An Analysis on Fuzzy Algorithm with Association Rule Mining Clustering on Crime Pattern

Prof. J. Mohana Sundaram #1, Dr. T. Karthikeyan #2
#1 Assistant Professor, Department of Computer Science
#2 Associate Professor, PG Department of Research,
# 1, 2 PSG College of Arts and Science, Coimbatore, Tamil Nadu, India,

ABSTRACT
The aim of crime pattern discovery is to hinder the police to prevent criminal activities. Fuzzy set theory exhibits immense potential for effective solving of the uncertainty in the problem. Fuzzy is flexible which will randomly initialized parameters, this make them particularly appropriate for criminal activity forecasting. ARM algorithm is used to save time which makes execution progress faster and quickly. Cluster is grouping of a number of similar things into single. It is collection of data objects.

Key words: Association Rule Mining, Fuzzy ARM, Crime Pattern Association Rule Mining (CPARM),

Corresponding Author: Prof. J. Mohana Sundaram

INTRODUCTION
The identification and tactical analysis of crime patterns is a primary responsibility of crime analysts at police agencies around the world. Crimes are the social nuisance and cost our society dearly in several ways. Crime analysis involves exploiting data about crimes to enable law enforcement to better apprehend crime and criminals and prevent crimes. The coupling between data mining and crime pattern detection is that data mining as an advanced analytical program may assist decision making on detecting crime and has the potential to solve the contradiction between effect and efficiency of crime detection. So appropriate field need to chosen to perform crime analysis and as data mining refers to extracting or mining knowledge from large amounts of data, data mining is used here on high volume crime dataset and knowledge gained from data mining approaches is useful and support police forces. In this paper, we study the application of fuzzy association rule mining for community crime pattern discovery. Fuzzy ARM is used to provide hidden knowledge in the crime analysis, supported by experimental results on the open-source crime data set. In this paper k means clustering technique of data mining used to extract useful information from the high volume crime dataset and to interpret the data which assist police in identify and analyse crime patterns to reduce further occurrences of similar incidence and provide information to reduce the crime. Crime is a behaviour disorder that is an integrated result of social, economic and environmental factors. In section 2 we discussed about Data Mining used in Crime pattern, Section 3 explained Clusters of Crime, Section 4 extends Crime Pattern Association Rule Mining, Section 5 shows Experimental Result and finally Section 6 explained about Conclusion.
DATA MINING USED IN CRIME PATTERN
We will look at how to convert crime information into a data-mining problem, such that it can help the detectives in solving crimes faster. We have seen that in crime terminology a cluster is a group of crimes in a geographical region or a hot spot of crime. Whereas, in data mining terminology a cluster is group of similar data points, this data points is called as possible crime pattern. Thus appropriate clusters or a subset of the cluster will have a one-to-one correspondence to crime patterns. In our case some of these clusters will useful for identifying a crime spree committed by one or same group of suspects. Given this information, the next challenge is to find the variables providing the best clustering. The automated detection of crime patterns, allows the detectives to focus on crime sprees first and solving one of these crimes results in solving the whole “spree” or in some cases if the groups of incidents are suspected to be identified in the different crime incidents. For instance, one crime site reveals that suspect has black hair, the next incident/witness reveals that suspect is middle aged and third one reveals there is tattoo on left arm, all together it will give a much more complete picture than any one of those alone. Nowadays the crime information’s are updated manually through files like excel sheet etc. We choose to use clustering technique over any supervised technique such as classification, since crimes vary in nature widely and crime database often contains several unsolved crimes. Therefore, classification technique that will rely on the existing and known solved crimes, will not give good predictive quality for future crimes.

DATA EXTRACTION
In proposed system, we extract data from single source in which the reports from http://www.tnpolice.gov.in/crimeprofile.html/. These data sources are somewhat similar and have common structure. They are from html files to xml documents and at times just plain text files with a format. Convert all the extracted date into a single data structure. This is the first major task in any data mining application. To make success in the extracted data we should deal with huge amount of inconsistencies because many area in the extracted data does not have geo coded with its latitude and longitudes. In the extracted data some values may be missed, for that we need to estimate or ignore that value based on their presence of importance so it will not effect on the models. After all these data source are gathered into a single database, it must have consistent format across the database, then we can start to make digitalize our learning and data mining algorithm to get the corresponding crime pattern.

