APPLICATION FOR ANALYSIS AND PREDICTION OF CRIME DATA USING DATA MINING

[1]ANISHA AGARWAL, [2]DHANASHREE CHOUGULE, [3]ARPITA AGRAWAL, [4]DIVYA CHIMOTE

[1,2,3,4]Sinhgad College of Engineering, Savitribai Phule Pune University
E-mail: [1]Agarwl.anisha93@gmail.com, [2]dhanuchougule11@gmail.com, [3]arpitaagrawal.pu@gmail.com,
[4]divvyachimote123@gmail.com

Abstract— Today, time is a concerning factor for sentencing criminals. Many a time a criminal released on bail may yet be a potential threat to the society, even after they have served their sentence. This threat can be reduced if a prediction analysis is done on the concerned person to determine if he is about to do the crime or not. This aspect can be beneficial both for law enforcement and the safety of our country. Data mining is an approach that can handle large voluminous datasets and can be used to predict desired patterns. Our sole users will be the police officers who from time to time shall be able to predict the possibility of the crime a criminal is probable to commence in the nearest future as well as which particular crime he will be committing. In this paper, we look at the use of frequent pattern mining with association rule mining to analyze the various crimes done by a criminal and predict the chance of each crime that can again be performed by that criminal. This analysis may help the law enforcement of the country to take a more accurate decision or may help in safeguarding an area if a criminal released on bail is very much likely to perform crime. We will concentrate on Apriori algorithm with association rule mining technique to achieve the result.

Keywords— Apriori algorithm, association rule mining, crime analysis, prediction.

I. INTRODUCTION

Technology has improved the world in a long run where it revolves around a large, voluminous data. If this data is applied to use in a right direction, then we can eliminate the existing flaws. Data mining helps in processing of large amounts of data and discovering hidden information. Our System is supposed to process huge amount of criminal records so that prediction can be made as per the past behavior of the criminal individuals. There is no need to add the fact that as the records in the system increases, the accuracy of the system also increases. The present existing system in India has no automated technology that will ease the efforts of the police department in looking over the files of the summoned cases. Present day, the information is searched manually over bulk of hard copies. This is a problem to the existing system, for which no solution has been introduced yet. Moreover, there is no such system that will enable to look over the potential criminals who may commit a particular crime again in future so that their sentence of punishment can be reviewed over. We are presenting a system that will overcome all the above drawbacks in best possible way. Here our paper focuses on the prediction technique. The

system will be predicting the crime that an individual criminal is likely to perform again in the future using Apriori algorithm with association rule technique. This will immensely help in seeking justice for the people who commit the crime accidently. Moreover, it will help in securing the area against the potential criminals who are predicted to be a threat by the system. In implementing the project, there is no 100% accuracy of the system. The same will depend upon the threshold values we as developer will provide for association rule technique in machine learning algorithms. We aim to design an application that will be accessed and available to the authorized users anytime and anywhere, along with the main functionality of prediction of the further crimes by individual criminal. Our goal is to design an effective system that will give accuracy of at most 80%.

II. EXISTING VS PROPOSED SYSTEM

Crimes in India have been in a predominant increasing rate. Finding out the crime patterns that a particular criminal is likely to perform will speed up the process of law system and reduce the crimes in society. There are following challenges in our proposed project:

- Ever Increasing size of crime information that has to be stored and analyzed.
- Problem of identifying techniques to analyze this data
- Methods and structures to be used for storing crime data.
- The inconsistent and incomplete data complicates analysis.

2.1 Existing System

- There is no existing system for this project. For the accessing of any information, one has to dig up piles of data manually.
- More time is needed in searching the required information.
- Bulks of hard copies are required to be referred.
- Access is not available anywhere.
- The increasing crime in our country is a big issue for us. To solve this issue, various

systems are developed but they cannot find the area where the crime will happen.

- But it is not possible to keep all the records in memory or in file because of large data.
- The large amount data maintenance is the big problem; searching and predicting a particular data manually is not possible from the files.

