May - June 2020
ISSN: 0193-4120 Page No. 3006-3010

Prediction and Pattern Identification of Crime Data using Data Mining Techniques

J Omana[1], B Gunasundari[2], C Kamatchi[3], I Mohan[4], R Thiagarajan[5]

[1,2,3,4,5]Assistant Professor, Prathyusha Engineering College
omanajayakodi@gmail.com[1], gunasundari.cse@prathyusha.edu.in[2], kamatchi.it@prathyusha.edu.in[3], mohan.it@prathyusha.edu.in[4], thiagarajan.cse@prathyusha.edu.in[5]

Article Info
Volume 83
Page Number: 3006-3010
Publication Issue:
May - June 2020

Article History
Article Received: *19 August 2019*
Revised: *27 November 2019*
Accepted: *29 January 2020*
Publication: *12 May 2020*

Abstract
Data Mining is used to extract knowledge information from a large amount of data. Due to the frequent occurrence of crime in India it is the need of the hour to detect crime and bring up a model that helps in prediction of crime. Hence crime analysis needs a systematic approach to identify various patterns occurrence. Hence data mining techniques are used to carry out this process in an accurate and effective manner. Due to the large available of techniques it is very mandatory to choose the appropriate method to carry out this process. This paper address the challenges and issues in different algorithm and provides the comparison in bringing up the suitable technique for crime model detection. It gives us a good amount of clarity in choosing up the appropriate techniques. Added to this some algorithms are proposed and found to be efficient when compared to other techniques.

Keywords: *Decision, Data Mining, Pattern identification, Crime Analysis*

1. Introduction

In current scenario when we see crime rate it keeps increasing day to day. When coming to the point of detecting and preventing crimes, it is really difficult since it is not in same sequence or at random manner. Due to advancement of technologies even the crime doing people are using high-tech tools to get escaped without being caught. To solve these types of issues various research studies have provided different techniques to sort out these crimes. By using these techniques investigation department will be able to detect the scenario and investigate the crime. When considering the crimes, murder, sexual abuse, stealing, rape etc have increased in due period. This situation is prevailing due to the case that either the crime doing people have lost the fear of punishment or investigation team needs more advancement in finding out the crimes well in advance. When it is difficult to predict who and all have involved in doing a particular crime it is said to be that probability prediction of crime occurrence can be done. But once the result is out, we are not sure of 100% accurate about the result and able to achieve security in fraudulent areas and alert zone. To bring up a powerful analytical tool it is mandatory to have more number of evaluations of crime data in detail.

Hence to provide solution for these problems mentioned above we can use data mining which can extract useful information and hidden patterns from a large amount of data. Using the extracted information we would be able to find the scenario of crime happening, pattern and frequency which helps us predict the crime in advance. By using these techniques we would be able to get benefitted in two ways: one is to solve the crime and investigate, get results fast and other one to do automatic crime detection. For our notice, only for few decades we found to use spatial data mining to be an apt solution for a large amount of data. The data related to the crime is not present s a whole in repository hence it is collected from various resources like web, blogs, news sites, social media etc. This data is been stored and labeled as crime database. After collecting the large amount of data, now our primary focus is to bring up an efficient, accurate pattern recognition tool for detecting crime. But in this type of process there are certain challenges and difficulties listed:

- Due to large amount of crime existing day to day we need to store and analyze.

May - June 2020
ISSN: 0193-4120 Page No. 3006-3010

- Also the data stored is incomplete and inconsistent, hence analysis of data is found to be difficult.
- There is also limitation in receiving data from the bodies of law.
- Depending on the data we have stored the accuracy is achieved.

Inspite of collecting the data and storing it, it is difficult to analyze and find the pattern of crime. Hence it takes so much of time in finding out the exact pattern of how the crime has occurred. Even though after detecting and coming up with certain patterns from crime dataset, when analyzing the new dataset if the pattern is found not to be fitting within the predicted pattern then this new pattern is classified to a different category. By setting up these different patterns it is easy to analyze and predict the crime scenario well in advance.

