

Suspect Vehicle Identification for Border Safety with Modified Mutual Information

Siddharth Kaza, Yuan Wang, and Hsinchun Chen

Department of Management Information Systems, University of Arizona
sidd@u.arizona.edu, ywang@email.arizona.edu,
hchen@eller.arizona.edu

Abstract. The Department of Homeland Security monitors vehicles entering and leaving the country at land ports of entry. Some vehicles are targeted to search for drugs and other contraband. Customs and Border Protection agents believe that vehicles involved in illegal activity operate in groups. If the criminal links of one vehicle are known then their border crossing patterns can be used to identify other partner vehicles. We perform this association analysis by using mutual information (MI) to identify pairs of vehicles that are potentially involved in criminal activity. Domain experts also suggest that criminal vehicles may cross at certain times of the day to evade inspection. We propose to modify the mutual information formulation to include this heuristic by using cross-jurisdictional criminal data from border-area jurisdictions. We find that the modified MI with time heuristics performs better than classical MI in identifying potentially criminal vehicles.

1 Introduction

In recent years border safety has been identified as a critical part of homeland security. The national strategy for homeland security [1] calls for the creation of “smart borders” that provide “greater security through better intelligence and coordinated national efforts.” In addition, the report also emphasizes that information sharing systems are the foundations to improve the nation’s infrastructure.

The Department of Homeland Security (DHS) monitors vehicles entering and leaving the country, recording their license plates with a date and time of entry using license plate readers. Customs and Border Protection (CBP) agents search vehicles for drugs and other contraband. These thorough checks are done for vehicles on watch lists (target vehicles) and on random vehicles as well. This process is time consuming and if the waiting times become too long, the flow of people, vehicles, and commerce is impaired.

CBP agents believe that vehicles involved in illegal activity (especially smuggling) operate in groups. If the criminal links of one vehicle in a group are known, then the group’s crossing patterns and frequency can be used to identify other partner vehicles. In a previous study [10] we found that the criminal associations of vehicles crossing the border may be recorded in local law enforcement databases in border-area jurisdictions. However, Customs and Border Protection does not always have access to criminal records of vehicles and sometimes lacks the methods to perform this analysis.

We perform this association analysis by using mutual information (MI) to identify pairs of vehicles crossing together and potentially involved in criminal activity. Our previous study [7] had found that the use of MI may be a promising solution to this problem. In this paper we do an evaluation of MI in this problem domain and also attempt to modify the measure to incorporate domain heuristics. Domain experts (CBP agents, police detectives and analysts) suggest that groups of criminal vehicles may cross at certain times during the day to try and evade inspection. It is difficult to identify these heuristics with border crossing information alone since it does not contain clear indications of criminal history or possible intent. We use law enforcement information from border-area jurisdictions to identify times that criminal vehicles prefer and incorporate this knowledge in the MI formulation.

This study attempts to answer the following questions:

- Can law enforcement information from border-area jurisdictions be used to identify target vehicles at the border?
- How can we include domain heuristics to enhance the performance of mutual information?

In the next section we discuss background information and previous studies using mutual information. Section 3 presents the research testbed and explains the research design. Experimental results are shown and discussed in Section 4. Section 5 concludes and presents future directions.

2 Literature Review

In this section we review previous studies that have used association mining and mutual information in various domains. We also briefly discuss the challenges of using information from multiple data sources.

2.1 Information from Multiple Sources

In order to explore the criminal links of border-crossing vehicles it is necessary to extract data from multiple sources. To triangulate information about a vehicle, all the instances of the vehicle across datasets have to be reconciled, which is a challenging task. Matching of entities and their relationships is a task that is hampered by problems that include [4]: *name differences*: similar entities in different databases have different names, *missing and conflicting data*: incomplete data or different values in different sources, and *object identification*: lack of global identifiers.

We use the BorderSafe information sharing and analysis framework [10] for accessing information from multiple datasets. These datasets include border crossing and local law enforcement records. Information on border crossing vehicles is located in local law enforcement datasets using their license plates and issue authorities (states). This enables us to extract the criminal histories for border crossing vehicles. More details about the datasets and their use are presented in Section 3.

