© https://fti-tn.net/publications Future Technologies and Innovations (FTI) Proceedings: 4th international conference on computer applications and information security (iccais'2021) / March 19 / 2021/ Tunisia: https://fti-tn.net/iccais-2021-list-of-papers

Survey of Crime Data Analysis Using the Apriori Algorithm

Sara Khalid Al-Osaimi
I.T. Specialist, Information Technology Department
Naif Arab University for Security Sciences
Kingdom of Saudi Arabia
salosaimi@nauss.edu.sa

Isra Al-Turaiki
Associate Professor, Information Technology Department
College of Computer and Information Sciences
King Saud University
Kingdom of Saudi Arabia
ialturaiki@ksu.edu.sa

***Abstract*— Law enforcement authorities need to use data-driven strategies to prevent and detect crimes. However, the volume of data generated every day is rising, making the task of processing and archiving very difficult. Data mining techniques have been successfully applied in many areas to determine patterns. In recent years, data mining has been used to analyze past crime data from various sources with the goal of finding crime patterns and trends. In this paper, we look at how crime datasets can be interpreted using frequent pattern mining, in particular using the Apriori algorithm.**

***Keywords*—**Apriori Algorithm, Crime Analysis, Crime Data, Data Mining.

I. INTRODUCTION

Crime is one of the major concerns for communities and governments worldwide. Crime is not only an economic burden on countries, but also has many social impacts. Law enforcement authorities employ various strategies to detect and prevent crime. However, with rapid advancements in technology, crime types and complexity are changing. There is a need to analyze crime records to infer patterns that lead to better understanding and thus to better decision-making, but manual inspection of crime data is not possible due to its magnitude and complexity. Recently, there has been an interest in the use of data mining for crime analysis from researchers in a wide range of fields. Data mining refers to the analysis of large datasets to extract hidden patterns, correlations, and relationships [1] as a basis for improved decision-making. Data mining has been successfully applied in many areas, such as education, health, and business[2]. Recently, data mining has being used to analyze crime data[3] [4] [5], which has great potential to help criminal investigators focus on the most important information in crime data. Mining a large number of crime records allows law enforcement authorities to discover crime trends. It increases efficiency in solving crimes and improves crime prediction. Data mining techniques can be used in crime analysis to improve law enforcement effectiveness and to provide a better understanding of how changes in society affect criminal behaviors. However, it is not an easy task. In general, the steps for analyzing crime datasets are shown in Figure 1.

Today, crime records are stored in digital format. They are multi-dimensional and huge in size. Adding to the complexity, methods and formats for reporting crimes change over the years. The records may also have missing values, making the process of analyzing it more difficult and time consuming. There are many techniques for data mining. The most widely used techniques are classification, clustering, and frequent pattern mining. Clustering is the task of dividing a set of data objects into groups based on similar features. Each group is then called a cluster. Clustering is sometimes referred to as "unsupervised learning" since no previous knowledge is provided to direct the clustering process. In contrast with clustering, classification takes a supervised learning approach since class labels are given to train the algorithm.

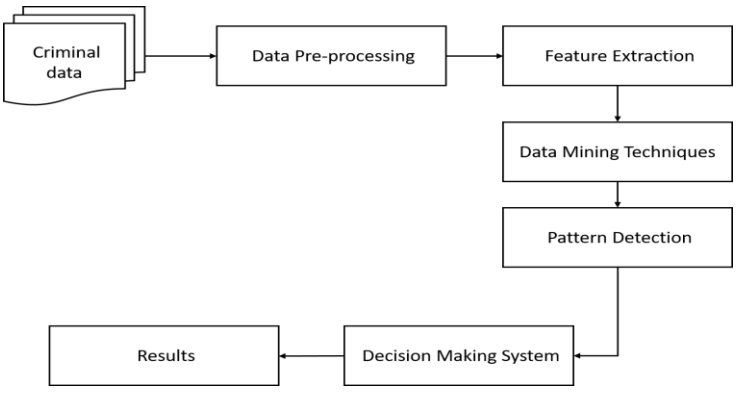

Figure 1. Crime Data Analysis Process

The process of frequent pattern mining discovers relationships between features. Then, IF-THEN rules are generated. There are many algorithms for this purpose, such as the Apriori algorithm[6], a well-known algorithm for frequent pattern mining. Apriori was designed to discover frequent patterns and association rules in a large database of sales transactions. The rules generated by the algorithm can be given in ordinary language, which can be used by police officers to help improve their decision-making, a crime prevention strategy [6]. In this paper, we focus on research that has been done on analyzing crime data using the Apriori algorithm. The rest of the paper is organized as follows: Section II presents the basics of the Apriori algorithm. In Section III, we review the literature on analyzing crime data using Apriori algorithm. In Section IV, we discuss the reviewed work. Section V concludes the survey.

