Cyber Crime Classification and Analysis of Various Indian States and Union Territories During 2011 to 2022 Using Data Mining Techniques

Oinam Tomba Singh

Associate Professor of Statistics Dhanamanjuri University – Manipur - 795001

*Corresponding author’s: Oinam Tomba Singh

### Abstract

This paper attempts to identify the performance and to classify the cybercrime rate in Indian states during the period of 2022 with twenty-seven states and seven union territories. The secondary source of data were collected from National Crime Records Bureau in India. Initially, the researchers selected 27 parameters related to crime. After data mining few variables are discarded from the analysis. The discarded variables are considered as outliers. The remaining twenty parameters were considered for further analysis, the variables like hacking the system, sexual exploitation, business, personal emotion, purchase of illegal drugs, spreading piracy, motives of blackmailing, etc. In this connection the main objectives are (i) to identify in which states crime rate is more, (ii) to identify which factors influence more for cybercrime, and (iii) to classify the performance based on the k-means clustering techniques. The results are classified and are labelled as High Cyber Crime Rate States (HCCRS), Moderate Cyber Crime Rate States (MCCRS) and Low Cyber Crime Rate States (LCCRS)

### Keywords:

Cyber Crime, Classification, Orange Data mining, Factor Analysis and k-means clustering techniques.

---

1. Introduction

This research paper addresses state and national security, which has become significantly more important since the attacks on the Indian Parliament and Mumbai, as well as local and international terrorism, social media crime, etc.

To stop future attacks and hacking of the government networks, the Indian security agencies are actively gathering local and international intelligence data. In response, local security forces have been encouraged to keep a closer eye on cybercrime activity throughout all Indian states and union territories. Given that India has the second-largest population in the world after China, law enforcement and intelligence gathering in all Indian states have significant challenges in reliably and effectively obtaining crime data and analyzing the massive volume of data. Digital India is encouraged and implemented by the current administration. Nowadays, the majority of urban and rural residents use the internet, mobile devices, net banking, and other sources to conduct online transactions. Due to heavy network traffic and frequent online transactions that produce vast amounts of data, only a small part of cybercrime can be detected.

Data mining is a powerful tool that enables criminal investigations who may lack extensive training as data analysis to explore large amount of data quickly and efficiently [1]. The computation can process thousands of commands in seconds, saving valuable time. In addition, installing and training software often costs less level of errors than human analysis, especially those who work extensive hours. This research paper analyses the Indian states cybercrime data using Orange data mining software and conclude with results and suggestions.

2. Literature Review

Numerous scholars have recently examined data on cybercrime and shared their own opinions based on their databases. In order to research crime databases and other related fields, applied datamining techniques are used. These approaches are mostly concerned with classification, clustering, data reduction, social networking analysis, etc. [2] The other strategies advocate using log files as historical
information to seek for relationships based on how frequently incidents occur [3]. The government typically creates institutions like courts, prosecutors, and police that are in charge of upholding law and order in their respective nations. Controlling the frequency and occurrence of crimes is the responsibility of these agencies and other associated entities. Crime prevention plans must be developed and put into action by the crime prevention organizations [4].

**Database and Cyber Crime Parameters**

In this section, a discussion of the database and the cybercrime parameters selected for the analysis using data mining techniques are presented.

**Database**

The cybercrime data considered for this study is published by National Crime Records Bureau, which covers major crime records during the period of 2011-2022 in all the Indian states. Out of 27 variables after applying data mining techniques few parameters excelled from the database and the remaining 20 variables are considered for the present analyses. Few variables are listed in the following Table (Table 1).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Variables Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>X_1</td>
<td>Personal Revenge</td>
</tr>
<tr>
<td>X_2</td>
<td>Emotional Motives</td>
</tr>
<tr>
<td>X_3</td>
<td>Extortion</td>
</tr>
<tr>
<td>X_4</td>
<td>Fraud/illegal</td>
</tr>
<tr>
<td>X_5</td>
<td>Sexual Exploitation</td>
</tr>
<tr>
<td>X_6</td>
<td>Political Motives</td>
</tr>
<tr>
<td>X_7</td>
<td>Developing own Business</td>
</tr>
<tr>
<td>X_8</td>
<td>Spreading Piracy</td>
</tr>
<tr>
<td>X_9</td>
<td>Motives of Blackmailing</td>
</tr>
</tbody>
</table>

**The Crime Parameters**

Cybercrime parameters are simple and easy to understand. Many researchers have used to analyze some of the aspects of the cyber condition and performances. Recently, cybercrime parameters are used to find natural groups in large databases using Factor analysis and k-mean clustering techniques.

**3. Materials And Methods**

This Three data mining tools are applied for Indian cybercrime data, viz., Factor analysis, k-mean clustering technique and classification methods which are used to assess the cybercrime rate based on data mining techniques.

