AOI: 10.11234.2024.0008
ACT2024: Special Proceedings of Applied Computer Technology.

Offenders crime series identification based on connectivity and probabilistic supervised algorithms

Riyam Patel[1], Borra Sivaiah[2], Punyaban Patel[3],
Saroj Kumar Meher[4], Bibhudatta Sahoo[5]
[1]Department of Computer Science, Tandon School of Engineering,
New York University, Brooklyn, New York, USA.
[2]Department of Computer Science and Engineering,
CMR College of Engineering & Technology, Kandlakoya, Hyderabad, India,
[3]Department of Computer Science and Engineering-Cyber Security,
CMR College of Engineering & Technology, Kandlakoya, Hyderabad, India,
[4]Department of Systems Science and Informatics Unit,
Indian Statistical Institute, Bangalore Centre, Bangalore, India
[5]Department of Computer Science and Engineering, National Institute of Technology, Rourkela, India.

Abstract. Statistical clustering technique can be used by crime analysts to the generate suspected list of unsolved crimes, locate crime clusters which have committed by the same person or group of persons, forecasting future events, & develop offender profiles. In this paper, the log-bayes' factor and maximum posterior probability has been used as a similarity measures for solving an unsolved case. Since the offender is known only for a fraction of crimes, in this article the proposed approach is partially semi-supervised. It employs the crime attributes, along with spatial and temporal locations, to describe the offender. It is possible to link and compare crimes using a single link, average link, and a complete link strategy. It employs the agglomerative hierarchal-based clustering method for making crime clusters, based on log Bayes factor and Bayesian clustering model which uses the maximum posterior probability as a similarity measure, for unsolved crime identification. Naive Bayes model calculates the Log-Bayes factor, which helps investigators uncover unsolved murders that are linked to one another. The Naïve Bayes classifier outperforms than the Agglomerative Hierarchical Clustering(AHC) as it uses the log Bayes factor.

Keywords: Agglomerative Hierarchical Clustering, Bayesian Model-based Clustering, Log Bayes Factor, Euclidean Distance, Crime.

1 Introduction

The threat of criminality to our society is significant. There are different types of crime occur on a regular basis. Possibly, it's growing and spreading at a rapid and wide pace. Regardless of the size of the city or town, criminal activity can occur. It can be charged with robbery, murder, rape, assault, and other offences, including false imprisonment, kidnapping, and homicide.

Crime is a broad term and it can be characterized into different forms. This includes property crimes (burglary, theft, and shop lifting), violent crimes (homicide, kidnapping, and sexual assault), and so on Crime frequency can vary based on the time of day / week, and even at different time of year. Geographic location also plays a large role, as there are both high-risk and low-risk areas regarding crime occurrence [1]. As the crimes are increasing, there is a need to develop technique to control it. In order to reduce criminal activity, the police department required to work at faster.

Crime analysis is a critical part of criminology that focuses on studying behavioural patterns and tries to identify the indicators of such events. Crime prevention is complicated. Due to the variety of crimes, motives, ramifications, handling methods, and prevention approaches. Due to these complications and various attributes, crime prediction and clustering has become a powerful and widely used techniques

With the growing shift toward technology and advancements in artificial intelligence (AI), Machine Learning (ML) techniques could reduce this effort by quickly analyzing large amounts of data to extract crime patterns.

Crime datasets might include location-based attributes like where the crime happened and neighbourhood information like unemployment, household income, population, etc. Other datasets contain attributes related to the crime itself, crime type, day of the week, weapon used, victim information, etc. [2]. There are even some data sets that combine both crime and neighbourhood data.

For finding the location of a potential crime, the crime trends are enormously effective. Using the past data, it is possible to forecast what steps will be implemented in the future to prevent crimes. It's difficult to identify recurring criminal activity in a given area. Data mining is one of the emerging fields that can handle enormous amounts of data [3].

Grouping crimes committed by the same offender is the purpose of criminal linkage analysis. The identification of a common perpetrator for a group of crimes is required by a number of crime modelling methods. The locations from a criminal's crime sequence are needed for geographic profiling in order to ascertain their final location [4]. Predicting where future crimes will take place can be improved by analysing the connected sequence of recent occurrences, which can reveal the preferences of the perpetrators in terms of site selection [5].

The goal of pairwise case linking is to determine if a collection of crimes is committed by the same person. In the real world, it is more typical to link crimes committed by a single individual or group of individual. The Log-Bayes factor and maximum posterior probabilities plays a role in clustering to find the similarity between solved crimes and unsolved crimes.

1.1 Motivation of research

Researcher have developed many techniques for solving an unsolved crime. Solving an unsolved crime is going to be very difficult for the police. Clustering technique plays an important role in the crime series identification. K-means algorithm uses the Euclidean distance as a similarity measure which shows poor performance because of lower robustness: small perturbations in the input space will lead to diverse clustering results since labels are absent in the unsupervised clustering task.