2.2 Proposed System

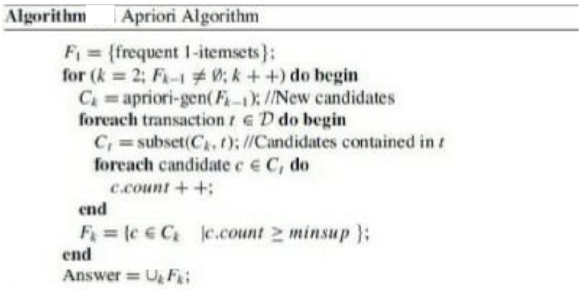
- In the proposed system, we are introducing the application which will predict the crime that criminals can do in the future.
- This prediction is based on attributes like criminal record, education, occupation, friend circle, family background and other various factors.

We will store all the previous record of criminal and by mining his previous record we can calculate the possibility and prediction of the crime he is likely to perform.

III. APRIORI ALGORITHM

Apriori algorithm is an approach for frequent itemset generation. Generally 2k-1 itemsets, excluding null set are generated for k items in dataset. Value of k can be very large in practical applications which will generate exponentially large itemsets. For finding frequent itemsets we need to count the support for candidate itemset. Apriori algorithm provides an effective way to reduce the number of candidate itemsets. Apriori principle states that “If an itemset is frequent, then all of its subsets must also be frequent”.

For example, if itemset {1,3,4} is a frequent, then all of its subsets that is {1},{3},{4},{1,3},{1,4},{3,4} will also be frequent. If an itemset is infrequent then none of its supersets is frequent.

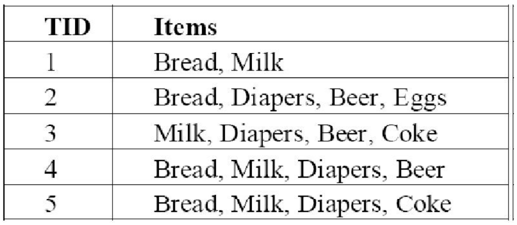


```
Algorithm    Apriori Algorithm
$F_1$ = {frequent 1-itemsets};
for ($k = 2$; $F_{k-1} \neq \emptyset$; $k++$) do begin
    $C_k$ = apriori-gen($F_{k-1}$); //New candidates
    foreach transaction $t \in \mathcal{D}$ do begin
        $C_t$ = subset($C_k$, $t$); //Candidates contained in $t$
        foreach candidate $c \in C_t$ do
            $c.count++$;
    end
    $F_k = \{c \in C_k \mid c.count \geq minsup\}$;
end
Answer = $\cup_k F_k$;
```

Fig 1. Pseudocode for frequent itemset generation of Apriori algorithm

Example:

Table 1. Transactions



TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

We assume that the support threshold is 60%, which is equivalent to a minimum support count equal to 3.

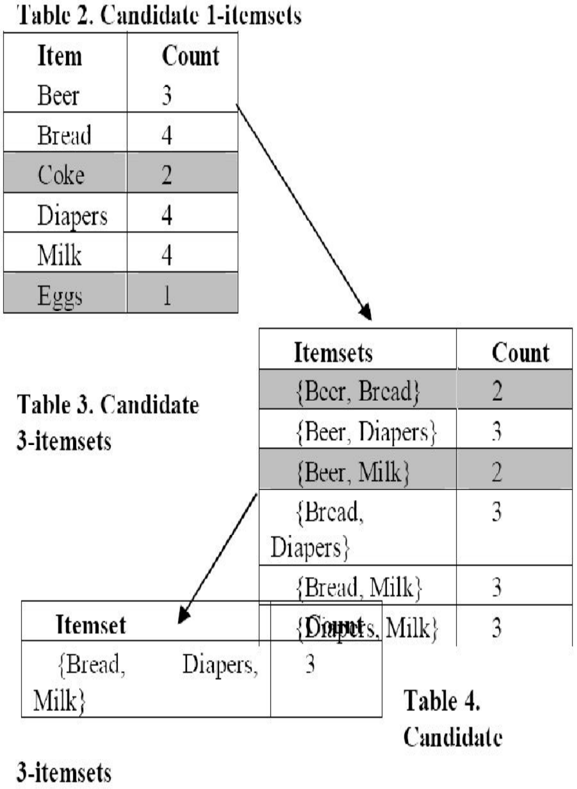


Table 2. Candidate 1-itemsets

Item	Count
Beer	3
Bread	4
Coke	2
Diapers	4
Milk	4
Eggs	1

Table 3. Candidate 3-itemsets

Itemsets	Count
{Beer, Bread}	2
{Beer, Diapers}	3
{Beer, Milk}	2
{Bread, Diapers}	3
{Bread, Milk}	3
{Diapers, Milk}	3

