Volume 5, Issue 8, August 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Drug Trafficking Suspect Prediction Using Data Mining

Chiranjeevi C B, Revathy R

Assistant Professor, Department of CSE,

J.N.N. Institute of Engineering, Chennai, Tamil Nadu, India

Abstract— In malevolence of the plodding surge in the number of speculative studies on trafficking crime, emphasis is infrequently placed on the solicitation of data mining to crime prevention. This study provides deeper understanding and assessment of the assistances of information technology for the empathy of smuggling crime. This study focuses on smuggling of drugs. The data source is the complete record of drugs such as cannabis, opioid, cocaine or amphetamine-type stimulant (ATS) group leaving and returning to ports in the Airport region. This paper essay applies both ID3 and C 4.5 to classify and predict criminal comportments in smuggling. At the same time, it shows the difference between C 4.5 and human inspection (HI), also the difference between ID3 and HI. This study inaugurates prototypes for drugs of different tonnage and operation purposes that can provide law enforcers with flawless verdict criteria. It is needed to hypothesis different prototypes for drugs to achieve the actual cases in the reality since smugglers will use different kinds of ship for different smuggling purposes. The study results show that the application of data mining techniques to smuggling drugs attains an average precision of 80%, and the application of ID3 and C 4.5 to smuggling drugs can achieve an average precision of 65%, both of which are of significantly higher efficiency levels compared with the current human inspection (HI) method. This study suggests the value of using an both ID3 and C 4.5 model to obtain good identification performance for different drug types as well as average savings of 92% on the manpower loading. Information technology can greatly help to increase the probability of seizing smuggling drugs. Nowadays, public administration information is saved electronically however is not employed well. In fact, it can increase the administrative efficiency by proper use of electronic data. In this study, for example, we expect better use of the data stored in the database to establish an identifying model of smuggling. Applying the automatic identification mechanism, it is useful to reduce the probability of smuggling crime. We will look at ID3 and C 4.5 clustering with some enhancements to aid in the process of identification of crime patterns. We applied these techniques to real crime data from a world drug report and validated our results. We also use semi-supervised learning technique here for knowledge discovery from the crime records and to help increase the predictive accuracy. We also developed a premium scheme for features here to deal with limitations of various out of the box clustering tools and techniques. This easy to implement data mining framework works with the geospatial plot of crime and helps to improve the productivity of the detectives and other law enforcement officers. It can also be applied for counter terrorism for homeland security.

Keywords— crime patterns, government information application, Crime data mining, smuggling predictions, ID3, C4.5

I. INTRODUCTION

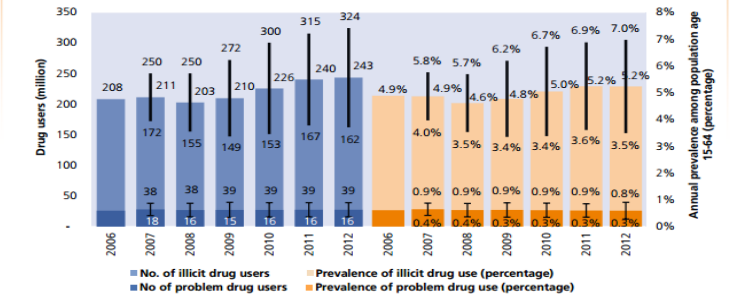
Universally, it is assessed that in 2012, some 243 million people (range: 162 million-324 million) analogous to some 5.2 per cent (range: 3.5-7.0 per cent) of the world population aged 15-64 had used an illicit drug — mainly a substance fitting to the cannabis, opioid, cocaine or amphetamine-type stimulant (ATS) group — at least once in the previous year. Although the degree of illegal drug use among men and women differs from country to country and in terms of the substances used, normally, men are two to three periods more likely than women to have used an illegal substance. While there are varying regional leanings in the amount of illicit drug use, overall global occurrence of drug use is considered to be steady. Similarly, the extent of difficult drug use, by regular drug users and those with drug use complaints or need, also remains stable, at about 27 million people (range: 16 million-39 million). With deference to the different groups of materials, there has been growth in opioid and cannabis use then 2009, while the use of opiates, cocaine and ATS (excluding "state") has either continued stable or followed a decreasing tendency. However, not all countries behavior national journals on drug use, and most nations that do so conduct them only sometimes, once each three to five years.