The police departments spend a great deal of time and resources detecting crime trends and predicting and preventing them. With the growth technology and advancements in artificial intelligence (AI), and Machine Learning (ML) techniques could reduce the effort by quickly analysing large amounts of data to extract crime patterns. The agglomerative hierarchical and probability-based clustering methods plays an important role in analysing crime groups.

Many AI methods have been studied to reduce crime and protect people in different countries. This predictive machine learning models can predict future crimes, behaviours, etc.

Machine learning will enable police departments to optimize their resources by finding locations based on time, type, or any other factors. Moreover, analysing crime records could reveal more information about the social structure of communities. Thus, government agencies and decision and policy-makers can better identify age groups, ethnic groups, etc. to prevent related issues. The above problems motivated us to carry out the underlying research.

1.2 Research Objective

The objectives of this paper are;
(i) To develop model for crime series data using agglomerative hierarchical clustering along with log bayes factor for unsolved crime identification.
(ii) To develop a model using Naïve Bayes clustering for find the unsolved crime using solved crimes maximum posterior probability as a similarity measure.

1.3 Challenges

The major challenges in crime analysis are

1) The identification of a common perpetrator for a group of crimes is required by a number of crime modelling methods.
2) The locations from a criminal's crime sequence are needed for geographic profiling in order to ascertain their final location.

3). Predicting where future crimes will take place can be improved by analysing the connected sequence of recent occurrences, which can reveal the preferences of the perpetrators in terms of location selection.

1.4 Contributions

The detailed analysis of the various performance measures obtained in the simulation study using crime and offender datasets, leads to the following contribution of the paper.
(i)We implemented agglomerative hierarchical crime clustering using log Bayes factor for unsolved crime identification. We used only average link in this paper to group scores.
(ii) We implemented the Naïve Bayes model based crime clustering using the posterior grouping probabilities and the crime similarity are calculated based on Maximum posterior probability.
(iii)The Naïve Bayes model based crime clustering is better than agglomerative hierarchical crime clustering using log Bayes factor because it doesn't waste time for searching in the dendrogram for unsolved crimes.

1.5. Highlights

The key highlights of this research are
(i) Crime Series Identification
(ii) Crime Series Clustering
This paper has been organised as where the Section 1: Introduction, Section 2: Related works, Section 3: Proposed Methodology Section 4: Matrices used for performance Measures, Section 5 Experimental Result and Analysis, and Section 6: Concluded the paper.

2 Related Works

In recent years, data mining has seen substantial growth in the field of crime literature. It has also established itself as an indispensable tool for enhancing one's impression and bringing the crime rate down. This in-depth study focuses on the methodologies and tools that have been utilised in previous examinations of data mining in criminal cases, and it does so in a variety of contexts. It is laid out with the assistance of images and is segmented into a wide variety of categories. The investigation into mining data pertaining to criminal activity has uncovered a few challenges.

Many of the crime prediction methods were developed for generic crimes and situations, where different models were used and tested in crime prediction to determine the most effective one relative to the provided dataset. However, some methods have been developed for particular crime types or categories, other researchers have focused on performing a comparative analysis between the different learning model types. Recent studies have examined crime prediction methods. For example, [2], provided a survey that explores data mining methods for crime prediction based on different crime prediction factors, such as socioeconomic, spatial-temporal, demographic, and geographic attributes.

The author [6] has conducted a criminal analysis by applying k-means clustering on the crime dataset with the help of the fast miner tool. De Bruin et al. [7] developed a new distance measure for comparing and clustering of persons built on their physiognomies. They also presented a framework for crime trends using this measure.

The authors address the construction of a Visual Interactive Malaysia Crime News Retrieval System (i-JEN) in [8], where they also cover the approach, user research and goals, the system architecture, and future plans for the system. Their goals were to construct crime-based events, investigate the use of crime-based events in improving classification and clustering, develop an interactive crime news retrieval system, efficiently interpret crime news, integrate them into a usable and robust system, and evaluate usability and system performance. The study will also help understand criminal data usage in Malaysia.

Manish Gupta et al. [9] described the Indian police's e-governance systems. They also suggest a query-based crime analysis interface to aid police. They proposed an interface to retrieve meaningful information from the NCRB's large crime database and locate crime hot spots using crime data mining techniques such as clustering. It also offered an interface for the FBI's large crime database. Using Indian criminal case files, the interface's usefulness was demonstrated.

K. Zakir Hussain et al. [10] utilized data mining and a simulation model to add human experience to clustering. The author [11] used a qualitative and quantitative approach using K-means clustering data mining on a crime dataset from New South Wales, Australia, where they observed high crime rates in multiple cities. They used a Sydney crime dataset, one of the world's most dangerous cities.

