Datamining in Cyber Crime Analysis and Prediction

Deepthi .K[1], Aruna .A .S[2]

[1]Asst. Professor, Department of Computer Science, College of Engineering, Kozhikkode, Kerala
[2]Asst. Professor, Department of Computer Science, Vadakara College of Engineering, Vadakara, Kozhikkode, Kerala

Abstract: ***Crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime. Our system can predict regions which have high probability for crime occurrence and can visualize crime prone areas. With the increasing advent of computerized systems, crime data analysts can help the Law enforcement officers to speed up the process of solving crimes. Using the concept of data mining we can extract previously unknown, useful information from an unstructured data. Here we have an approach between computer science and criminal justice to develop a data mining procedure that can help solve crimes faster.***

1. Introduction

Crime cannot be predicted since it is neither systematic nor random. Even though we cannot predict who all may be the victims of crime but can predict the place that has probability for its occurrence.

Final results cannot be assured of 100% accuracy but the results shows that our application helps in reducing crime rate to a certain extent by providing security in crime sensitive areas. For generating powerful crime analytics tool we have to collect crime records and evaluate it. it is only within the last few decades that the technology made spatial data mining a practical solution for wide audiences of Law enforcement officials which is affordable and available. The available criminal data or records is limited we are collecting crime data from various sources like web sites, news sites, blogs, social media, RSS feeds etc.

These data are used as a record for creating a crime record database. The important challenge we are facing is developing a better, efficient crime pattern detection tool to identify crime patterns effectively. Important challenges are:

- Increase in crime information that has to be stored and analyzed.
- Analysis of data is difficult since data is incomplete and inconsistent.
- Limitation in getting crime data records from Law Enforcement department.
- Accuracy of the program depends on accuracy of the training set.

Finding the patterns and trends in crime is a challenging factor. To identify a pattern, crime analysts takes a lot of time, scanning through data to find whether a particular crime fits into a known pattern. If it does not fit into an existing pattern then the data must be classified as a new pattern. After detecting a pattern, it can be used to predict, anticipate and prevent crime.

The reason for choosing this method is that we have only data about the known crimes we will get the crime pattern for a particular place. So the classification technique that rely on the existing and known solved crimes, will not give good predictive quality for future crimes. Behaviour of crimes change over time, so in order to be able to detect newer and unknown patterns in future, clustering techniques work better.

There are steps in doing Crime Analysis:

1) Data Collection
2) Classification
3) Pattern Identification
4) Prediction
5) Visualization

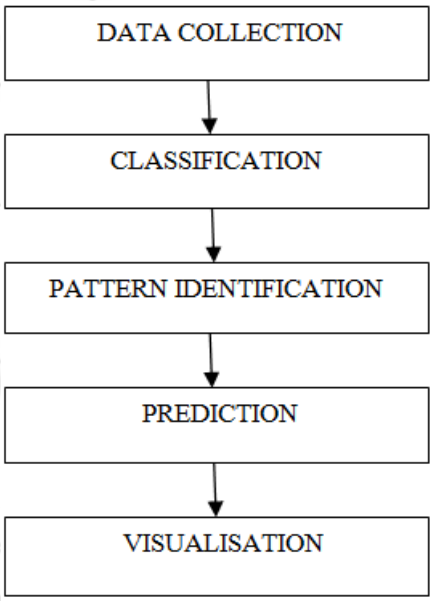

Figure 1: Steps in Crime Analysis

2. Related Work

Most of the countries like England, Cambridge Police Department have done a similar one named Series Finder for finding the patterns in burglary. For achieving this they used the modus operandi of offender and they extracted some crime patterns which were followed by offender. The algorithm constructs modus operandi of the offender. The M.O. is a set of habits of a criminal and is a type of behaviour used to characterize a pattern.

3. Methodology

A. Data Collection

In data collection step we are collecting data from different web sites like news sites, blogs, social media, RSS feeds etc. The collected data is stored into database for further process. Since the collected data is unstructured data we use

MongoDB. Crime data is an unstructured data since the no of field, content, and size of the document can differ from one document to another the better option is to have a schema less database. And absence of joins reduces the complexity.

Other benefit of using an unstructured database is that:

• Large volumes of structured, semi-structured, and unstructured data.
• Object-oriented programming that is easy to use and flexible.

The advantage of NoSQL database over SQL database is that it allows insertion of data without a predefined schema. Unlike SQL database it not need to know what we are storing in advance, specify its size etc.

