A Bayesian Network Model for Analysis of the Factors Affecting Crime Risk

ROONGRASAMEE BOONDAO, VATCHARAPORN ESICHAIKUL
and NITIN KUMAR TRIPATHI
School of Advanced Technologies
Asian Institute of Technology
P.O. Box 4, Klong Luang, Pathumthani 12120
THAILAND

Abstract: - Crime is an inevitable societal problem and is a major problem of the nation. Effective crime control requires accurate prediction for decision making in crime control planning. This research analyses the factors affecting crime risk in the Bangkok metropolitan area, Thailand, using a Bayesian Network. The factors considered in this study are classified into five main groups: variables describing population, variables describing crime location factors, variables describing types of crimes, variables describing traffic, and variables describing the environment. Due to the uncertainty and incomplete nature of the variables, Bayesian Network theory is used to analyse the data since it is well suited to dealing with noisy and incomplete crime data. From the result of the model, the factors that considerably affected crime risk in the Bangkok Metropolitan Area are environment factors, type of crime factors, crime location factors, traffic factor and population factors, in that order. An empirical study on the predictive accuracy performance of the model is included. The Receiver Operating Characteristic analysis was used to test the model and the results show the model performed well.

Key-Words: - Bayesian Network, Crime Pattern Analysis, Crime Factors Analysis, Receiver Operating Characteristic (ROC) analysis

1 Introduction

A Bayesian Network [1] is a method of reasoning using probabilistic inference by representing processes or problems with a combination of causal relations between variables and conditional probability tables relating cause and effect variables. Such networks have become popular within the artificial intelligence, probability and uncertainty community as a decision support tool. Various methods have been used in crime analysis, especially Neural Network as described in [2, 3, 4]. Bayesian Network has been used in criminal profiling [5], drug crime knowledge management [6] and legal reasoning [7]. This paper describes the use of a Bayesian Network for analyzing the factors affecting crime risk, especially, murder cases in the Bangkok Metropolitan Area, Thailand. The Bayesian Network model was developed by expert elicitation and crime theory and it learned using a machine learning software, Hugin Researcher 6.3 [8]. There are five groups of factors: population, crime location, types of crimes, traffic, and environment. These set of variables were used to analyse the factors affecting crime risk. The results of the analysis are expected to be used for crime control planning.

In this paper we first describe the relevant Bayesian Network theory, followed by the methodology development and then examine the predictive accuracy of the Bayesian Network model by means of receiver operating characteristics (ROC) analysis.

2 Theoretical Background

2.1 Bayes' theorem

Bayesian Networks [9, 10] are known by names such as causal graphs, causal networks, belief networks and probabilistic networks. Bayesian Networks are directed acyclic graphs: they are "directed" in that the connections between nodes are "one-way" and they are "acyclic" because they cannot include loops. Network nodes represent random variables and arcs represent the direct probabilistic dependencies between variables. Each node has a conditional

probability table (CPT) which indicates the probability of each possible state of the node given each combination of parent node states. The tables of root nodes contain unconditional prior probabilities.

2.2 The basic concept of Bayesian Networks

One of the main benefits of Bayesian Networks is that they allow a probability distribution to be decomposed into a set of smaller distributions. The independence semantics associated with the network topology specify how to combine these local distributions to obtain the complete joint-probability distribution over all the random variables represented by the nodes in the network. The relationship between network topology and independence is captured by a property called d-separation [10].

3 Methodology Development

3.1 Data preparation

The data were collected from the National Statistical Office of Thailand, the Royal Thai Police, the Bangkok Metropolitan Administration and the Ministry of Transportation. In this research, data from January 2000 to December 2003 was used. The variables are classified into five groups: population, crime location factors, types of crimes, traffic, and environment see Fig. 1.