The author [12] has used a number of different clustering algorithms, such as K-Means clustering and agglomerative clustering, on the Stop, Question and Frisk Report Database, which is maintained by the New York City Police Department (NYPD). This was done in order to analyse the location of crimes and people who were stopped, based on the reason that they were stopped, with the goal of lowering the city's overall crime rate. Our statistical and visual analysis lead us to conclude that the K-Means algorithm is the most effective clustering method. This is due to the K-Means algorithm's advantageous characteristics, which ensure that the models are beneficial.

The authors [13] have clustered the crime data for total cognizable crimes using the fuzzy C-Means clustering algorithm. These crimes include kidnapping, murder, theft, burglary, cheating, crime against women, and other similar offences.

The authors [14] used K-Means clustering, Agglomerative clustering, Density Based Spatial Clustering with Noise (DBSCAN), and algorithms for clustering activities based on predefined cases of six cities in Tamilnadu, Chennai, Coimbatore, Salem, Madurai, Tirunelveli, and Tiruchirappalli, with a total of 1760 instances and nine attributes spanning 2000-2014 to represent the instances. Cities Each clustering algorithm's performance is compared using precision, recall, and F-measure. This approach will help Tamilnadu police officers forecast and identify criminal activities more accurately, lowering the state's crime rate.

Hazarika et al. [15] plotted the geo-location of each site and utilised it to perform geo-spatial grouping in k-means with distance measures. The performance of each distance matrix is analysed to determine the best measure for location-based clustering in a major city like Delhi.

K-means algorithm presents a superior way to forecasting the results, with an emphasis on predicting places with higher crime rates and age groups with more or less criminal inclinations. This approach was developed by Krishnendu.et.al [16]. There are a few different clustering algorithms that can be used for criminal analysis and pattern prediction; however, these algorithms do not fulfil all of the requirements. When it comes to predicting future events, the K means algorithm performs far better than its competitors.

Bharathi, S., et al. [17] trained the proposed system on supervised data from Tamil Nadu using online data. In the testing phase, K-Medoids should be used to find the cluster closest to the test crime. After that, thy used similarity to identify the suspects.

The K Means-based behaviour identification system, which was developed by Min, Xing, and others [18], is a tool that can differentiate fraudulent activities and locate fraudulent phone numbers. The approach for extracting the behaviour characteristics, reducing the dimensions of the features using principal component analysis, and using grid search to determine the right clustering settings.

The conventional clustering approach that is used for the analysis of crime data to identify potential criminals yields results that are only moderately accurate. The authors [19] suggested a new method for similarity measure that they dubbed Segmented Multiple Metric Similarity Measure (SMMSM), with the goal of improving the accuracy of similarity measure. This was done in order to tackle the problem. Using this approach, the qualities are segmented into their respective categories based on the significance of the relationship between them. The new metric is scalable with respect to the dimensionality of the data, and it works just as well with categorical data as it does with numerical data. In addition to that, it demonstrates that the accuracy of this method is superior to that of other metrics.

The author [20] has investigated the Auto Regression Techniques in order to accurately forecast the crime with a low amount of error for crime time series data. This is accomplished by determining the relationship between the characteristics of crimes. It is helpful in detecting similar crime trends across a variety of crime locations, which is necessary for the selection of criminal sites. Because of this, the ARIMA (Auto Regressive Integrated Moving Average) model is able to reduce the error that is produced by the predictive model, and as a consequence, it pinpoints the offender spot in advance.

Traditional clustering methods are inaccurate. To increase similarity measure accuracy, the authors [21] proposed the Segmented Multiple Metric Similarity Measure (SMMSM). In this measure, attributes are grouped by similarity importance, and compensatory relationships do not occur between attribute values in different groups.

To better understand how criminal activity is distributed, the authors [22] proposes a model-based clustering algorithm based on the E-M algorithm, initialised by K-means clustering with geodesic distance classification to estimate the model parameters, and then compare it to the classical E-M algorithm, initialised with hierarchical clustering. When compared to the standard E-M and K-means clustering algorithms, the model-based clustering of the E-M algorithm combined with the K-means clustering algorithm produced the same classification in some cases, demonstrating its efficacy as a fast and stable converger with low probability of uncertainty by classifications. Researchers looking to model big spatial features in data mining can benefit from a combination of model-based clustering techniques, such as those used in a hotspot analysis of criminal activity.

In order to speed up the process of prioritising suspects in criminal investigations, the authors [23] created and evaluated a new spatial profiling tool. In order to rank suspects based on their proximity to and nature of these locations in relation to an input crime, the Geographic Profiling Suspect Mapping and Ranking Technique (GP-SMART) maps suspects' activity locations available in police records, such as their home addresses, family members' home addresses, prior offence locations, non-crime incident locations, and other contacts with police. GP-SMART outperformed baseline methods (approximating current algorithms) that ranked suspects based on the closeness of their activity locations (or home addresses) to the input crime, thanks to the new incorporation and distinction of many various types of activity location.

