Original Research Paper

Volume-8 | Issue-4 | April-2018 | PRINT ISSN No 2249-555X

Agricultural Science

COGNIZABLE OF CYBER CRIME IN INDIAN STATES: A STATISTICAL DATA MINING APPROACH

G. Manimannan Department of Mathematics, TMG College of Arts and Science, Chennai.

R. Lakshmi Priya* Department of Statistics, Dr. Ambedkar College Arts College, Chennai *Corresponding Author

N. Manjula Devi Department of Statistics, DRBCCC Hindu College, Chennai

ABSTRACT This paper attempts to identify the performance and to classify the cyber crime rate in Indian states during the period of 2015 with twenty seven states and seven union teritories. The secondary source of data were collected from National Crime Records Bureau in India. Intially, 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 remaing 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 cyber crime, and (iii) to classify the performance based on the k-means clustering techniques. The results are classifed 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, Indian States, Orange Data mining, Factor Analysis and k-mean clustering techniques.

INTRODUCTION

This research paper concern about State as well Nation security has increased significntly since the Indian parlimentry attack, Mumbai attack, local and cross border terrorism, social media crime, etc. The Indian secruity agenciens are acively collecting domesting and foreign inteliigence to prevent the future attacks and hacking the government networks. These efoorts have in turn stimulated local security force, to more closely observed cyber criminal activities in all indian states and union territories. A major challenge facing all indian law enforcement and intelligence gathering accurately and efficently collect the crime data and analyzing the huge volumes of data, because india second largest country in population.

The current government to motivate and implementing digital india. Now days most of the urban and rural population using internet, mobiles, net banking and other source bying and selling online transctions. Deteting cyber crime can likewise be difficult because busy network traffic and frequent online transactions generate large amount of data, only little portion of which illegal activities. Data mining is a powerful tool that enables criminal investigations who may lack extensive training as data analysis to explore large amount of data quicly and efficiently[1]. The computation can process thousands of commands in seconds, saving valuable time. In addition, installing and traing software often costs less level of errors than human analysis, esspecially those who work extensive hours .In this research paper to analyse the indian states cyber crime data using Orange data mining software and conclude the results and suggestions.

REVIW OF LITERATURE

In recent days many researchers to analyse the cyber crime data and disccuss their own views based on their database. Crime data mining have been made though dataming techniques. Applied datamining techniques to study crime database and other related areas, which is mainly concernened clasffification, clustering, data reduction, social networking analysis, etc.[2] The other method to propose to employ to log files as history data to search relationship by using the frequency of occurrence of incidents[3]. The governments frequently set up organizations such as courts, prosecutions and police, which are responsible for the maintenance of law and order in their respective country. These agencies and other related organizations are responsible to control the rate and occurrence of crimes. The crime prevention agencies need to issue and implement crime prevention strategies[4].

DATABASE AND CYBER CRIME PARAMETERS

In this section, a discussion of the database and the cyber crime parameters selected for the analyses using data mining techniques are presented in following sections. The main objectives of this study (i) to assess cyber crime rates using classification method, (ii) To identify the hidden pattern using Principal Component Analysis and (iii) To visualize the assessment of cyber crime parameters.

DATABASE

The cyber crime data considered for this study is published by National Crime Records Bureau, which covers major crime records from the year 2013 to 2015 in all the Indian states. Out of 27 variables after applied data mining techniques few parameters excelled from the database the remaining 20 variables are considered for the present analyses[5]. Few variables are listed in the following Table 1.

Table 1. Variable and variable Names

Variables	Variables Name
X_1	Personal Revenge
X_2	Emotional Motives
X_3	Extortion
X_4	Fraud/illegal
X_5	Sexual Exploitation
X_6	Political Motives
X_7	Developing own Business
X_8	Spreading Piracy
X_9	Motives of Blackmailing

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

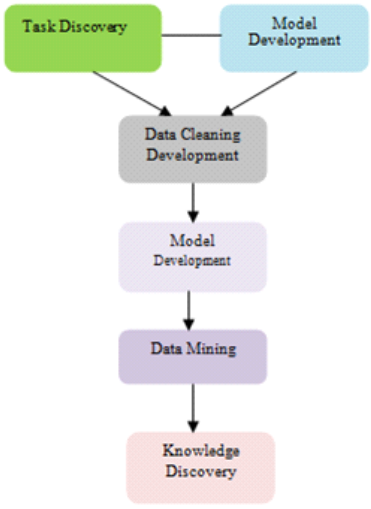
METHODOLOGIES

Three data mining tools are applied for Indian cyber crime data and they are, factor analysis, k-mean clustering technique and classification methods to assess the cyber crime rate based data mining techniques.

