REVIEW

A Review of Data Mining Applications in Crime

Hossein Hassani1*, Xu Huang2, Emmanuel S. Silva3, and Mansi Ghodsi1

1Institute for International Energy Studies, Tehran 1967743711, Iran
2Faculty of Management, Bournemouth University, Fern Barrow, Poole BH125BB, UK
3Fashion Business School, London College of Fashion, University of the Arts London, London WC1V7EY, UK

Received 23 September 2014; revised 25 February 2016; accepted 28 February 2016
DOI:10.1002/sam.11312
Published online 21 April 2016 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Crime continues to remain a severe threat to all communities and nations across the globe alongside the sophistication in technology and processes that are being exploited to enable highly complex criminal activities. Data mining, the process of uncovering hidden information from Big Data, is now an important tool for investigating, curbing and preventing crime and is exploited by both private and government institutions around the world. The primary aim of this paper is to provide a concise review of the data mining applications in crime. To this end, the paper reviews over 100 applications of data mining in crime, covering a substantial quantity of research to date, presented in chronological order with an overview table of many important data mining applications in the crime domain as a reference directory. The data mining techniques themselves are briefly introduced to the reader and these include entity extraction, clustering, association rule mining, decision trees, support vector machines, naive Bayes rule, neural networks and social network analysis amongst others.

© 2016 Wiley Periodicals, Inc.

Keywords: big data, data mining, crime, review

1. INTRODUCTION

Crime has evolved rapidly over time, with criminals now exploiting the latest in technology not only to commit crimes but also to evade being captured. Criminal activity is no longer limited to the streets and back alleys in our neighborhoods. The Internet, which connects the entire world, is also a thriving playground for the more sophisticated criminals in the modern age. Building upon acts of terror, such as the 9/11 terrorist attacks and use of technology to hack into the most secure defense databases, the need for new and effective methods of crime prevention is increasingly significant [1]. The evolution of ‘Big Data’, which demands novel approaches towards the effective and accurate analysis of the growing volumes of crime data, has been a major challenge for all law enforcement and intelligence-gathering organizations [2].

* Correspondence to: Hossein Hassani (hassani.stat@gmail.com)

© 2016 Wiley Periodicals, Inc.

It is in this backdrop that data mining is described as a powerful tool with great potential to help criminal investigators focus on the most important information hidden within the ‘Big Data’ on crime [3]. Data mining as a tool for crime analysis is recognized as a comparatively new and highly sought after area of research [4]. This is not surprising as data mining itself is a relatively new and rapidly evolving subject, and those interested in the historical and modern definitions of data mining are referred to [5] as data mining is not concerned with estimation and tests or prespecified models but with discovering models through an algorithmic search process exploring linear and nonlinear models, explicit or not.

Alongside the increasing use of the computerized systems to track crimes, computer data analysts have begun helping law enforcement officers and detectives speed up the process of solving crimes [6] and predict crimes in advance. The popularity in the application of many data mining techniques are further influenced by the increasing availability
of Big Data and its ease of use for people who lack data analysis skills and statistical knowledge [7]. As identified by many authors, access to data plays an important role in the effectiveness of data mining in crime, but problems arise as access is hindered by privacy concerns [2,8].

This paper is aimed at providing a concise review of the data mining applications used for identifying and preventing crime over the years. It is expected that this informative review paper will be useful for introducing Data Mining techniques to crime researchers and investigators in addition to supporting and encouraging future research into developing data mining for crime analysis. In order to enable such use, the review has been organized so that interested parties could easily refer to this article alone to apprise themselves on research that has already been conducted to date and the resulting outcomes that have been attained. The main contribution of this paper is two-fold as it not only captures a majority of the significant data mining applications in crime by classifying these based on different types of techniques but also presents a concise introduction into each of the relevant data mining techniques that have been exploited for mining crime. Moreover, the review also includes, in tabular format, a summary of data mining applications in crime that can act as a quick reference guide for researchers.

According to ref. [9] the main crime-related data mining techniques are clustering, association rule mining, classification and sequential pattern mining. In line with the findings in ref. [9], our research uncovered that the following data mining techniques are most frequently adopted for crime analysis. These include entity extraction, clustering, association rule mining, decision trees, support vector machines, naive Bayes rule, neural networks and social network analysis.

The remainder of this paper is organized such that a review of the applications of data mining for crime analysis is presented in Section 2 in chronological order along with more specific explanations on the implementation of data mining techniques. The paper concludes in Section 3, whilst introductions to various data mining techniques are presented in the Appendix.

2. DATA MINING APPLICATIONS IN CRIME

In this section, we present a summary of the Data Mining applications used to detect and prevent crime. In comparison to traditional data mining techniques, the advanced techniques focus on both structured and unstructured data to detect patterns [2,10]. With the emergence of Big Data, most existing systems use a combination of data mining techniques (which have been introduced in the appendix) in order to obtain more precise and accurate extractions. In terms of crime, as a broad range of research disciplines, crime analysis can encompass a wide range of crime activities, from simple violation of civic duties to internationally organized crimes [4]. According to ref. [9], complex conspiracies are often difficult to unravel because information on suspects can be geographically diffused and span long periods of time; detecting cyber crime can likewise be difficult because busy network traffic and frequent online transactions generate large amounts of data, whilst only a small portion would actually relate to illegal activities.

In order to provide a clear view of data mining applications in crime, Table 1 is provided as a reference directory. This table summarizes information based on the data mining technique used and provides information relating to software, regions and purpose of the underlying applications. A recent review of data mining techniques used for financial accounting fraud detection up until 2011 can be found in ref. [11] and are therefore not reproduced here.

2.1. Entity Extraction

Entity extraction can be defined as the process of extracting metadata from unstructured text documents.1 Chau et al. [8] proposed a neural network-based entity extractor, which applied named-entity extraction techniques to identify useful entities from police narrative reports in 2002. They highlighted the importance of valuable information stored as text objects in criminal-justice data (i.e., the free-text police narrative reports), which are considered unstructured data. Unlike the information from structured data, these unstructured data cannot be easily accessed and used by investigators or detectives. In ref. [8], four major named-entity extraction approaches are briefly summarized and listed accordingly: lexical lookup [98], rule-based [99], statistic-based [100] and machine learning [98,101,102]. Similar to most existing information extraction systems, the entity extractor proposed by Chau et al. [8] consists of more than one of these listed approaches. More specifically, it combines lexical lookup, machine learning and minimal hand-crafted rules. By applying the neural network-based entity extractor in 36 reports randomly selected from the Phoenix Police Department database for narcotic-related crimes, the precision rates of extracting entities of persons and narcotic drug are 74.1% and 85.4%, with recall rates of 73.4% and 77.9%, respectively [8].