The authors [24] has proposed an effective multi-module method for predicting crime using deep learning techniques which has two modules: Feature Level Fusion and Decision Level Fusion. The first module employs temporal-based Attention LSTM, Spatio-Temporal based Stacked Bidirectional LSTM, and Fusion model. The proposed model outperforms numerous other well-known models.

The authors used Auto Regression Techniques [25] to forecast crime with minimal error for time series data by determining the association between crime variables. The experimental results enhance prediction accuracy.

A crime forecasting model was developed [31,33,34], based on Spearman's Correlations and a clustering technique (DBSCAN), which captures significant groupings in a geospatial dataset. A Multi-Input Hidden Markov Model (MI-HMM) machine learning framework was developed to train the dataset then the results were used to make a Maximum a Posteriori (MAP) decision over the possible state of crime for the next month. This novel model, MI-HMM-MAP, was used to predict the density of crime including criminal hot spots over time. The model was evaluated using real-world dataset. Findings show an average of 72.5% accuracy and 81.7% correctness.

3 Proposed Methodology

The model for the proposed approach is shown in Figure 1.

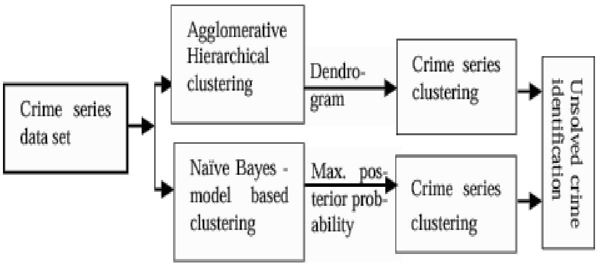

Crime series data set
Agglomerative Hierarchical clustering
Naïve Bayes - model based clustering
Dendro-gram
Max. pos-terior prob-ability
Crime series clustering
Crime series clustering
Unsolved crime identification

Figure 1. Proposed approach for crime series clustering

In this model, the Agglomerative Hierarchical clustering and Naïve Bayes-model based clustering techniques are applied in the crime series data set after pre-processing. Both clustering methods are used for crime series clustering and unsolved crime identification. The agglomerative hierarchical clustering creates dendrogram using single, average, and complete linkage approaches. It links crimes using Euclidian distances. The Naïve Bayes model based clustering approach uses the log Bayes factor as a distance measure to form crime series clusters and it also used for unsolved crime identification. The Naïve Bayes approach is better than agglomerative approach as it uses maximum posterior probability.

3.1 Agglomerative Hierarchical Crime Series Clustering [26]

Hierarchical clustering, an algorithmic approach that methodically constructs a hierarchy of clusters, is a technique that can be used to analyse a crime series. Each observation is initially placed in its own cluster using the agglomerative method. The two clusters that are next to one another in terms of proximity are combined to create a new, larger cluster. This process is repeated until all of the observations either fit into the same cluster or until a stopping condition is met.

This algorithm requires both the pairwise similarity between two sets of observations in order to perform correctly. To determine how similar two groups of observations are to one another, there are three fundamental approaches. These are called single link, average link and the complete link. We used only average link because it uses the group scores.

The Figure 2 shows the dendrogram of Average Linkage that how similarity scores between clusters are represented as well as the complete path to a solution. One can determine the number of crime series contained in the data as well as their makeup by using a dendrogram to split the crimes into distinct clusters and cutting them at a specific level of similarity. See Figure 1 for the results of an average linkage agglomerative clustering using log Bayes factors on 490 crimes committed 1507convicted individuals. The Figure 3 and Figure 4 shows the Dendrogram of Single Linkage and Complete Linkage vs. log Bayes Factor respectively.

