Classification of Criminal data using J48 Algorithm

Article · August 2014

CITATIONS
12

READS
1,851

2 authors:

Nitin Sakhare
Vishwakarma Institute Of Information Technology, Pune
16 PUBLICATIONS **70** CITATIONS
SEE PROFILE

Swati Joshi
Sinhgad Technical Education Society
6 PUBLICATIONS **23** CITATIONS
SEE PROFILE

IIJDWM

Journal homepage: www.ifrsa.org

Classification of Criminal Data Using J48-Decision Tree Algorithm

Nitin Nandkumar Sakhare
PG Student, Sinhgad College of Engineering, Pune
Swati Atul Joshi
Associate Professor, Sinhgad College of Engineering, Pune

ABSTRACT

In recent years, amount of crime has become a serious problem in most of the countries. Now a day's criminals have maximum use of modern technologies to commit the crime and investigators on the other hand are still using traditional manual processes, for example First Information Report (FIR) to analyze the suspect of the crime. Crime investigation becomes time consuming and tedious task with existing investigation techniques. Large amount of information is collected during investigation process. Extracting only useful information from it is a challenging task if done manually. Crime investigation process needs to be faster and efficient. Crime is an interesting domain where data mining plays an important role in terms of prediction and analysis. This paper presents a detailed study on a J48 - decision tree algorithm for predicting a crime suspect and overall analysis of crime data. This study also helps investigators for better prediction and classification of crime and criminal data.

Keywords- Crime, Data Mining, Decision Tree, Classification, Suspect

1. MOTIVATION

Security is one of the major concerns for any country across the world. The need for strengthening of national security is seriously considered after the terrorist attack on 11th September 2001 [2]. As most of the security agencies are working on preventing future crimes, the challenge of analyzing a large amount of this crime data has become a major problem for them. Data mining is a powerful technique that helps investigators who generally lack extensive training as data analysts to explore large crime databases quickly and efficiently. Classification plays an important role in depicting the crime patterns and analyzing these crime patterns in less amount of time.

2. LITERATURE REVIEW

Table 1 indicates the range of possible criminal activities and yet includes many more. Criminal Identification is the ultimate aim of analyzing crime and criminal information. In this, identification of person who is most likely to commit the crime could be done. An analysis of crime and criminal information includes monitoring crime pattern, evidences found at the crime scene, particular behavior of a person committing a crime and other features. All this analysis helps crime investigators in finding possible suspects of crime accurately with much less investigation time. Various systems that have been developed for to fulfill above requirements are discussed below:

Table 1: Crime types and law enforcements [2]

Crime Type	Law enforcement
Traffic Violations	Speedy driving, causing property damage or personal injury in collision, driving under the influence of drug and alcohol, hit and run
Theft	Robbery, burglary
Fraud	Misappropriation of assets, corruption, Money laundering, insurance fraud,
Sex crime	Sexual abuse, rape, sexual assault, child pornography, prostitution
Drug offenses	Possessing, distribution and selling illegal drugs
Violent crimes	Murder, half murder, armed robbery, forced rape

Regional Crime Analysis Program (ReCAP) is the result of requirement of database management system and a geographic information system by USA law enforcement to support crime analysis. In ReCAP data mining plays an important role and it supports a geographic information system. The combination of GIS and DBMS allows analyst to concentrate on important data instead of redundant one. Results from geo-graphic data mining can be applied immediately in crime analysis process to identify when and where the crime occurs. With the ReCAP system knowledge from data analysis can be presented in the form of geo-graphic map [4].

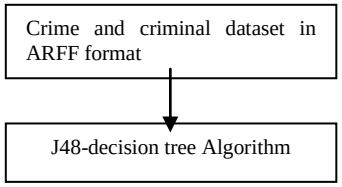
COPLINK is another crime investigation project which is developed by Arizona University in association with police department. This project basically extracts entities from police records. Extracted entities include frequency, seriousness, duration and nature have been used to compare the similarity between pairs of criminals by a new distance measure and then cluster the data [6].

Crime and Criminal Tracking Network System (CCTNS) project is under implementation. This project will bring the connectivity of 16000+ police stations and 6000+ higher police stations. Main aim of this project to retrieve all the information related particular criminal and investigation progress at a single click. This project also includes the training program for police officers at various levels [11].

3. DATASET GENERATION

For building the database field work approach is employed. Under this field work, police station visit, meeting with different criminal investigators is expected. The main advantage of this field work is to create criminal database with real attributes.

Database is created in Attribute Relation File Format (ARFF). ARFF format supports variety of data types for attributes such as real, numeric, etc. We can also pass direct values for attributes. Once attributes and their data type are defined, it is very easy to add records. ARFF file format does not have any size constraint. We can add n number attributes and n number of instances. ARFF file format ensures the first level of data independence or logical data independence. When any of the data mining algorithms operates on a subset of attributes of a relation, the conceptual structure of the dataset does not get affected when new attributes and instances are added to the same relation. Dataset of 20 attributes and 1000 criminal records is created [1].