In order to detect what crimes may or may not have been committed by the same group of individuals, in ref. [12], a distance measure is proposed with a four-step paradigm and adaptation of the probability density function to extract entities from a collection of documents. It is then used to

1 http://www.dataversity.net/entity-extraction-and-the-semantic-web/
<table>
<thead>
<tr>
<th>Data Mining Techniques</th>
<th>References</th>
<th>Key Techniques or Softwares</th>
<th>Specific Tested Regions</th>
<th>Purpose and Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity Extraction</td>
<td>[8,10,12–20]</td>
<td>Named Entity Extraction (lexical lookup, rule-based, machine learning and land-crafted rules), SPSS LexiQuest, Natural Language Processing, Investigative Interviewing Technique, Principles of the Cognitive Interview, General Architecture for Text Engineering System, Semantic Inferentialism, Part-of-Speech Tagging, LBJ Tatter Algorithm, Conditional Random Field</td>
<td>Phoenix [8], Taiwan [17], Arabin [18], Malaysian [19], (New Zealand, Australia and India) [20]</td>
<td>Extract valuable information especially from unstructured text data (i.e., Person(also Name by Different Languages), Address, Location, Time, Vehicle, Nationality, Phone, Gender and Race, Crime Type, Personal Property, Organization, Narcotic Drug, Suspect Descriptions).</td>
</tr>
<tr>
<td>Cluster Analysis</td>
<td>[6,21–34]</td>
<td>GIS (Geographic Information System), Self-Organizing Map, Hierarchical Clustering Technique, Partitioning Clustering Technique, Concept Space Technique, Co-occurrence Analysis Algorithm, Link Analysis Technique (Shortest Path Algorithms, Priority-First-Search and Two Tree Priority-First-Search), K-Means Clustering, Heuristic Approach</td>
<td>United Kingdom [35], Tucson [26,36], Phoenix [37], Netherlands [27], Northamptonshire [28], Tucson [38], India [34,39], Italy [29], Nigeria [30]</td>
<td>Detect crime hot spots; automatically identify associations from existing crime data and weight relationships to detect the strongest association among all possible pairs of crime related entities; overcome the challenges like ‘information overload, high search complexity and heavy reliance on domain knowledge’ [38].</td>
</tr>
<tr>
<td>Association Rule</td>
<td>[40–45]</td>
<td>Apriori Algorithm, Distributed Association Rule Mining (Survey in [46]), Transformed Categorical Similarities, Dynamically Adjusted Weights, Outlier Score Function, Distributed High Order Text Mining, Temporal Association Rule, Incremental Mining, Frequent Pattern Growth</td>
<td>Richmond (Virginia) [40,41], Hong Kong [43], USA [44]</td>
<td>Link crime incidents, narrow down possible suspects, provide informative association among criminal entities or items, discover crime patterns.</td>
</tr>
<tr>
<td>Classification Techniques</td>
<td>[10,32,47–77]</td>
<td>Iterative Dichotomiser 3 Algorithm, C4.5 Algorithm, STAGE Algorithm, CART, Hunt’s Algorithm, Deceptive Theory</td>
<td>Greek [48], USA [50,51,54,56,74], Asian [62], Dutch [63], Brazil [66], Malaysia [76], San Francisco [77]</td>
<td>Efficient detection of specific criminal activities among large-sized data sets; Categorize crime data; Predict crime hot spots.</td>
</tr>
<tr>
<td>Social Network Analysis</td>
<td>[78–97]</td>
<td>K-core, Core/periphery Ratio; Measure of Centrality, Closeness and Betweenness; Center Weights Algorithm; Borgatti’s Key Player Approach;</td>
<td>Canada [81], United States [82], Richmond (Virginia, US) [96]</td>
<td>Provide analyses of functions, structures and the interaction measurements, which, in the crime domain, detect and illustrate relationships and connections of criminal entities (including individuals, groups, events, organizations, etc.); identify key members and interaction patterns between sub-groups in criminal networks; depict the suspects and their connections to other individuals or artifacts (like phone numbers, bank accounts).</td>
</tr>
</tbody>
</table>
transform a high-dimensional vector table into input for a police-operable tool. By applying this proposed distance measure, the authors employed the SPSS LexiQuest text mining tool [103] to form a table of all the entities included in each investigation. Thereafter, the transformation stage compared the investigations on common entities, the adaptation of symmetrical distance measure and probability density function with the normal distribution, and finally, a two-dimensional representation of the distances between all possible couples of investigations is obtained in order to help the crime analysts and investigators attain an overall clear understanding of all ongoing investigations.

An Online Crime Reporting System was developed in ref. [104] to extract relevant crime information from witness narratives and also to generate additional questions based on the extracted information. The proposed system combined natural language processing and an investigative interviewing technique (based on the cognitive interview principles) with the adoption of the General Architecture for Text Engineering System as the information extraction tool. By evaluating the performance on the suspect description module, an overall recall rate of 70% and 100% precision was achieved.

Ku et al. [13,14] recalled that the goal of applying information extraction techniques in crime analysis is to help investigators extract crime-related information quickly and effectively. They developed an online reporting system that is based on an information extraction technique and combined natural language processing with insights from the cognitive interview approach to obtain more information from witnesses and victims. According to ref. [13], this proposed system combines information extraction and principles of the cognitive interview with the purpose of efficiently gathering more valuable information from those victims and witnesses who are too scared or embarrassed to report crime incidents. As this system also encourages the use of natural language, it not only makes the process of reporting crime easier but also enables the gathering of more information. A large lexicon that combines a rule-based system is developed to extract crime-related entities by triggering this proposed system to ask questions according to the principles of the cognitive interview. The performances of the proposed system show significantly high precision rates (94% for police narratives and 96% for witness narratives) and recall rates (85% for police narratives and 90% for witness narratives).

