2021 International Conference on Computational Performance Evaluation (ComPE), North-Eastern Hill University, Shillong, Meghalaya, India. Dec 1–3, 2021

Association Rule Mining on Crime Pattern Mining

1st Suman Roy
dept. of IT
Kaziranga University
Jorhat, India
cs16msiit011@kazirangauniversity.in

2nd Ripunjoy Bordoloi
dept. of IT
Kaziranga University
Jorhat, India
cs16msiit017@kazirangauniversity.in

3rd Kayboy Jyoti Das
dept. of IT
Kaziranga University
Jorhat, India
cs16msiit044@kazirangauniversity.in

4th Santosh Kumar
dept. of IT
Kaziranga University
Jorhat, India
cs16msiit055@kazirangauniversity.in

6th Monoj Kumar Muchahari
dept. of IT
Kaziranga University
Jorhat, India
0000-0002-6758-5811

***Abstract*—Recognizing similar offenses during a criminal investigation is a very important goal for crime analysts. The use of pattern-finding has the potential to help crime specialists in discovering new patterns of criminal activity. Over the past few years, association rule mining is implemented to analyze crime data from actual crime datasets to find similar patterns and trends in crime. This paper has worked with the (Frequent Pattern) FP-Growth algorithm to determine identical crime data patterns. Observational results are presented that will help crime experts predict crime and determine the maximum chances of corruption in a particular area.**

***Index Terms*—Association rules, similar crime patterns, FP-Growth algorithm**

I. Introduction

Gradually, the crime rates are increasing in such a way that it is becoming a global threat to humankind. We cannot predict the crime easily since it is neither orderly nor systematized. Crime experts are facing many challenges while dealing with crime. Frequent pattern mining plays a very important role in association rule mining. The FP-growth algorithm is the foremost famed algorithm that may be used for Frequent Pattern mining. Many researchers are trying to find a way of dealing with crime and improving existing algorithms that are used to find similar crime patterns from real crime data sets, for public welfare and crime experts to achieve their goal and encouraging public safety.

Association rule mining is a method to find out the relations between variables from a large data set. It is one of the main concepts of machine learning. It is designed to discover strong rules from a large data set. Association rule is having: - An antecedent (if) and consequent (then). Ex: - "If bread, then butter." Here, if bread is purchased, there is a high chance to buy butter. So, antecedent – Bread, Consequent – Butter

FP Growth algorithm generates the item sets according to the minimum support. It uses a suffix tree (FP-tree) structure to encrypt transactions without generating candidate sets explicitly. FP Growth algorithm is an improvement to the Apriori Algorithm. It doesn't scan the whole data sets like the Apriori algorithm, it only scans the data set twice. It takes less execution time compared to the Apriori algorithm.

Random forest is a supervised learning algorithm. It is one of the widely used algorithms because it can be used for both classification and regression tasks. The idea of the bagging method is that it is a combination of learning models that increases the entire result. Random forest algorithm takes less time for training as compared to other algorithms. It can easily predict output with high accuracy, even for the large dataset it runs efficiently. Using the Random Forest algorithm, we have found an accuracy of 63.45% by using 30% of the data as TEST data.

Logistic regression is that the applicable multivariate analysis to conduct once the variable quantity is divided (binary). Like all regression analyses, logistical regression may be a prophetical analysis. logistical regression is employed to explain knowledge and to elucidate the link between one dependent binary variable and one or additional nominal, ordinal, interval, or ratio-level freelance variables. therefore, logistical regression is that the correct form of analysis to use once we are operating with binary knowledge. we all know we tend to ar managing binary knowledge once the output or variable quantity is divided or categorical; in different words, if it fits into one among 2 classes (such as "yes" or "no", "pass" or "fail", and so on). By using the Logistic regression, we have found an accuracy of 54.04% by using 25% of data as TEST data.

The Organization of paper is as follows: section II discusses Literature study, section III Methodology, section IV Results and Conclusion is explained in section V.