B. Classification

We use algorithm called Naïve Bayes for classification which is a supervised learning method as well as a statistical method for classification. Naive Bayes classifier is a probabilistic classifier which when given an input gives a probability distribution of set of all classes rather than providing a single output. The algorithm classifies a news article into a crime type to which it fits the best. By the word classification what we get is "What is the probability that a crime document D belongs to a given class C?

The advantage of using Naive Bayes Classifier is that it is simple, and converges quicker than logistic regression. Compared to other algorithms like SVM (Support Vector Machine) which takes lot of memory the easiness for implementation and high performance makes it different from other algorithms. Also in case of SVM as size of training set increases the speed of execution decreases.

probability P(A) * P(B/D) * P(C/D) * P(E/D) where P(C/D)=O.

So the estimated probability result always gives zero which leads to uncertainty in results. To avoid this condition we add + I to the count of every zero value classes to achieve uniform distribution. Test results shows that Naive Bayes shows more than 90%accuracy since it categorise each words as tokens and removing frequent words like "the", "and", "of' etc which improves accuracy. A word is automatically terminated if it occurred fewer times or less than 3 times. Figure 2 shows a sample pseudo code of Naive Bayes algorithm.

We are also integrating the concept of Named Entity Recognition (NER) in the crime articles. NER also known as Entity Extraction finds and classifies elements in text into predefined categories such as the person names, organizations, locations, date, time etc. So by using this concept in crime article we can get more details related to crime like victim and
Offender names, location of crime, date, time

I) Input: NAVI MUMBAI: The bike borne chain snatchers targeted two women pedestrians in Sanpada and Panvel onMay 6, 2014, Tuesday and robbed their gold ornaments. While, 60-year-old woman's gold chain worth Rs 20,000 was snatched by the bike's pillion rider around 3.45 pm, while she was walking on the street near HDFC bank in sector-14, Sanpada, yet another woman from Khalapur was targeted bythe pillion rider while she was walking along the road near old Thane naka in Panvel.


```
{
"nerList" : [
{
“location”:”vashi”
},
{
“location”:”MUMBAI”
},
{
“location”:”Sanpada”
},
{
“location”:”Sanpada”
},
{
" location" : "Panvel"
},
{
"date": "May 6,2014"
}
{
"date": "Tuesday"
}
}
```


Figure 3: A sample output of NER

For example: Seema said she would come i.e. here "she" refers to person "Seema". Likewise we are extracting all referenced entities in a text. Below example shows the working of Conference concept.

2) Input: E.g.: A pillion bike rider snatched away a gold mangalsutra worth Rs 85,000 of a 60-year-old woman pedestrian in sector 19, Kharghar on Friday. The victim, Shakuntala Mande, was walking towards a vegetable outlet around 9.40am, when a bike came close to her and the pillion rider snatched her mangalsutra. A robbery case has been registered at Kharghar police station.

C. Pattern Identification

Third phase is the pattern identification phase where we have to identify trends and patterns in crime. Forfinding crime pattern that occurs frequently we are using Apriori algorithm. Apriori can be used to determine association rules which highlight general trends in thedatabase. The result of this phase is the crime pattern for a particular place. In relation with each location we take the attributes of that place like VIP presence, weather attributes, area sensitivity, notable event, presence of criminal groups etc. When gathering the general crime pattern for a place, when a new case arrives and if it follows the same crime pattern then we can say that the area has a chance for crime occurrence.

Figure 2: (TABLE1) DECISION TREE FOR DELHI

Area sensitivity	Notable event	VIP presence	Criminal group	crime
Yes	Yes	Yes	No	yes
Yes	Yes	No	Yes	no
No	No	No	Yes	no
Yes	No	No	No	no
Yes	Yes	Yes	yes	yes
No	Yes	No	No	no