Pinheiro and Furtado [15] proposed the semantic inferentialism-based natural language processing model to extract crime information from unstructured text. In addition, this framework particularly provided the adaptation of the collaborative environments on the Web. The evaluation of this framework was conducted on 100 crime-related texts on the Web, and a precision rate of 87% was achieved for extracting the crime scene alongside a 72% precision rate for extracting the type of crime.

In ref. [16], an expanded entity phrase is defined as the key component for extracting an entity. Effective performance is conducted by combining part of the speech-based template matching and ontology-driven national language processing. The preliminary results by implementation of this proposed approach on free text law enforcement data outperformed the named-entity extraction technique and reported over 80% precision and recall rates in general.

Moreover, Yang et al. [17] used the entity extraction technique combined with part-of-speech (POS) tagging for criminal information analysis and relationship visualization. They reported their method as an efficient and effective term-relationship mining technique, which showed great performance even with the most complex case of Chinese POS tagging during its implementation on criminal information data from Taiwan.

In ref. [18], a rule-based Arabic named entity recognition (NER) system is presented to identify and classify named entities in Arabic crime text. It comprises of three modules in the pre-processing stage: sentence splitting, tokenization and part-of-speech tagging; grammatical rules, patterns and gazetteer are also considered in the process of the named entities identification stage. The proposed system achieved an overall 91% precision rate and 89% recall rate when tested on the corpus of Arabic crime documents from newspapers. Recently, Alkaff and Mohd [19] used named entity recognition with gazetteers and rule-based extraction for extracting information on nationalities from crime news in Malaysia.

In ref. [20], NER (LBJTatter algorithm is chosen due to the best performance in comparison to several algorithms) was used with conditional random field (a machine-learning approach to classify a crime location sentence in an article) to extract crime information found in online news articles. Both comparisons of two newspapers in New Zealand and comparisons of crime location extraction across countries (Australia and India) were conducted with an overall 80–90% accuracy for tests on New Zealand newspapers and generally about 75% accuracy for cross-country scenarios.

### 2.2. Cluster Analysis

Cluster analysis is a technique used to group observations where observations within each group are similar, and the groups are different from each other. Cluster analysis was combined with the geographic information system (GIS) for detecting crime hotspots in [25], where the authors evaluated the performances of different clustering techniques and stated the concept of cluster significance.
as the standard to define a spot as ‘hot’ or not in crime analysis.

Adderly and Musgrove [35] combined clustering techniques and self-organizing maps to model the behavior of sex-offenders in UK. Their results indicated that crimes within a single cluster have strong similarities and may contain crimes that have been committed by the same offender.

The Coplink system is recognized as one of the successful implementations of the clustering technique for crime data mining [2,26,36,37]. According to Hauck et al. [26], Coplink applies the concept space as its underlying structure to identify relationships between suspects, victims and other relevant data in order to accelerate criminal investigations and enhance law enforcement efforts. To elaborate, the concept space algorithm is stated as a statistics-based algorithmic technique that can automatically compute the strength of relationships between each possible pair of concept descriptors identified in a document collection [26]. As most criminal-justice data are properly structured in the existing reports, the sources that a concept space can be derived from are located. All the terms in this located concept space are then filtered and indexed by an appropriate co-occurrence analysis algorithm. In ref. [26], terms are categorized into five main types: Person, Organization, Location, Crime and Vehicle, and according to feedback from crime analysts and detectives, Coplink can provide valuable leads by discovering links between different types of terms and can provide the crucial advantage of operating efficiency in the investigating process.

In ref. [105], a link analysis technique is proposed to identify the strongest association paths between entities in a criminal network. In comparison to the existing link analysis tools used in crime analysis (for example, Anacapa charting system [106], Netmap [107], Analyst’s Notebook [108], Watson [109] and Coplink Detect [26]), the authors employed shortest-path algorithms and logarithmic transformation of the link weights to identify the shortest path; therefore, the strongest associations can be detected among the crime-related entities.

De Bruin et al. [27] describes a tool that extracts important factors that play a role in the analysis of criminal careers from databases and creates digital profiles for all offenders. Their method compares all individuals on these profiles using a new distance measure and clusters them accordingly, which, in turn, yields a visual clustering of these criminal careers and enables the identification of classes of criminals [27]. The clustering technique was employed again in [12] in an investigation process to identify the offenses that were committed by the same group of criminals.

K-means clustering was used in [28] to assess the performance of crime scene investigators. By modeling their performance, a clear and objective rationale is provided for further development as a viable path for improving crime investigation. Schroeder et al. [38] proposed the Crimeless Explorer system, which combined several techniques, including co-occurrence analysis, shortest-path algorithm and the heuristic approach, to provide a better link analysis to assist crime investigations. In terms of enhancing investigation effectiveness for Indian police, Gupta et al. [39] proposed a crime analysis tool with an interactive query-based interface to help the police in crime investigations. Furthermore, the clustering technique is adopted to find crime hotspots based on the National Crime Record Bureau database in India.

More recently, in ref. [29], cluster analysis was used to identify the effect of various economic factors on crime in Italy, while in ref. [30], a GIS is complemented with multivariate cluster analysis for assessing property crime in Nigeria. Self-organizing map and multilayer perceptron neural networks are employed with the clustering technique in ref. [31] for crime analysis and matching. Sukanya et al. [32] applied the clustering technique with GIS to identify criminals and crime hotspots by analyzing the spatial pattern. K-means clustering was used in combination with the rapid miner tool in ref. [33] for crime analysis of offences recorded in England and Wales. A spatial-temporal analysis combining GIS and the clustering technique was adopted in ref. [34] for day-to-day crime forecasting in India.

2.3. Association Rule Mining

Association rule mining is a method that exploits the relationships among observations for uncovering crucial information hidden within Big Data. In ref. [21], the basic and advanced concepts and algorithms of association analysis are provided with detailed calculation and examples. In ref. [110], an association rule mining algorithm is proposed by using the sales data of customer transactions obtained from a large retailing company. Association rule mining is designed for applications where observations consist of transactions, and a subset of the available items appears in each transaction. Later, in ref. [111], two new algorithms are presented and compared with the known algorithms, and significant improvements are recorded by both new algorithms and their combined AprioriHybrid algorithm.