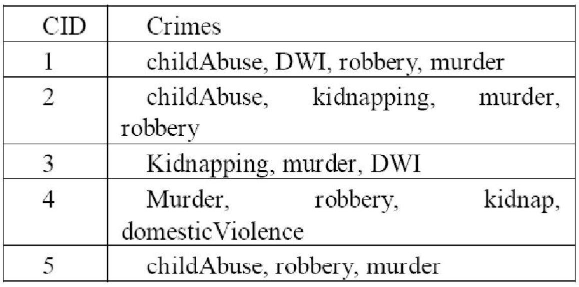
Itemset	Count
{Bread, Diapers, Milk}	3

Table 4. Candidate 3-itemsets

Performance of apriori algorithm is better than Brute force strategy. Brute force produces candidates as
${}^6C_1 + {}^6C_2 + {}^6C_3 = 6 + 15 + 20 = 41$
On the other hand ,Apriori produces
${}^6C_1 + {}^4C_2 + 1 = 6 + 6 + 1 = 13$

IV. IMPLEMENTATION OF APRIORI ALGORITHM

Table 5.Crime Transactions



CID	Crimes
1	childAbuse, DWI, robbery, murder
2	childAbuse, kidnapping, murder, robbery
3	Kidnapping, murder, DWI
4	Murder, robbery, kidnap, domesticViolence
5	childAbuse, robbery, murder

We assume that the support threshold is 60%, which is equivalent to a minimum support count equal to 3.
In tables 6, 7, and 8, the number of candidate itemsets generated is as follows:
${}^6C_1 + {}^4C_2 + 1 = 6 + 6 + 1 = 13$
Frequent itemsets among them are: ({childAbuse}, {robbery}, {murder}, {kidnapping}, {childAbuse, robbery}, {childAbuse, murder}, {robbery, murder}, {robbery, kidnapping}, {childAbuse, robbery, murder})

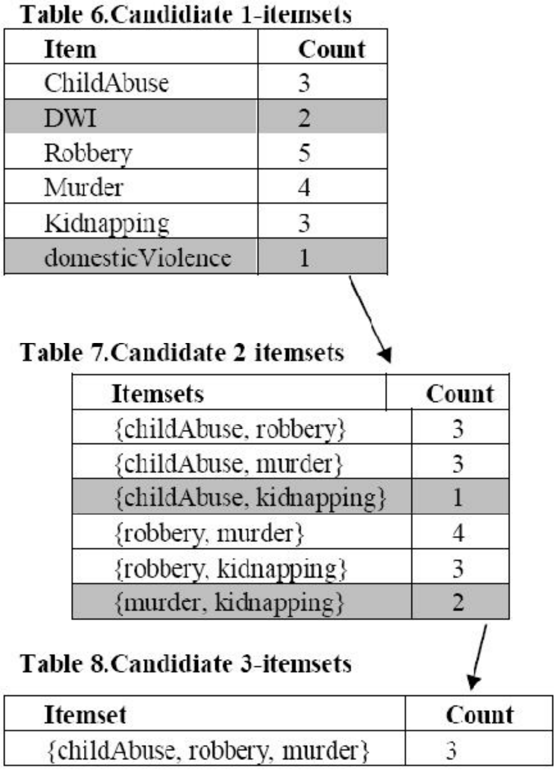


Table 6.Candidiate 1-itemsets

Item	Count
ChildAbuse	3
DWI	2
Robbery	5
Murder	4
Kidnapping	3
domesticViolence	1

Table 7.Candidate 2 itemsets

Itemsets	Count
{childAbuse, robbery}	3
{childAbuse, murder}	3
{childAbuse, kidnapping}	1
{robbery, murder}	4
{robbery, kidnapping}	3
{murder, kidnapping}	2