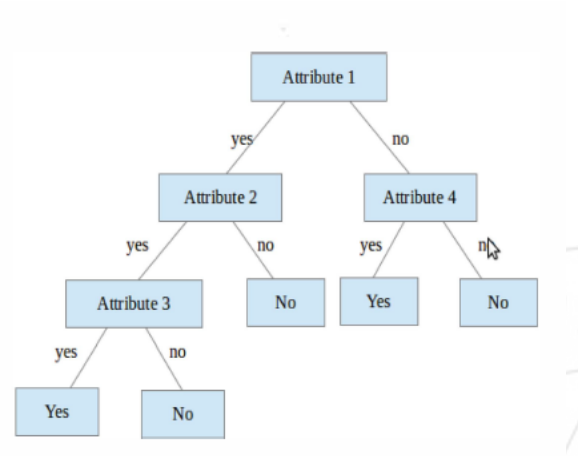

Figure 5: E.g. of a decision tree

The working of decision tree seems to be little confusing but it's really easy. We can select a variety of plant species. We classify them according to order, genus, species etc. Besides this, we have to classify them into a common category as shrubs and trees. If a new species is identified then we have to classify this into any of the two categories. The classification is based on its characteristics i.e. we have a set of questions to check whether it satisfies the conditions. When the above condition is satisfied then we check the next case and if the first condition itself is not satisfied then there is no need to check the rest. So the series of questions and their answers can be organized in the form of a decision tree. The tree has three types of nodes:

- A Root node, which has incoming edges and zero or more outgoing edges.
- Internal nodes, each of which has one incoming edge and two or more outgoing edges.
- Leaf node or end node, each of which has exactly one incoming edge and no outgoing edges.

This supervised machine learning technique builds a decision tree from a set of class labeled training samples and by using this tree, tests the new samples. It can be considered as a predictive model which uses a set of binary rules to calculate the class value. The tree determines:

- Which variable to split at a node.
- Decision to stop or split.
- Assign terminal nodes.

E. Visualization

The crime prone areas can be graphically represented using a heat map which indicates level of activity, usually darker colors to indicate low activity and brighter colors to indicate high activity. Below figure is an example of a heat map .Below Figure shows the regions that has high probability for crime occurrence. The advantages of using heat maps over other representational mechanisms are:

- Numeric and category based color images.
- Gradient color range.
- Analyze only the data we want.
- Out of range data is automatically discarded.

Due to this reason, by knowing about the probable regions we can prevent crimes by taking preventive mechanisms like night patrolling, fixing burglar alarms, fixing CCTV camera etc.

4. Future Works

A. Criminal Profiling

Besides this a new concept called criminal profiling which helps the crime investigators to record the characteristics of criminals. The advantage of doing criminal profiling is that:

- To provide crime investigators with a social and psychological assessment of the offender;
- To evaluate belongings found in the possession of the offender.

For the purpose of this we have to analyse the criminal backgrounds and criminal records for collecting the maximum criminal data. So the maximum details of each criminal are collected from criminal records. i.e. when crimes like burglary occurs in a certain place then from reports like FIR we get the offender details and their modus operandi(mode of operation).After getting these details we can know about the criminals with these behaviour. So sifting through each crime record after a particular crime occurrence is tedious task. So instead we can use some visualization mechanisms to represent the criminal details in a human understandable form. For representing criminal data we use a graph database called Ne04j. We can suggest a data model and query language that integrates an explicit modeling and querying of graphs smoothly into a standard database environment.

B. Snatching

We are concentrating more on crimes like snatching to get more details related to it like crime location, time, date, crime type(which type of snatching), victim and offender names etc.

Currently we are getting crime details like:

1) Name of person (victims, offenders)
2) Location
3) Organization
4) Type of crime (whether murder, robbery)
5) Subcategories of crime type (for snatching there are other categories like chain snatching, purse snatching etc)
6) Type of vehicle offender used.
7) Whether any weapons used.

8) Time of incident
9) Date
10) Incident summary
11) Criminal groups involved

5. Conclusion

In this paper we have tested the accuracy of classification and prediction based on different test sets. Classification is done based on the Bayes theorem which showed more than90% accuracy. Using this algorithm we trained numerous news articles and build a model. For testing we are inputting some test data into the model which shows better results. Our system takes factors/attributes of a place and Apriori algorithm gives the frequent patterns of that place. The pattern is used for building a model for decision tree. Corresponding to each place we build a model by training on these frequent patterns. Crime patterns cannot be static since patterns change over time. By training means we are teaching the system based on some particular inputs. So the system automatically learns the changing patterns in crime by examining the crime patterns. Also the crime factors changeover time. By sifting through the crime data we have to identify new factors that lead to crime.

References

[1] ieeexplore.ieee.org
[2] http://www.cs.sunysb.edul-cse634/lecture notes/07apriori.pdf
[3] www.google.com
[4] Wikipedia contributors.(l2 May 2014 at 19:05.), Series Finder.[online]