal, etc.,. The pre- processed result is further given to the classification algorithm or the clustering algorithm based on the requirements. The requirements may be anything from selecting the crime prone areas to predicting the criminal based on the previous crime records.

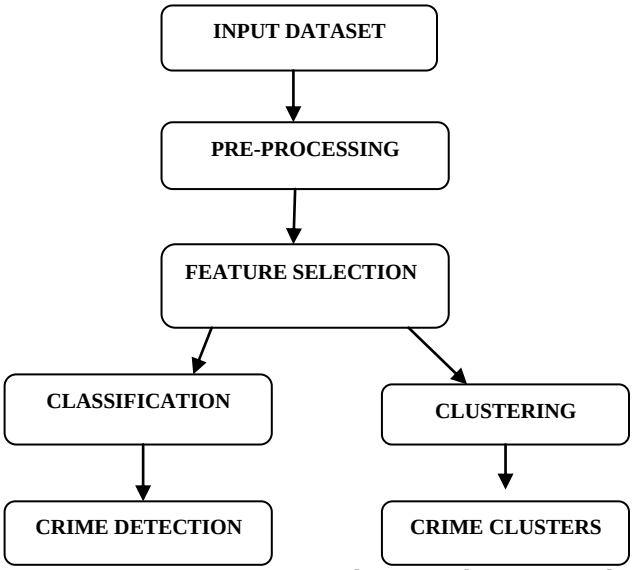

Fig 2. Crime Detection and Crime Clusters Based on input dataset.

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

A. Traffic Violation and Border Control: Police Eyes 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 background 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. 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, Thongsatapornwatana and Chuenmanus proposed the suspect vehicle detection system using association rule to analyze the vehicles with forged license plate crossing the check point that potentially involved the criminal activity. Reference has applied association analysis by using mutual information (MI) and modified the MI formulation with the time heuristic 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 analysed 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 cannot be extract useful hidden knowledge.

B. Violent Crime: Reference 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 result, 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. To improve the accuracy of clustering technique, the segmented multiple metric similarity measure (SMMSM) is proposed that used to find the crime suspects.

C.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. 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. It is 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, 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. has focused on data collection and text extraction which this 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 handcrafted rules. Text extraction from police reports to identify significant entities or useful entities can enhance the crime detection systems.

D.Cyber crime: For the detection and prevention on cyber crime, It has presented comparing the performance of the event ontology method as the apriori knowledge 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 also proposed. 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 proposed an improved ID3 algorithm, an enhanced feature selection method and an attribute-importance factor to classify e-mails as either may be-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 Marketing or Newsletter Sender Reputation System (FMNSRS) is developed from applying of classification method called as sender reputation algorithm with the centralized user feedback database. This framework can classify the unwanted emails and prevent the recipients from attackers or spammers.

IV. ISSUES AND CHALLENGES ON CRIMES

The summaries of research gaps and challenges in crime are described as follows.

A.Data Collection and Integration: In the crime analysis processes, input data is very important to use 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, different 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 analysing and extracting hidden knowledge from large volumes of data. The methods may be useful to collect and integrate data such as entity extraction or grouping and filtering method.

B.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 always 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 challenge 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. The crime model should be able to predict and detect the criminal behaviours.

C.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 on performance is developing the detecting algorithms to increase the crime detection accuracy although the crime patterns are always changing or the crime data increases continuously.

D.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 visualization methods such as geometric projection, image- based visualization technology, pixel-oriented visualization methods, distortion techniques, etc.

V.CONCLUSION

Crime are characterized which change over time and increase 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.

REFERENCES

[1] S. Sathyadevan, M. Devan, and S. Surya Gangadharan, "Crime analysis and prediction using data mining," in Networks Soft Computing (ICNSC), 2014 First International Conference on, Aug 2014, 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, 2006, pp. 308–318.

[4] V. Vaithiyanathan, K. Rajeswari, R. Phalnikar, and S. Tonge, "Improved apriori algorithm based on selection criterion," in Computational Intelligence Computing 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), 2011 Third Pacific- Asia Conference on, July 2011, pp. 1–3.

[6] S. Kaza, T.Wang, H. Gowda, and H. Chen, "Target vehicle identification for border safety using mutual information," in Intelligent Transportation 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, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on, Dec 2006, pp. 601–605.

[8] N. Sasaki, R. Nishimura, and Y. Suzuki, "Audio watermarking based on association analysis," in Signal Processing, 2006 8th International Conference on, vol. 4, Nov 2006.

[9] A. Ben Ayed, M. Ben Halima, and A. Alimi, "Survey on clustering methods: Towards fuzzy clustering for big data," in Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of, Aug 2014, pp. 331–336.

[10] A. Thammano and P. Kesisung, "Enhancing k-means algorithm for solving classification problems," in Mechatronics and Automation (ICMA), 2013 IEEE International Conference on, Aug 2013, 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, 2002, 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 component wise extrapolation for accelerating data mining from large incomplete data sets with the em algorithm," in Data Mining, 2006. ICDM '06. Sixth International Conference on, Dec 2006, 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 2003 Annual National Conference on Digital Government Research, ser. dg.o '03. Digital Government Society of North America, 2003, 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. Siddhichai, "Police eyes: Real world automated detection of traffic violations," in Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2013 10th International Conference on, May 2013, 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 detection 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, 2015, 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), 2011 IEEE International Conference on, June 2011, 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), 2011 IEEE 11th International Conference on, Dec 2011, pp. 779–786.

[20] G. Yu, S. Shao, and B. Luo, "Mining crime data by using new similarity measure," in Genetic and Evolutionary Computing, 2008. WGEC '08. Second International Conference on, Sept 2008, pp. 389–392.