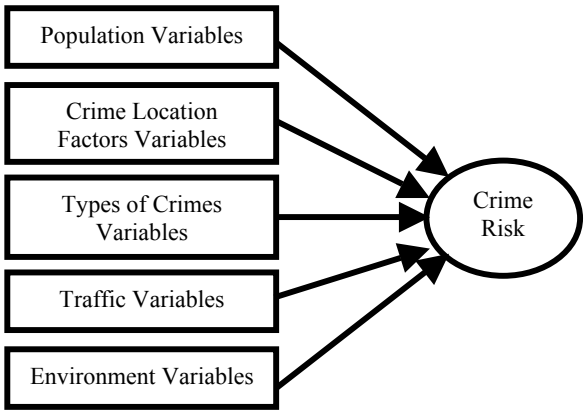

Fig.1 The set of variables for analysis of the factors affecting crime risk

- The variables describing population are population density (pop_density), population gender (pop_gender), male population (pop_male), female population (pop_female), population income (pop_income), population career (pop_primary, pop_secondary, pop_tertiary) and population age (infant, young, working, old).
- The variables describing crime location factors (crime_location_factors) are number of nightclubs (nightclubs), number of shopping centers (shopping_centers), number of movie theatres (movie_theatres), number of banks (banks) and number of hotels (hotels).
- The variables describing types of crimes are number of murder cases (murder), number of robbery cases (robbery), and number of rapes (rape).
- The variable describing traffic is traffic volume (traffic_volume).
- And variables describing the environment are number of low standard housing areas (low_standard_housing), number of drug-sale areas (drug_sales), and lighting on the roads of the Bangkok districts (lighting).

These data were used in recognizing the crime data pattern. The variables have five states: Very_low, Low, Medium, High, and Very_high. As an example, Table 1 shows some values in the partial Conditional Probability Table for the variable murder and the variable drug_sales. From Table 1, the probability of murder (very_low) given drug_sales (very_low) is 1.0.

Table 1 The partial Conditional Probability Table of Murder given Drug_sales

Drug_sales	***Murder***				
	Very_low	Low	Medium	High	Very_high
Very_low	1	0.5	0.2	0.2	0.2
Low	0	0.5	0.2	0.2	0.2
Medium	0	0	0.2	0.2	0.2
High	0	0	0.2	0.2	0.2
Very_high	0	0	0.2	0.2	0.2

3.2 Processes for analysis of the factors affecting crime risk

Processes were constructed to systematically analyse the factors affecting crime risk. The processes can be summarized as follows (See Fig. 2):

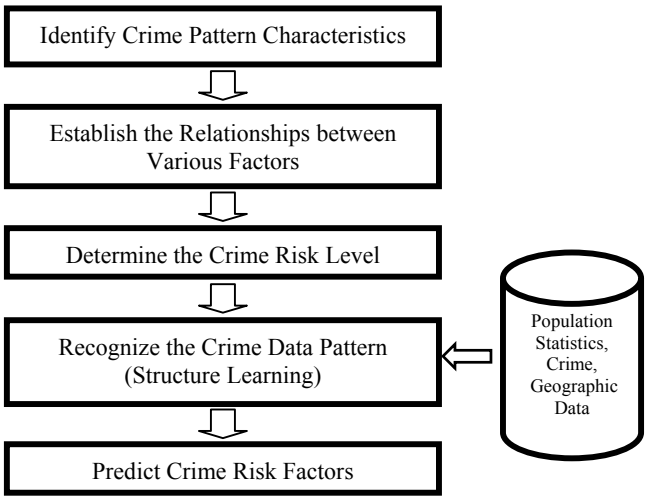

Fig.2 The Processes for analysis of the factors affecting crime risk

1. Identify crime pattern characteristics. This step was derived from studying crime theory and consulting with police officers responsible for crime investigation and suppression. The purpose of this step was to identify the factors that were related to high crime risk in each district.

2. Establish the relationships between various crime factors. We have to consider which factors influence what and determine the relationships among them.

3. Determine the crime risk level. After the establishment of the relationships, we need to decide the crime risk level for each factor. The results of steps 2 and 3 were derived from expert elicitation in the field of crime control.

4. Recognize the crime data pattern using structured learning. In this step, specialized software [8] was used to recognize the crime data pattern.