2.2 Association Rule Mining

Inferring associations between items in a database was motivated by decision support problems faced by retail organizations [14]. Retail stores needed information on which items their customers were likely to buy together. The problem spawned a method in data mining known as association rule mining. An association rule is a relationship of the form $A \rightarrow B$, where A is the antecedent item-set and B is the consequent item-set. The antecedent and consequent item-sets can contain multiple items. $A \rightarrow B$ holds in a transaction set T with confidence ‘ c ’ if $c\%$ of transactions in T that contain A , also contain B . $A \rightarrow B$ holds with support ‘ s ’ if $s\%$ of transactions in T contain both A and B . To find associations between two item-sets, the association mining procedures identify all relationships (rules) that have support and confidence greater than user-specified thresholds.

Association rule mining has been applied in many domains including ‘market basket’ data [2], web log analysis (to identify online user behavior) [11], network intrusion detection [8], and gene regulatory network extraction (to identify cause-effect relationships between genes) [3].

2.3 Mutual Information

Mutual Information is an information theoretic measure that can be used to identify interesting co-occurrences of objects. The earliest definition of MI was given by Fano [6]. It was defined as the amount of information provided by the occurrence of an event (y) about the occurrence of another event (x). They formulated it as:

$$I(x; y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

Intuitively, this concept measures if the co-occurrence of x and y ($P(x, y)$) is more likely than their independent occurrences ($P(x).P(y)$). This formula is referred to as the classical mutual information in the rest of the paper. Classical MI can be considered a subset of association rule mining with 1-item antecedent and 1-item consequent item-sets.

The MI measure has been applied to problems in many domains. It works well for phrase extraction from text documents. This is because text documents can be considered as a set of events (words), and the probability of the occurrence of a word can be calculated over the entire document. Previous studies in this area have used MI to study association between words in English texts and identify commonly occurring phrases [5]. It has also been used for key phrase extraction in Chinese texts [12].

Pantel et al. [13] used MI to match database columns containing similar information. In the bioinformatics domain, MI has been used to extract protein motif patterns from sequences [15]. However, the above studies have not modified the classical MI measure to include domain heuristics.

Work on extending or modifying the classical MI measure to add domain heuristics includes studies in natural language processing: Magerman and Marcus [9] modified the MI measure (bi-gram) to include n-grams and bioinformatics: Wren [16] extended the measure to calculate transitive MI scores for biological associations.

Border-crossing records can be considered as a stream of text (license plates) ordered by the time of crossing. So, MI can be used to identify frequent co-occurrence between a pair of vehicle crossings. If one vehicle in the pair has a criminal record, some inferences may be made about the second vehicle if they cross together frequently. In a previous study [7] we found that the time of crossing may be an important heuristic for improving the performance of MI in this domain. We propose to use conditional probability to include these domain heuristics in the MI formulation (Section 3.1.5).

3 Research Testbed and Design

The testbed for this study includes datasets obtained from the Tucson Police Department (TPD) and Pima County Sheriff's Department (PCSD). Data from these agencies is referred to as police data throughout this paper. In addition, we also use data from the Tucson Customs and Border Protection (CBP). These datasets were provided to us through the BorderSafe project funded by the Department of Homeland Security. The TPD and PCSD datasets include information on police incidents over 15 years (1990-2005). These incidents include individuals and vehicles that are involved in illegal activity in southern Arizona. A summary of these datasets is shown in Table 1.

Table 1. Key statistics of TPD and PCSD data

	TPD	PCSD
Date Range	1990 – 2005	1990 – 2004
Recorded Incidents	3.3 million	2.18 million
Vehicles	800, 656	520, 539

CBP data includes information on vehicles crossing the border between Arizona and Mexico at six ports of entry. This data includes the license plate, state, date, port, and time for crossings between 2003 and 2005. Details of this dataset are shown in Table 2.