DATA MINING TECHNIQUES

Data mining regarded as the non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data has been popularly treated as synonym to Knowledge Discovery in Databases (KDD). In the present context, data mining exhibits the structural patterns by applying techniques namely, factor analysis and **k**-means clustering. Such structures identified from the data are presented to the user, which is the final phase of data mining. Although data mining is a new term, the technology is not. In general, a knowledge discovery process mainly consists of an iterative sequence; it is depicted in the following diagram.

Figure 1. Data mining iterative sequence



Mining also enables the company owners to determine the impacts of sales, customer's satisfaction and corporate profits to place their company's performances in perspective. The data mining and knowledge presentation processes are the most important steps in mining process, which reveal new and hitherto unknown structural patterns present in the data[6].

ORANGE DATA MINING
k-MEANS CLUSTERING ALGORITHMS

The Orange data mining widget applies k-means clustering algorithm to the cyber crime data and outputs a new data set in which the cluster index is used as a class attribute. The original class attribute, if it exists, then it is moved to Meta attributes. Scores of clustering results for various k are also shown in the widget. The following k-mean clustering algorithm is applied to classify cyber crime data.

Step 1: Select the number of clusters using distance measure with their centroids. The measures of distances are to calculate 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 iteration within each algorithm run) can be set manually.

Step 7: 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 8: If *Run on every change* is ticked, the widget will commit changes automatically.

ORANGE DATA MINING ALGORITHMS

From the oragnge data mining software, a schema is drawn with utmost care as per the research requirement. The step by step conststruction of the schema is given below and represented in figure 1.

Step 1: Select file wideget and loaded your database in the form of file like .tab and.xls format.

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

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

Step 4: Finally, Double click file widget, data table widget, selected row widget, k-means clustering widget one by one. All these widget 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 cyber crime with the label of all districts in the study period[7].

In the following sections the results are interpreted based on cyber crime data (5). To explore the widget with the following schema is depicted in Figure 2.

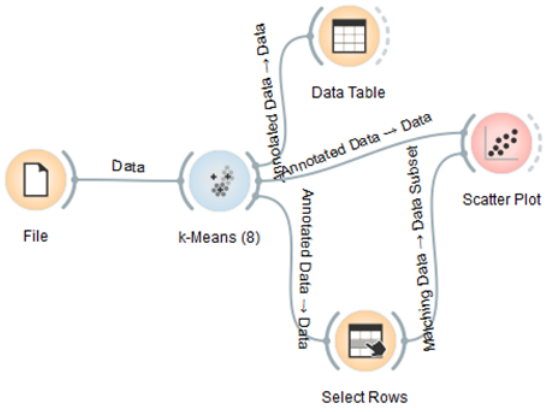


Figure 2. Orange Data Mining Schema for k-mean Clustering Method

FACTOR ANALYSES

Different factor analysis methods are used to test the stability of cyber crime patterns for the study period. Although there are several techniques of data and variable reduction, factor analysis is by far the most frequently used method. Like any other data reduction method, factor analysis reduces the variable space under consideration to a smaller number of patterns that retain most of the information contained in the original data matrix. In the present context, principal component analysis is first initiated to ascertain the structural patterns through a linear combination of the cyber crime parameters of Indian states. However, in factor extraction method the first *m* number of factors that explained 85% of variance are considered as significant. Both orthogonal rotations, such as Varimax and Quartimax rotations, are used to measure the similarity of a variable with a factor by its factor loading. In factor analysis, the focus is centered on the parameter in the factor model that estimated values of the common factor[8].

k-MEANS CLUSTERING ALGORITHM

Many data mining applications make use of clustering techniques in classification problems. In the present study, a non-hierarchical clustering algorithm suggested by MacQueen, also known as ***unsupervised classification*****,** is chosen as no presumption are made regarding the group structures present in the database. The **k**-means clustering is a technique in applied statistics that discovers acceptable classes[9].

This process partitions or groups the data set into mutually exclusive groups such that the members of each group are as close as possible to one another, and different groups are as far as possible. Generally, this technique uses Euclidean distance measure computed on 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 cyber crime 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 cyber crime parameter.**RESULTS AND DISCUSSION**In factor analysis the researcher discussed in both Varimax and Quartimax criterion of orthogonal rotation have been used for the pruned data, consolidated by pruning algorithm for different values of **k**, where number of Classes are identified as **3** that had meaningful interpretations. The results obtained under both the methods of factor analysis are very similar but the varimax rotation provided relatively better clustering of cyber crime parameters. Factor analysis revealed consistently five factors in the study period that explained 85 percent of total variation in the data with eigen values little less than or equal to unity (Table 3, Figure 2). From this analysis we observed that the clustering of cyber crime variables is unstable during the study period. In the original database is slight changes are encountered due to statistical variations.