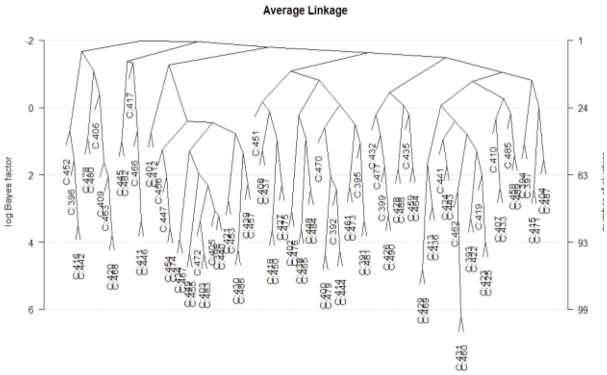


Figure 2. Dendrogram of average linkage vs. log Bayes Factor

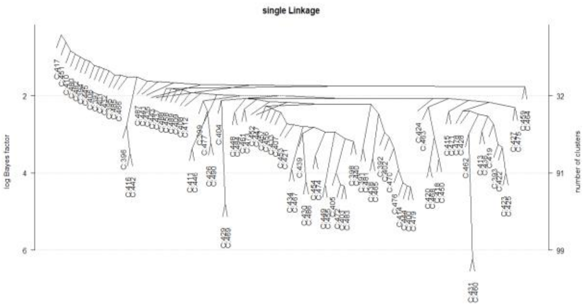


Figure 3. Dendrogram of Single linkage vs. log Bayes Factor

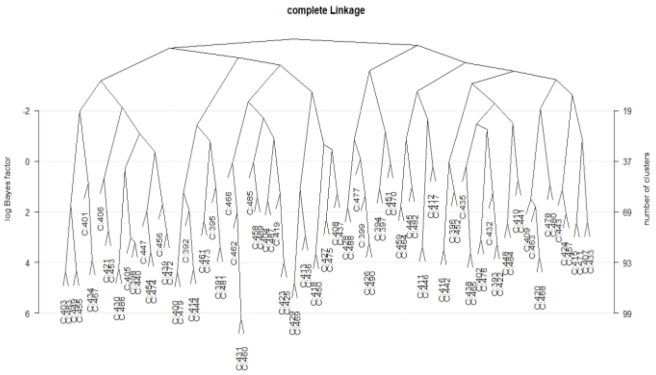


Figure 4. Dendrogram of Complete linkage vs. log Bayes Factor

```
Algorithm: Clustering Crime Data Using Agglomerative Hierarchical Analysis

Initialize:
Step 1: Use the log Bayes factor to determine the degree of similarity between each crime that has ever occurred.
Step 2: Classify each open case separately (or into existing clusters if that is an option).
Iterate:
 Step 3: Determine the degree of similarity (using either a single, complete, or average linkage) between each cluster.
Step 4: Combine the two groups with the most shared characteristics.
Step 5. Repeat 3-4 until there is a single cluster or stopping criterion is met.
```

3.2 Hierarchical Based Crime Series Linkage

The collection and processing of data pertinent to each specific instance of criminal conduct is one of the first phases in the case linking process [27]. Processing forensic evidence is one example of this. Other examples include extracting and categorising criminal behaviours or physical descriptions from an event report, geocoding the crime scene's address, and geocoding its coordinates. Let the resulting vector of crime variables, which we will refer to as Vi, represent crime i.

These criminal offence statistics will include information about the perpetrator (if it is known), the crime, and the location of the crime (for example, spatial location, timing, crime type, criminal behaviour, and victim characteristics). As a consequence of this, the data on crime will almost certainly consist of a mixture of categorical, discrete, continuous, and missing valued variables.'

3.3 Bayesian Model-Based Clustering Approaches [20]

It is a supervised Bayesian model which is based on partially supervised clustering technique. To link crimes in a series, this strategy is used. Because it uses spatiotemporal crime locations as well as criminal qualities that explain the perpetrator's manner of operation, and because the offender is known for a portion of the occurrences.

Complex features that are typically seen in crime data can be easily accommodated by the hierarchical model. Missing data, interval censored event times, and a mix of continuous and discrete variables are some of these components. Additionally, it is capable of offering uncertainty estimates for each parameter of the model, as well as the relative importance of each feature in the model. Furthermore, the model produces posterior grouping probabilities that let analysts only act on model output when it is absolutely necessary to do so.

Direct estimation of the Bayes factor is something that can be accomplished with only a handful of different evidence variables. On the other hand, simplifying models might be of assistance in improving estimations when there are a growing number of evidence variables. Recognizing that pairwise case linking is, at its core, a problem of binary classification could be helpful in selecting an efficient modelling technique [29, 30,32].

The Naïve Bayes model is built for crime analysis by using the formula (1) and cases linked and unlinked using formula (2). Given a set of crime pairings x, two spatial variables (X, Y), two time variables (DT. FROM and DT.TO), and three category variables (MO1, MO2, and MO3).

The formula used in Naïve Bayes model is

$$Y \sim spatial + temporal + tod + dow + MO1 + MO2 + MO3 \qquad (1)$$

The equation (6) is used for building the Naive Bayes clustering model and the equation (2) is used for crime linkage is given by:

$$type = \begin{cases} 1, linkage \\ 0, otherwise \end{cases} \qquad (2)$$

4 Performance Measures

The performance evaluation metrics used in this paper are presented below.

(i) Bayes Factor [20]

Let, D be the crime database, and there are two hypothesis called h_1 and h_2 . The priori probabilities are $pr(h_1)$ and $pr(h_2) = 1 - pr(h_1)$, and posteriori probabilities $pr(h_1|D)$ and $pr(\ h_2|$D). the odds scale, which can be expressed as

$(odds = probability/(1 - probability))$.