5. Predict crime risk factors. The details of the model will be explained in the next section.

3.3 The Bayesian Network model

The model was developed based on the crime pattern analysis of Brantingham and Brantingham [11] and theory of crime control through environmental design [12, 13]. Pattern theory focuses attention and research on the environment and crime, and insists that crime locations, characteristics of such locations, the movement paths that bring offenders and victims together at such locations, and people's perceptions of crime locations are significant objects for studies. Pattern theory synthesizes its attempt to explain how changing spatial and temporal ecological structures influence crime trends and patterns. The model was constructed and tested using specialized software [8], which was also used to analyse the relationships within the data. From 150 samples, the data were divided into 2 groups: 70% of all data used for learning and 30% of all data used for testing.

The model was constructed, and the probabilistic values were calculated and stored in the Conditional Probability Table (CPT). The Expectation Maximization (EM) algorithm [14] was used for learning the data. EM is an iterative optimization method to estimate some unknown parameters Θ, given measurement data U. However, we are not given some "hidden" nuisance variables J, which need to be integrated out. In particular, we want to maximize the posterior probability of the parameters Θ given the data U, marginalizing over J:

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \sum_{\mathbf{J} \in \mathcal{J}^n} P(\Theta, \mathbf{J}|\mathbf{U}) \qquad (1)$$

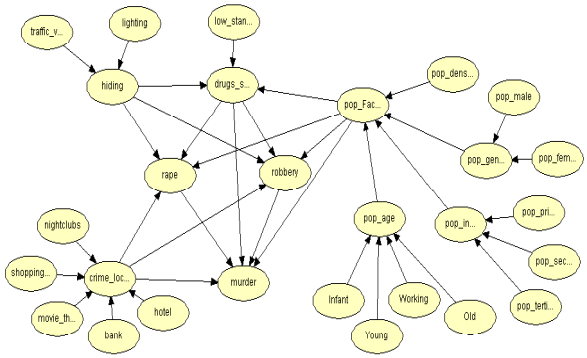

Fig.3 A Bayesian Network model for analysis of the factors affecting crime risk

3.4 The result of applying the model

According to the result of the model, the factors that considerably affected crime risk (measured by the expected increase in murder rate) in the Bangkok Metropolitan Area are environment, types of crimes, crime location, traffic and population, in that order. Environment factors have the highest value (11.98%) when compared to other factors. Therefore, they were found to have the greatest effect on the expected increase in murder rate. This

can be seen from the figures in Table 2. Of the environmental factors, the number of drug-sale areas in a district had a value of 14.53%. This means that it has the most powerful influence on the expected increase in murder rate. By concentrating on the elimination of the drug trade the government could greatly reduce the murder rate.

Table 2 The posterior probability of murder rate

Factors	Murder Rate
Population	4.61%
- Population density	4.78%
- Population gender	4.36%
- Population income	4.47%
- Population career	5.12%
- Population age	4.30%
Traffic volume	5.02%
Type of crime	10.86%
- Number of robbery cases	11.55%
- Number of rapes cases	10.17%
Crime location factors	6.73%
- Number of nightclubs	6.72%
- Number of shopping centers	6.97%
- Number of movie theatres	6.73%
- Number of banks	6.24%
- Number of hotels	6.97%
Environment factors	**11.98%**
- Number of low standard housing areas	14.02%
- Number of drug-sale areas	**14.53%**
- Lighting	7.40%

Table 3 The Correlation Coefficient of the murder variable relative to the other variables

Factors	Correlation Coefficient (r)
Population	0.244
- Population density	0.030
- Population gender	0.283
- Population income	0.296
- Population career	0.554
- Population age	0.091
Traffic volume	0.250
Type of crime	0.494
- Number of robbery cases	0.470
- Number of rapes cases	0.519
Crime location factors	0.510
- Number of nightclubs	0.503
- Number of shopping centers	0.438
- Number of movie theatres	0.529
- Number of banks	0.601
- Number of hotels	0.528
Environment factors	0.738
- Number of low standard housing areas	0.589
- Number of drug-sale areas	**0.617**
- Lighting	-0.468