Table 2. Key statistics of CBP border crossing data

Recorded crossings	10.7 million
Number of vehicles	1.7 million

3.1 Research Design

Prior to presenting the research design we need to define the terms *criminal vehicle* and *police contact*. A criminal vehicle is a vehicle that has been suspected, arrested, or has a warrant (with its occupant) for crimes that include narcotics (sale, possession, etc.), violence (homicide, aggravated assault, armed robbery, etc.), larceny and theft (property, vehicles, etc.), and other serious crimes in the TPD/PCSD datasets. Police detectives and analysts consider these crimes and roles (suspect, arrestee) as strong

indications of involvement in criminal activity. A vehicle that has had a police contact is one that is recorded in the law enforcement databases; this may be for serious crimes (as listed above) or for other activities that may include forgery and counterfeiting, suspicious activity, and others. Vehicles with police contacts are also referred to as *potentially criminal* vehicles in this paper. These definitions are used in the description of the design and the evaluation process.

To identify interesting pairs of vehicles that cross the border together we use the *time of crossing* as a heuristic to enhance mutual information. The time of crossing heuristic suggests that vehicle pairs that cross during certain times of the day/night are more interesting than others. Domain experts (CBP agents, police detectives and analysts) and our previous study [7] suggest that criminal vehicles regularly cross at odd times during the night. The mutual information measure modified to include the time heuristic is referred to as *MIT* (Equation 2, shown in Section 3.1.5) and classical mutual information (without heuristics) is referred to as *MIC* (Equation 1, shown in Section 3.1.4).

Fig. 1 shows the research design and the process of utilizing information from multiple sources, heuristic calculation, and identification of potential target vehicles at the border. Different parts of the figure are explained in the following sub-sections.

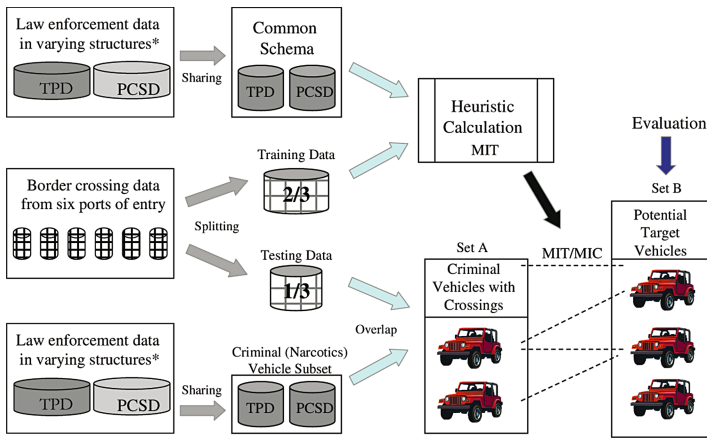


Fig. 1. Research design and process

3.1.1 Heuristic Calculation

The TPD and PCSD datasets were consolidated by transforming them to a single schema [10]. The common schema contained information on all vehicles that had police contacts along with information on the incidents that they were involved in. This was done to simplify access to multiple sources of information. To evaluate the performance of MIT and MIC, the CBP border crossing records were divided into training and testing datasets. This was done using a 2/3 – 1/3 hold out procedure. The training dataset contained 6.5 million ($\approx 2/3$ of total) crossing records from March 2004 to November 2004.

To calculate the time heuristic the day was divided into six time periods corresponding to office travel (5am to 10am), travel for lunch (10am to 2am), night time (8pm to 12pm, 12pm – 5am), and others. These time periods were defined with the help of domain experts. For each of these time periods the ratio of vehicles with police contacts to the total number of crossings was calculated. This value was used to inform the mutual information score between vehicles in a given time period (as shown in Section 3.1.5).

3.1.2 Testing

The testing data contained 3.5 million ($\approx 1/3$ of total) crossings from November 2004 to June 2005. Police data and the border crossings in the testing dataset were used to identify two sets of vehicles:

Set A: 140 criminal vehicles that had been arrested or suspected for narcotics sale in the TPD/PCSD jurisdiction since January 2003.

Set B: All the border crossing vehicles crossing *within one hour* of vehicles in *Set A* at the same port and in the same direction (i.e., both vehicles are either entering the U.S. or leaving it).

MIT and MIC were calculated between vehicles in *Set A* and *Set B*. The vehicles with high scores were considered potential target vehicles.