In ref. [40], association methods are employed for applications in law enforcement. The proposed approach automatically looks for similarities by using a new total similarity measure with information theoretic-based weights among attributes of robbery records data from the Richmond Police Department in order to associate incidents possibly committed by the same or group of criminals. Thereafter, Lin and Brown presented a new association
method that combined outlier score function and tested it on the same robbery data from the Richmond Police Department in ref. [41]. The new outlier-based method outperforms the similarity-based method, with promising results in providing more helpful information for police officers.

A distributed higher-order association rule mining algorithm is proposed by Li et al. [112] with example law enforcement data, which overcomes the requirements of knowledge of a global schema and the distribution of data being horizontal or vertical. In ref. [42], association rule mining is used for detecting suspicious e-mails by identifying unusual and deceptive communication in e-mails with the implementation of an Apriori Algorithm. This proposed technique was designed to assist the investigators in efficiently obtaining information and taking effective actions to reduce or prevent criminal activities. Ng et al. [43] introduced an incremental algorithm for maintaining temporal association rules with numerical attributes by using the negative border method. Their new algorithm has been implemented to support the discoveries of crime patterns in a district of Hong Kong, and the results showed significant improvements.

Fuzzy association rule mining was introduced in ref. [44] as a novel means for knowledge discovery in the crime domain, supported by experimental results on open-source communities and crime datasets [113] in USA. The most significant result is that crime patterns are discovered and also consistent across all regions, subsets of regions and all states, which proves the capability and potential benefit of implementing this proposed technique for crime analysis and police force investigations. A review of more applications of association rule mining for crime detection up until 2014 can be found in ref. [45] and are therefore not reproduced here.

2.4. Classification Techniques

As one of the most fundamental and significant data mining techniques, Classification techniques are used for classifying observations based on some significant rules/attributes that are discovered from the database. In terms of applying classification techniques in crime data mining, many implementations combined more than one specific type of classification technique. Therefore, the review that follows is classified by each technique in chronological order, and depending on circumstances, those complex combination cases are not reproduced.

2.4.1. Decision Trees

The classification technique, also known as decision trees, is used in ref. [47] for detecting suspicious e-mails and reported over 95% accuracy in correctly classifying e-mails in a large-sized dataset. The model of deceptive theory is applied to the dataset of e-mails, and the decision tree is generated via the ID3 algorithm. Kirkos et al. [48] explored and compared the performances of different classification techniques for auditors in detecting firms that issue fraudulent financial statements. Three models (decision trees, neural networks and Bayesian belief networks) are evaluated by dealing with the identification of factors associated with fraudulent financial statements on datasets from 76 Greek manufacturing firms. In ref. [49], decision trees, neural networks, support vector machines and stochastic boosting are utilized for hotspot detection in an urban development project dataset, which contains 1.4 million cases, 14 predictors and a binary response variable. Decision trees are adopted for predicting crime reporting in ref. [50] by identifying the variables that influence whether a crime is reported from survey data obtained via the Bureau of Justice Statistics of USA, while in ref. [51], classification techniques like decision trees, neural network, naive Bayes and support vector machines are applied and compared for predicting crime hotspots and crime forecasting. In ref. [52], a reduction strategy of attribute is combined with the decision trees algorithm for analyzing criminal behavior. Naive Bayesian, rule-based classification and C4.5 decision trees are exploited in ref. [53] for detecting auto insurance fraud. The decision trees technique is particularly compared with other data mining techniques for automotive insurance fraud detection in ref. [54]. In ref. [55], a decision tree-based classification model is used for discovering crime patterns and predicting future trends. Recently, in ref. [56], classification algorithms are evaluated for crime prediction where they find decision trees outperforming a naive Bayesian algorithm in accurately predicting the crime category for different states in USA. An improved decision tree algorithm based on the Maclaurin-Priority Value First method is used in ref. [57] for computer crime forensics, and this improved algorithm outperformed the commonly used ID3 algorithm in terms of both efficiency and accuracy. More recently, a variety of classification techniques were evaluated for detecting insurance fraud in ref. [58] where an algorithm for determining the relationship between classification models and different types of insurance fraud data was proposed.

2.4.2. Neural Networks (NN)

Artificial neural networks (ANNs), decision trees and logistic regression are used in ref. [59] for uncovering lies from 371 statements of different types of crimes. In ref. [60], logistic regression and ANNs are applied to identify smuggling vessels, and the findings indicate that the ANN reports a higher accuracy than logistic regression.
Multilayer perceptron (MLP), Decision trees, support vector machines, genetic programming (GP) and probabilistic neural networks were compared in ref. [61] for detecting phishing e-mails from a dataset of 2500 e-mails, whilst in ref. [62], decision trees, ANNs and support vector machines were compared for discriminating between those charged and not charged for initial juvenile offending. Logistic regression, boosting, SVM, neural networks, K-nearest neighbor and a variety of other data mining techniques are evaluated for predicting recidivism in ref. [63].

2.4.3. Support Vector Machines (SVM)

The SVM classification has been used to identify the sources of e-mail spamming based on the sender’s linguistic patterns and structural features [10, 64]. An SVM model is successfully used in [64] to aid author identification forensics via mining e-mail content. SVM is used for crime hotspot prediction in ref. [65], where they find it outperforming neural networks and spatial auto-regression-based approaches. In ref. [66], SVM is exploited for crime scene classification using a database comprising of 400 crime scenes. The authors find SVM-based approaches performing better than Multilayer perceptron neural networks. Liu et al. [67] used SVM to help with identifying digital evidence relating to computer crimes via outlier detection. In ref. [68], SVM was used to help crime investigators produce a set of possible interpretations for domain-relevant concepts through kernel-based relation extraction. Wang et al. [69] used the SVM technique for predicting criminal recidivism and compared their results with those obtained from logistic regression and neural networks. Salem et al. [70] evaluated the performance of SVM for detecting identity theft using the Schonlau dataset, and whilst they found one-class SVM models to be practical in accurately locating identity theft in the general case, the results were not satisfactory where specific user profiles fit the average user profile. Advanced fee fraud activities are detected using both SVM and random forests (RF) in ref. [71], where they find SVM outperforming RF2. Moreover, an SVM model is used alongside logistic regression and RF for detecting credit card fraud in [72], where the performances are evaluated on real world transaction data from international financial institutions. The SVM technique, combined with the AdaBoost algorithm, is used in ref. [73] for cyber-crime detection and prevention based on a Facebook dataset. Alwee et al. [74] exploited a hybrid support vector regression (SVR) model in combination with ARIMA and particle swarm optimization for forecasting property crime rates in USA and found it outperforming forecasts from the individual models, while later in [75], SVR was employed for forecasting property crime rates along with grey rational analysis. The behavior of criminals are analyzed using SVM in ref. [76] for Malaysia, where the ratio of police to population is extremely low at 3.6 to 1000. More recently, it has been shown in ref. [77] that an SVM model can be used to provide live predictions of crime in urban areas based on Twitter data related to an urban subarea from within the city of San Francisco.