Therefore, somewhat than looking at the year-to-year change, it is more expressive to take a longer-term viewpoint. Also, year-on-year changes in a country's commonness rate have only a small impact on a region's overall occurrence unless they happen in a country with a large people. For 2012 data, updated incidence guesses are available for 33 countries, mostly nations of Western and Central Europe and North America, on behalf of nearly 12 per cent of the worldwide population aged 15-64. Therefore, the leanings and global yearly estimations of general drug use and of unlike substances imitate only the changes in or review of the evaluations for those countries and districts.

Drug use and gender

Nearly all drug use educations specify that men are more probable than women to usage drugs such as opiates and cannabis. Though the gender gap psychiatrists when data on the misappropriation of drugs are careful. In five recently plotted countries (Australia, United States of America, Spain, Urban Afghanistan, and Pakistan), the illegal use of drugs is more mutual among men than women, but the non-medical use of medicinal drugs is nearly corresponding, if not developed among women (see figure 3). Attractive together the shared estimates of those five examinations, the dishonest use of treatments is especially different for the two genders, as nearly half the women with past-year drug use had used medications, related with only one third of men.

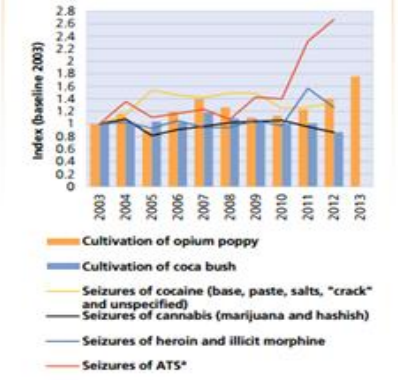
Fig. 1. **Global trends in drug use, 2006-2012**



Source: Estimates based on the UNODC annual report questionnaire.

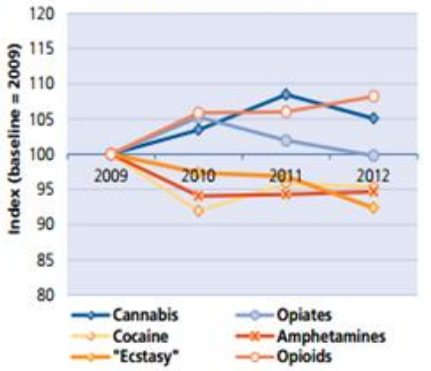
1 This is based on the prevalence rates of any drug use among males and females reported to the United Nations Office on Drugs and Crime (UNODC) by Member States through the annual report questionnaire.

Trend in main indicators of drug supply and drug supply reduction, 2003-2013



Source: Seizure data: annual report questionnaire supplemented by other official sources.
Cultivation data: UNODC estimates based on national illicit crop monitoring systems supported by UNODC supplemented by other official data.

Fig. 2. **Trends in the prevalence of use of different drugs, 2009-2012**



Source: Estimates based on the UNODC annual report questionnaire.

Table 1. **Global estimates of users of different drugs, 2012**

	Number of users (millions of users)			Prevalence (percentage)		
	Best estimate	Low	High	Best estimate	Low	High
Cannabis	177.63	125.30	227.27	3.8	2.7	4.9
Opioids	33.04	28.63	38.16	0.7	0.6	0.8
Opiates	16.37	12.80	20.23	0.35	0.28	0.43
Cocaine	17.24	13.99	20.92	0.37	0.30	0.45
ATS	34.40	13.94	54.81	0.7	0.3	1.2
"Ecstasy"	18.75	9.4	28.24	0.4	0.2	0.6

Source: Estimates based on the UNODC annual report questionnaire.

II. LITERATURE SURVEY

Kondaveeti, A. Runger, G. Huan Liu Rowe, J. suggested Large networks of sensors are used to detect intrusions and provide security at the borders of the United States. Sensor signals are used to detect possible intrusions such as illegal immigration traffic in drugs, weapons, and smuggled goods at specific targeted geographic locations. GIS systems can be used to capture, store and analyze this location based intervention data. Using a GIS system, a spatial database can be generated from the sensor intervention data which can take into account relevant geographic information in the vicinity of the sensed interventions. Important geographic features that are close to the intervention locations such as: plateaus, hills, valleys or roadways can be extracted and added to the analysis using ArcGIS. GIS techniques alone cannot reveal meaningful hidden information within geographic data. We have developed an integrated approach involving data mining and GIS techniques to extract patterns and trends in geographic data that can aid and inform analysis. Our approach uses both spatial and association data mining techniques. Spatial data mining is the process of discovering previously unknown, interesting and potentially useful patterns from spatial datasets. Applying association rule mining to

the spatial data can reveal additional important spatial relationships and help determine the relevance and importance of the sensor data. Spatial association rule mining was used to discover patterns in the intervention data, such as linking a sensed intrusion with a potentially hidden location such as a canyon, to infer a high probability of illegal traffic or immigration.