3.1.3 Evaluation

The potential target vehicles identified were evaluated by measuring their overlap with police datasets. This was done by measuring the number of vehicles with police contacts that were contained in the set of potential target vehicles. The number of potentially criminal vehicles identified by MIT and MIC were compared to each other to ascertain the performance of the modified measures. Since the aim of CBP is to target potentially criminal vehicles, a greater number of such vehicles in the target vehicle set indicates a higher quality result.

3.1.4 Classical Mutual Information (MIC) Formulation

The classical mutual information score between any two vehicles is defined as:

$$MIC(A, B) = \log_2 \frac{P(A, B)}{P(A)P(B)} \quad (1)$$

Here A is a vehicle in *Set A*, and B is a vehicle in *Set B*. $P(A)$ and $P(B)$ are the probabilities of the vehicles A and B crossing the border, these are calculated from the border crossing datasets. $P(A, B)$ is the probability of B crossing within one hour of A , this is calculated based on the number of times A and B are seen crossing together.

3.1.5 Mutual Information with Time Heuristics (MIT)

In the MIT formulation, we use conditional probability to modify the definition of $P(A)$, $P(B)$, and $P(A, B)$.

$P'(A)$: Probability that vehicle A crosses the border and has a police contact.

$P'(B)$: Probability that vehicle B crosses the border and has a police contact.

$P'(A, B)$: Probability that vehicles A and B cross the border together and have police contacts.

Thus, a high $MI'(A,B)$ indicates that the vehicles are likely to cross the border and potentially commit crimes together.

Given this, we can now modify the classical MI formulation to include the time heuristic: Let $P_c(a)$ be the probability that vehicle ‘ a ’ has contact with the police, and $P_b(a)$ be the probability that ‘ a ’ crosses the border. The probability of vehicles with police contacts crossing during the six time periods is calculated using historical information in the police databases. So, we can obtain $P_c(V|t)$, which is the probability that any vehicle V in time period t ($1 \leq t \leq 6$) will have a contact with the police.

Now, by definition of $P'(A)$,

$$P'(A) = \sum_{t=1}^6 P[(A_b \text{ and } A_c) | t]$$

In the above equation A_b refers to vehicle A crossing the border, and A_c refers to vehicle A having contact with the police. This equation reduces to

$$P'(A) = \sum_{t=1}^6 P_b(A | t) P_c(V | t)$$

since the probability that a vehicle crosses the border and having police contact are independent (so they are multiplied to obtain $P'(A)$). In addition, A is replaced by V in the second term since the probability that a vehicle in time period t has a police contact is the same for all vehicles in that time period. So the above process basically utilizes historical information (about crime) in the police datasets as a weight to modify $P'(A)$. Similar derivations can be used to obtain $P'(B)$, $P'(A,B)$, and thus $MIT(A,B)$ as shown in the following equations:

$$\begin{aligned}
 P'(B) &= \sum_{t=1}^6 P_b(B | t) P_c(V | t) \\
 P'(A,B) &= \sum_{t=1}^6 P[((AB)_b \text{ and } (AB)_c) | t] = \sum_{t=1}^6 P_b((AB) | t) P_c(V | t) P_c(V | t) \\
 MIP(A,B) &= \log_2 \frac{P'(A,B)}{P'(A)P'(B)} \tag{2}
 \end{aligned}$$

4 Experimental Results

To ascertain whether law enforcement information can be used to identify potential criminal vehicles, we first measured the overlap between border-crossing vehicles and police records in border-area jurisdictions. There were 45,091 border crossing vehicles that had police incident records in TPD/PCSD datasets. The number suggests that many vehicles crossing the border have incidents recorded in local law enforcement databases. This is a positive sign since it allows us to identify target vehicles at the border by exploring their criminal links. The existence of the overlap is also important for the calculation of heuristics based on law enforcement information.

4.1.1 Temporal Patterns of Border Crossings

Studying the temporal patterns of border crossings helps better understand the crossing activity. Fig. 2(a) shows the time distribution of border crossings for all vehicles entering and leaving the country over six time periods. Each slice of the pie shows the percentage of the total crossings that take place in the respective time period. It can be seen that a majority (about 65%) of the border crossings occur due to work and lunch/dinner related traffic during working hours (approximately 6am to 8pm). The chart also shows that about 37% of all crossings take place during the night or after dark (approximately 7pm to 6am).