II. The Apriori Algorithm

Apriori, first introduced by Agrawal [7] in 1994, is a well-known algorithm for association rule mining in transactional databases. It is based on a *candidate generation* principle with an iterative process consisting of two main steps. As shown in Figure 2, the algorithm starts by counting the frequency of each 1-itemset. A whole scan of the database is required to accomplish this task. Then, infrequent itemsets, with a frequency less than a predefined threshold, called *support*, are eliminated. The *support* measure is defined as the percentage of transactions in the database containing an itemset and is calculated as:

$$Support(A) = \frac{count\ (A)}{|D|} (1)$$

where count(A) is the number of transactions containing itemset A, and |D| is the total number of transactions in the database. In the second step, candidate 2-itemsets are generated from frequent 1-itemsets by a process called *joining*. Some 2-itemsets undergo elimination based on the Apriori property, which states that *all non-empty subsets of frequent itemsets must be frequent*. This step is called *pruning*. The algorithm continues repeating these two steps until no more itemsets can be generated. Apriori's main advantage that the use of the Apriori property reduces the search space. The Apriori property ensures that subsets of a frequent itemset must also be frequent. When itemsets of size (k+1) are generated by joining k-itemsets, the (k+1) itemsets that do not satisfy the Apriori property are pruned.

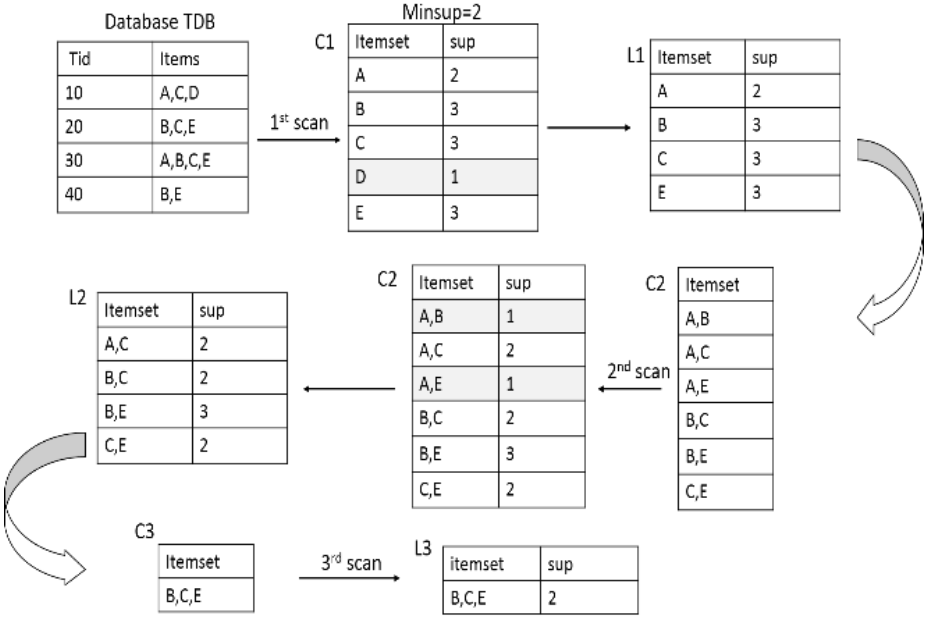

Figure 2. Candidate Generation in Apriori Algorithm

Association rules are then generated from the frequent itemsets. The *confidence* of a rule IF A THEN B, or A→B, measures the percentage of transactions in the database that contain A and also contain B. Confidence is calculated as follows:

$$Confidence(A \rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} (2)$$

The details of Apriori are shown in Algorithm 1.

