ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Crime Investigation using Data Mining

S.R.Deshmukh[1], Arun S. Dalvi[2], Tushar J .Bhalerao [3],

Ajinkya A. Dahale [4], Rahul S. Bharati [5],Chaitali R. Kadam[6]

Professor, Department of Computer Engg, College of Engineering, Kopargaon, India [1]

Student of BE Computer Engg, College of Engineering, Kopargaon, India [2,3,4,5,6]

Abstract: Crime rate is increasing very fast in India because of increase in poverty and unemployment. With the existing crime investigation techniques, officers have to spend a lot of time as well as man power to identify suspects and criminals. However crime investigation process needs to be faster and efficient. As large amount of information is collected during crime investigation, data mining is an approach which can be useful in this perspective. Data mining is a process that extracts useful information from large amount of crime data so that possible suspects of the crime can be identified efficiently. Numbers of data mining techniques are available. Use of particular data mining technique has greater influence on the results obtained. So the performance of three data mining techniques { J48, Nave Bayes and JRip will be compared against sample crime and criminal database and best performing algorithm will be used against sample crime and criminal database to identify possible suspects of the crime. Data mining is a process of extracting knowledge from huge amount of data stored in databases, data warehouses and data repositories. Clustering is the process of combining data objects into groups. The data objects within the group are very similar and very dissimilar as well when compared to objects of other groups**.**

Keywords: Criminal database, Crime investigation, Data Mining, J48, JRip, and Naïve Bayes.

I. INTRODUCTION

In India, police department is the largest unit for preventing crimes, maintaining law orders, rules and peace throughout the country. However problem with the Indian police is that they are still using the traditional manual process such as First Information Report (FIR) to keep and analyse the records crime and criminals. On the other hand criminals use more sophisticated technologies to commit the crime. In 2011, 34305 cases of murder, 31385 cases of half murder, 24206 cases of rape, 44664 cases of kidnapping and around 600000 cases of robbery, theft and dacoit were recorded. Total numbers of crime cases recorded were 325575. This is only the statistics of recorded crimes. Crime rate is increasing very fast in India and we are becoming unsafe. If we perform necessary calculations on the above statistics, police department has to handle near about 6400 cases per day. In order to prevent crimes police officers have to identify evidences for those cases.

II. LITERATURE SURVEY

Crime is basically “unpredictable" event. It is not constrained by space and time. It entirely depends on human behaviour. There can be huge rang of crime activities, for example, from illegal driving to terrorism attacks. Various activities performed by criminal generate large amount of information and again this information can be present in variety of formats. Because of this analysis of crime data becomes very difficult. Data mining is a useful process for extracting important information from large amount of data. In modern era criminals use more advance technologies to commit the crime, on the other hand there is inadequate use of technology in crime prevention and criminal identification. Since large data and more complex queries need to be processed, a more powerful system is required for the analysis of crime data. ReCAP, COPLINK, ICIS, Crime Criminal Information System (CCIS), Crime and Criminal Tracking Network System (CCTNS) and a lot of such systems have been developed and are in use for making the crime investigation process easier. These systems have used different data mining techniques for the analysis of crime data.

III. DETAILS OF DISSERTATION WORK

A. Comparison with Existing System

Multi agent based approach has been used for classification of crime data and identifying possible suspects. Naïve Bayes classifier is trained with only three crime categories-robbery, burglary and theft, on the other hand we have covered many other crime categories and trained classifiers accordingly. Multi agent approach also requires training of eleven different agents such as crime scene agent, place agent, resource agent etc. This is time consuming. This system neither ensures classification accuracy nor does it comment on performance of naïve Bayes algorithm. To overcome this issue we have been using two more data mining techniques J48 and JRip. Performance of these three algorithms will be compared using simulation tools such as WEKA and best performing algorithm will be used for generating classification rules.

B. Proposed Methodology

Analysis of large amount of crime and criminal information has been a challenging field for researchers. Along with the analysis of crime and criminal data, a criminal identification system is proposed that will identify the suspect of the crime based on the various attributes and theevidences found atthe crime scene. The performance of J48, Naïve Bayes and JRip data algorithms

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

will be compared in reference with various parameters such as accuracy of classification, correctly and incorrectly classified instances, precision, recall, true positive rate, false positive rate etc. The main reason behind this performance comparison is to get most accurate classification.