From the Bayes' theorem, we obtain

$$pr(h_k|D) = \frac{pr(D/h_k)pr(h_k)}{pr(h_k)} \qquad (3)$$

$$posterior\ odds = Bayes\ factor \times prior\ odds \qquad (4)$$

The similarity between crimes p an q is $(p, q) = log\big(BF(p, q)\big)$,where $BF(p, q)$ is the estimated Bayes factor(BF) for linkage [20].

(ii) Euclidean distance [28]

The formula for calculating the Euclidean distance is as follows:

$$d_{euclidian}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (5)$$

Where, x and y are two vectors of length n.

$d_{euclidian}$= Euclidian distance

(iii) Pearson Correlation distance (d_{PCor}) [28]

To calculate correlation-based distance, simply subtract the correlation coefficient from 1 to get the answer.

(iv) Pearson Coefficient [21]

The Pearson correlation can be defined as

$$Pearson(x,y) = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}} \quad (7)$$

Where, μ_x and μ_y are the means for x and y respectively.

5 Experimental Result and Analysis

We consider crimes data set with 490 observations of 8 variables and offenders data set with 1507 observations of 2 variables. The crimes and offenders data sets collected from R Studio. we computed Log-Bayes factor using naïve Bayes model and implemented agglomerative hierarchical clustering and Naïve-Bayes clustering in R. We divided the crimes in two categories called solved crime and unsolved crime. The crimes are clustered using three approaches such as single, average, or complete links. The largest score from each group is used in single linkage, the smallest score is used in complete linkage, and the average score is used as the group score in average linkage.

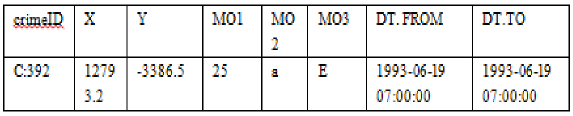
5.1 Dataset Used

For the purposes of crime series clustering, we used data pertaining to both incidents and offenders. We analysed 1507 criminals based on two criteria and 490 criminal acts based on eight variables. The training data comprised 70 percent of the total samples, and the testing data comprised 30 percent of the total samples. The naive Bayes model, which was utilised in the process of estimating the log Bayes factor. The first step is to collect all of the unsolved crimes and then to execute an agglomerative hierarchical crime grouping.

(A) Unsolved crime identification using log-Bayes factor

We extracted solved and unsolved crimes from the crimes data set. The most similar crime series to the unsolved crime is found in Table 1 with crime ID C:392.

Table 1. Unsolved Crime with Crime ID C:392



crimeID	X	Y	MO1	MO 2	MO3	DT. FROM	DT.TO
C:392	1279 3.2	-3386.5	25	a	E	1993-06-19 07:00:00	1993-06-19 07:00:00

The top 6 scores using Log-Bayes factor are shown in Table 2.

Table 2. The top 6 scores using log-Bayes factor

Sl. No.	Group No.	Average Link	Single Link	Complete Link
1	8	3.543836	3.543836	3.543836
2	9	3.479244	3.885876	3.171058
3	10	3.479244	3.885876	3.171058
4	11	3.479244	3.885876	3.171058
5	154	3.182368	3.182368	3.182368
6	12	3.178842	3.178842	3.178842

It is clear from the Table 2 that the unsolved crime is most similar to the crime(s) in crime group 8 with an average linkage log Bayes factor of 3.54. The crimes and offenders associated with this groups are shown in Table 3.

Table 3. Crime series with crime group 8

Crime ID	Index	CS	Offender ID	Date and Time	Group no.
C:139	139	8	O:105	1993-06-19 01:42:30	8

The unsolved crime is related to multiple crime series with crime group 9 is given in table 4.

Table 4. Multiple crime series with crime group 9

CrimeID	Index	CS	offenderID (Person)	Date and Time	Group No.	CG
C:144	144	9	O:106	1993-06-20 at 03:27:00	9	41
C:163	163	9	O:106	993-06-20 at 13:15:00	9	41
C:145	145	9	O:106	1993-06-20 at 01:30:00	9	41
C:164	164	9	O:106	1993-06-20 at 13:15:00	9	41
C:165	165	9	O:106	1993-06-20 at 12:45:00	9	41
C:166	166	9	O:106	1993-06-20 at 12:45:00	9	41

The crime IDs for crimes "C:408" "C:417", and "C:464"with strongest linkage are given in figure 4.