CLUSTERS OF CRIMES (TAMIL NADU)
We will look at some of our contributions to this area of study. We will show a simple clustering example here. Let us take an oversimplified case of crime record.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Murder</td>
<td>1644</td>
<td>1630</td>
<td>1521</td>
<td>1273</td>
<td>1365</td>
<td>7433</td>
</tr>
<tr>
<td>2</td>
<td>Murder for Gain</td>
<td>123</td>
<td>105</td>
<td>102</td>
<td>89</td>
<td>74</td>
<td>493</td>
</tr>
<tr>
<td>3</td>
<td>Robbery</td>
<td>1144</td>
<td>662</td>
<td>495</td>
<td>450</td>
<td>437</td>
<td>3188</td>
</tr>
<tr>
<td>4</td>
<td>Burglary</td>
<td>4221</td>
<td>3849</td>
<td>3717</td>
<td>3300</td>
<td>3738</td>
<td>18825</td>
</tr>
<tr>
<td>5</td>
<td>Theft</td>
<td>15712</td>
<td>15019</td>
<td>13217</td>
<td>13651</td>
<td>15851</td>
<td>73450</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>22844</td>
<td>21265</td>
<td>19052</td>
<td>18763</td>
<td>21465</td>
<td>103389</td>
</tr>
<tr>
<td>Year wise total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>103389</td>
</tr>
</tbody>
</table>

Table 1 Major crime trends in Tamil Nadu, India
A crime data analyst or detective will use a report based on this data sorted in different orders, usually the first sort will be on the most important characteristic based on the detective’s experience. We look at table 1 with a simple example of crime list. The type of crime is Murder, Murder of Gain, Robbery, Burglary, Theft and it will be the most important attribute. The aim here is that we can use data mining to detect much more complex patterns since in real life there are many attributes or factors for crime and often there is partial information available about the crime. In a general case it will not be easy for a computer data analyst or detective to identify these patterns by simple querying. Thus clustering technique using data mining comes in handy to deal with enormous amounts of data and dealing with noisy or missing data about the crime incidents.

**Fig 1**: Tamil Nadu Political Map.

**ABOUT DATA SET**

We collected data from single source such as gathered data from disparate sources such as http://www.tnpolice.gov.in/crimeprofile.html. The data set is an indication of crime which occurred over a year from different districts of the state. The crime data was organized in xml format. The xml files needed to be preprocessed to meet the mark up and encoding standards. A total of 103389 records of crime in the entire crime data set. The start date field was not present in all records. Group A of the crime data includes the following categories: murder, murder for gain, robbery, burglary, theft. Group A comprises about 95% of the entire dataset. The records span from a time frame of 2005 to 2009 and are distributed across 32 districts in one state. The distribution per year and its type is shown in fig2 and fig3 respectively.

**Year Wise Statistical Report**

![Year Wise Statistical Report](image)

**Fig 2**: Distribution of crime across year
Thus from the above two graphs we can see that, the crime records are distributed mainly in the period 2005-2009 around the state. Each crime record has a type and a time associated with it and also the area in which the crime happened. There were a total of 32 districts from which data is extracted. Then for each type of crime, basket transactions were created. For each window of size $W_k = 1$ week, all the crimes that occurred during that calendar week were recorded as single transactions. Since we aim to model only the relationship between different types of crime, we converted the continuous counts of each crime type to a binary attribute defining the presence or absence of a crime during that window. Fig. 1 gives the distribution of crime data year wise. We can clearly see that the years 2005-2009. Also, Fig 3 shows the distribution of the crime incidents across the type of. Thus we can see that certain categories like burglary and theft have insignificant numbers. This could be mostly due to missing data. Thus we ignored such insignificant types of crime and only concentrated on the major types such as robbery, burglary, murder etc., When looking at the distribution of crime by type, the crime record was largest for theft (71%) and second largest (18%) for burglary.

CRIME PATTERN ASSOCIATION RULE MINING (CPARM)

As mentioned before, each crime record in the data set has a crime type, a time of occurrence and location (district) of the respective crime. There were a total of 32 districts from which data is collected. For each county, timing window is used, which is of size $W_k = 1$ week. Also, the basket transactions were created using the data set from 2005-2009. For each window, all the crimes that occurred during a calendar week were recorded as a single transaction. Since we aimed to model only the relationship between different types of crime, we converted the discrete counts of each crime type to a binary attribute defining the presence or absence of a crime within that window. Crime types like burglary, theft and murder were always present for few districts. Hence, they were removed from the transaction for those districts in order to yield more specific rules. In the next run, those crime types were grouped as “theft" so that we could mine for interesting patterns that make intuitive sense. Pattern for Association Rule is given as $R_1$. Consider “$Cr$” as “Crime”. So $R_1$: {Cr1, Cr2, Cr3} and {Cr4}.