Table 8.Candidiate 3-itemsets

Itemset	Count
{childAbuse, robbery, murder}	3

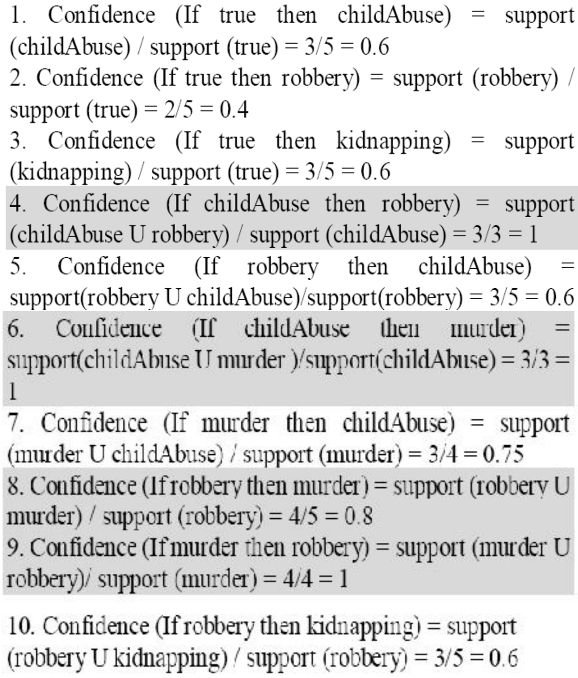
V. RULES GENERATION

Rules generation can be done by association rule mining with the help of support and confidence.

If there is an expression in the form of X → Y, where X and Y are disjoint datasets, then Support determines how often a rule is applicable to a given dataset and Confidence determines how frequently items in Y appear in transactions that contain X.

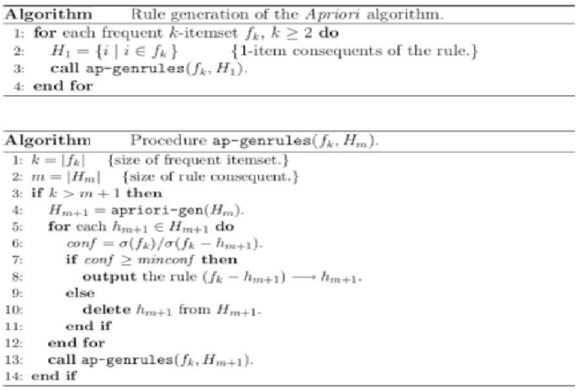
Support, S(X → Y) = support (X U Y) / N

Confidence, C(X → Y) = support (X U Y) / support (X)



1. Confidence (If true then childAbuse) = support (childAbuse) / support (true) = 3/5 = 0.6
2. Confidence (If true then robbery) = support (robbery) / support (true) = 2/5 = 0.4
3. Confidence (If true then kidnapping) = support (kidnapping) / support (true) = 3/5 = 0.6
4. Confidence (If childAbuse then robbery) = support (childAbuse U robbery) / support (childAbuse) = 3/3 = 1
5. Confidence (If robbery then childAbuse) = support(robbery U childAbuse)/support(robbery) = 3/5 = 0.6
6. Confidence (If childAbuse then murder) = support(childAbuse U murder)/support(childAbuse) = 3/3 = 1
7. Confidence (If murder then childAbuse) = support (murder U childAbuse) / support (murder) = 3/4 = 0.75
8. Confidence (If robbery then murder) = support (robbery U murder) / support (robbery) = 4/5 = 0.8
9. Confidence (If murder then robbery) = support (murder U robbery)/ support (murder) = 4/4 = 1