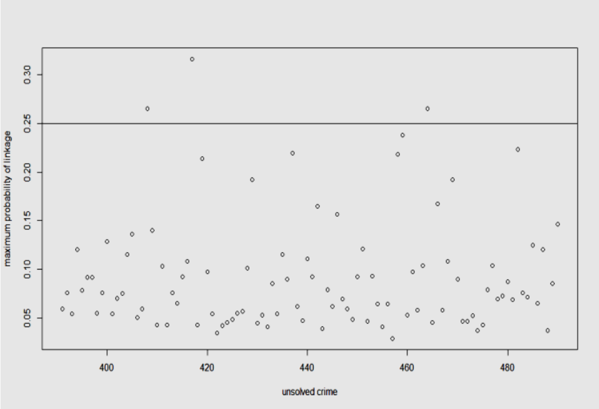


Figure 4. Bayesian Model-Based Clustering

(B) Unsolved crime identification using Bayesian Model-Based Clustering

The most probable crimes linked to the unsolved case C:417 are listed by Bayesian Model-Based Clustering model. Because the first two crimes C:15 and C:26 have been solved, it is possible that the person or people who committed C:417 are also responsible for C:15 and C:26. The two crimes C:459 and C:446, have no group IDs as shown in Table 5. This means they are unsolved crimes. The posterior probability help crime analysts determine whether to continue an investigation further only if the link is strong adequate.

Table 5. Posterior probabilities using Bayesian Clustering

Sl.No.	Index	Probability	CrimeID	CG
1	81	0.3160	C:15	46
2	197	0.2870	C:26	146
3	459	0.2095	C:459	NA
4	446	0.1160	C:446	NA
5	2	0.0475	C:10	2
6	482	0.0365	C:482	NA

6 Conclusion

In this paper, existing AI-technologies can perform well in crime prediction and clustering. These technologies can accurately forecast crime. It improved efficiency especially in the application of crime identification. The unsolved crimes have been solved using two popular agglomerative hierarchical crime series clustering and Naïve Bayes model based clustering techniques. The hierarchal crime series clustering uses the log Bayes factor as a similarity measure for solving unsolved crimes and Bayesian clustering model uses the maximum posterior probability for solving unsolved crimes. The results showed that Navies model clustering techniques better than the agglomerative hierarchical clustering techniques as it uses log Bayes factor as a similarity measure which is probability based.

Our goal was to help crime analysts by developing statistical methods that would provide them with the ability to make more accurate judgments and predictions regarding the linkages between crimes. It is possible for the interpretable statistical models to place linkage analysis on a foundation that is more scientific and dependable, to increase consistency, to lower the cognitive load of the analyst, and to improve the chance of using linkage analysis as evidence in legal proceedings. one of the future directions is adding more location information to the crime to benefit and improve the detection of crime location.

References

1. Fatima Dakalba et. al. : Artificial intelligence & crime prediction: A systematic literature review : Social Sciences & Humanities Open, Science Direct, Elsevier, vol.6(2022).

2. Saravanan, P., Selvaprabu, J., Arun Raj, L., Abdul Azeez Khan, A., Javubar Sathick, K. : Survey on crime analysis and prediction using data mining and machine learning techniques. Lecture Notes in Electrical Engineering, pp.435–448, Vol. 688(2021).

3. Wayne Petherick : Applied Crime Analysis: A Social Science Approach to Understanding Crime, Criminals, and Victims . 1st Edition - June 12(2014).

4. Rossmo, D. : Geographic profiling. CRC Press (2000).

5. Bernasco, W. and Nieuwbeerta, P. : How Do Residential Burglars Select Target Areas?. A New Approach to the Analysis of Criminal Location Choice . British Journal of Criminology, 45, pp.296–315(2005).

6. Jyoti Agarwal, Renuka Nagpal, Rayna Sehgal : Crime Analysis using K-Means Clustering . International Journal of Computer Applications, Volume 83, Number 4, 2013. DOI: 10.5120/14433-2579

7. De Bruin,J.S.,Cocx,T.K,Kosters,W.A.,Laros,J., Kok,J.N : Data mining approaches to criminal carrier analysis . Proceedings of the Sixth International Conference on Data Mining (ICDM"06), Pp. 171-177(2006).

8. Nazlena Mohamad Ali et.al. ," Visual Interactive Malaysia Crime News Retrieval System", Part II, LNCS 7067, Springer, pp. 284–294 (2011).
DOI: 10.1007/978-3-642-25200-6_27

9. Manish Gupta1, B. Chandra1 & M. P. Gupta1 : Crime Data Mining for Indian Police Information System(2007).

10. K. Zakir Hussain, M. Durairaj, G. Rabia Jahani Farzana : Application of Data Mining Techniques for Analyzing Violent Criminal Behavior by Simulation Model . International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 2, No. 1 (2012). ISSN: 2249-9555

11. Joshi, A., Sabitha, A. S., Choudhury, T. : Crime Analysis Using K-Means Clustering. 3rd International Conference on Computational Intelligence and Networks (CINE), (2017). doi:10.1109/cine.2017.23