Correlation is significant at the 0.01 level

Further analysis leads to the determination of the Correlation Coefficient of the murder variable relative to the other variables. It was calculated by Pearson's method [15]. The Correlation Coefficient (See Table 3) between the murder variable and environment variables; number of low standard housing areas, number of drug-sale areas, and lighting on the roads of the Bangkok districts have r=0.738, r = 0.589, r = 0.617 and r = -0.468 respectively and the correlation is significant at the 0.01 level. These results indicated that there is a high positive relationship between the number of murder cases and environment. There is a moderate positive relationship between the number of murder cases and the number of low standard housing areas and the relationship between the number of murder cases and the number of drugs sales areas is relatively high. In contrast, there is a negative relationship between the number of murder cases and lighting on the roads. If the lighting is low, the number of murder cases is higher and if the lighting is high, the number of murder cases is lower. The Correlation Coefficient between the murder variable and crime location factors, type of crime, traffic volume and population variables have r = 0.510, r = 0.494, r = 0.250 and r = 0.244 respectively and the correlation is significant at the 0.01 level. These results mean that there is a positive relationship between the number of murder cases, crime location factors, types of crime, traffic volume and population variables. If these variables are high, the number of murder cases is also high.

From the posterior probability of murder rate and the Correlation Coefficient results, we can see that the environment factors have the greatest effect on the expected increase in murder rate and have a relatively high correlation to the murder variable.

4 Experimental Results

4.1 Test method

In order to evaluate the prediction accuracy of the model, the Receiver Operating Characteristic (ROC) analysis [16, 17] was used. ROC analysis comes from statistical decision theory and was originally used during World War II for the analysis of radar images. From the computer science point of view, ROC analysis has been increasingly used as a tool to evaluate discriminate effects among different methods.

The ROC curve relies heavily on notions of sensitivity and specificity and these values depend on the specific data set. It has the sensitivity plotted vertically and the reversed scale of the specificity on

the horizontal axis. The scale of the horizontal axis is also called the false positive rate. The sensitivity and specificity, and therefore the performance of the model, vary with the cut-off. It simply looks at the area under the ROC curve. A value of the area under the ROC curve close to 1.0 indicates an excellent performance in terms of the predictive accuracy.

4.2 Predictive accuracy of the Bayesian Networks

It should be noted that for forecasting purposes by using ROC, the accuracy required should be greater than 0.5. Therefore, the value of 0.77 (See Fig. 4) obtained for the area under the ROC curve for the murder variable indicated a good performance of the model in terms of its predictive accuracy. This accuracy suggests that this machine learning technique can be used to analyse crime data and help in crime control planning.

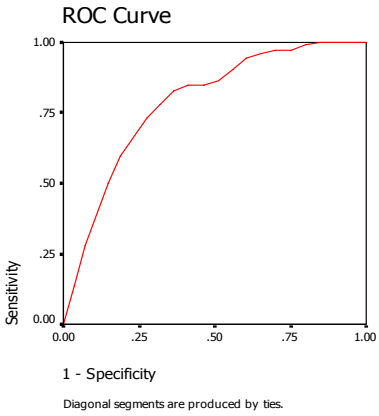

Fig.4 Receiver operating characteristic curve of murder variable

5 Conclusions

In this study, some factors that affect the crime risk in the Bangkok Metropolitan Area, Thailand were analysed. The factors considered in this research were classified into five groups. These groups were population, crime location, crime type, traffic and environment. From the result of the model, the factors that considerably affected crime risk in the Bangkok Metropolitan Area are environment, types of crimes, crime location, traffic and population, in that order. The area under the ROC curve for the model is 0.77.

