A Survey on Data Mining Techniques for Predicting Crime Patterns & Suspect Prediction

Vinayashree G[1] R. Anitha[2]
[1,2]Department of Computer Science and Engineering
[1,2]National Institute of Engineering (NIE), Mysuru, Karnataka, India

Abstract— In recent age the data mining is data analyzing techniques helps to analyze crime data that are previously stored from various sources to find patterns and trends in crimes and can easily find the suspects. In additional, the accuracy and time of tracing are robust while data mining technique is indulged. However, there are many data mining techniques. In order to increase efficiency of crime detection and for predicting the suspects, it is necessary to select the data mining techniques suitably. This paper reviews the literatures on various data mining applications, especially applications that applied to solve the crimes and predict the suspects. Survey also throws light on research gaps and challenges of crime data mining. In additional to that, this paper provides insight about the data mining for finding the crime patterns and suspects who had made the crimes that are used appropriately and to be a help for beginners in the research of crime data mining.

Key words: Data Mining; Crime Patterns; Data Analysis: Suspects: Data Mining Techniques

I. INTRODUCTION

In this survey, we have various aspects of past and recent research directions in crime analytics. Data is the vital thing that decides about the decision of various aspects in all fields from financial, medical, marketing, demographic and scientific, the list goes on. Data mining is the powerful technology to analyse the data skilfully from different perspectives and compress it to useful information [1]. The finding of the cold spot and hot spot is the challenging task in crime analysis. Crime is the act that harms the public, increases the violence, demolishes the assets and denies the respect to people. Distribution of Crime is not even across the globe. The huge available data and hands-on experience helps for investigation [1].

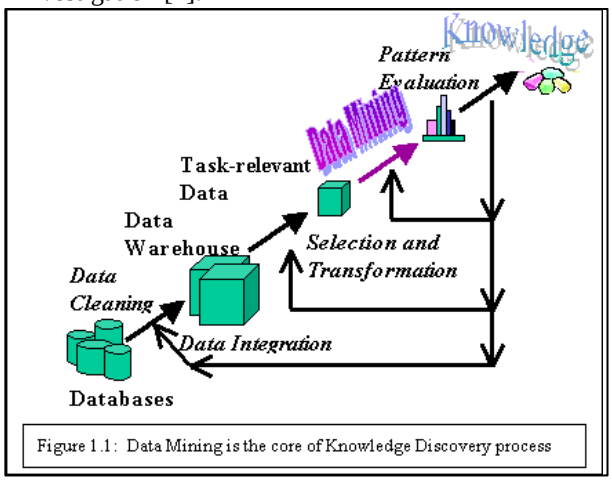

Fig. 1: Phases of Data Mining

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

1) Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
2) Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
3) Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
4) Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
5) Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
6) Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
7) Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

Proposed system is applicable in the field of crime. Proposed system includes modeling of crimes for finding suitable algorithms to detect the crime, precise detection, data preparation and transformation, and processing time. Proposed system identifies crime behavior, crime predicting, precise detection, and managing large volumes of data obtained from various sources. Proposed system is an automation for complaints registration, crime pattern prediction based on the previous crime details collected from various sources [11].

II. DATA MINING FUNDAMENTALS

Data mining is widely known as Knowledge Discovery in Database, refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques are used to analyze the large volumes of data to discover the hidden patterns and relationships helpful in decision making for predicting the crime pattern and the suspects. While data mining and knowledge discovery in database are frequently treated as antynoms, data mining is actually part of the knowledge discovery process.

Data mining is the analysis process used to analyze the large data source to find crime patterns and criminals related to that particular crime. To extract the hidden information, there are important factors for analysis as follows: 1) The data used for analysis require the accuracy and sufficiency. 2) Knowledge and experiences of specialists. The knowledge results obtained from data mining processes are used to guide in decision making and to solve the crime

patterns. The data mining diagram is shown in Figure. 2. In the data mining, the analyzing techniques are explained in the following sub-sections[7].