C. Selection of data Mining Algorithm

Many data mining algorithms are available. Selection of suitable data mining algorithms is a big challenge. Following criteria are considered for selecting data mining algorithms

- Simplicity in understanding and interpretation;
- Intelligence in identifying useful information;
- Addition of possible scenarios with an ease;
- High information gain;
- Simulation results on other datasets;

J48, Naive Bayes and JRip strongly fitted for above mentioned criteria.

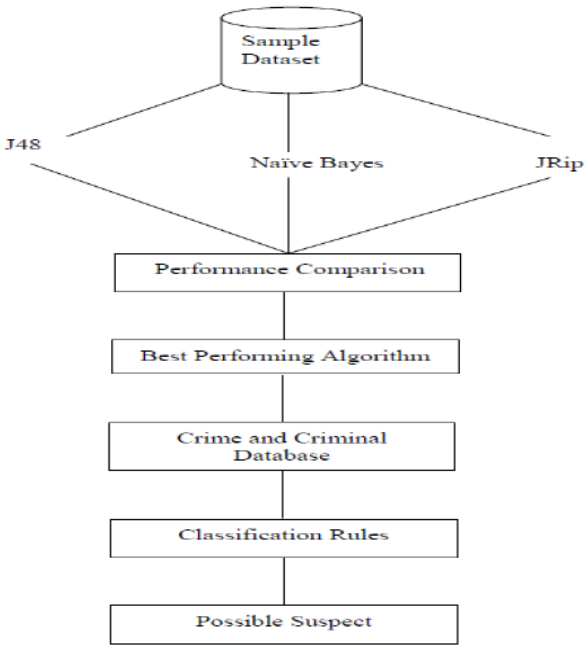

Fig. 1 Process Block Diagram

IV. ALGORITHMS

A. J48

1. A flow-chart-like tree structure
Internal node denotes a test on an attribute
Branch represents an outcome of the test
Leaf nodes represent class labels or class distribution
2. Decision tree generation consists of two phases Tree construction
At start, all the training examples are at the root
Partition examples recursively based on selected attributes
Tree pruning
Identify and remove branches that reflect noise or outliers
3. Use of decision tree: Classifying an unknown sample
Test the attribute values of the sample against the decision tree.

Output: A Decision Tree for "buys computer"

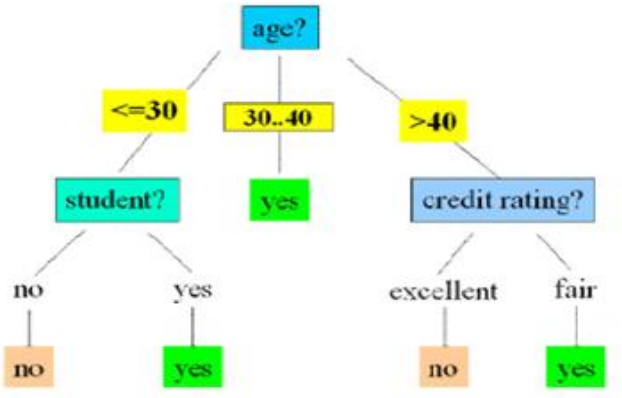

Fig.2Decision Tree

Algorithm for Decision Tree Induction

1. Basic algorithm (a greedy algorithm)
Tree is constructed in a top-down recursive divide-and conquer manner. At start, all the training examples are at the root Attributes are categorical (if continuous-valued, they are discredited in advance)
Examples are partitioned recursively based on selected1 attributes Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)
2. Conditions for stopping partitioning
All samples for a given node belong to the same class. There are no remaining attributes for further partitioning majority voting is employed for classifying the leaf
There are no samples left.

Extracting Classification Rules from Trees

1. Represent the knowledge in the form of IF-THEN rules.
2. One rule is created for each path from the root to a leaf.
3. Each attribute-value pair along a path forms conjunction.
4. The leaf node holds the class prediction
5. Rules are easier for humans to understand.

Example

IF age = "<=30" AND student = "no" THEN buys computer = "no"
IF age = "<=30" AND student = "yes" THEN buys compute = "yes"
IF age = "31…40" THEN buys computer ="yes"
IF age = ">40" AND credit rating = "excellent"THEN buys computer = "yes"
IF age = ">40" AND credit rating = "fair" THEN buy computer = "no"

B. JRip

JRip (RIPPER) is also a decision tree algorithm. This classifier is basically a propositional rule learner. It is also called as Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It is based on forming association rules with reduced error pruning (REP). With this classifier training data is split into a growing set and a pruning set. First, an initial rule set is formed using some heuristic method. This initial rule set can be overlarge. This overlarge rule set can be simplified using pruning

DOI 10.17148/IJARCCE.2015.4306

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

operators. Typical pruning operators would be to delete any single condition or any single rule. At each stage of simplification, the pruning operator is selected so that there is greatest reduction of error on the pruning set. Simplification ends when applying any pruning operator would increase error on the pruning set. This algorithm works in two stages.