Algorithm 1 Classic Apriori

```
k = 1
Freq_k = frequent 1-itemsets.
while Freq_k ≠ ∅ do
    for all pairs of patterns p_i and p_j ∈ Freq_k do
        p_new ← join(p_i, p_j)in lexicographic order
        Add p_new to Cand_{k+1}
    end for
    scan the database to count the support ∀c ∈ Cand_{k+1}
    for all c ∈ Cand_{k+1} do
        if c.support ≥ minimum support then
            Add c to Freq_{k+1}
        end if
    end for
    k=k+1
end while
return ⋃_k Freq_k
```

III. Related work

In this section, we focus on the applications of association rule mining and how the Apriori algorithm is used for crime analysis. In [8], the authors used the Apriori algorithm to mine fuzzy association rules for crime patterns. The proposed solution starts with data preprocessing. Then, fuzzification is applied to convert crisp values to fuzzy values. Apriori was applied with post-pruning of fuzzy rules. The resulting fuzzy rules identified crime patterns that were consistent across the U.S., as well as patterns specific to some regions. For example, patterns based on 3188 rules with at least 60% confidence were present in all regions. All rule consequents contain Assaults (Low), Assaults (Medium), Robberies (Low), Murders (Low), or Violent Crimes (Low). None of the rule consequents contain High for any of the major crime categories.

Crime data mining can also be applied to unstructured data. [9] used textual data from offenders, such as email messages, blogs, and chats on social networks. Frequent pattern mining using Apriori was applied to discover criminals' communities. The study aimed to fill the gap between criminal network mining and unstructured textual data.

Also, [10] used the Apriori algorithm to generate rules from crime datasets based on frequent occurrence of patterns to help security decision makers in Libya to take preventative action. The model helped use Libyan crime datasets collected manually from Libyan police departments for various purposes, such as finding relationships and possible explanations for crimes, discovering patterns and trends, making predictions, mapping criminal networks, and identifying possible suspects. They used a small data set of crimes and criminal attributes to discover rules. The model gives excellent results in analyzing crime datasets, but a huge amount of historical data is needed to create and test the model.

Solving crimes usually takes time, which delays the arrests of offenders. In [11], the researchers designed a system that predicts high-crime areas in India on specific days using unstructured real data collected from various websites, such as news sites, blogs, social media, RSS feeds. To define entities such as places they used entity extraction. Then they used a Naive Bayes algorithm [12] to classify data according to crime types, which gave 90% accuracy. Then, the Apriori algorithm produced frequent patterns of places from factors/attributes input into the system, yielding expected crime patterns of particular places. They used a decision tree to simplify decision making for prediction and visualized results using a heat map. After analyzing historical crime data to identify crime-prone regions, they found that crimes such as robbery, murder, highway robbery, and burglary are higher in less-inhabited regions lacking sufficient security, whereas crimes such as arson and vandalism are likely to occur when there is a notable event happening or a VIP presence.

Pereira and Brandao [13] introduced a novel model called "ARCA" that filters and extracts datasets from external data to a "data warehouse." The work consists of two main tasks: first, the ARCA filters and extracts datasets from a real dataset from the Brazilian government, then the Apriori algorithm is used to detect relationships. Since the ARCA is tested on real datasets, it is useful for agencies to prevent or reduce crime. For example, it generated a confidence of 0.95 that gunpoint robberies do not result in injuries.

The authors of [14] proposed a web-based XQuery application to analyze campus crimes in the U.S. Using a real dataset from the U.S. Department of Education, a database was constructed from 23 documents originally in Excel and Word formats. The Apriori algorithm was then used to mine XML data to discover relationships based on two user-selected parameters, min-support and min-confidence. The website allows the user to enter a search criterion to inquire about the safety level of given location. The tool displays rules based on frequent item sets that satisfy a user-supplied minimum confidence, for example Drug→Weapon with confidence 0.8.

In [15], the authors proposed a framework that uses two algorithms for frequent pattern mining: Apriori and FP-Growth [16]. The main goal of the proposed framework was to discover hidden information to help in criminal analysis and to support criminal arrest planning. The framework consists of three steps: data collection, generating association rules, and storing the resulting rules. Comparison of accuracy between Apriori and FP-Growth showed that FP-Growth has better performance.

In [17], the Apriori algorithm was used to analyze features of criminal records based on a real dataset, National Incident-Based Reporting System (NIBRS), which includes 5 million crime incidents in the U.S. from 2013. The main purpose of their work is to predict unknown characteristics of a specific case, for example, offender profile, crime weapon, victim profile. Experimental results showed that the rules created by Apriori are useful for criminal analysis because they were extracted from a large dataset. In addition to the most general association rules created in this study for all crime types, more-

specific association rules can be created according to crime type, offender profile, victim profile, geographical location, and other strong rule attributes, for example, Pre-Rule: Sex of Victim: Female, Race of Offender: White, Ethnicity of Victim: Not Hispanic or Latino, Post-Rule: Race of Victim: White, Location Type: Residence/home, and Confidence: 0.674.