II. Literature Review

Survey in [1], mentions various frequent pattern mining and rule mining algorithm which can be applied to crime pattern mining. This paper explains the concepts of Frequent Pattern Mining and three important approaches that are candidate generation approach, without candidate generation, and the vertical layout approach. It also explains various frequent

978-1-6654-3656-4/21/$31.00 ©2021 IEEE

pattern algorithms and how they can be applied to different areas particularly in crime pattern detection. This paper surely helps to get a clear idea about the application of frequent pattern mining algorithms in various areas. Satyadevan et al. [2], analyze numerous crime patterns and trends in crime. The system will predict a region that features a high chance for crime incidence. Authors have an associate approach between engineering and criminal justice to develop an information mining procedure that may facilitate resolving crimes quicker. The authors focus principally on crime factors daily, to spot a pattern, crime analysts take loads of your time, scanning through knowledge to search out whether or not a selected crime fits into an identified pattern. If the info doesn't work into an associated existing pattern, then the information should be categorized as a replacement pattern and when police investigate the pattern, it will be accustomed to anticipate, predict, and stop crime.

Yuki et al. [3] analyzed the pattern by taking crime datasets from the Chicago Police Department's CLEAR (Citizen Enforcement Analysis and Reporting) system, using Random Forest. The main motive of this paper is to use algorithms on these datasets to classify the type of crime occurring based on time and location which achieved 95.99% accuracy. Omowunmi et al. [4] used Crime Transaction Database including various crimes that happen concerning their locations. They used the FP-Growth algorithm to find similar patterns of crimes in CTDB. They proposed an automatic threshold selection method (ATS) for MST selection in the pruning phase of the FP-Growth model. They further presented a Pattern-pattern paradigm using Pattern-pattern similarity, which is capable of identifying subtle frequent crime pattern trends in the generated frequent crime.

In [5], Naïve Bayes Algorithm is used to classify categorical data. The random forest and K-Mean's algorithm are also used for classification and clustering. The decision tree and neural network-based classification algorithms are used for classifying criminal data. They used Apriori Algorithm to determine the patterns of crimes and the association rule is used to identify the particular place in which the crime rate has increased. They used an area chart and scatter plot for a graph representation of predicted crimes based on the location.

[6] used the Apriori algorithm to recognizes mutual implications among criminal occurrences, retrieving relevant information on criminal behavior implications among criminal occurrences, retrieving relevant information on criminal behavior. They used the crime dataset as an external source of information for the ARCA approach, which processes and loads the external data to the DW, considering the ARCA multidimensional data model, filters and extracts relevant datasets from DW, and finally uses the Apriori algorithm to discover association rules from DW datasets.

Chen et al. [7] projected an inquiry paper on Time, Place, and procedure an easy Apriori rule Experiment for Crime Pattern Detection. This paper aims to resolve the matter of characteristic potential serial sinning patterns mistreatment antecedently underutilized attributes from police-recorded crime knowledge. To attain criminal offense processing procedure extracts three variables in police recorded crime event data: Time, setting, and procedure. In their analysis work, they used the Apriori rule for implementation. They run the associate degree Apriori rule as to whether important patterns can be known from a police-recorded bicycle larceny dataset and demonstrate that bicycle larceny patterns within the bicycle larceny dataset.

Vladimir et al. [8] incorporated two data discovery techniques, clustering, and association rule mining, into a fruitful preliminary tool for the invention of Spatio-temporal patterns. This tool is an Associate in Nursing autonomous pattern detector to reveal plausible cause-effect associations between layers of purpose and space knowledge. They tend to demonstrate the approach to a brand new kind of analysis of the Spatio-temporal dimensions of records of criminal events.

[9] makes a case for the basics of association rule mining and what is more, derives a general framework. supported this we tend to describe today's approaches in context by commenting on common aspects and variations. at the moment we tend to completely investigate their strengths and weaknesses and do many runtime experiments. It seems that the runtime behavior of the algorithms is far additional similar on be expected.

Analyzing numerous crime patterns and trends in crime is mentioned in [2]. The system will predict regions that have a high likelihood of crime incidence and might visualize crime-prone areas. exploitation construct knowledge of data mining Authors extracts antecedently unknown helpful information from Associate in Nursing unstructured data. During this paper, rather than specializing in the reason behind crime incidence like the criminal background of political enmity, etc. authors focus principally on prime factors of every day. Finding the patterns and trends in crime may be a difficult issue. to spot a pattern, crime analysts take plenty of your time, scanning through information to seek out whether or not a selected crime fits into an acknowledged pattern. If it doesn't work into the Associate in Nursing existing pattern then the information should be classified as a brand new pattern. when sleuthing a pattern, it is accustomed to predict, anticipate and stop crime.