FUZZY ASSOCIATION RULE MINING

The above rule suggests that there is a relationship between $Cr1$, $Cr2$, $Cr3$ and $Cr4$. To be simpler, if $Cr1$, $Cr2$ and $Cr3$ occur, then there is a high chance that $Cr4$ occurs within the same week. The different measures for rules interestingness are support, confidence, coverage, strength, lift and leverage.

Support, $s(P, Q) = \frac{\sigma(P \cap Q)}{N}$
Confidence, \( c(P \cap Q) = \frac{\sigma(P \cap Q)}{\sigma(X)} \)
Coverage, \( cv(P \cap Q) = \frac{\sigma(P)}{N} \)
Strength, \( st(X \cap Q) = \frac{\sigma(Q)}{\sigma(P)} \)
Lift, \( l(X \cap Q) = \frac{N^*\sigma(P \cap Q)}{\sigma(P) \cdot \sigma(Q)} \)
Leverage, \( lv(X \cap Q) = N^*\sigma(P \cap Q) - \{\sigma(P) \cdot \sigma(Q)\} \)

We experimented with all the above mentioned measures and selected support, confidence, leverage and strength as final measures. The number of burglary and theft were very large compared to the other crime types. The association mining algorithms generated many rules that were redundant and hence pruned on the basis of confidence. For example if the algorithm generates rules
\[
\begin{align*}
R1: \{C_{r1}, C_{r2}, C_{r3}\} & \{C_{r4}\} \\
R2: \{C_{r2}, C_{r3}\} & \{C_{r4}\} \\
R3: \{C_{r1}\} & \{C_{r4}\}
\end{align*}
\]
All the rules having the same confidence implies that they give the same information and hence are redundant and should be pruned. The rules were pruned based on confidence, support, list and leverage. The interpretation of the generated association rule is as follows. The rules were generated using a window size of week and the order of occurrence of the individual crime types were not taken into account. Hence the association rules do not give a casualty equation but only gives a correlation between the two events. For example the rule \( P \cap Q \) implies that they appear together but does not imply that the crime sequence \( P \) leads to the crime \( Q \).

**EXPERIMENTAL RESULTS**

A set of fuzzy membership functions was defined for each of the 5 attributes and each attribute was fuzzified. The fuzzified data constituted the input to Fuzzy Apriori. Fuzzy Apriori was run with confidence 60% and with different supports depending on the data subset. All membership functions for attributes 1-36 were selected as antecedents, and the following membership functions for attributes 37-40 were selected as consequents: Murder, Murder for Gain, Robbery, Burglary, Theft.

As the number of communities within each region differs, minimum supports were calculated for each region based on the support of a rule occurring the same number of times within that region. This unifies support across regions for reliable comparison of rule measures, and facilitates discovery of consistent rules across Coimbatore City, Tamil Nadu, India.

**Support:**

\[
\begin{align*}
[Murder] \& [Murder for Gain] \rightarrow [Robbery] \\
Support = 7.14, Confidence = 0.534, Lift = 4.97
\end{align*}
\]

**Relative Support:**

\[
\begin{align*}
[Murder for Gain] \& [Robbery] \rightarrow [Burglary] \\
Support = 0.534, Confidence = 4.97, Lift = 18.35
\end{align*}
\]

**Confidence:**

\[
\begin{align*}
[Robbery] \& [Burglary] \rightarrow [Theft] \\
Support = 4.97, Confidence = 18.35, Lift = 68.31
\end{align*}
\]

**Lift:**

\[
\begin{align*}
[Burglary] \& [Theft] \rightarrow [Murder] \\
Support = 18.35, Confidence = 68.31, Lift = 7.14
\end{align*}
\]

These discovered rules represent patterns that are of interest to law enforcement officials.
CONCLUSION
The aim of the study is to Analysis on Fuzzy Algorithm with Association Rule Mining Clustering for Crime Pattern Discovery in this regard we provide the algorithm explanation with its technique, the dataset can be transferred in the form of graph using mat lab. From the experimental results fuzzy association rule mining presents the best accuracy specified by the graph. We have shown via exploratory comparisons in the term of Fuzzy ARM and ARM Clustering Algorithms due to the nature of this data set. Through the implementation of this graph method in matlab. It is used to demonstrate the important phase to enhance the mining quality. Fuzzy association rule mining has proven useful for this crime application, and has utility for other crime-related data sets Results presented in this paper suggest that further analysis is required to gain a closer understanding of crime at both the ARM and Fuzzy ARM. Crime patterns were discovered which are consistent across all regions, subsets of regions, and all states. The attributes of interest were computed to measure their occurrence per 100K population, so as to remove the element of community and state size during the rule generation process.

REFERENCES