Stage 1: Building Stage

Initially an empty rule set is formed.

RS= {}

In building stage again there are two phases.

1.1 Growing phase

In growing phase, one or more conditions can be added each time to generate 100% accurate rule. The procedure tries every possible value of each attribute and selects the condition with highest information gain.

1.2 Pruning phase

Pruning phase is used for simplification. In this phase one or more condition is deleted to simplify the rule set.

Stage 2: Optimization stage

Optimizing the rule set means to refine the rule set so that it operates as smoothly and efficiently as possible. Optimization stage is important because rule set generated can be much larger and we may not require all the rules or the rules generated may not be adequate. Optimization can also be done using growing and pruning phase.

C. Naïve Bayes

Naïve Bayes is a statistical classifier. It is used to predict class membership probabilities. Here a membership probability means the probability that a given tuple belongs to a particular class. Naïve Bayes classification is based on independence assumption. This classifier assumes that the presence or absence of particular feature is independent of presence or absence of any other future. For example a person can be considered as a suspect if he is 6 feet tall, has round face and white hair. Naïve Bayes classifier works on each feature independently and concludes that probably it is a suspect. Classification is done using parameter estimation such as calculating mean, variance. However parameter estimation can also be done using method of maximum likelihood. Naïve Bayes classification is based on Bayes‟ theorem with naïve (strong) class conditional independence. Class conditional independence means the effect of an attribute value on a given class is independent of the values of other attributes. The equation for Bayes‟ theorem is given as follows:

$$P(A|B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Where,

- $P(A)$ is the prior probability of A. It is "prior" in the sense that it does not take into account any information about B.
- $P(A|B)$ is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.
- $P(B|A)$ is the conditional probability of B given A. It is also called the likelihood.
- $P(B)$ is constant.

V. MATHEMATICAL MODEL

1. Input Set

I= {I1, I2, I3, I4}

Where,

I1=Admin Details.
I2=Algorithm Selection.
I3=Crime Details.
I4= Feedback.

2. Process Set

P= {P1, P2, P3, P4, P5, P6, P7}

Where,

P1=Registration.
P2=Log In.
P3=Select Algorithm.
P4=Give Criminal/Crime Details.
P5=Process Input.
P6=Display Result.
P7=Feedback.

3. Output Set

O= {O1, O2}

Where,

O1=Details of Criminal Information.
O2=Receive Feedback.

VI. CONCLUSION

Traditional crime investigation processes require a lot of skilled man power and paperwork. There is lack in use of technology for sensitive domain like crime investigation. So crime investigation has become a time consuming process. Data mining is the process of extracting useful information or knowledge from large data sources. Large amount of information is collected during crime investigation process and only useful information is required for analysis. So data mining can be used for this purpose. Selection of particular data mining technique has greater influence on the results obtained. This is main reason behind the performance comparison and selection of best performing data mining algorithm.

.

ACKNOWLEDGMENT

It gives me a great pleasure to present the paper on “Crime Investigation Using Data Mining”. I would like to express my gratitude to Prof. D. B. Kshirsagar, Head of Computer Engineering Department and Prof. S.R.Deshmukh sir and Prof. Saiprasadsir for their kind support and valuable suggestions.

REFERENCES

[1] Malathi A., Dr. S.Santosh Baboo, "Algorithmic Crime Prediction Model Based on the Analysis of Crime Cluster;" GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY, Vol. 11, Issue 11 Version 1.0 July 2011.

[2] Nitin Sakhare, PG Student, Sinhgad College of Engineering, Pune 41, Swati Joshi, Associate Professor, Sinhgad College of Engineering, Pune 41," Information Using Data Mining;", Third Post Graduate Symposium on cPGCON2014 Organized by department of Computer Engineering, MCERC Nasik.

[3] Anshul Goyal and Rajni Mehta"Performance Comparison of Naive Bayes and J48 Classi_cation Algorithms;" International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11 (2012).

[4] Revatthy Krishnamurthy, J.Satheesh Kumar, "Survey of Data ining Techniques on Crime Data Analysis;" International Journal of Data Mining Techniques and Applications; Vol.1, Issue 2, December 2012.