The result from this analysis can be used to help in crime control planning and environmental design to prevent crime. Based on data from the study, the environmental factor, the number of drug-sale areas in a district, had the most powerful influence on the expected increasing murder rate. By concentrating on the elimination of the drug trade the government could greatly reduce the murder rate. The government can apply the model to a specific area to analyse the crime problems. For instance if we know that an area in the district has high crime rate as a result of environment factors, the government can solve the problem by redesigning the environment of the area to make it safer by cutting a new road or building a new shopping mall or sport field or increasing the number of patrol officers in that area. This will increase the number of people moving around which will make the area safer. In addition, the model would be useful to apply for new practical problems. For example it could be used to analyse the effect after cutting a new road or building a new shopping mall to see how the environment had been changed and how it affected the people in the area. The government can use the analysis result to solve problems in other districts that have similar problems and environments.

In a further study, this model will be applied to geographic profiling to determine the most probable area of residence of the offender.

Acknowledgements

We thank Hugin Expert for providing us the use of the Hugin Researcher software.

References:

[1] E. Charniak, Bayesian Networks without Tears, *AI Magazine*, Vol. 12, 1991, pp.50-63.

[2] M. Strano, A Neural Network Applied to Criminal Psychological Profiling: An Italian Initiative, *International Journal of Offender Therapy and Comparative Criminology,* Vol. 48, No. 4, 2004, pp.1-9.

[3] H. Chen, H. Atabakhsh, T. Peterson, J. Schroeder, T. Buetow, L. Chaboya, C. O'Toole, M. Chau, T. Cushna, D. Casey and Z. Huang, COPLINK: Visualization for Crime Analysis, *The National Conference on Digital Government Research 2003.* http://www.digitalgovernment.org/dgrc/dgo2003/cdrom/PAPERS/, (Accessed June 15 2004)

[4] J. J. Corcoran, I. D. Wilson and J. A. Ware, Predicting the Geo-Temporal Variations of Crime and Disorder, *International Journal of Forecasting,* Vol. 19, 2003, pp. 623-634.

[5] S. Ferrari and K. Baumgartner, *Bayesian Networks for Criminal Profiling,* Laboratory for Intelligent Control, Duke University. http://www.fred.mems.duke,edu/projects/

[6] P. C. Pendharkar and R. Bhaskar, A Hybrid Bayesian Network-Based Multi-Agent System and a Distributed Systems Architecture for the Drug Crime Knowledge Management, *International Journal of Information Technology & Decision Making,* Vol. 2, No. 4, 2003, pp. 557-576.

[7] P. E. M. Huygen, *Use of Bayesian Belief Networks in Legal Reasoning,* Computer/Law Institute, Vrije Universiteit Amsterdam. http://pubs.cli.vu/pub67.php, (Accessed July 5 2004)

[8] Hugin, *Hugin Tutorials*, 2004. http://developer.hugin.com/tutorials, (Accessed February 15 2004)

[9] F. V. Jensen, *An Introduction to Bayesian Networks*, Springer, 1996.

[10] R.E. Neapolitan, *Learning Bayesian Networks*, Pearson Prentice Hall, USA., 2004.

[11] P. L. Brantingham and P. L. Brantingham, Notes on the geometry of crime, *Environmental Criminology*, Wavelend Press Inc., USA., 1991.

[12] C. R. Jeffery, *Crime Prevention through Environmental Design*, Beverly Hills : SAGE, USA., 1971.

[13] W. M. Rhodes and C. Conly, "The criminal commute: A theoretical perspective" , *Environmental Criminology*, Wavelend Press Inc., USA., 1991.

[14] F. Dellaert, *The Expectation Maximization Algorithm*, College of Computing, Georgia Institute of Technology, Technical Report number GIT-GVU-02-02, 2002. http://www.cc.gatech.edu/~dellaert/em-paper.pdf, (Accessed September 2 2004)

[15] A. Bryman and D. Cramer, *Quantitative Data Analysis with SPSS for Windows: A Guide for Social Scientists,* London: Routledge, 1997.

[16] L. K. Westin, *Receiver operating characteristic (ROC) analysis*, Umea University, Sweden, 2002.

[17] Rockit, *ROC Analysis*, Department of Radiology, University of Chicago, 2004. http://xray.bsd.uchicago.edu/cgi-bin/roc_software.cgi, (Accessed September 9 2004)