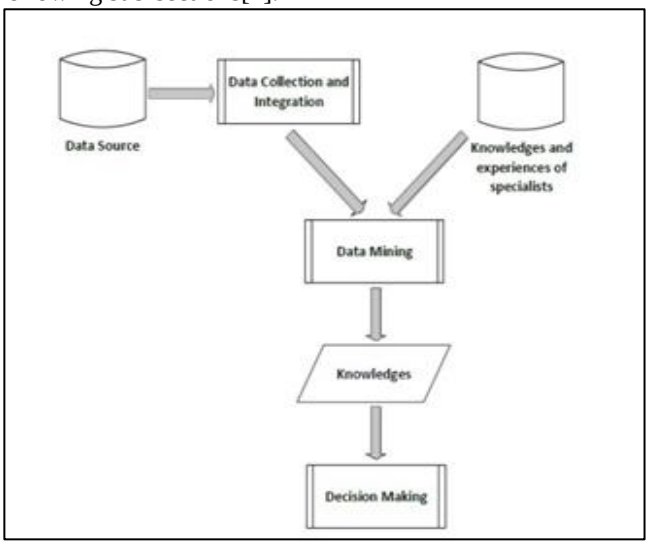

Fig. 2: Data mining Fundamentals

A. Classification:

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This technique is supervised learning method that used to assign objects to one of many pre-determined categories. The algorithms of classification have been widely applied to the several problems that include many various applications. For example, it helps in detecting the suspect vehicles and intruders, the prediction of heart disease, the categorizing the document, etc. The basic concept of classification is described as the following: A collect data, also known as an input data, is used to process in a classification task. Each record consists of the attribute set and a class label. The class label is pre-determined category. A collect data is divided into two sets. 1) Train set is partitioned randomly that is used to create a classification model, also known as a classifier, to predict the class of the new unknown record. 2) Test set is a remaining set that is used to evaluate the performance of the classification model. For building the classification models, there are many systematic approaches such as: decision tree, nearest neighbor, Bayes' Theorem and neural network, etc [8].

B. Clustering:

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters achieves simplification. It models data by its cluster. This is one of the data analyzing technique in unsupervised type. This technique is used to divide the same data into the same group and the different data into the other group. The clustering techniques have a variety of concepts. The use of clustering techniques depends on applied fields [9]. For the simple and effective clustering techniques, there are several algorithms such as K-means, Hierarchical Clustering and Expectation-Minimization

C. Decision Trees:

Decision tree is tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include the Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). The main component consists of root node, internal nodes and leaf or terminal nodes. The root node is the top of tree which is chosen from important attribute in data set and used to separate records. The internal nodes contain the attributes that is test condition. Each internal node has dissimilar characteristics. The leaf or terminal nodes represent the class label or the result of prediction. In decision process, the root node is first considered by comparing the test condition. The test outcome determines the next appropriate branch to consider the next internal nodes. The internal node is considered steadily until a leaf node is reached [5]. The class label or the result of prediction is assigned to the record.

D. Association rule:

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. This technique is unsupervised learning method it helps to find the hidden knowledge in unlabeled data. It helps to solve the issues if the learners get the unlabeled example data. In additional, association rule can discover the interesting co-occurrences of objects in large data sets. In the basic of association rule, the rule consists of two parts. 1) The antecedent, which is on the left side or called the left hand side (LHS). 2) The consequent, which is on the right side or called the right hand side (RHS). A form of general association rule is LHS ! RHS, where LHS and RHS are disjoint item-sets. Apriori algorithm is used to help prune the candidates explored during frequent item-set generation to reduce the processing time. Apriori algorithm needs to scan the all item-sets. So, it uses a long period of time as well. Reference [4] proposed the improved Apriori algorithm by using the compressed database algorithm for association rule mining to reduce the amount of time needed to read data from the database. Apriori designed and developed by [5] has improved Apriori algorithm to find the effective association rule and to reduce the amount of processing time. Additionally, there are many techniques that helps to analyze associations between two item sets more effectively such as mutual information concept [3], [6], association bundle [7], audio watermarking [8], etc.

III. THE DATA MINING TECHNIQUES FOR ANALYZING CRIME PATTERNS

Nowadays, the various data mining techniques are used for different objectives such as: criminality, science, finance and banking, email filtering, healthcare and other industries [14].

According to Software Engineering the approach adopted to develop this application is the Iterative waterfall Model. The iterative waterfall Model is a systematic approach that begins at the feasibility study phase and progress through analysis, design, coding, testing, integration and maintenance. Feedback paths are there in each phase to its preceding phase as show in the fig to allow the correction of the errors committed during a phase that are detected in later phase [2].

In this paper, system predicts crime patterns based on the past crime details collected from various sources. Here

system uses "*Association Rule Mining*" to analyze previous crime data and to extract the crime trends. Another module predicts the suspects based on some constraints/ attributes. Attributes depends on the type of the crime. For example: for "*chain stanching*" crime type we can consider the attributes such as location, time, item type, criminal height, color, appearance, type of vehicle used, criminal dress type, hair style etc.. Here we make use of data mining technique called as "*Classification rules*" for suspects prediction.