Bansal et al. [10] uses Association rule mining in extracting patterns that occur frequently within a dataset for Grievous Crimes against women. The author used two Association rule algorithms i.e., Apriori and predictive Apriori algorithm for implementation. The authors compare both results by using WEKA. They found that the Apriori algorithm performs better than the predictive Apriori algorithm. The result assures all the questions as the age group of men is 20 to 24, age of girls is 16 to 22. Through this data, they will take certain actions towards society.

The authors in [11] proposed a paper on analyzing crime data in Kenya using data mining techniques and R software. The authors used the k-means algorithm, mapping, and Apriori algorithm to analyze how different crimes are related and how they occur. Through k-means they found that robbery and

stealing have a strong relationship.

III. Methodology

This section describes the functioning, underlying principles, and applications of the FP-Growth algorithm approach to crime pattern mining. This involves using a descriptive statistical approach, as a strategy for the pruning phase in the TFP-Growth model, and adopting a Pattern-pattern based paradigm to identify subtle crime pattern sequences.

A. Data Collection

Data sets are collected from various sources. The reports are gathered from the department of police records. Data collection is done at the end of the year at police reports which are made available by NCRB. The records were also collected from resources like new sites, blogs, social media, National Crime Record Bureau (NCRB), and the police department. The data collected will be in many formats such as text, graph images which are called unstructured data and relational data, and also as semi-structured data. The collected data can be stored in a database for further processing. Table I shows some of the links used for data collection:

TABLE I: Parameters

Crime Dataset Repository	URL
Kaggle	https://www.kaggle.com
Data.gov in	https://data.gov.in
National Crime Record Bureau	https://ncrb.gov.in
Data world	https://data.world/datasets

B. Data Classification

The classification technique is used to identify the pattern of the crime data. The concept of Named Entity Recognition is used to classify all the elements in the text into some predefined categories like names, locations, and so on. The random forest algorithm is also used for classification and clustering. The classification of crimes is different kinds: Murder, Rape, Theft, Traffic Violation, Kidnapping, Cybercrime, Assault, Trespassing, and Vandalism.

C. Pattern Identification

Identification of patterns is used to find the trends and patterns in crime using various data mining technologies like classification and clustering. Clustering techniques are applied to existing and known crimes. FP-Growth algorithm determines the crime pattern and an association rule is used to identify the particular place in which the crime rate has increased. Pattern identification avoids the crime that occurs in a particular place by providing security, CCTV, fixing alarms, etc. It helps for improving the capacity of the private investigators and other law enforcement officers. It can be adapted for counter-terrorism for homeland security.

D. Prediction

Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome. It is used to state the probability of an event occurring in the future. Prediction can be done with the techniques like decision trees, naïve Bayes, and linear regression. With the help of predicted values, crime experts can predict crime and determine the maximum chances of corruption in a particular area.

IV. Results and Discussion

From Fig. 1, it can be seen that in the 0-5 hours of the day, all crimes fall, whereas, in 10-15 hours, there is an increase in the crime rate. We can say that 10-15 hours of the day is where the crime rate is too high therefore one should be more alert in that time for crime prevention.

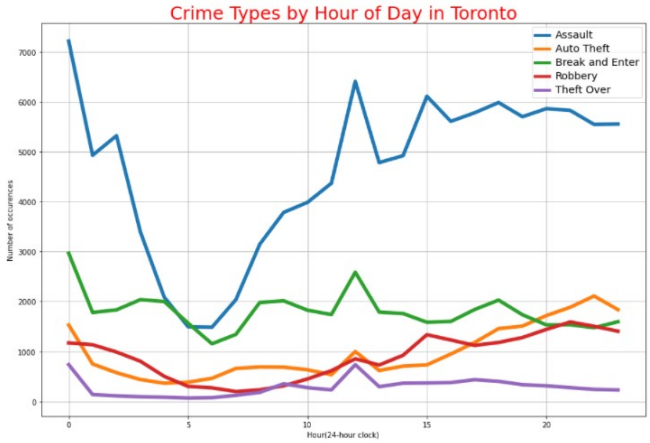

Fig. 1: Graphical data of crime types

Fig. 2 shows the heat map of a major crime in months. It can be seen that Assault has the maximum number of occurrences every month whereas Auto theft and theft over has the least number of occurrences every month.