**Data Mining Techniques**

Data mining is often used interchangeably with Knowledge Discovery in Databases (KDD), which is the non-trivial process of extracting implicit, previously undiscovered, and potentially beneficial knowledge from data. Data mining in this situation uses the techniques of factor analysis and k-means clustering to show structural trends. The final stage of data mining involves presenting the user with the structures that were discovered in the data. Data mining is a relatively new concept, but the technology is not. The following diagram illustrates the iterative sequence that makes up a knowledge discovery process in general.
In order to put their company's results in perspective, mining also enables business owners to ascertain the effects of sales, customer satisfaction, and corporate profitability. The most crucial phases of the mining process are data mining and knowledge presentation, which make new and previously undetected structural patterns in the data visible [6].

**Orange Data Mining**

**k-means Clustering Algorithms**

The cybercrime data is subjected to the K-means clustering technique by the orange data mining widget, which produces a new data set with the cluster index utilized as a class attribute. The original class attribute is transferred to the meta-attributes section, if it still exists. The widget also displays scores for various k for clustering results. The data on cybercrime is categorized using the k-mean clustering algorithm below.

**Step 1:** Select the number of clusters using distance measure with their centroids. The measures of distances are calculated using arithmetic means of clusters.

**Step 2:** Select initialization method, k = 2, 3, 4.

**Step 3:** k-Means++, first center is selected randomly, subsequent are chosen from the remaining points with probability proportioned to squared distance from the closest center.

**Step 4:** Random initialization, the clusters are assigned randomly at first and then simplified with further iterations.

**Step 5:** Re-runs (how many times the algorithm is run) and maximal iterations (the maximum number of iterations within each algorithm run) can be set manually.

**Step 6:** The widget outputs a new data set with appended cluster information. Select how to append cluster information (as class, feature or meta-attribute) and name the column.

**Step 7:** If Run on every change is ticked, the widget will commit changes automatically [5].

**Orange Data Mining Algorithms**

From the orange data mining software, a schema is drawn with utmost care as per the research requirement. The step-by-step construction of the schema is given below and represented in figure 2.

**Step 1:** Select file widget and load your database in the form of file in .tab and .xls format.

**Step 2:** Select a data table widget and connect to the file widget, then file widget is connected to distance widget.

**Step 3:** Select k-means clustering widget and connect to selected row and table widget.

**Step 4:** Finally, Double click file widget, data table widget, selected row widget and k-means clustering widget one by one. All these widgets assign their output and display their results in the report and output window.

**Step 5:** Open scatter widget which shows the two-dimensional k-means clustering results of cybercrime with the label of all districts in the study period [7].

In the following sections the results are interpreted based on cybercrime data (5). To explore the widget the following schema is depicted (Figure 2).
Factor Analysis

Various types of factor analysis are utilized to examine the stability of cybercrime trends across the study period. Although there are several methods for reducing data and variables, factor analysis is by far the most used approach. Factor analysis lowers the variable space under examination to fewer patterns while retaining the majority of the data from the original data matrix, much like any other data reduction technique. In the current situation, principal component analysis is initially started to identify structural patterns using a linear combination of the Indian state-specific cybercrime factors.

However, the top m factors, which account for 85% of the variance, are regarded as significant when using the factor extraction approach. By looking at a variable's factor loading, orthogonal rotations like Varimax and Quartimax rotations can be used to assess how comparable two variables are. In factor analysis, the parameter of the factor model that estimates values of the common factor is the main point of interest [8].

**k-Means Clustering Algorithm**

Clustering methods are frequently used in classification issues in data mining applications. Since no assumptions are made about the group structures existing in the database, the non-hierarchical clustering approach proposed by MacQueen—also known as unsupervised classification—is used in the current work. In applied statistics, the k-means clustering method is used to identify appropriate classes [9].

With each group's members as close to one another as possible and the members of other groups as far apart as possible, this method divides or groups the data set into mutually exclusive groups. This method typically computes Euclidean distance using variables.

Algorithms

**Pruning Method**

In order to remove the outliers, a method to prune the data for each of the study period is described below:

**Step 1:** Factor analysis is initiated to find the structural pattern underlying the data set.

**Step 2:** k–means analysis is used to partition the data set into k–clusters using cybercrime parameters as input matrix.

**Step 3:** Repeat Steps 1 and 2 until meaningful groups are obtained, by removing outliers in each cycle, where an outlier is a group with only a few cybercrime parameters.

3. Results and Discussion

The number of Classes is determined as 3 that had relevant interpretations, and the researcher discusses both the Varimax and Quartimax criterion of orthogonal rotation that have been applied for the pruned...
data, aggregated by pruning method for different values of k. Both techniques of component analysis produced extremely comparable results, although the varimax rotation offered a comparatively stronger clustering of cybercrime parameters. Five factors were consistently found using factor analysis to account for 85% of the overall variation in the data across the research period, with eigenvalues close to or equal to unity (Table 3). This research shows that the clustering of the variables related to cybercrime is unstable during the study period. In original database slight changes are encountered due to statistical variations.