Fig. 2(b) shows the time distribution of border crossings by vehicles with police contacts. Each slice of the pie shows the percentage of total crossings by such vehicles that took place in the respective time-period. For instance, 27% of all the border crossings by police contact vehicles took place between 8pm and midnight. The chart suggests that a large percentage (about 48%) of crossings by these vehicles take place after dark. MIT incorporates this information to assign more weight to time periods with high percentage of crossings by vehicles with police contacts. The weights also discount work travel related periods since they have a lower percentage of such crossings. Such information can also be used by CBP to increase or decrease enforcement in certain time periods.

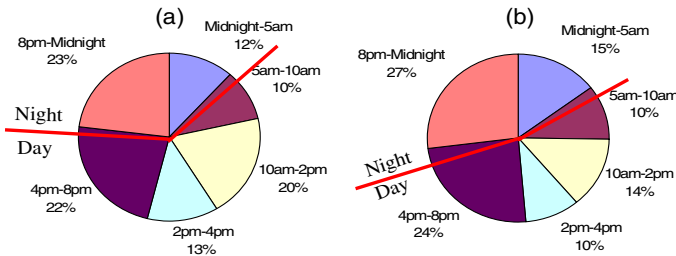


Fig. 2. Temporal distribution of crossings - (a) percentage of total crossings (b) percentage of crossings with police contacts

4.1.2 Comparative Evaluation of MIT and MIC

Mutual information scores (MIT and MIC) were calculated for 230,000 pairs of vehicles (the first vehicle from *Set A* and the second from *Set B*). To compare the two measures, the number of police contact vehicles identified by each was counted. The results are shown in Fig. 3. On the X-axis are top-*n* pairs (*n* ranging from 10-2500) of vehicles ordered by their MIT and MIC scores. On the Y-axis is the number of vehicles with police contacts identified by the two measures. For instance, three vehicles of the top-100 vehicles identified by MIT had previous police contacts. As can be seen in the figure MIT consistently identified more potentially criminal vehicles (vehicles with prior police contacts) than MIC.

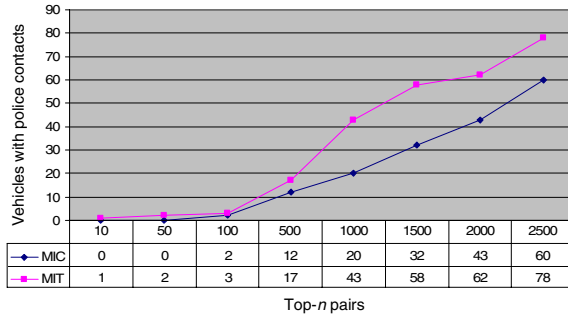


Fig. 3. Number of vehicles with police contacts identified by MIT and MIC

Even though the top- n pairs contained a few potentially criminal vehicles, they also contained other vehicles that had no past criminal records. This might not look promising in other domains, but has positive connotations in this one. It suggests that many of the vehicles postulated to be criminal by the algorithms were not known to have police records before. So the new measure can be used to identify potentially new criminal vehicles that can be targeted at the border. The low number of police contacts might also be a result of properties of the datasets. A more accurate evaluation of the algorithm is possible if a larger dataset was available for training and testing. We commit this to future work.

6 Conclusions and Future Directions

Exploring the criminal links of border crossing vehicles in local law enforcement databases can be used to enhance border security. In this study we used mutual information to identify pairs of border crossing vehicles that may be involved in criminal activity. We found that mutual information may be used to identify potential target vehicles at the border. In addition, we concluded that the mutual information measure modified to include domain heuristics like time of crossing performs better than classical mutual information in the identification of potentially criminal vehicles.

In the future, we plan to incorporate other domain heuristics like port of crossing, traffic at the port of entry, and makes of vehicles in the mutual information formulation. In addition we plan to use larger datasets for training and testing of the new measures. We also plan to design a more comprehensive evaluation scheme (including cross-validation) to test the effectiveness of the modified measures as compared to classical mutual information.