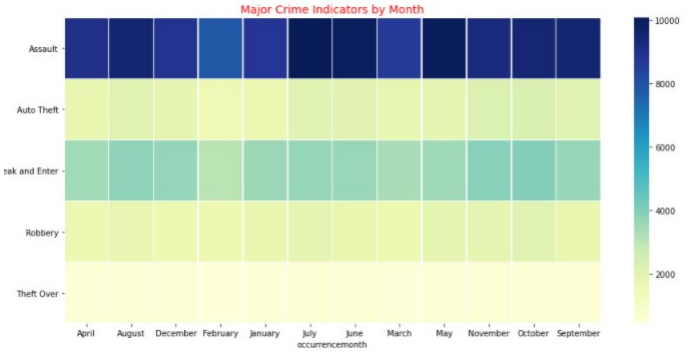

Fig. 2: Heat map of major crimes

Fig. 3 shows the treemap of the top 5 items predicted by the algorithm. Result depicts that Assault has the maximum chance to occur whereas Bay Street Corridor has a minimum chance to occur.

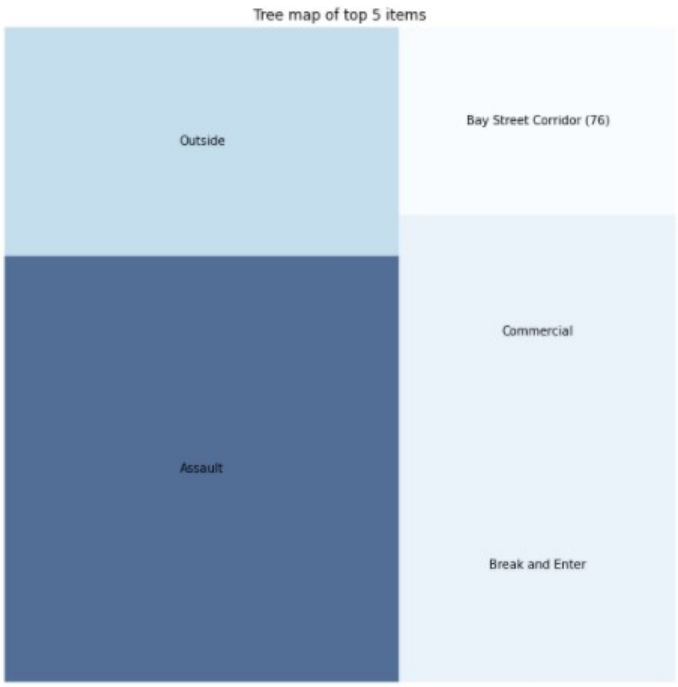

Fig. 3: Treemap of the top 5 items

The main crime pattern prediction by the FP-Growth algorithm can is shown in Fig. 4. Generating antecedents and consequents, antecedents are something that is found in the database more frequently whereas consequents are something that is found in combination with antecedents. The Fig. 4 prediction indicates that Assault has the most chances to be occurred in outside with the antecedent support of 0.65, consequent support of 0.35, and minimum support of 0.30.

V. Conclusion

Crime analysis and pattern prediction are extremely important now-a-days. The combination of facts such as the extensive growth of terrorism and the lack of a truly secure system makes it an important field of research. We have applied Data mining in the context of law enforcement and crime-related problem. Using the Random Forest algorithm, we have found an accuracy of 63.45% by using 30% of the data as TEST data. The support value is found to be 0.7 at max. Antecedent and consequent values were also calculated and found that minimum antecedent support value to be 0.1 and the maximum antecedent value was found to be 0.9, the same range of values for the minimum and maximum consequent supports. A wide range of people, society regions, and the world are affected by crime, so with the help of crime prediction, it helps the people stay away from those who commit crimes in the districts and also aware travelers to select better places to travel with the low crime rate in that are. The result of this analysis helps users in understanding the range of available crime data mining techniques and technologies.

References

[1] D. Usha and K. Rameshkumar, "A complete survey on application of frequent pattern mining and association rule mining on crime pattern mining," *International Journal of Advances in Computer Science and Technology*, vol. 3, no. 4, 2014.

[2] S. Sathyadevan, M. Devan, and S. S. Gangadharan, "Crime analysis and prediction using data mining," in *2014 First International Conference on Networks & Soft Computing (ICNSC2014)*. IEEE, 2014, pp. 406–412.