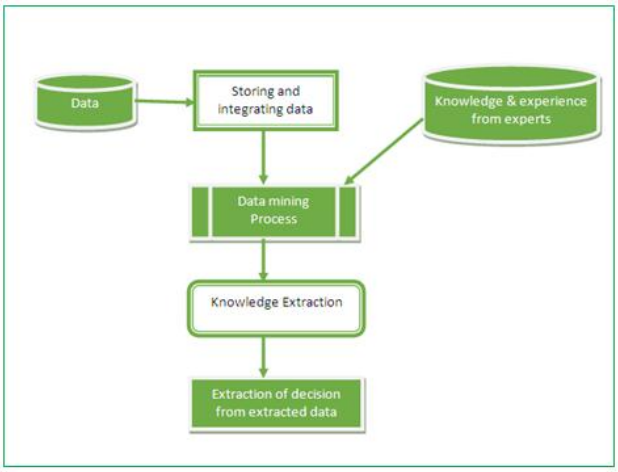

Figure 1: Data Mining Process

In existence different clustering algorithms are used to group various crimes together. Normally we detect certain features of crime person from each scene, say that suspect is found to be youth, then other will describe having tattoo in right arm, other theft describe he is middle age etc. According to the current scenario these are identified by the investigation team by analyzing different photos taken at crime scene and it is done manually. Main consideration for this technique is that we will be able to arrive crime pattern of that particular incident. By considering the classification algorithms we won't be able to get accurate results since it will be able to predict happened crimes and existing one comparison. But we consider in real world the crime method and strategy of crime is supposed to differ at certain period of time. So the clustering techniques seem to be more prominent than using the existing classification method.

2. Methodology

The various classification step considered in successful analysis and prediction of crime patterns are listed as follows:

i. Data Collection
ii. Classification of data based on features
iii. Identification of Pattern
iv. Prediction of crime
v. Visualization of data patterns

All these steps are explained in detail in below sub-section. Each one can be implemented by various data mining algorithms.

A. Data Collection

Large amount of data is collected from website, news site, blogs, RSS feed, social media etc. The data collected are stored in database and process is been carried out. The data collected from various sites are large and unstructured. It has no labels such as no of fields, parameters; features etc differ from one crime to the other. The collected data are schema less and does not have a defined structure. Hence database that suits unstructured data is used. These data do not have joins since the scene rate differs and due to this there is complexity of pattern detection.

B. Classification

Various classification algorithms are present and it is carried out by researchers. The survey of these algorithms is listed in the table below. Detailed analysis of these algorithms are done by which the challenges and issues are identified in using up crime data. Naïve bayes is a supervised algorithm and can be used for statistical classification. This classifier is used to find the probability of the different classes present in the dataset. When a new crime record arrives it finds out which class suits its best and fits into them. This algorithm is very simple and suits the best when compared to other classification techniques and found to be quick compared to even logistic regression.

The next algorithm considered is SVM(Support vector Machine) which brings easiness of implementation, taking large amount of memory and is found to have large difference in performance wise when compared to other algorithms. But one thing considered here is when the size of training set is large the execution speed automatically decreases that is it is said to be inversely proportional. Comparison is made with Naïve Bayes classifier and the algorithm of its working is stated as follows:

Algorithm 1: Pseudocode

1. Training data set is taken and considered as TD, consist of class X and Y
2. Probability calculation for class X is done=>No of objects in class X / Total no. of objects
3. Probability calculation for class Y is done=>No of objects in class Y / Total no. of objects
4. To find pi, the whole total is considered, word frequency of each class

pa = word frequency of class A
pb= word frequency of class B

5. Next is to find the conditional probability of each class

May - June 2020
ISSN: 0193-4120 Page No. 3006-3010

P(word1 / Class X) = wordcount / pi(X)
P(word2 / Class Y) = wordcount / pi(Y)
P(word1 / Class X) = wordcount / pi(X)
P(word2 / Class Y) = wordcount / pi(Y)
……
P(wordn / Class Y) = wordcount / pi(Y)

6. Avoiding frequency problems , use uniform distribution
7. To classify the new crime record based on the probability P(E/W)
8. The record is assigned to the class having highest probability.

C. Pattern Identification

Third step of identification is where we need to identify patterns and section happened in crime. Different algorithms are considered to find the pattern identification. Association mining algorithms are considered under this section. One of the algorithms considered is Apriori Algorithm. This algorithm generates the rules associated to the crime scene. These rule set helps in exhibiting the pattern of database association. Based on the crime scene location, attributes are considered which includes the location, weather condition, type of crime, evidences etc. Then based on the rules generated the crime pattern are identified. When the case of new crime is received if it is considered to be the same pattern then that area is said to be getting the same pattern of crime and area is declared to be sensitive crime zone.