2.5. Social Network Analysis (SNA)

Social network analysis (SNA) is a method built on the analysis of social structure of observations for identifying significant information. Sparrow [78] summarized existing concepts of network analysis in the applications of the crime analysis domain with detailed explanations and comparisons. Wang and Chiu [79] applied SNA and identified two transactional network structure measurements, k-core and core/periphery ratio, to detect the online auction inflated-reputation traders from regular accounts. Significant results proved that SNA can act as an effective indicator to distinguish criminal accounts and possibly prevent and reduce problematic transactions and online auction frauds. Qin et al. [80] proposed an analysis of the structure of the Global Salafi Jihad network with SNA and Web structural mining. Results showed that the proposed technique can be an effective tool to identify key members in a terrorist network and, therefore, help authorities develop efficient and effective disruptive strategies and measures.

In ref. [81], the SNA technique was used to investigate organized criminal groups by applying it to an outlaw motorcycle gang operating in Canada. Ressler [82] introduced SNA as a new type of intelligence for homeland security in USA, which can provide important information on the unique characteristics of terrorist organizations, understand terrorist networks and form the basis for a more effective countermeasure to net war. In ref. [83], SNA was employed on the Enron corporation’s e-mail archive, and it proved to be able to analyze and catalog patterns of communications between entities in an e-mail collection to extract social standing. The authors stated that this technique makes it possible to view a snapshot of a corporate community and effectively determine the real relationships and connections between individuals.

Chau and Xu [84] applied the web mining and network analysis technique to study hate groups in online blogs. Their proposed approach successfully identified and analyzed a selected set of hate groups (which contained 820 bloggers) on Xanga. Nair and Sarasamma [85] performed SNA alongside fuzzy theory with the aim of modeling
multi-modal social networks. A new fuzzy binary operation was proposed to satisfy the requirements of a fuzzy consolidation operator. Liu et al. [86] proposed an enhanced shortest-path algorithm, combing SNA to mine the core member of a terrorist group. Wang and Chiu [77] used two SNA indicators, k-core and center weights algorithms, to form a recommendation system that can suggest the risks of collusion associated with an account for online auction site users. The results are promising, with 76% detection accuracy on real world ‘blacklist’ data and sound evidence that the approach can provide effective warnings several months ahead of the official release of blacklists.

In ref. [88] the authors discussed the solution for the case of privacy-preserving SNA in collaboration with multiple investigators. The proposed algorithm, compute metrics, can operate the computation and analysis without knowing the personally identifiable data with high privacy guarantees. Chau and Xu [89] again addressed the importance of studying the emergence of cyber communities in blogs with a combination of web mining and SNA techniques. In ref. [90], the authors designed a simulation e-mail system based on personality trait dimensions to model the traffic behavior of e-mail account users and employed this with SNA to identify key members of a criminal group. Chen [91] shows the applicability of SNA for homeland security data mining with case studies based on gang/narcotic networks, US extremist networks, Al-Qaeda member networks and international Jihadist websites and forum networks.

Fard and Ester [92] proposed a framework for automated network data analysis and deduction from multiple social networks by converting a transaction dataset and applying association mining and statistical methods. The proposed method varies from previous work and combined the game theory concept in a multi-agent model with the purpose of building P2P applications for the police force to detect relationships between criminals and narrow down possible suspects. In ref. [93], the authors proposed an SNA-based model for targeting criminal networks. It is built on Borgatti’s key player approach with modifications on incorporating the relative strength of actors as well as the strength of the relationships binding network actors. Chiu et al. [94] employed SNA for identifying internet auction fraud. They performed experiments on data from the Yahoo Auctions website with comparisons of different types of internet auction accounts and achieved promising results pertaining to prediction accuracy.

In ref. [95], Carrington classified the applications of SNA in the crime domain into three areas (the influence of the personal network on ego’s delinquency or crime, the influence of neighborhood networks on crime in the neighborhood and the organization of criminal groups and activities) and provided detailed explanations together with significant theories and literature. The SNA technique was used as a tool for helping crime analysts and detectives in police forces develop interdiction strategies in ref. [96], with an example case study focusing on the Richmond City Police Department. Detailed examples of implementations are provided to show how the proposed approach can assist police in understanding complex behavioral motivations of offenders, strategically hot-spotting people of interest and developing stronger inter-jurisdictional working relationships [96]. More recently, in ref. [97], SNA was used to detect and analyze hidden activities in social networks. Both regression-based and Min Cut-based algorithms are adopted and compared with datasets from several sources.

3. CONCLUSION

This review paper begins with a brief introduction into the evolution of crime and the importance of data mining for crime detection, analysis and prevention. Given the vast amount of funding that is allocated for defense-related expenditure around the globe, it is clear that considerable savings could be attained through the correct application of data mining techniques, which are primarily aimed at uncovering hidden relationships in Big Data. Following thorough research, we are able to present a list of data mining techniques as the most frequently adopted at present for crime analysis. These include entity extraction, clustering, association rule mining and classification techniques like decision trees, support vector machines, naive Bayes rule, neural networks and social network analysis.

As opposed to providing a review of applications alone, this paper goes a step further and provides a concise introduction to these data mining techniques (see Appendix) as well as a brief summary of the specific functions of each technique in crime analysis and thereby enables readers from different backgrounds to obtain a much richer understanding of the underlying process whilst guiding them to relevant articles for a more detailed description. Table 1 particularly functions as a useful resource or ‘quick guide’, which summarizes the data mining applications in crime, providing useful information not only on the software and purpose but also the regions that have been tested.