In [18] presented a framework to discover how different crimes are related in Kenya. The proposed model was based on a real dataset of crime figures for Kenya between 2012 and 2015. The framework has the following steps: First, the most frequent crimes in all counties was mapped. Then, records in each cluster were grouped by a k-means clustering algorithm without any prior knowledge. The results of this step are clusters with different degrees of linear relationships between offenses and counties. The last step is to determine relationships with crime. The results of this step are the associations found using the Apriori algorithm. Based on the analysis, they found crime figures decreased from 2012 through 2014, but an increase was noted in 2015. The crime category with the highest crime figures throughout the period under study was 'other offenses against persons' followed by 'breaking' while 'stealing' and 'dangerous drugs' followed at a close margin. 'Offenses involving tourists' had the least crime figures throughout the years. Kiambu, Nakuru, Nairobi, and Meru counties had the highest crime figures in 2015. Table 1 summarizes the studies reviewed in this paper.

IV. Discussion

The aim of this survey paper has been to investigate and define the current state of crime data analysis using the Apriori algorithm from 2010 through 2020 around the world. We found that Apriori has not been used to analyze crime data in Arabic countries except for Libya [10]. None of these studies used parallel implementation with the Apriori algorithm and all of them used real datasets to test their models. The accuracy of the results was affected whether datasets were real or sample and dataset size. In some cases, datasets are needed to apply text pre-processing (e.g. [8], [9], [11], [13], [15], [17]). When the results are visualized it makes it more understandable and easy to read (e.g. [14], [18]). Moreover, we found there are no public historical crime databases for scientific use except in the U.S.A. [19], where various crime datasets have been used in many studies, such as [8], [14], and [17].

TABLE I. SUMMARY OF WORKS ANALYZING CRIME DATA USING THE APRIORI ALGORTHIM

Ref #	Country of dataset	Algorithm used in the study	Tool used	Dataset size	Collected manually (Y/N)	Need preprocessing (Y/N)	Features of the proposed model	Limitations of the proposed model
[10]	Libya	Simple K-means clustering, and Apriori	Weka, and Excel	350 record with 7 attributes	Y	Y	Arabic dataset.	Small dataset (350 crime records) affects the accuracy of the results.
[8]	U.S.A.	Fuzzy Association rules	--	128 attributes are computed to measure per 100 K population for each state.	N	Y	Big data approach	The results wrote as a text and figures, which is difficult to relate figure with text. I.e. they are put 4 graphs together then they explain each one which is take time understand.
[9]	Canada	Apriori	--	50 K files.	Y	Y	Tested and used in real life at digital forensic team of law enforcement unit in Canada.	They did not write the results of crime analysis.
[11]	India	Naïve Bayes, Decision tree, heat map and Apriori	--	--	Y	Y	1. They compare between algorithms that shows Naïve Bayes is 90% accurate. 2. Easy to understand. 3. They use new concept called "criminal profiling" which helps the crime investigators to record the characteristics of criminals.	1. They create the database, which affect the decisions on criminal events. 2. If they consider a particular state/region, it will be more helpful.
[13]	Brazil	Apriori	--	--	Y	Y	They use a novel crime data mining approach.	The results wrote as a text, which is difficult to relate figure with text.

[14]	U.S.A.	Apriori	Novel tool implemented using X-Query, C#.net, and SQL for server	--	N	N	1) The dataset is updated every year. 2) Web-based data mining tool which visualize the results. 3) The tool has "Extract reports" function	They focus on the tool more that analyze crime data.
[15]	Thailand	Apriori, and FP-Growth	Weka	--	Y	Y	Comparison of accuracy and efficiency between used algorithms.	The results of analyze crime data were not available.
[17]	U.S.A.	Apriori	implemented using Python, and analyzed using SPSS	48 different crime and contain 5 million record	N	Y	1. Big data, which makes results more accurate. 2. They add some recommendations based on their results	--
[18]	Kenya	K-means clustering, Mapping and Apriori Algorithm	Analyzed using Shiny app	In 2015, it is more than 28 thousand record.	N	N	1. Results visualization. 2. They use more than one technique. 3. They add recommendations based on the results.	• They did not compare the results of data mining techniques.