For eg, 150 crime news is taken and considered. Apriori mines the frequent pattern of crime in each area wise. If a same pattern of crime has occurred in that place then it is found to be that similar crimes can occur in that area. Several attributes are considered for identification of patterns

- Attribute 1, attribute 2, attribute 4, attribute 5
- Attribute 1, attribute 3, attribute 4, attribute 5
- So if the crime occurs like the same said above then it is found to have probability of the occurrence in the area again.

D. Prediction

Next phase is the prediction. Decision tree algorithm can be used to do this process. This algorithm is very simple when compared to other prediction methods and is easy to interpret. This brings out decisions that are better about variables. Region wise based on the data decision tree model is arrived. When the date of crime and attributes are given as input to prediction technique then area of crime prone is detected. The tree generated has nodes which fall into three types. They are

- A Root node, no incoming edges and 1 or more outgoing edges.
- Internal node, one incoming edge and more than one outgoing edges.
- Leaf or end node, 1 incoming and 0 outgoing edge.

This technique is said to be predictive model and considers the generated binary rules that is used to find out the value of the class. By using the tree generated we will be able to determine the splitting value of node, whether to proceed or stop and assign terminal values for nodes.

E. Visualization

The areas which are of high alert can be represented visually by different graphs denoting the activity level with a different color, say higher probability with darker color and low area with less color. Sample image below is represented with regions having higher probability.

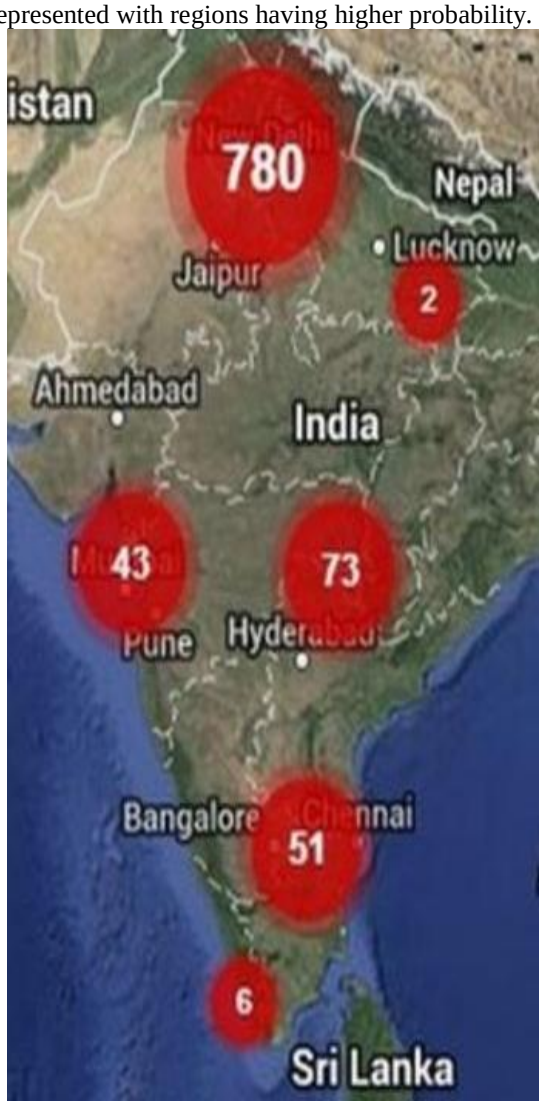

Figure 2: Crime prone area – graph highlighted

3. Issues and Challenges Identified

The overall comparison of various algorithms done in each of the research paper along with the challenges are compared and discussed here. As said earlier, previous sections denote one of the best algorithm that gives out the better results. The summary table is given below:

May - June 2020
ISSN: 0193-4120 Page No. 3006-3010

Table 1: Comparison summary of various algorithms

Research Work	Algorithm used	Method	Issues	Challenges
[1]	Apriori Algorithm	Extract features from database and crime patter analysis	Detecting is not appropriate	Need accurate detection
[6]	Neural Network	Data collection and extraction from police bodies	Model detection and visualization is not available	Need improvisation in crime model and visualization
[7]	Sender reputation algorithm	Classification of emails, received from offenders	No model for crime	Need model for crime attacks
[10]	Cox regression	Relationship is done using Social Network Analysis.	Data collection is not proper and no visualization	Data to be collected to improve visualization
[12]	Crawler, Document Classifier	Data extracted and analyzed for frequency and crime type.	Crime prediction and model is not created	Create model which improves the performance

4. Conclusion

In this paper, the classification of accuracy and level of prediction is done for various data set in comparison of different algorithms. Classification algorithm is said to receive 92% of accuracy and it is achieved through Naïve Bayes algorithm. When considered for generating rule set to predict the high prone area, among various association rule mining algorithm Apriori is said to be performing well giving out better results. Based on these rule sets a pattern is identified and model is built using decision tree. Based on each place the model is build by segregating the training data. Considering the crime patterns it is not static and is said to differ over a period of time. Through these training set we can train the system to act as a model.