[3] J. Q. Yuki, M. M. Q. Sakib, Z. Zamal, K. M. Habibullah, and A. K. Das, "Predicting crime using time and location data," in *Proceedings of the 2019 7th International Conference on Computer and Communications Management*, 2019, pp. 124–128.

[4] O. Isafiade, A. Bagula, and S. Berman, "A revised frequent pattern model for crime situation recognition based on floor-ceil quartile function," *Procedia Computer Science*, vol. 55, pp. 251–260, 2015.

[5] S. Vijayarani, E. Suganya, and C. Navya, "A comprehensive analysis of crime analysis using data mining techniques," 2020.

[6] B. L. Pereira and W. C. Brandão, "Arca: Mining crime patterns using association rules," in *11th International Conference Applied Computing. Porto*, 2014, pp. 159–165.

[7] P. Chen and J. Kurland, "Time, place, and modus operandi: a simple apriori algorithm experiment for crime pattern detection," in *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, 2018, pp. 1–3.

[8] V. Estivill-Castro and I. Lee, "Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data," in *Proc. of the 6th International Conference on Geocomputation*. Citeseer, 2001, pp. 24–26.

[9] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, "Algorithms for association rule mining—a general survey and comparison," *ACM sigkdd explorations newsletter*, vol. 2, no. 1, pp. 58–64, 2000.

[10] D. Bansal and L. Bhambhu, "Usage of apriori algorithm of data mining as an application to grievous crimes against women," *International Journal of Computer Trends and Technology*, vol. 4, no. 19, pp. 3194–3199, 2013.

[11] S. M. Wainana, J. N. 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, no. 1, p. 20, 2020.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Commercial)	(Assault)	0.25	0.65	0.25	1.000000	1.538462	0.0875	inf
1	(Assault)	(Commercial)	0.65	0.25	0.25	0.384615	1.538462	0.0875	1.218750
2	(Break and Enter)	(Woburn (137))	0.25	0.10	0.10	0.400000	4.000000	0.0750	1.500000
3	(Woburn (137))	(Break and Enter)	0.10	0.25	0.10	1.000000	4.000000	0.0750	inf
4	(Break and Enter)	(Apartment)	0.25	0.10	0.10	0.400000	4.000000	0.0750	1.500000
5	(Apartment)	(Break and Enter)	0.10	0.25	0.10	1.000000	4.000000	0.0750	inf
6	(Assault)	(Other)	0.65	0.15	0.10	0.153846	1.025641	0.0025	1.004545
7	(Other)	(Assault)	0.15	0.65	0.10	0.666667	1.025641	0.0025	1.050000
8	(Outside)	(Assault)	0.35	0.65	0.30	0.857143	1.318681	0.0725	2.450000
9	(Assault)	(Outside)	0.65	0.35	0.30	0.461538	1.318681	0.0725	1.207143
10	(Outside)	(Bay Street Corridor (76))	0.35	0.20	0.20	0.571429	2.857143	0.1300	1.866667
11	(Bay Street Corridor (76))	(Outside)	0.20	0.35	0.20	1.000000	2.857143	0.1300	inf
12	(Bay Street Corridor (76))	(Assault)	0.20	0.65	0.15	0.750000	1.153846	0.0200	1.400000
13	(Assault)	(Bay Street Corridor (76))	0.65	0.20	0.15	0.230769	1.153846	0.0200	1.040000
14	(Outside, Bay Street Corridor (76))	(Assault)	0.20	0.65	0.15	0.750000	1.153846	0.0200	1.400000
15	(Outside, Assault)	(Bay Street Corridor (76))	0.30	0.20	0.15	0.500000	2.500000	0.0900	1.600000
16	(Bay Street Corridor (76), Assault)	(Outside)	0.15	0.35	0.15	1.000000	2.857143	0.0975	inf
17	(Outside)	(Bay Street Corridor (76), Assault)	0.35	0.15	0.15	0.428571	2.857143	0.0975	1.487500
18	(Bay Street Corridor (76))	(Outside, Assault)	0.20	0.30	0.15	0.750000	2.500000	0.0900	2.800000
19	(Assault)	(Outside, Bay Street Corridor (76))	0.65	0.20	0.15	0.230769	1.153846	0.0200	1.040000

Fig. 4: Generating antecedents and consequents