12. Alkhaibari, A. A., Ping-Tsai Chung : Cluster analysis for reducing city crime rates . IEEE International Conference on Long Island Systems, Applications and Technology Conference (LISAT) (2017). doi:10.1109/lisat.2017.8001983

13. B. Sivanagaleela, S. Rajesh : Crime Analysis and Prediction Using Fuzzy C-Means Algorithm . 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (2019). doi:10.1109/icoei.2019.8862691

14. Sivaranjani, S., Sivakumari, S., Aasha, M. : Crime prediction and forecasting in Tamilnadu using clustering approaches . International Conference on Emerging Technological Trends (ICETT) (2016). doi:10.1109/icett.2016.7873764

15. Hazarika et.al. : Cluster analysis of Delhi crimes using different distance metrics . IEEE International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp.565–568, Chennai, India, (2017). doi:10.1109/ICECDS.2017.8389500

16. S.G. Krishnendu, P.P. Lakshmi, L. Nitha : Crime Analysis and Prediction using Optimized K-Means Algorithm. Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp.915–918, India (2020).
doi: 10.1109/ICCMC48092.2020.ICCMC-000169

17. S. T. Bharathi, B.Indrani, M. Amutha Prabakar : A supervised learning approach for criminal identification using similarity measures and K-Medoids clustering. IEEE International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kerala State, India, pp. 646–653, (2017). doi:10.1109/ICICICT1.2017.8342639

18. Min, Xing & Lin, Rongheng : K-Means Algorithm: Fraud Detection Based on Signalling Data . IEEE World Congress on Services, San Francisco, CA, USA, pp. 21–22, (2018). doi:10.1109/SERVICES.2018.00024

19. Guangzhu Yu, Shihuang Shao Bing Luo : Mining Crime Data by Using New Similarity Measure . Second International Conference on Genetic and Evolutionary Computing, pp.389-392, IEEE Computer Society, Information and Technology College, Donghua University, Shanghai, China(2008).

20. John W. Lau & Peter J.: Bayesian Model Based Clustering Procedures . Department of Mathematics, University of Bristol, Bristol, UK, June 7(2006).

21. Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, Teh Ying Wah,: A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. PLOS ONE | December 11(2015). DOI:10.1371/journal.pone.0144059

22. Simon Kojo Appiah et. al., : A model-based clustering of expectation–maximization and K-means algorithms in crime hotspot analysis", Research in Mathematics, Vol. 9, No. 1, 2073662(2022),

23. Sophie Curtis-Ham et. al.: A New Geographic Profiling Suspect Mapping and Ranking Technique for crime investigations: GP-SMART. Journal of Investig Psychol Offender Profil, Vol.19, Pp.103–117(2022).

24. Nowshin Tasnim et. al.: A Novel Multi-Module Approach to Predict Crime Based on Multivariate Spatio-Temporal Data Using Attention and Sequential Fusion Model. IEEE Access, Vol. 10, Pp.48009-48030 (2022).

25. Yadav, R., & Kumari Sheoran, S. : Crime Prediction Using Auto Regression Techniques for Time Series Data. 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), 22-25 November (2018). doi:10.1109/icraie.2018.8710407

26. Michael D. Porter: A Statistical Approach to Crime Linkage", October 10(2014).

27. Woodhams, J., Bull, R., Hollin, C.R., :Case Linkage. In: Kocsis, R.N. (eds) Criminal Profiling", Humana Press (2007). https://doi.org/10.1007/978-1-60327-146-2_6

28. https://www.datanovia.com/en/lessons/clustering-distance-measures/

29. Bennell, C. and Canter, D. : Linking commercial burglaries by modus operandi: Tests using regression and ROC analysis. Science & Justice, Vol.42, pp.153–164, (2002).

30. Yu, G., Shao, S., Luo, B.,: Mining Crime Data by Using New Similarity Measure. Second International Conference on Genetic and Evolutionary Computing, IEEE Computer Society, pp.389-392 (2008). doi:10.1109/wgec.2008.125

31. Devon L. Robertson, Wayne S. Goodridge,: Predicting density of serious crime incidents using a Multiple-Input Hidden Markov Maximization a posteriori model, Machine Learning with Applications, Vol.7(2022).

32. Jeffery T. Walker, Grant R. Drawve,: Foundations of Crime Analysis: Data, Analyses, and Mapping", 1st Edition(2018).

33. Dawei Qiu et.al. :Crime Type Identification Using High-Order Deep Residual Network with Multiple Attention Algorithm, Applied Artificial Intelligence, Tylor $& Francis, Vol. 38, No.1(2024).

34. Dileep Kumar Kadali et. al.: Uncertain crime data analysis using hybrid approach, Springer Nature, Discover Artificial Intelligence, Vol.5, Article no. 15, (2025).