Acknowledgements

This research was supported in part by the NSF Digital Government (DG) program: “COPLINK Center: Information and Knowledge Management for Law Enforcement” #9983304, NSF Knowledge Discovery and Dissemination (KDD) program:

"COPLINK Border Safe Research and Testbed" #9983304, NSF Information Technology Research (ITR) program: "COPLINK Center for Intelligence and Security Informatics Research - A Crime Data Mining Approach to Developing Border Safe Research" #0326348, and Department of Homeland Security (DHS) through the "BorderSafe" initiative #2030002.

We thank our BorderSafe project partners: Tucson Police Department, Pima County Sheriff's Department, Tucson Customs and Border Protection, ARJIS (Automated Regional Justice Information Systems), San Diego Super Computer Center (SDSC), SPAWAR, Department of Homeland Security, and Corporation for National Research Initiatives (CNRI). We also thank Homa Atabakhsh and Hemanth Gowda of the AI Lab at the University of Arizona, Tim Petersen and Chuck Violette of the Tucson Police Department, and Ron Friend of Tucson Customs and Border Protection for their contributions to this research.

References

1. National Strategy for Homeland Security. Office of Homeland Security. (2002)
2. R. Agrawal, T. Imielinski, and A. Swami: Mining Association Rules between Sets of Items in large Databases. In: Proc. of ACM SIGMOD Conference on Management of Data (1993)
3. D. Berrar, W. Dubitzky, M. Granzow, and R. Eils: Analysis of Gene Expression and Drug Activity Data by Knowledge-Based Association Mining. In: Proc. of Critical Assessment of Microarray Data Analysis Techniques (CAMDA '01) (2001)
4. I.-M. A. Chen and D. Rotem: Integrating Information from Multiple Independently Developed Data Sources. In: Proc. of 7th International Conference on Information and Knowledge Management, Bethesda, Maryland (1998)
5. K. W. Church and P. Hanks: Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16 (1990) 22-29
6. R. M. Fano: *Transmission of Information*. MIT Press, Cambridge, MA (1961)
7. S. Kaza, T. Wang, H. Gowda, and H. Chen: Target Vehicle Identification for Border Safety using Mutual Information. In: Proc. of 8th International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria (2005)
8. W. Lee and S. J. Stolfo: Data Mining Approaches for Intrusion Detection. In: Proc. of 7th USENIX Security Symposium (1998)
9. D. M. Magerman and M. P. Marcus: Parsing a Natural Language using Mutual Information Statistics. In: Proc. of Eight National Conference on Artificial Intelligence (1990)
10. B. Marshall, S. Kaza, J. Xu, H. Atabakhsh, T. Petersen, C. Violette, and H. Chen: Cross-Jurisdictional Criminal Activity Networks to Support Border and Transportation Security. In: Proc. of 7th International IEEE Conference on Intelligent Transportation Systems, Washington D.C. (2004)
11. B. Mobasher, N. Jain, E. H. Han, and J. Srivastava: Web mining: Pattern discovery from world wide web transactions. Department of Computer Science, University of Minnesota. Minneapolis, Technical Report (1996)
12. T. Ong and H. Chen: Updateable PAT-Tree Approach to Chinese Key Phrase Extraction Using Mutual Information: A Linguistic Foundation for Knowledge Management. In: Proc. of Second Asian Digital Library Conference, Taipei, Taiwan (1999)

13. P. Pantel, A. Philpot, and E. Hovy: Aligning Database Columns using Mutual Information. In: Proc. of The 6th National Conference on Digital Government Research (dg.o), Atlanta, GA (2005)
14. M. Stonebraker, R. Agrawal, U. Dayal, E. Neuhold, and A. Reuter: The DBMS Research at Crossroads. In: Proc. of The VLDB Conference, Dublin (1993)
15. T. Tao, C. X. Zhai, X. Lu, and H. Fang: A study of statistical methods for function prediction of protein motifs. *Applied Bioinformatics* 3 (2004) 115-124
16. J. D. Wren: Extending the Mutual Information Measure to Rank Inferred Literature Relationships. *BMC Bioinformatics* 5 (2004)