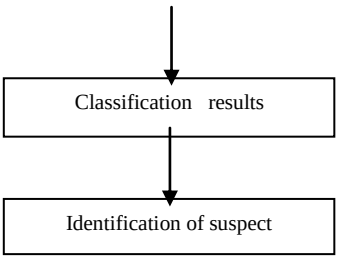


Figure 1 Block Diagram of the research work

Figure 1 shows the block diagram of the research work. In this, crime and criminal dataset is built in ARFF format. Dataset has 1000 records. After building the dataset J48-decision tree algorithm is evaluated against it to classify the crime data. Results of the algorithm are measured in terms of correctly classified instances, confusion matrix, TP rate, FP rate, Precision, Recall, F_measure and MCC. Classification results are used for identifying if particular person is suspect or not.

4. CLASSIFICATION USING J48-DECISION TREE ALGORITHM

Classification technique is used to find the common properties among different attributes of crime and criminal dataset and organizes them into different predefined classes. Classification is a data mining technique which is often used for predicting crime patterns and it also can reduce the time required to identify these crime patterns. However classification technique requires proper training and testing data because a certain amount of missing values may limit the prediction accuracy. [2] For example, we can apply classification in crime application that given all past records of criminals who commit the crime, predict which current criminals are probably to commit the crime in the future. In this case, we divide the personal records into two classes- suspect and innocent. Then data mining algorithm, here J48-decision tree to classify the person into either of the class. Dataset of 1000 records is divided as training dataset and testing dataset (700:300 respectively). Training dataset is used to build the classification model and testing dataset is used to validate that model.

J48-decision tree algorithm

J48-decision tree algorithm is a simple yet widely used classification technique. Decision tree has a flow chart like tree structure. This tree structure has three types of nodes.

1. A root node that has no incoming edges and zero or more outgoing edges
2. Internal nodes, each of which has exactly one incoming edge and two or more outgoing edges.

Internal nodes denote test on attributes.

3. Leaf nodes, each of which has exactly one incoming edge and has no outgoing edges.

Leaf nodes denote the class labels.

Use of the decision tree consists of two phases.
1. Building the tree
2. Validating the tree
Building the tree
Initially all the training records are at the root.
The input to this algorithm consists of the training records E and the attribute set F. The algorithm works by recursively selecting the best attribute to split the data and expanding the leaf nodes until stopping criterion is met.
Algorithm
TreeGrowth(E,F)
if stopping_cond(E,F) = true then
leaf = createNode().
leaf.label = Classify(E).
return leaf
else
root = createNode().
root.test_cond = find_best_split(E,F).
let V = {v|v is a possible outcome of root.test_cond}.
for each v ε V do
E_v = {e | root.test_cond(e) = v and e ε E}.
child = TreeGrowth(E_v , F)
add child as descendent of root and label the edge (root-child) v.
end for
end if
return root

Details of the algorithm are given below

The createNode() function builds the decision tree by creating a new node. A node in the decision tree has a either test condition, denoted as node.test_cond, or a class label denoted as node.label
The find_best_split() function determines which attribute should be selected as the test condition for splitting the training records.
The Classify() function determines the class label to be assigned to a leaf node.
The stopping_cond() function is used to terminate the tree-growing process by testing whether all the records have either the same class label or the same attribute values. After building the decision tree, a tree pruning can be done to reduce the size of the tree.
Classifying a test record is simple once a decision tree is built. Starting from the root node, test condition is applied to the record and appropriate branch is followed based on the outcome of the test. This will lead to either another internal node or to the leaf node. The class label associated with leaf node is then assigned to the record [21].

5. CHALLENGES FOR BUILDING THE J48-DECISION TREE [21]

1. Splitting the training records
Each recursive step of the tree growing process must select an attribute test condition to divide the records into smaller subsets. To implement this step, the algorithm must provide a method for specifying the test condition for different attribute types as well as an objective measure for evaluating the goodness of each test condition.
2. Stopping of splitting procedure
A stopping condition is needed to terminate the tree growing process. A possible strategy is to continue expanding a node until either all the records belong to the same class or all the records have identical attribute values. Although both conditions are sufficient to stop any decision tree algorithm, other criteria can be imposed to allow the tree growing process to terminate earlier.

6. RESULTS

Here the classification result of J48 – a decision tree is shown using parameters such as correctly classified instances, true positive rate, false positive rate, precision, recall, F_measure, MCC.
Confusion matrix
Confusion matrix contains information about actual and predicted classifications done by a classification system [17].

Table 2: Confusion Matrix for J48 algorithm

A	B	Classified as
173	52	Suspect
12	63	Innocent

Instances correctly classified by J48-decision tree algorithm are 236 (173+63).
True positive (TP)
If the outcome from a prediction is p and the actual value is also p, then it is called as true positive.
It is calculated as diagonal element / sum of relevant row
For class suspect
J48 classifies 173 suspects as true positives out of 225 so for class suspect and J48 algorithm true positive rate becomes 0.769.
For class innocent
J48 classifies 63 innocent as true positives out of 75 so for class innocent and J48 algorithm true positive rate becomes 0.840
False positive (FP)
If the outcome from a prediction is p and the actual value is n, then it is called as false positive.
It is calculated as diagonal element / sum of relevant row

For class suspect
J48 misclassifies 12 suspects as false positives out of 75 so for class suspect and J48 algorithm false positive rate becomes 0.160.
For class innocent

J48 classifies 52 innocent as false positives out of 225 so for class innocent and J48 algorithm false positive rate becomes 0.231.