This review itself is unique as it captures over 100 applications and is the most up-to-date and thorough review of data mining applications in crime to date. We find strong evidence based on the number of applications to conclude that classification techniques are the most popular form of data mining in crime. This is interesting as there exists a major difference between the findings from data mining in official statistics [5] and our findings here. We notice that SVM, neural networks and association rule mining are seldom used to mine official statistics,
whereas in crime data mining, these methods are extremely popular and well exploited. This further highlights the institutional and cultural differences that exist between national statistical institutes and organizations engaged in analyzing crime.

The increasing applications of data mining techniques alongside the emergence of Big Data also points toward the need for more training and investment in educating and empowering youth with knowledge on the advantages, developments and practical use of data mining techniques. In countries such as the UK, it is encouraging to see universities now offering not only units but also courses on data mining, which will most certainly have a positive impact on future generations as well as lead towards an exponential increase in data mining applications. We believe that this review paper can help researchers around the globe easily identify existing research so as to develop more complex and improved techniques for mining crime data in the future.

**APPENDIX**

**A. DATA MINING TECHNIQUES IN CRIME**

This appendix aims to give the reader an introduction to the data mining techniques that have been used for crime analysis.

**A.1 Entity Extraction**

Entity extraction is a process that depends on the availability of extensive amounts of clean input data for the identification of particular patterns such as text, images or audio materials that can then provide basic information for crime analysis [2,10]. The Message Understanding Conference (MUC) series, which coordinates multiple research groups and government agencies [115], has been the major forum for researchers in this area, where they meet and compare the performance of their entity extraction approaches [116] and seek to improve the technologies [115]. As concluded in ref. [8], the main approaches for entity extraction are lexical-lookup, rule-based, statistic-based and machine learning. Below, we provide some basic introductions to the major approaches, and in doing so, we mainly follow ref. [8].

**Lexical lookup** Most entity extraction systems maintain hand-crafted lexicons that contain lists of popular names for entities of interest. These systems work by looking up phrases in text that match the items specified in their lexicons. See, for example, [98].

**Rule-based** The rule-based system relies on hand-crafted rules to identify named entities. The rules may be structural, contextual or lexical [99].

**Statistic-based** Statistical-based systems often use statistical models to identify occurrences of certain cues of particular patterns for entities in texts. A training dataset is needed for these systems to obtain the statistics. The statistical language model in ref. [100] is an example of statistical-based entity extraction systems.

**Machine learning** Instead of human-created rules, the machine learning system relies on what is termed as machine learning algorithms to extract knowledge and identify patterns from texts. Few examples are neural networks, decision trees [101], hidden Markov models [102] and entropy maximization [98].

**A.2 Cluster Analysis**

According to ref. [21], cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships. The key objective of this process is to find groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups [21]. The fundamental concept of clustering is the distance measurement among objects. Euclidean distance, Manhattan distance and Minkowski distance are the most widely adopted distance measures for cluster analysis. In order to identify different types of structures or clusters in the data, different algorithms have been developed to date, which can be mainly classified into hierarchical methods and partitioning methods based on how the subsets are organized [24]. Other methods, except the two major approaches, include the density-based method, grid-based method, model-based method and constraint-based method. For the sake of briefly introducing the process of cluster analysis without covering all available algorithms to date, we introduce one specific algorithm of cluster analysis below by following ref. [22]. Note that alternatives exist for each step due to the significant developments in cluster analysis. More details can be found in refs. [21–24].

Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}^l$, which we wish to cluster into $k$ subsets, the first stage will be the ordination of the data (includes steps 1–4) followed by the second stage, which focuses on clustering (includes steps 5–6):

1. From the affinity matrix $A \in \mathbb{R}^{n \times n}$ defined by $A_{ij} = \exp(-\|s_i - s_j\|_2^2 / 2\sigma^2)$ if $i \neq j$, and $A_{ij} = 0$. Note that $\|s_i - s_j\|$ refers to the normalized distance between $s_i$ and $s_j$.
2. Define $D$ as the diagonal matrix whose $(i,i)$-element is the sum of $A$’s $i$-th row, and construct the matrix $L = D^{-1/2}AD^{-1/2}$.
3. Find $x_1, x_2, \ldots, x_k$, the $k$ largest eigenvectors of $L$ (chosen to be orthogonal to each other in the case of repeated eigenvalues) and form the matrix $X = [x_1x_2\ldots x_k] \in \mathbb{R}^{n \times k}$ by stacking the eigenvectors in columns.
4. Form the matrix $Y$ from $X$ by re-normalizing each of $X$’s rows to have unit length (i.e., $Y_{ij} = X_{ij}/(\sum_i X_{ij}^2)^{1/2}$).
5. Treating each row of $Y$ as a point in $\mathbb{R}^k$, cluster them into $k$ clusters via $K$-means or any other algorithm (that attempts to minimize distortion).
6. Finally, assign the original point $s_i$ to cluster $j$ if and only if row $i$ of the matrix $Y$ was assigned to cluster $j$.

**A.3 Association Rule Mining**

Association rule mining was initially proposed by Agrawal et al. [110] as a method of discovering interesting co-occurrences in supermarket data. The association rule is defined as a technique for investigating the possibility of the simultaneous occurrence of data [117]. Yun et al. [117] suggested an association rule discovery technique that is able to identify significant, rare data associated with specific data in a way that the rare data occurs simultaneously with the specific data more frequently than the average co-occurrence frequency in the database. We follow ref. [111] and
briefly introduce the association rule mining technique below. For a more detailed explanation, see ref. [110].

Let \( I = \{i_1, i_2, ..., i_m\} \) be a set of literals, called items. Let \( D \) be a set of transactions, where each transaction \( T \) is a set of items such that \( T \subseteq I \). Associated with each transaction is a unique identifier, called \( TID \). We say that a transaction \( T \) contains \( X \), a set of some items in \( I \), if \( X \subseteq T \). An association rule is an implication of the form \( X \Rightarrow Y \), where \( X \subseteq I \), \( Y \subseteq I \), and \( Y \cap X = \emptyset \).