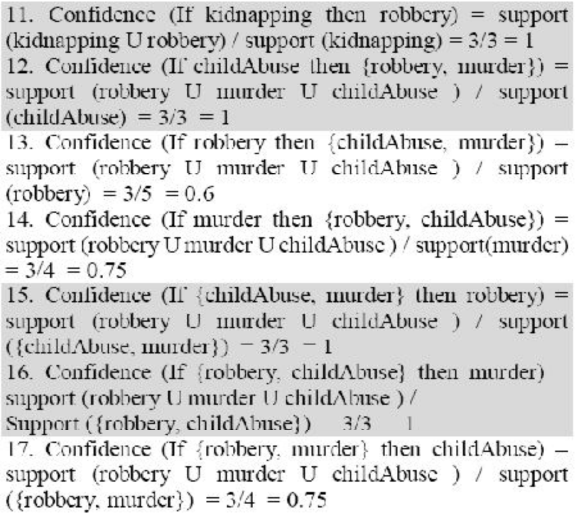
10. Confidence (If robbery then kidnapping) = support (robbery U kidnapping) / support (robbery) = 3/5 = 0.6



```
Algorithm    Rule generation of the Apriori algorithm.
for each frequent $k$-itemset $f_k$, $k \geq 2$ do
    $H_1 = \{i \mid i \in f_k\}$        {1-item consequents of the rule.}
    call ap-genrules($f_k, H_1$).
end for
```

```
Algorithm    Procedure ap-genrules($f_k, H_m$).
$k = |f_k|$    {size of frequent itemset.}
$m = |H_m|$    {size of rule consequent.}
if $k > m + 1$ then
    $H_{m+1}$ = apriori-gen($H_m$).
    for each $h_{m+1} \in H_{m+1}$ do
        $conf = \sigma(f_k)/\sigma(f_k - h_{m+1})$.
        if $conf \geq minconf$ then
            output the rule $(f_k - h_{m+1}) \longrightarrow h_{m+1}$.
        else
            delete $h_{m+1}$ from $H_{m+1}$.
        end if
    end for
    call ap-genrules($f_k, H_{m+1}$).
end if
```

Figure 2.Rules Generation for Apriori



11. Confidence (If kidnapping then robbery) = support (kidnapping U robbery) / support (kidnapping) = 3/3 = 1
12. Confidence (If childAbuse then {robbery, murder}) = support (robbery U murder U childAbuse) / support (childAbuse) = 3/3 = 1
13. Confidence (If robbery then {childAbuse, murder}) = support (robbery U murder U childAbuse) / support (robbery) = 3/5 = 0.6
14. Confidence (If murder then {robbery, childAbuse}) = support (robbery U murder U childAbuse) / support(murder) = 3/4 = 0.75
15. Confidence (If {childAbuse, murder} then robbery) = support (robbery U murder U childAbuse) / support ({childAbuse, murder}) = 3/3 = 1
16. Confidence (If {robbery, childAbuse} then murder) support (robbery U murder U childAbuse) / Support ({robbery, childAbuse}) 3/3 1
17. Confidence (If {robbery, murder} then childAbuse) = support (robbery U murder U childAbuse) / support ({robbery, murder}) = 3/4 = 0.75

Assuming Confidence threshold C_0 to be 80%, following 8 association rules is found:

1. If ChildAbuse then robbery
2. If ChildAbuse then murder
3. If robbery then murder
4. If murder then robbery
5. If kidnapping then robbery
6. If childAbuse then {robbery, murder} – 1 & 2 combined generates this rule
7. If {childAbuse, murder} then robbery
8. If {robbery, childAbuse} then murder

Table 9.Association Rules Generated

Association Rule	Confidence
If ChildAbuse then robbery	100%
If ChildAbuse then murder	100%
If robbery then murder	80%
If murder then robbery	100%
If kidnapping then robbery	100%
If childAbuse then {robbery, murder}	100%
If {childAbuse, murder} then robbery	100%
If {robbery, childAbuse} then murder	100%

So, let’s suppose if a query comes for criminal with id 3, our system will generate the output that the criminal may commit robbery in the near future.