Precision

Precision is the fraction of retrieved instances that are relevant.

Precision is basically the measure of exactness or quality.

It is calculated as diagonal element / sum of relevant column.

For class suspect precision value is 0.935.

For class innocent precision value is 0.548.

Recall

Recall is the fraction of relevant instances that are retrieved.

Recall is basically the measure of completeness.

Recall is same as true positives for both classes.

F_measure

F_measure is used when both Precision and Recall are important to measure accuracy.

F_measure= 2*precision*recall/(precision+recall) .

For class suspect F_measure is 0.844 and for class innocent F_measure is 0.663.

Mathew's Correlation Coefficient (MCC)

MCC is a factor which indicates accuracy of final class predictions for bi-variant classification problem. Its value ranges from -1 to +1. -1 value indicates the complete disagreement between observed and predicted values, 0 indicates random prediction and +1 indicates perfect prediction. Value of MCC in our classification problem is 0.542.

Table 3: Classification Results

Factor\Class	Suspect	Innocent
TP Rate	0.769	0.840
FP Rate	0.160	0.231
Precision	0.935	0.548
Recall	0.769	0.840
F_measure	0.844	0.663
MCC	0.542	0.542

7. CONCLUSION

Traditional crime investigations processes are time consuming and require much effort. These processes lack in use of technology for crime investigation. Large amount of information is collected during crime investigation process and only useful information is required for analysis.

Data mining is the process of extracting useful information or knowledge from large data sources. J48-decision tree is widely used classification technique which can be effectively used serve the necessary purpose. Results have shown that J48-decision tree algorithm has good classification accuracy and classification rules can be used to predict the suspect of the crime.

8. FUTURE SCOPE

As a future extension of this study, a web based criminal identification system is proposed which will give information of criminal at a single click from any place. This system will also serve the purpose of crime information sharing and investigators can access the criminal information from any place.

REFERENCES

[1] Nitin Sakhare, Swati Joshi, "Criminal Identification System Based On Data Mining", 3rd ICRTET, ISBN No. 978-93-5107-220-1, 28-30 March 2014.

[2] Hsinchun Chen, Wingyan Chung, Jennifer Xu, "Crime Data Mining: A general framework and some examples", IEEE computer society, 2004

[3] Anshul Goyal, Rajni Mehta, "Performance comparison of Naïve Bayes and J48 classification algorithm", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11 (2012).

[4] Faith Ozgul, Claus Atzenbeck, Ahmet Celik, Zeki Erdem, "Incorporating data sources and methodologies for crime data mining"; IEEE International Conference, 2011.

[5] Chung-Hsien Yu, Max W. Ward, Melissa Morabito, Wei Ding, "Crime Forecasting Using Data Mining Techniques"; 11th IEEE International Conference on Data Mining workshops, 2011.

[6] P.Thongtae, S.Srisuk, "An Analysis of Data Mining Applications in Crime Domain" IEEE 8th International Conference on Computer and Information Technology workshops, 2008.

[7] Sara Hajian, Josep Domingo-Ferrer, Antoni Martinez-Balleste, "Discrimination Prevention in Data Mining for Intrusion and Crime Detection"; the IEEE International Conference, 2011.

[8] Devesh Bajpai, "Emerging Trends in Utilization of Data Mining in Criminal Investigation: An Overview"; JECET; June-August 2012; Vol.1.No.2, 124-131

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

[10] Neha Gohar Khan, V.B.Bhagat, "Effective Data Mining Approach for Crime-Terrorpattern Detection using Clustering Algorithm Technique"; International Journal of Engineering Research and Technology, Vol. 2, Issue 4, April 2013.

[11] Prof. H. N. Renushe, Prof. P. R. Rasal, Prof. A. S. Desai," Data mining practices for effective investigation of crime", International Journal of Computer Applications in Technology, Volume 3, May 2012, pp. 865-870.
[12] Kaushik H. Raviya, Biren Gajjar, "Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA.
[13] //www.ncrb.gov.in
[14] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition.
[15] Satish Kumar ,"Neural Network: A classroom approach" McGraw Hill Education, Second Edition
[16] Dr.S.P. Gupta, "Statistical Methods"
[17] Ian H. Witten, Eibe Frank, "WEKA – Machine Learning Algorithm in Java"
[18] www.sigkdd.org/kddcup/
[19] http://www.sans.org/securityresources/idfaq/data_mining.php
[20] https://en.wikipedia.org./wiki/Decision_Tree_classifier
[21] Pang-Ning Tan, Vipin Kumar, Michael Steibach, "Introduction to Data Mining" Pearson Publication.