V. Conclusion

The rapid increase in crime cases has led to the development of new crime analysis techniques. One of the common techniques used is data mining, which plays an important role in solving crimes, predicting crimes, or discovering offenders. It allows hidden patterns to be uncovered to help law enforcement officers to make better decisions. Crime analysis is a valuable tool for public security. This will be reflected positively and efficiently on the safety of citizens. The aim of this article is to include a brief and concise overview of a number of scholarly publications focused on Apriori algorithm applications in crime data analysis. Despite the fact that there have been several reports on the subject, little research work has been undertaken to utilize data mining for analyzing crime datasets in the Middle East. In terms of efficiency, the greater the number of association rules generated by Apriori, the more time is needed for data processing. According to that, for computation time, parallel computing is one of the important future trends to make the data analysis work for big data, and consequently the technologies of cloud computing, Hadoop, and map-reduce will play important roles for big data analytics. In future work, we plan to build a framework to analyze crime data in Saudi Arabia using parallel implementation of data mining techniques with the help of what we presented here.

References

[1] H. Jiawei, K. Micheline, and P. Jian, *Data Mining: Concepts and Techniques*, First. Morgan Kaufmann is an imprint of Elsevier, 2000.

[2] M. K. Gupta and P. Chandra, 'A comprehensive survey of data mining', *International Journal of Information Technology*, vol. 12, no. 4, pp. 1243–1257, Dec. 2020, doi: 10.1007/s41870-020-00427-7.

[3] D. Das and M. Nayak, 'Crime Pattern Detection Using Data Mining', in *Intelligent Data Analytics for Terror Threat Prediction*, John Wiley & Sons, Ltd, 2021, pp. 221–236.

[4] P. Saravanan, J. Selvaprabu, L. Raj, A. Khan, and J. Sathick, 'Survey on Crime Analysis and Prediction Using Data Mining and Machine Learning Techniques', 2020, pp. 435–448.

[5] U. Thongsatapornwatana, 'A survey of data mining techniques for analyzing crime patterns', in *2016 Second Asian Conference on Defence Technology (ACDT)*, Jan. 2016, pp. 123–128, doi: 10.1109/ACDT.2016.7437655.

[6] A. Rakesh and S. Ramakrishnan, 'Fast Algorithms for Mining Association Rules', 1994, pp. 487–499.

[7] R. Agrawal, T. Imieliński, and A. Swami, 'Mining Association Rules between Sets of Items in Large Databases', in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, 1993, pp. 207–216, doi: 10.1145/170035.170072.

[8] A. Buczak and C. Gifford, 'Fuzzy association rule mining for community crime pattern discovery', *Proceedings of the Acm Sigkdd Workshop on Intelligence and Security Informatics*, Jan. 2010, doi: 10.1145/1938606.1938608.

[9] R. Al-Zaidy, B. Fung, and A. Youssef, *Towards discovering criminal communities from textual data*. 2011, p. 177.

[10] Zakaria, Z. Zubi, and A. Mahmmud, *Using Data Mining Techniques to Analyze Crime Patterns in the Libyan National Crime Data*. 2013.

[11] S. Sathyadevan, M. S. Devan, and S. S. Gangadharan, 'Crime analysis and prediction using data mining', in *2014 First International Conference on Networks Soft Computing (ICNSC2014)*, Aug. 2014, pp. 406–412, doi: 10.1109/CNSC.2014.6906719.

[12] J. H. Friedman, 'On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality', *Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 55–77, Mar. 1997, doi: 10.1023/A:1009778005914.

[13] B. Pereira and W. Brandão, *ARCA: mining crime patterns using association rules*. 2014.

[14] R. Surve, M. Lu, and J. Dai, 'XQuery Data Mining Tool for Campus Security', in *2015 IEEE International Conference on Information Reuse and Integration*, Aug. 2015, pp. 193–196, doi: 10.1109/IRI.2015.38.

[15] R. Lawpanom and W. Songpan, 'Association Rule Discovery for Rosewood Crime Arrest Planning', 2016, pp. 1025–1032.

[16] J. Han, J. Pei, and Y. Yin, 'Mining Frequent Patterns without Candidate Generation', *SIGMOD Rec.*, vol. 29, no. 2, pp. 1–12, May 2000, doi: 10.1145/335191.335372.

[17] M. Sevri, H. Karacan, and M. Akcayol, 'Crime Analysis Based on Association Rules Using Apriori Algorithm', *International Journal of Information and Electronics Engineering*, vol. 7, pp. 99–102, May 2017, doi: 10.18178/IJIEE.2017.7.3.669.

[18] S. Wainana, J. Karomo, R. Kyalo, and N. Mutai, 'Using Data Mining Techniques and R Software to Analyze Crime Data in Kenya', *International Journal of Data Science and Analysis*, vol. 6, p. 20, Jan. 2020, doi: 10.11648/j.ijdsa.20200601.13.

[19] 'data.world', 2021. https://data.world/datasets/crime.