More specifically, in ref. [21], the strength of an association rule can be measured in terms of its support and confidence, in which support determines how often a rule is applicable to a given dataset, while confidence determines how frequently items in \( Y \) appear in transactions that contain \( X \). The formal definitions of these measures are as follows, in which \( N \) refers to the number of transactions:

\[
\text{Support}(\sigma(X \Rightarrow Y)) = \frac{\sigma(X \cup Y)}{N}. \quad \text{(A1)}
\]

\[
\text{Confidence}(c(X \Rightarrow Y)) = \frac{\sigma(X \cup Y)}{\sigma(X)}. \quad \text{(A2)}
\]

Therefore, it can be concluded that support indicates that the rule \( X \Rightarrow Y \) holds in the transaction set \( D \) with confidence \( c \) if \( \% \) of transactions in \( D \) contain \( X \cup Y \); and confidence indicates that the rule \( X \Rightarrow Y \) holds in the transaction set \( D \) with confidence \( c \) if \( \% \) of transactions in \( D \) that contain \( X \) also contain \( X \cup Y \). Given a set of transactions \( D \), the problem of association rule mining is to generate all association rules in the database that have support and confidence greater than the user-specified minimum support (called \textit{minsup}) and minimum confidence (called \textit{minconf}), respectively [111]. There are many algorithms for association rule mining, among which, the most well-known and widely-used Apriori algorithm was proposed in refs. [110,111] as a seminal algorithm for identifying support using candidate generation. Wu et al. [118] stated that the Apriori algorithm is one of the most popular data mining approaches to find support from a transaction dataset and derive association rules. Other algorithms, like direct hashing and pruning (DHP), FP-growth and H-mine, can be found with more details in [119–122].

### A.4 Classification Techniques

As one of the most fundamental and significant data mining techniques, classification is briefly defined in ref. [21] as the task of assigning objects to one of several predefined categories. The aim of classification rule mining is to discover a small set of rules in the database to form an accurate classifier [123]. The formal description of the algorithm underlying the classification rule is listed below by mainly following ref. [21].

The input data for a classification task can be regarded as a collection of records that contains the instances that are characterized by a tuple \((X, y)\) (note that \( X \) is the attribute set, \( y \) is a special attribute, also known as category). Therefore, classification can be explained as the task of learning a target function \( f \) that maps each attribute set \( X \) to one of the predefined class labels \( y \) (see Figure A1). In order to evaluate the performance of a classification model, the counts of correct and incorrect predicted test records are summarized in a table named a confusion matrix. Then, the performance metric, which summarizes the information in the confusion matrix into a single number, provides an easier way for comparing the performance of different classification models [21].

### Decision Trees

Decision trees ([124–126]) are applied to accomplish the classification task by giving a series of carefully crafted questions about the attributes of the test record [21]. When an answer is achieved, it will be followed by a question until the category of this attribute is concluded. All the series of questions and possible answers are carefully predefined as well as the process repeated to all subsets of the tree. The classification task is established at the same time a decision tree is constructed. Note that the training records are split based on an attribute test that optimizes some certain criterions; additionally, it is possible that many different decision trees can be conducted from the same given set of cases. The final choice depends on the research and the individual circumstances. Many algorithms have been developed for getting the most reasonably optimal decision tree with good accuracy in a timely manner. For example, CART [124], C4.5 [125], ID3 [126], Hunt’s Algorithm [127], SLIQ [128] and SPRINT [129].

Among all the effective algorithms, Hunt’s algorithm is the basis of many existing decision tree-induction algorithms, by which the decision tree is constructed via recursively partitioning until each path ends in a pure subset [21]. The process of Hunt’s algorithm is listed below, mainly following [21]. In addition, one example problem of predicting whether a loan applicant will succeed in repaying her loan obligations or become delinquent is adopted for applying Hunt’s algorithm and inducing decision trees in Figure A2. Let \( D_t \) be the set of training records that are associated with node \( t \), and \( y = y_1, y_2, ..., y_r \) be the class labels.

#### Step One

If all the records in \( D_t \) belong to the same class \( y \), then \( t \) is a leaf node (which refers to a node that has exactly one incoming edge and no outgoing edges), and it is labeled as \( y \).

#### Step Two

If \( D_t \) contains records that belong to more than one class, an attribute test condition is selected to recursively partition the records into smaller subsets. A child node is created for each outcome of the test condition, and the records in \( D_t \) are distributed

---

Statistical Analysis and Data Mining: The ASA Data Science Journal DOI:10.1002/sam
to the children based on the outcomes. The algorithm is then recursively applied to each child node.

**Support Vector Machines**

Support vector machines (SVM) were initially introduced in ref. [130] as a training algorithm for classification, which divides the objects into two classes by using an optimal separating hyperplane that minimizes the classification error (see Figure A3). We briefly present the theory underlying SVM below by following [130].

Consider the simplest problem of separating a set of training vectors $D$ belonging to two separate classes:

$$D = \{(x^1, y^1), ..., (x^t, y^t)\}, \text{ where } x \in \mathbb{R}^n, y \in \{-1, 1\}. \quad (A3)$$

The hyperplane that geometrically separates the surfaces of two classes is defined as $<w, x > + b = 0$, where $w$ is the weight vector and $b$ refers to the bias. Additionally, the parameters are constrained by $\min \{<w, x^i > + b \} = 1$ [130]. A separating hyperplane in such canonical form must satisfy the following constraint: $y^i [<w, x^i > + b] \geq 1$, where $i = 1, ..., t$.