VI. SOME OTHER AREAS

1. MARKETING RESEARCH Association rules can be a very powerful tool in the field of marketing. The purchase pattern of the customers can be used to generate the frequent itemsets by using support and then generating the rules based on these itemsets using confidence. This can help in the early diagnosis of marketing strategy by arranging the likely purchased products together say keeping bread and butter together in shopping mall.
2. MEDICAL FIELD In areas of medical, this analysis algorithm can be used to predict the disease of a concern
patient by evaluating the symptoms attributes of the patient. This consecutively may help in knowing about newly introduced symptoms and the disease that are prone to be faced by people. With the help of symptoms, we can generate interesting rules which can help in early diagnosis of diseases.
3. EDUCATIONAL INSTITUTES Apriori algorithms can be applied on educational data to understand the knowledge and performance of the students. We can apply these rules on student logs to generate interesting rules. These rules can help in improving the performance of the students and improve the quality of curriculum. The marks obtained by the students can help in knowing their interest in a particular course by using support and confidence. This also helps the faculty to understand the knowledge, interest and performance of the students.
4. BANKING SECTOR the data mining techniques and its algorithms may help in the prediction of credit card fraud, risk management and marketing. Detection and prevention of fraud is very difficult. Thus, the statistical data of customers in bank will help in detecting any new suspicious activity by comparing them with the usual behavior patterns of customers.
5. ENVIRONMENTAL HAZARDS these techniques can also be used in dealing with the environmental hazards like tsunami, earthquakes, flood in advance. Early prediction of disasters will help in saving lives and property. The rules can be generated based on the climatic conditions, terrain textures and other relevant data. Present conditions will be compared with the rules and hence, the mishap can be avoided by rehabilitating the people to other place and warning them as early as possible.

CONCLUSION AND FUTURE SCOPE

Along with the present scope of our project, which is prediction of the crime an individual criminal is likely to commit, we can also predict the estimated time for the crime to take place as a future scope. Along with this, one can try to predict the location of the crime. We will test the accuracy of frequent-itemsets and prediction based on different test sets. So the system will automatically learn the changing patterns in crime by examining the crime patterns. Also the crime factors change over time. By shifting through the crime data we have to identify new factors that lead to crime. Since we are considering only some limited factors full accuracy cannot be achieved. For getting better results in prediction we have to find more crime attributes. Our software predicts the crime an individual criminal is likely to perform.We will use Apriori Algorithm with association rule mining for this purpose. This will determine the next crime a criminal is about to commit.

REFERENCES

[1] "Crime Analysis using K-Means Clustering" JyotiAgarwalMtech CSE Amity University,NoidaRenukaNagpal Assistant Professor Amity University ,Noida RajniSehgal Assistant Professor Amity University,Noida-2013.

[2] "An Enhanced Algorithm to Predict a Future Crime using Data Mining" Malathi. A Assistant Professor Post Graduate and Research department of Computer Science, Government Arts College, Coimbatore, India. Dr. S. SanthoshBaboo Reader, Post Graduate and Research Department of Computer Science, D.G Vaishnav College Chennai, India-2011.

[3] "Association Rule Mining and Frequent Pattern Mining Applications On Crime Pattern Mining: A Comprehensive Survey" V.R. Sadasivam, Dr K Duraisamy, R Mani Bharathi, K.S. Rangasamy College of Technology, Tiruchengode, India.

[4] "Execution of Apriori Algorithm of Data Mining directed towards tumultuous crimes concerning Women", DivyaBansal, LekhaBhambhu, J.C.D. College of Engineering and Technology, G.J.U. University of science and technology, India

[5] "Datamining Techniques to Analyze and Predict Crimes" S.Yamuna, N.SudhaBhuvaneswari 1 M.Phil (CS) Research Scholar School of IT Science, Dr.G.R.Damodaran College of science Coimbatore 2 Associate Professor MCA, Mphil (cs), (PhD) School of IT Science Dr.G.R.Damodaran College of science Coimbatore-2012.

[6] "A Survey on Crime Data Analysis of Data Mining Using Clustering Techniques" Shiju Sathyadevan, Devan M.S Amrita Center for Cyber Security Amrita Vishwa Vidyapeetham, Amritapuri, Kerala, India Surya Gangadharan. Amrita Vishwa Vidyapeetham Amritapuri, Kerala, India-2014.

★★★