Table 2: k-mean clustering result for cybercrime data (Statewise)

Table 3: Rotated factor component matrix for cybercrime data (Statewise)

<table>
<thead>
<tr>
<th>Cyber Crime Variables</th>
<th>Component</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satisfaction</td>
<td></td>
<td>.966</td>
<td>-.136</td>
<td>.155</td>
<td>-.024</td>
<td>.097</td>
</tr>
<tr>
<td>Extortion</td>
<td></td>
<td>.948</td>
<td>-.018</td>
<td>.128</td>
<td>.057</td>
<td>.223</td>
</tr>
<tr>
<td>Country</td>
<td></td>
<td>.948</td>
<td>.161</td>
<td>.148</td>
<td>.097</td>
<td>.129</td>
</tr>
<tr>
<td>Causing Disrepute</td>
<td></td>
<td>.777</td>
<td>.121</td>
<td>.192</td>
<td>.431</td>
<td>.232</td>
</tr>
<tr>
<td>Sexual</td>
<td></td>
<td>.733</td>
<td>.512</td>
<td>.174</td>
<td>.093</td>
<td>.308</td>
</tr>
<tr>
<td>Financial</td>
<td></td>
<td>.725</td>
<td>.325</td>
<td>.229</td>
<td>.458</td>
<td>.221</td>
</tr>
<tr>
<td>Business</td>
<td></td>
<td>.571</td>
<td>.361</td>
<td>.411</td>
<td>.251</td>
<td>-.234</td>
</tr>
<tr>
<td>Insult</td>
<td></td>
<td>.035</td>
<td>.906</td>
<td>.126</td>
<td>.007</td>
<td>.181</td>
</tr>
<tr>
<td>Illness</td>
<td></td>
<td>.128</td>
<td>.879</td>
<td>.007</td>
<td>-.033</td>
<td>-.275</td>
</tr>
<tr>
<td>Farud</td>
<td></td>
<td>.242</td>
<td>.862</td>
<td>.245</td>
<td>-.008</td>
<td>.252</td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td>-.012</td>
<td>.741</td>
<td>.002</td>
<td>.180</td>
<td>.273</td>
</tr>
<tr>
<td>Drugs</td>
<td></td>
<td>.043</td>
<td>.616</td>
<td>-.083</td>
<td>.508</td>
<td>.287</td>
</tr>
<tr>
<td>Disruba</td>
<td></td>
<td>-.028</td>
<td>.155</td>
<td>.871</td>
<td>.383</td>
<td>-.025</td>
</tr>
<tr>
<td>Blackmail</td>
<td></td>
<td>.269</td>
<td>.146</td>
<td>.868</td>
<td>.098</td>
<td>.301</td>
</tr>
<tr>
<td>Piracy</td>
<td></td>
<td>.332</td>
<td>-.088</td>
<td>.862</td>
<td>-.166</td>
<td>.130</td>
</tr>
<tr>
<td>Political</td>
<td></td>
<td>.407</td>
<td>.172</td>
<td>.545</td>
<td>.262</td>
<td>.272</td>
</tr>
<tr>
<td>Steal community</td>
<td></td>
<td>.078</td>
<td>.509</td>
<td>.212</td>
<td>.740</td>
<td>.018</td>
</tr>
<tr>
<td>Personal</td>
<td></td>
<td>.294</td>
<td>-.168</td>
<td>.128</td>
<td>.729</td>
<td>-.007</td>
</tr>
<tr>
<td>Emotional</td>
<td></td>
<td>.236</td>
<td>.210</td>
<td>.237</td>
<td>-.007</td>
<td>.799</td>
</tr>
</tbody>
</table>
With only 3 clusters to take into account, there are 3 classes. Assigning initial group labels to cybercrime data comes after factor analysis in the data mining process, and is followed by iterative discriminant analysis using each of the three suggested methodologies. Although the results for each method of the study periods processed by the suggested algorithms were included, only the summary statistics are reported in Table 2 after the application of discriminant analysis achieved by zero percent misclassification and a sample scatter diagram are shown in figures 3 and 4.

Finally, the two methods achieved three clusters based on cybercrime data and are labelled as High Cyber Crime Rate States (HCCRS), Moderate Cyber Crime Rate States (MCCRS) and Low Cyber Crime Rate States (LCCRS). In addition, the cybercrime data provides the same results over the study period using data mining tools like, Neural Network Classification, Self-Organizing Map, Support Vector Machine, Expectation Maximization (EM) Algorithm, DBASCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, etc.

4. Conclusion
The Orange data mining program effectively displays the outcome visually. In this work, data on cybercrime are estimated utilizing orange k-means clustering, factor analysis, and classification techniques. The approaches of grouping and classification produced perfect results. In this survey, states with high rates of cybercrime include Uttar Pradesh, followed by Karnataka, while those with moderately high rates include Maharashtra, Rajasthan, and Telangana. The remaining states have low rates of cybercrime. In the Union Territory, Delhi has the highest rate of cybercrime; Chandigarh has
the second-highest rate, while the remaining areas have the lowest rates. Using the Varimax rotation approach, factor analysis identified five components that accounted for 85% of the overall variation. Personal and financial factors, criminal factors, piracy and drug issues, theft and personal factors are among the factors. To generate a set of three classifications of cybercrime data for any given year, an analysis of the findings is being done.

References:
7. Manimannan G. et.al. (2017), An Evaluation of Agriculture Productivity Indices using Orange Data Mining Techniques,