Having decided to consider only **3** cluster number of classes to be 3, performing factor analysis, the next stage in data mining process is to assign initial group labels to the year 2016 followed by the two different suggested methods. Inspite of incorporating the results for each method for the study periods processed through the proposed algorithms, only the summary statistics are reported in *Table 2*.

Finally the two methods achieved us three clusters based on cyber crime data and k-means clustering methods 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 cyber crime data get 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,

CONCLUSIONS

The application of Orange data mining shows the result visually in a very effective manner. This study determined on estimation of cyber crime data using Orange k - means clustering methods and factor analysis methods. In clustering methods are achieved hundred percent results. The results showed that the cyber crime data classified as three categories and they are labelled as High Cyber Crime Rate States are, Uthra Pradesh followed by Karnataka, Moderate Cyber Crime Rate States are, Maharastra, Rajasthaan and Telungana, rest of states are Low Cyber Crime Rate. In Union Territory High Cyber Crime Rate in Delhi, Moderate Cyber Crime Rate Chandigarh and rest of the union territory fall Low Cyber Crime Rate. Factor analysis conceived five factors with 85 percent of total variation using Varimax rotation method. The factors are personal and financial factors, illegal factor, piracy and drugs factor, Steal and Personal factors. An overview of the results is under investigation to obtain a set of three classifications of cyber crime data for any given year.

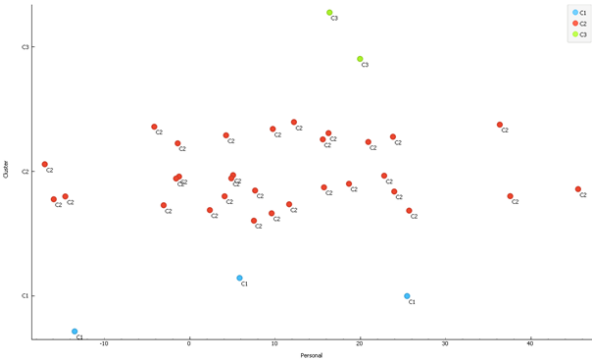
Table 2. k-mean clustering ressult for cyber crime data (Statewise)

	Cluster	States	Personal	Emotional	Financial	Extortion	using Disrepu	Satatisfaction	Farud	Insult	Sexual	Political	country	ommuni	Disrubt	Drugs	Business	Piracy	Illness	Steal	lackma	Others
15	C3	Maharashtra	21.000	18.000	682.000	17.000	35.000	1.000	354.000	234.000	113.000	5.000	39.000	0.000	2.000	4.000	17.000	1.000	6.000	5.000	16.000	625.000
22	C3	Rajasthan	15.000	23.000	287.000	36.000	34.000	7.000	31.000	1.000	33.000	0.000	13.000	0.000	0.000	5.000	2.000	2.000	0.000	0.000	9.000	451.000
25	C3	Telangana	0.000	0.000	148.000	0.000	5.000	0.000	15.000	13.000	11.000	0.000	0.000	0.000	2.000	0.000	3.000	0.000	0.000	0.000	16.000	474.000
12	C2	Karnataka	19.000	21.000	894.000	16.000	74.000	0.000	74.000	39.000	34.000	8.000	14.000	3.000	7.000	3.000	21.000	6.000	0.000	7.000	34.000	173.000
27	C2	Uttar Pradesh	29.000	34.000	1154.000	171.000	112.000	186.000	95.000	2.000	139.000	9.000	115.000	2.000	2.000	0.000	30.000	59.000	0.000	1.000	41.000	27.000
1	C1	Andhra Prad...	17.000	11.000	66.000	1.000	3.000	1.000	98.000	38.000	14.000	6.000	1.000	0.000	9.000	0.000	9.000	98.000	0.000	0.000	68.000	96.000
2	C1	Arunachal Pr...	0.000	0.000	0.000	0.000	2.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	4.000
3	C1	Assam	34.000	31.000	18.000	18.000	0.000	2.000	155.000	66.000	61.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	16.000	82.000
4	C1	Bihar	44.000	36.000	123.000	0.000	0.000	0.000	14.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	10.000	15.000
5	C1	Chhattisgarh	7.000	2.000	10.000	0.000	23.000	0.000	1.000	25.000	10.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	25.000
6	C1	Goa	0.000	2.000	4.000	0.000	2.000	2.000	1.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	4.000
7	C1	Gujarat	1.000	0.000	163.000	0.000	31.000	3.000	6.000	3.000	8.000	1.000	8.000	0.000	1.000	0.000	6.000	0.000	0.000	0.000	1.000	10.000
8	C1	Haryana	2.000	0.000	20.000	1.000	0.000	0.000	35.000	5.000	17.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	1.000	1.000	8.000	131.000
9	C1	Himachal Pr...	0.000	13.000	0.000	0.000	1.000	0.000	3.000	4.000	0.000	0.000	0.000	0.000	0.000	0.000	5.000	1.000	0.000	0.000	0.000	23.000
10	C1	Jammu & Ka...	0.000	0.000	10.000	0.000	0.000	0.000	0.000	3.000	3.000	0.000	1.000	4.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	12.000
11	C1	Jharkhand	2.000	0.000	76.000	4.000	0.000	5.000	34.000	0.000	3.000	0.000	0.000	0.000	1.000	0.000	34.000	0.000	4.000	0.000	9.000	8.000
13	C1	Kerala	34.000	5.000	46.000	3.000	22.000	0.000	31.000	33.000	32.000	12.000	0.000	0.000	0.000	0.000	1.000	3.000	0.000	0.000	11.000	57.000
14	C1	Madhya Pra...	6.000	3.000	16.000	5.000	17.000	4.000	15.000	42.000	3.000	1.000	2.000	0.000	0.000	0.000	2.000	0.000	0.000	0.000	1.000	114.000
16	C1	Manipur	0.000	0.000	0.000	0.000	0.000	0.000	2.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	3.000
17	C1	Meghalaya	1.000	4.000	4.000	1.000	4.000	0.000	9.000	12.000	5.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	4.000	12.000