IV. PROPOSED SYSTEM

A. Problem Statement:

Determining the crime patterns is a major challenge in today's world to reduce the crimes and to take the precautionary measures to avoid crimes.

Association (or relation) is probably the better known and most familiar and straightforward data mining technique. Here, we make a simple correlation between two or more items, often of the same type to identify patterns.

For example, Market-basket analysis, where we track people's buying habits, we might identify that a customer always buys cream when they buy strawberries, and therefore suggest that the next time that they buy strawberries they might also want to buy cream.

1) Existing System:

No system to predict crimes and their patterns. We have many software and tools to maintain crime details, police stations details, their employee details, compliant details etc.

Web based complaints management system, child abuse complaints management system are the existing system, these online based applications will permit the user to post the complaints online but no prediction of crime patterns, so that we can take precautionary measures to avoid crimes. Limitations are:

- Stores crime data and retrieves the same
- No extraction of useful information
- No extraction of crime patterns
- Lack of user satisfaction
- Less Efficient

B. Crime Pattern Pridiction:

Prediction of crime is a great aid to the administration in order to curb the crime incidences. Prediction is stating probability of an event in future period time.

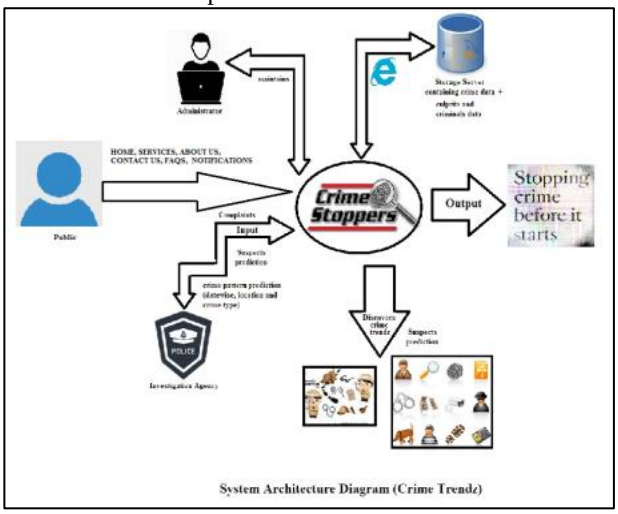

Fig. 3: system architecture

We choose to use Apriori technique over any unsupervised technique such as classification, since crimes vary in nature widely and crime database often contains several unsolved crimes and also the unpredicted suspects. From figure 3 therefore, classification technique that will rely on the existing and known solved crimes, will not give good predictive quality for future crimes. Thus, in order to be able to detect newer and unknown patterns in future, clustering techniques work better.

1) Association Rule Mining:

Apriori Algorithm:

1) STEP 1: Scan the Crime data set and determine the support(s) of each item.
2) STEP 2: Generate L1 (Frequent one crime item set).
3) STEP 3: Use Lk-1, join Lk-1 to generate the set of candidate k - crime item set.
4) STEP 4: Scan the candidate k crime item set and generate the support of each candidate k –crime item set.
5) STEP 5: Add to frequent crime item set, until C=Null Set.
6) STEP 6: For each crime item in the frequent item set (L) generate all non-empty subsets.
7) STEP 7: For each non empty subset determine the confidence. If confidence is greater than or equal to this specified confidence .Then add to Strong Association Rule.

For example if the crime datasets of Mysuru city is taken about 5 days item ste then it can be showed as:

City Name - Mysuru (A,B,C,D and E are crime types)
Minimum Support = 50%
Minimum Confidence = 80%
Item set : A, B, C, D, and E (crime types)

Tid	Crime Types
1	A,C,D
2	A,C,E
3	A,B,C,E
4	B,E

Table 1:

2) Frequent Item Set (L):

Items	Support
A	75%
B	50%
C	75%
E	75%
AC	75.00%
AE	50%
BE	50%
CE	50%
ACE	50%

Table 2:

3) Strong Association Rule:

- {B} -> {E}
- {CE} -> {A}
- {AE} -> {C}
- {A} -> {C}
- {C} -> {A}

C. Suspect Pridiction:

For suspect prediction implementation, we need a Training set and Pattern. In the training set, we have to find the overall probability of the class that we are going to predict the

suspects. After finding the probability of the class we need to find the probability for all the given attributes in the training set.