Kowalska, K. Adv. Technol. Centre, BAE Syst., Bristol, UK, Peel, L proposed a model of normal vessel behaviours is useful for detecting illegal, suspicious, or unsafe behaviour; such as vessel theft, drugs smuggling, people trafficking or poor sailing. This work presents a data-driven non-parametric Bayesian model, based on Gaussian Processes, to model normal shipping behaviour. This model is learned from Automatic Identification System (AIS) data and uses an Active Learning paradigm to select an informative subsample of the data to reduce the computational complexity of training. The resultant model allows a measure of normality to be calculated for each newly-observed transmission according to its velocity given its current latitude and longitude. Using this measure of normality, ships can be identified as potentially anomalous and prioritised for further investigation. The model performance is assessed by its ability to detect artificially generated AIS anomalies at locations around the United Kingdom. Finally, the model is demonstrated on case studies from artificial and real vessel data to detect anomalies in unusual tracks.

In this survey, Anomaly detection for sea surveillance - unsupervised clustering of normal vessel traffic patterns is proposed and implemented, where patterns are represented as the momentary location, speed and course of tracked vessels. The learnt cluster models are used for anomaly detection in sea traffic. The Gaussian Mixture Model is used as cluster model and a greedy version of the Expectation-Maximization algorithm is used as clustering algorithm. The models have been trained and evaluated using real recorded sea traffic. A qualitative analysis reveals that the most distinguishing anomalies found in the traffic are vessels crossing sea lanes and vessels traveling close to and in the opposite direction of sea lanes. In order to detect complex anomalies involving multiple vessels and/or behavior that develop over time, a more sophisticated pattern model should be developed. Yet, the generality of the proposed model is stressed, as it is potentially applicable to other domains involving surveillance of moving objects

III. DISCUSSION & RESEARCH RESULT

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains. The ID3 algorithm begins with the original set S as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates the entropy H(S) (or information gain IG(A)) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set S is then split by the selected attribute to produce subsets of the data. The algorithm continues to recur on each subset, considering only attributes never selected before. Recursion on a subset may stop in one of these cases:

Every element in the subset belongs to the same class (+ or -), then the node is turned into a leaf and labelled with the class of the examples there are no more attributes to be selected, but the examples still do not belong to the same class (some are + and some are -), then the node is turned into a leaf and labelled with the most common class of the examples in the subset there are no examples in the subset, this happens when no example in the parent set was found to be matching a specific value of the selected attribute. Then a leaf is created, and labelled with the most common class of the examples in the parent set. Throughout the algorithm, the decision tree is constructed with each non-terminal node representing the selected attribute on which the data was split, and terminal nodes representing the class label of the final subset of this branch.

Suppose we want to use the ID3 algorithm to decide if the time ready to play ball. During an hour, the data are collected to help build an ID3 decision tree.

Table 2- Data Set S

Passengers	**Previous criminal offence cases**	**Locality**	**Note unfamiliar vehicles**	**Watch for packages being exchanged**	**Analysis of body language**	**Suspect**
Person 1	Yes	North America	No	Yes	Nervous	Yes
Person 2	No	Latin America and Caribbean	Yes	No	Cool	Yes
Person 3	Yes	Central Asia and Transcaucasia	No	No	Cool	Yes
Person 4	No	East and South-East Asia	No	No	Cool	No
Person 5	Yes	South-West Asia	No	No	Cool	Yes
Person 6	Yes	Near and Middle East	No	Yes	Nervous	Yes
Person 7	No	South Asia	No	No	Cool	No
Person 8	No	Eastern and South-Eastern Europe	No	No	Cool	No

Person 9	Yes	Western and Central Europe	No	No	Cool	Yes
Person 10	Yes	Oceania	No	No	Nervous	Yes
Person 11	No	North America	No	Yes	Cool	No
Person 12	No	Latin America and the Caribbean	No	No	Cool	No
Person 13	No	Central Asia and Transcaucasia	No	No	Cool	No
Person 14	No	East and South-East Asia	Yes	No	Cool	Yes
Person 15	No	South-West Asia	No	Yes	Cool	No