Table 3, Rotated factor component matrix for cyber crime data (Statewise)

Cyber Crime Variables		Component				
	Factor 1	Factor 2	Factor 3	Factor 4	Factor5	
Satisfaction	**.966**	-.136	.155	-.024	.097	Sexual and Financial Factor
Extortion	**.948**	-.018	.128	.057	.223	
Country	**.948**	.161	.148	.097	.129	
Causing Disrepute	**.777**	.121	.192	.431	.232	
Sexual	**.733**	.512	.174	.093	.308	
Financial	**.725**	.325	.229	.458	.221	
Business	**.571**	.361	.411	.251	-.234	
Insult	.035	**.906**	.126	.007	.181	Illegal Factors
Illness	.128	**.879**	.007	-.033	-.275	
Farud	.242	**.862**	.245	-.008	.252	
Others	-.012	**.741**	.002	.180	.273	
Drugs	.043	**.616**	-083	.508	.287	
Disrubt	-.028	.155	**.871**	.383	-.025	Piracy and Drugs Factor
Blackmail	.269	.146	**.868**	.098	.301	
Piracy	.332	-.088	**.862**	-.166	.130	
Political	.407	.172	**.545**	.262	.272	
Steal	.078	.509	.212	**.740**	.018	Steal Factor
Community	.294	-.168	.128	**.729**	-.007	
Personal	.236	.210	.237	-.007	**.799**	Personal
Emotional	.420	.196	.158	.113	**.758**	

Figure 3. Presentational Emotion data (State wise)



REFERENCES

[1]. Gupta G.K (2012), Introduction to Data Mining with Case Studies, PHI Learning Private Limited, New Delhi.

[2] Hsinchun Chen, Wingyan Chung, Yi Qin, et al. Crime Data Mining: An Overview and Case Studies. Proceeding of the 2003 annual national conference on Digital government research, Boston, M.A, 2003, pp 1-5. \

[3] T. Abraham and O. de Vel. Investigating profiling with computer forensic log data and association rules. Proc. Of the IEEE International Conference on Data Mining (ICDM'06), 2006, pp 11-18.

[4] H. F. Lin and J. M. Liang Event based ontology design for retrieving digital archieves on human religious self-help consulting. Proc. Of 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service, 2005 pp. 453-475.

[5]. Crime in India 2015 Compendium, National Crime Records Buraeu, India..

[6]. Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd ed, 443-491.

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

[8] Everitt (1980), Cluster Analysis, Halsted Press, Division of John Wiley and Sons, New York.

[9]. Anderson TW (1984). An Introduction to Multivariate Statistical Analysis, 2/e, John Wiley and Sons, Inc., New York.