The set of vectors is said to be optimally separated by the hyperplane if it is separated without error, and the distance $d$ between the closest vectors to the hyperplane is maximal, which can be computed by

$$d(w, b, x) = \frac{|<w, x^i > + b|}{||w||}. \quad (A4)$$

The optimal separating hyperplane is attainable via maximizing $\rho$ (the margin) subject to the constraints in the canonical form of the hyperplane

$$\rho(w, b) = \min_{x^i, y^i = -1} d(w, b, x^i) + \min_{x^i, y^i = 1} d(w, b, x^i) = \frac{2}{||w||}, \quad (A5)$$

and minimizes

$$\phi(w) = \frac{1}{2} ||w||^2. \quad (A6)$$

**Naive Bayes Rule**

The naive Bayes classifier was proposed in Langley et al. [131] and uses Bayes Rule to compute the probability of each class given the instance,

$$\begin{align*}
\mathbf{w} \cdot \mathbf{x} + b &= 0 \\
\mathbf{f}_2 &= +, - \\
\mathbf{f}_1 &= \text{classification variable, and let } c \text{ be the value of } C. \text{ Note that here, we assume that there are only two classes: + (positive class) or - (negative class).}
\end{align*}$$

According to Bayes Rule, the probability of an example $E$ being class $c$ is:

$$p(c \mid E) = \frac{p(E \mid c)p(c)}{p(E)}. \quad (A7)$$

where $E$ is classified as the class $C = +$ if and only if

$$f_b(E) = \frac{p(C = + \mid E)}{p(C = - \mid E)} \geq 1, \quad (A8)$$

where $f_b(E)$ is called a Bayesian classifier.

Assume that all attributes are independent given the value of the class variable, that is:

$$p(E \mid c) = p(x_1, x_2, ..., x_n \mid c) = \prod_{i=1}^{n} p(x_i \mid c). \quad (A9)$$

Then the resulting classifier is:

$$f_{ab}(E) = \frac{p(C = +)}{p(C = -)} \prod_{i=1}^{n} p(x_i \mid C = +). \quad (A10)$$

The function $f_{ab}(E)$ is so called a naive Bayesian classifier (or Naive Bayes). Therefore, the $p(x_1 \mid c), p(x_2 \mid c), ..., p(x_n \mid c)$ can be estimated from a training sample, and the posterior probability for each class can be calculated respectively. The class with the highest posterior probability will be the prediction class.

**Neural Networks**

As one of the most important tools for classification, neural networks has proved its high tolerance to noisy data and the ability to classify untrained patterns with convincing evidence through recently published research. According to ref. [134], neural networks are able to estimate the posterior probabilities, which provides the basis for establishing classification rule and performing statistical analysis [134–138]. Effective and successful performances have been achieved by neural networks in many real-world classification tasks [139]. The algorithm of neural networks is introduced below by mainly following [135].

Consider a general $M$-group classification problem in which each object has an associated attribute vector $x$ of $d$ dimensions. Note that $\omega$ denotes the membership variable that takes a value of $\omega_j$ if an object belongs to group $j$. Neural network for a classification problem can be viewed as a mapping function

$$F: \mathbb{R}^d \rightarrow \mathbb{R}^M, \quad (A11)$$

where $d$-dimensional input $x$ is submitted to the network, and an $M$-vectors output $y$ is obtained to make the classification decision. The network is built so that an overall error measure, such as

Statistical Analysis and Data Mining: The ASA Data Science Journal DOI:10.1002/sam
as the mean squared error, is minimized. According to ref. [135], the mapping function \( F: x \rightarrow y \) that minimizes the expected squared error \( E[y - F(x)]^2 \) is the conditional expectation of \( y \) given \( x \) \( F(x) = E[y | x] \).

Hence, the \( j \)th element of \( F(x) \) is given by

\[
F_j(x) = E[y_j | x] = 1 \cdot P(y_j = 1 | x) + 0 \cdot P(y_j = 0 | x) = P(y_j = 1 | x) = P(\omega_j | x).
\]

That is, the least squares estimate for the mapping function in a classification problem is exactly the posterior probability. The mean squared error function can be derived as

\[
\text{MSE} = \sum_{j=1}^{M} \int P(\omega_j | x) f(x) dx + \sum_{j=1}^{M} P(\omega_j | x) (1 - P(\omega_j | x)) f(x) dx.
\]

The first term of the right side, termed as the estimation error, is affected by the effectiveness of neural network mapping, while the second term is independent of neural networks and reflects the inherent irreducible error due to randomness of the data [135].

### A.5 Social Network Analysis

Social network data is now a significant resource for criminal investigations. According to Mena [140], social network analysis (SNA) is a data mining technique that reveals the structure and content found in a body of information by representing the same as a set of interconnected, linked objects. The basic structure of a social network consists of nodes that are connected to other related nodes by links. The connections between two nodes is called an edge, which indicates that two persons participated in a certain event [141]. The measurement techniques used in social network analysis are based on the principles of graph theory, which consists of various mathematical formulae and concepts for the study of pattern of lines [141] (for more details see, [143–145]). Among all the measurement techniques, degree, density and centrality are the most significant and widely used techniques.

**Degree** The ‘degree’ of any node of the network is defined as the number of other nodes to which it is directly linked [78]. It simply represents the scores or count of the number of ties with other actors in the network.

**Density** Density is defined as the number of edges in a portion of a social network to the maximum number of edges that theoretically make up the social network [104]. Assume \( C \) is the number of observed edges and \( n \) is the total number of nodes in the social network, then the formula of density is, \( \text{Density} = \frac{C}{\binom{n}{2}} \).

**Centrality** According to ref. [146], centrality is the measure of closeness of a node to the center of high activity in a network, which indicates the structural importance level of the node. The centrality can be evaluated by degree, closeness and betweenness: Degree centrality refers to the direct count of the number of connections a node has to other nodes; Closeness centrality is considering the overall closeness of a node to all other nodes in the network [147]; Betweenness centrality measures the extent of a node’s placement on the shortest path between other pairs of nodes in the network [145].

### REFERENCES


Statistical Analysis and Data Mining: The ASA Data Science Journal DOI:10.1002/sam


Statistical Analysis and Data Mining: The ASA Data Science Journal DOI:10.1002/sam


[54] A. Gepp, J. H. Wilson, K. Kumar, and S. Bhattacharya, Comparative analysis of decision trees vis-à-vis other computational data mining techniques in automotive insurance fraud detection, J Data Sci 10 (2012), 537–561.


[88] O. De-Vel, A. Anderson, M. Corney, and G. Mohay, Mining e-mail content for author identification forensics, ACM SIGMOD Rec 30 (4) (2001), 55–64.


[112] S. Li, T. Wu, and W. M. Potterenger, Distributed higher order association rule mining using information extracted from textual data, SIGKDD Explorations 7 (1) (2005), 26–35.


[121] J. Han, and J. Pei, and Y. Yin, Mining frequent patterns without candidate generation, ACM SIGMOD Record 29 (2) (1999), 175–186.