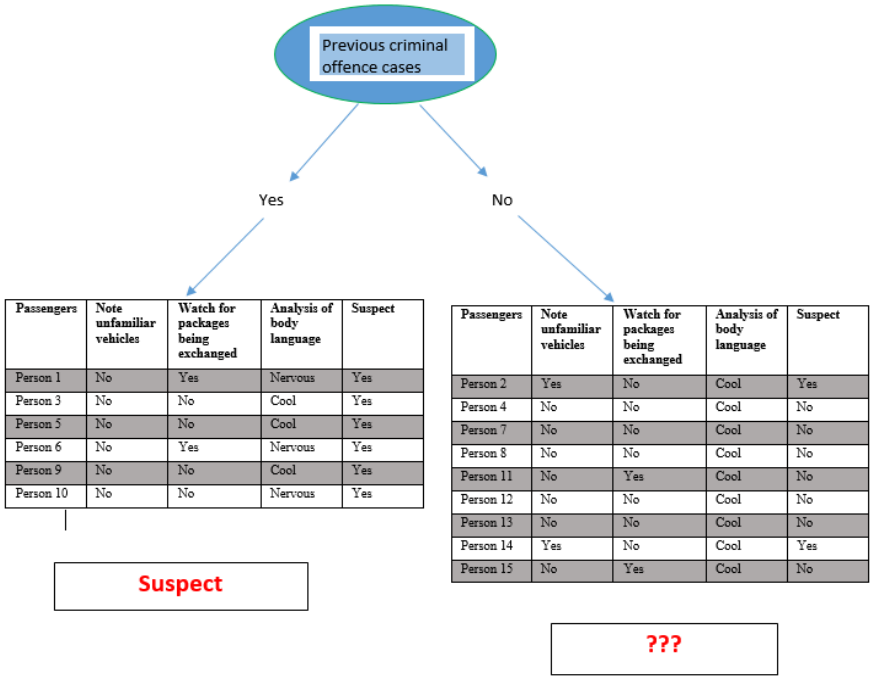


Fig 4 Root Node of ID3 & C 4.5 Decision Tree

Inputs: R: a set of non- target attributes, C: the target attribute, S: training data.
Output: returns a decision tree
Start
Initialize to empty tree;
If S is empty then
Return a single node failure value
End If
If S is made only for the values of the same target then
Return a single node of this value
End if
If R is empty then
Return a single node with value as the most common value of the target attribute values found in S
End if
D ← the attribute that has the largest Gain (D, S) among all the attributes of R
{dj j = 1, 2, ..., m} ← Attribute values of D
{Sj with j = 1, 2, ..., m} ←The subsets of S respectively constituted of dj records attribute value D
Return a tree whose root is D and the arcs are labeled by d1, d2, ..., dm and going to sub-trees ID3 (R-{D},

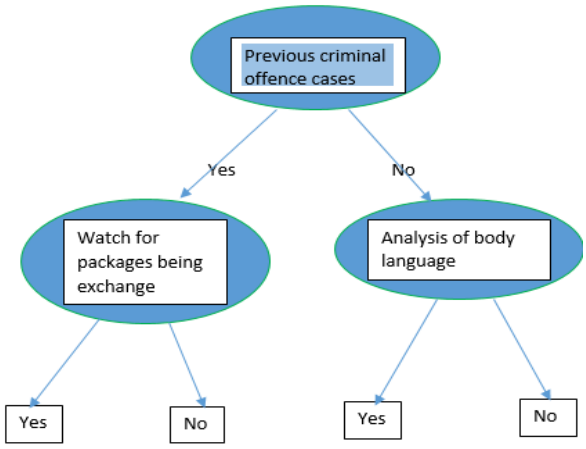
FormTree(T)

(1) ComputeClassF requency(T);
(2) if OneClass or F ewCases
return a leaf;
create a decision no de N;
(3) ForEach Attribute A
ComputeGain(A);
(4) N.test = A ttributeWithBestGain;
(5) if N.test is con tin uous
find Threshold;
(6) ForEach T' in the splitting of T
(7) if T' is Empty
Child of N is a leaf
else
(8) Child of N = F ormT ree(T 0);
(9) ComputeErrors of N;
return N

Program 2: Pseudo-code of the C4.5 Tree-Construction

C, S1), ID3 (R-{D} C, S2), .., ID3 (R-{D}, C, Sm)
End
Program 1: Pseudo-code of the ID3 Tree-Construction

Decision Rules:-



If Previous criminal cases= Yes then
If Watch for packages being exchanged = Yes
Then
Suspect = Yes ;
else
Suspect= No ;
Else if Previous criminal cases= Yes
Suspect = Yes ;
Else if Previous criminal cases= Yes then
If Analysis of body language =Yes
Suspect = No;
else
Suspect = Yes .