Whenever a new record comes in, the factors are considered and if finds to fit it with existing pattern then it is added else new factors are identified and pattern is sorted out. So for better prediction more number of attributes are considered. Even more accuracy can be retrieved when the segregation and pattern identification are done according to the specific regions. Time is said to be a foremost and important attribute since we not only detect only crime but also to be done at the exact and proper time.

5. Future Work

In future, criminal profiling can be done and incorporated into the database which can be considered as one more attribute for crime identification and prediction. It includes making a record of criminal with their characteristics. Their behavior is added and so hence each crime doing person will be holding a profile which holds full content of that person. Through this investigation team will get proper assessment of the crime doer and can compare the belongings received in the crime scene with the offender. Second attribute can be considered is type of snatching. When considering the attributes for this type we can be able to add much more details to the profile. Since the victim will be seen by one or more people around while the crime is done. Hence attributes considered are weapons, facial features, location, type of vehicle offender had, what type of crime how many people involved etc. By these parameter it will be helpful for the investigation team to get proper results at faster rate.

References

[1] S. Sathyadevan, M. Devan, and S. Surya Gangadharan, "Crime analysis and prediction using data mining," in Networks Soft Computing (ICNSC), 2014 First International Conference on, Aug 2014, pp. 406– 412.

[2] V. Vaithiyanathan, K. Rajeswari, R. Phalnikar, and S. Tonge, "Improved apriori algorithm based on selection criterion," in Computational Intelligence Computing Research (ICCIC), 2012

May - June 2020
ISSN: 0193-4120 Page No. 3006-3010

IEEE International Conference on, Dec 2012, pp. 1–4.

[3] C. Chu-xiang, S. Jian-jing, C. Bing, S. Chang-xing, and W. Yun-cheng, "An improvement apriori arithmetic based on rough set theory," in Circuits, Communications and System (PACCS), 2011 Third PacificAsia Conference on, July 2011, pp. 1–3.

[4] A. Kondaveeti, G. Runger, H. Liu, and J. Rowe, "Extracting geographic knowledge from sensor intervention data using spatial association rules," in Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on, June 2011, pp. 127– 130.

[5] Lafferty, McCallum, and Pereira (200 I); Sutton and McCallum (20 1 O). "http://aliasi.comilingpipe/demos/tutorial/classity/read -me.html [2010].

[6] M. Chau, J. J. Xu, and H. Chen, "Extracting meaningful entities from police narrative reports," in Proceedings of the 2002 Annual National Conference on Digital Government Research, ser. dg.o '02. Digital Government Society of North America, 2002, pp. 1–5. [Online]. Available: http://dl.acm.org/citation.cfm?id=1123098.1123138

[7] A. Kawbunjun, U. Thongsatapornwatana, and W. Lilakiatsakun, "Framework of marketing or newsletter sender reputation system (fmnsrs)," in Advanced Information Networking and Applications (AINA), 2015 IEEE 29th International Conference on, March 2015, pp. 420– 427.

[8] Li Zhang, Yue Pan, and Tong Zhang. Focused named entity recognition using machine learning. In Proceedings of the 27th Annual International ACM SIGlR Conference on Research and Development in Information Retrieval, SIGIR '04, pages 281-288, New York, NY, USA, 2004. ACM.

[9] L. Cunhua, H. Yun, and Z. Zhaoman, "An event ontology construction approach to web crime mining," in Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on, vol. 5, Aug 2010, pp. 2441–2445

[10] S. Kaza, D. Hu, H. Atabakhsh, and H. Chen, "Predicting criminal relationships using multivariate survival analysis," in Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains, ser. dg.o '07. Digital Government Society of North America, 2007, pp. 290–291.

[11] Wikipedia contributors.(l2 May 2014 at 19:05.), Series Finder. 412 [Online]. Available: http://en.wikipedia.org/wikilCrime_analysis

[12] sI. Jayaweera, C. Sajeewa, S. Liyanage, T. Wijewardane, I. Perera, and A. Wijayasiri, "Crime analytics: Analysis of crimes through newspaper articles," in Moratuwa Engineering Research Conference (MERCon), 2015, April 2015, pp. 277–282.