1) Step 1: Scan the crime dataset (storage servers)
2) Step 2: for each attribute a, calculate the gain [number of occurrences]
3) Step 3: Let a_best be the attribute of highest gain [highest count]
4) Step 4: Create a decision node based on a_best – retrieval of nodes[patient] where the attribute values matches with a_best.
5) Step 5: recur on the sub-lists [list of patient] and calculate the count of outcomes[Stages] – termed as sub nodes. Based on the highest count we classify the new node.

For Example
Crime Type - Robbery
Attributes(Clues) – C1,C2,C3 [m=3]
Subject (Suspects) – Anoop,Sujay [p=1/2=0.5]
Training Dataset

Name	C1(X,Y,Z)	C2(A,B,C)	C3(P,Q,R)	Culprit
Anil	X	A	P	Anoop
Kumar	X	B	Q	Anoop
Ajay	Y	B	P	Sujay
Naveen	Z	A	R	Anoop
Akash	Z	A	Q	Sujay

Table 3:

New Complaint Features – Akul C1-X,C2-A,C3-R
Which Anoop/Sujay - ?
Feature Count (X) in the dataset = 2
Feature Count (A) in the dataset = 3
Feature Count (R) in the dataset = 1
Sort();

Feature	Count
A	3
X	2
R	1

Table 4:

A – Anoop (2) & Sujay(1);This algorithm is based on single attribute values.
Output

Suspects	Priority
Anoop	2
Sujay	1

Table 5:

Thus with the higer priority will be the suspect .from this algorithm using the sing attribute we can obtain the culprit who had done the crime from the crime item set

V. Conclusion & Future Enhancement

Crime are characterized which change over time and increase continuously. The changing and increasing of crime lead to the issues of understanding the crime behaviour, crime predicting, precise detection, and managing large volumes of data obtained from various sources. Research interests have tried to solve these issues. However, these researches are still gaps in the crime detection accuracy. This leads to the challenges in the field of crime detection. The challenges include modeling of crimes for finding suitable algorithms to detect the crime, precise detection, data preparation and transformation, and processing time.

A. Future Enhancement:

- Online Complaint Registration (Public)
- Query Module (public can post queries to administrator)
- Use any of the valid Id like voter id , adhaar etc. as unique id for adding the complaints by the public.
- Security tools and encryption tools maintain the records that are very important and complex

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] T. Pang-Ning, S. Michael, and K. Vipin, Introduction to Data Mining, 1st ed. Pearson, 5 2005.
[3] S. Kaza, Y. Wang, and H. Chen, "Suspect vehicle identification for border safety with modified mutual information," in Proceedings of the 4th IEEE International Conference on Intelligence and Security Informatics, ser. ISI'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 308–318.
[4] V. Vaithiyanathan, K. Rajeswari, R. Phalnikar, and S. Tonge, "Improved apriori algorithm based on selection criterion," in Computational Intelligence Computing Research (ICCIC), 2012 IEEE International Conference on, Dec 2012, pp. 1–4.
[5] 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.
[6] S.Kaza,T.Wang,H.Gowda,andH.Chen,"Targetvehicleidentification for border safety using mutual information," in Intelligent TransportationSystems,2005.Proceedings.2005IEEE,Sept2005,pp.1141–1146.
[7] Revatthy Krishnamurthy,J. Satheesh Kumar, "SURVEY OF DATA MINING TECHNIQUES ON CRIME DATA ANALYSIS" , International Journal of Data Mining Techniques and Applications Vol 01, Issue 02, December 2012.
[8] Shyam Varan Nath Florida, "Crime Pattern Detection Using Data Mining", Proceedingsof the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2006 Workshops) (WI-IATW06).
[9] Priyanka Gera, Rajan Vohra, "City Crime Profiling Using Cluster Analysis" ,International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014.
[10] Malathi. A and Dr. S. Santhosh Baboo, "Enhanced Algorithms to Identify Change in Crime Patterns", International Journal of Combinatorial Optimization Problems and Informatics, Vol. 2, No.3, Sep-Dec 2011, pp. 3238, ISSN: 200.
[11] A Survey of Data Mining Techniques for Analyzing Crime Patterns Ubon Thongsatapornwatana Defence Technology Institute Nonthaburi, Thailand

ubon.t@dti.or.th 978-1-5090-2258-8/16/$31.00 ©2016 IEEE.

[12] I. 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.