Fig 5 Final Decision Tree

IV. CONCLUSIONS

Decision trees are simply retorting to a problem of insight is one of the few methods that can be presented quickly enough to a non-specialist audience data processing without getting lost in difficult to understand mathematical formulations. In this paper, we wanted to focus on the key elements of their construction from a set of data, then we presented the algorithm ID3 and C4.5 that respond to these specifications.

REFERENCES

[1] Johan Baltié, DataMining : ID3 et C4.5, Promotion 2002, Spécialisation S.C.I.A. Ecole pour l'informatique et techniques avancées.
[2] Benjamin Devéze & Matthieu Fouquin, DATAMINING C4.5 – DBSCAN, PROMOTION 2005, SCIA Ecole pour l'informatique et techniques avancées.
[3] E-G. Talbi, Fouille de données (Data Mining) -Un tour d'horizon -Laboratoire d'Informatique, Fondamentale de Lille, OPAC.
[4] Ricco Rakotomalala, Arbres de Décision, Laboratoire ERIC, Université Lumière Lyon 2, 5, av. Mendés France 69676 BRON cedex e-mail : rakotoma@univ-lyon2.fr
[5] Arbres de décision, Ingénierie des connaissances (Master 2 ISC).
[6] Thanh Ha Dang, Mesures de discrimination et leurs applications en apprentissage inductif, Thèse de doctorat de l'Université de Paris 6, spécialité informatique, juillet 2007.
[7] Vincent GUIJARRO.K, Les Arbres de Décisions L'algorithme ID3, Elissa, "Title of paper if known," unpublished.
[8] Rakotoarimanana Rija Santaniaina, Rakotoniaina Solofoarisoa, Rakotondraompiana Solofo, Algorithmes à arbre de décision appliqués à la classification d'une image satellite.
[9] J. Fürnkranz, Entscheidungsbaum-Lernen (ID3, C4.5, etc.) (V1.1,14.01.; neue Folie zu C4.5 Pruning)-- Site web : http://www.ke.tudarmstadt.de/lehre/archiv/ws0809/mldm
[10] Ankur Shrivastava and Vijay Choudhary ,Comparison between ID3 and C4.5 in Contrast to IDS Surbhi Hardikar, VSRD-IJCSIT, Vol. 2 (7), 2012, 659-667.
[11] A.P.Subapriya, M.Kalimuthu, EFFICIENT DECISION TREE CONSTRUCTION IN UNREALIZED DATASET USING C4.5 ALGORITHM, ISSN: 2230-8563.
[12] Anurag Upadhayay ,Suneet Shukla, Sudsanshu Kumar, Empirical Comparison by data mining Classification algorithms (C 4.5 & C 5.0) for thyroid cancer data set. ISSN:2249-5789 .
[13] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg. Top 10 algorithms in data mining, Knowl Inf Syst (2008) 14:1–37 DOI 10.1007/s10115-007-0114-2
[14] Introduction to Data Mining and Knowledge Discovery, Third Edition by Two Crows Corporation.
[15] G. COSTANTINI R. Nicole, Probabilité conditionnelle. Indépendance, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[16] CELINE ROBARDET, Data Mining, http://prisma.insa-lyon.fr/sc.
[17] P. Habermehl et D. Kesner, Algorithmes d'apprentissage, Programmation Logique et IA

AUTHORS

Chiranjeevi C B is Assistant Professor since 2013 in department of Computer Science and Engineering at JNN Institute of Engineering. He received the B.E. degree in Electronics and Communication Engineering in 2008 and M.E. degree in Computer Science and Engineering in 2013. He Obtained 41st University Rank in M.E (CSE), Anna University. He is a student member of CSI. His research interests include Data Mining, Web Technology and Embedded Programming.

Revathy R is Assistant Professor since 2013 in department of Computer Science and Engineering at JNN Institute of Engineering. She received her B.Tech. degree in Information Technology in 2010 and M.E. degree in Computer Science and Engineering in 2013. Her research interests include Data Mining, Database Technology and Computer Architecture.

