ILLINOIS INSTITUTE OF TECHNOLOGY

Crime and Victimization Risk Model (CVRM)*

OVERVIEW

Under a completed research grant from the U.S. Department of Justice, in cooperation with the Chicago Police Department (CPD), our research team at the Illinois Institute Technology (Illinois Tech) developed various mathematical techniques to analyze crime, including methods for crime mapping, forecasting, and risk modeling. This document explains a crime risk model that we developed as part of our research. Our Crime and Victimization Risk Model (CVRM) is a mathematical technique in the tradition of the statistical risk models that are used for other public-health issues.

To make the CVRM understandable to a general audience, it is described here first in plain English. Next it is described in full technical detail. Further detail will be provided in a scientific paper. This document describes an early version of the Chicago CVRM and the most recent one.

EXPLANATION OF THE CVRM

Q: What is a risk model?

A: A risk model is a statistical technique to estimate the chances that something will happen. For example, a person who smokes is demonstrably at elevated risk for lung cancer. Similarly, if an individual has been shot recently on multiple occasions, his or her risk for being shot again in the future is substantially increased. Increasingly, violence is being viewed as a public health issue, just like smoking, so similar approaches for identifying and reducing risks are now being investigated.

Q: What is the Crime and Victimization Risk Model?

A: The CVRM is a mathematical technique, defined by a set of mathematical procedures. It uses a small subset of information from an individual's crime records to assess the risk that the individual might be involved as a victim or arrestee in a shooting or homicide in the next 18 months. The CVRM uses these crime data to assemble risk factors that are used in a risk calculation. The output of the CVRM is a number that reflects an individual's level of risk relative to others. The higher the number, the greater the risk. A high number does not necessarily mean that the individual is a threat to the community. For example, it can be the case that the individual is at elevated risk of victimization, and this may be due to involvement in non-violent crime incidents. The purpose of risk models such as ours is to identify at-risk individuals so that various forms of assistance can be offered to them with the aim of changing the dynamic and avoiding tragic outcomes. Many outreach programs are following this approach. Our CVRM is designed to assist these programs in prioritizing their efforts and making best use of limited resources.

Q: What is a risk factor?

A: A risk factor is a piece of information that can be used to assess the chances that something will happen. For example, smoking is a risk factor for lung cancer. Being shot is a risk factor for being shot again.

This research project was supported by Award No. 2011-IJ-CX-K014 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this document are those of the authors and do not necessarily reflect those of the Department of Justice, the Chicago Police Department, or the Illinois Institute of Technology.

Q: What information goes into the CVRM risk factors?

A: The CVRM automatically identifies a small number of risk factors that it finds to be truly relevant, and only these factors are used in the risk assessments. The CVRM is never permitted to access personal information such as race, gender, ethnicity, place of residence, or family relationships, nor does the CVRM use other data sources that we consider improper for this purpose, such as social media, telephone data, or video surveillance. The model excludes any information that is more than four years in the past, which is considered irrelevant to a person's risk today. The model also gives greatest weight to incidents in the past few months, with less importance assigned to older events.

The following pieces of information are used by the CVRM for Chicago:

Current version:

- Incidents in which an individual was a shooting victim
- Age at the time the individual was most recently arrested
- Incidents in which an individual was a victim of aggravated battery or assault
- Slope of a line showing the trend in involvement in crime incidents as victim or arrestee
- Violent crime incidents for which an individual was arrested
- Incidents in which an individual was arrested for unauthorized use of a weapon

Previous version:

- Incidents in which an individual was a victim of aggravated battery or assault
- Age at the time the individual was most recently arrested
- Violent crime incidents for which an individual was arrested
- Incidents in which an individual was a shooting victim
- Incidents in which an individual was arrested for narcotics charges
- Slope of a line showing the trend in involvement in crime incidents as victim or arrestee
- Incidents in which an individual was arrested for unauthorized use of a weapon
- Affiliation with a gang (yes or no)

NOTE: The changes in the list of risk factors between the two models resulted from a change in the mathematical data scaling (normalization) method that was employed, as explained later. The current version and previous version have the same accuracy, but the current one uses fewer risk factors, therefore it is a little easier to interpret.

Q: What does this tell us about crime risk?

A: Not surprisingly, those at greatest risk for future violence have already been the victim of a shooting or other violent crime and have specific patterns to their arrest history, especially arrests relating to violent crimes and weapons charges. As one might expect, young people are at greater risk than older people. Our research has found that inclusion of narcotics arrests and gang affiliation have only marginal impact on the results beyond what the model discerns from the other risk factors; therefore, in the current CVRM version, these two risk factors have been omitted. Of course, narcotic arrests and gang affiliation can indeed contribute to risk, and may be very important factors to consider in other contexts. However, in this context, the other risk factors are already sufficient to accurately capture the statistical risk.

Q: How does the CVRM digest the information listed above into risk factors?

A: The CVRM turns the raw crime data listed above into risk factors through the following steps:

1. The dates of occurrence of each crime incident relating to one of the risk factors listed above are assembled, but only for dates within the most recent four-year period. Anything prior to that is considered irrelevant.
2. The CVRM "learns" to weight the crime incidents in which an individual has been involved according to how old the incidents are. For example, a crime incident taking place yesterday has

a very significant impact, while the impact of incidents in the past drops by roughly half as you go back in time. The model figures out automatically how to do this weighting in such a way as to make the risk assessments as accurate as possible.

3. The CVRM creates a preliminary risk factor by adding up weighted contributions from the crime incidents in an individual's past. A crime that took place yesterday contributes a value of 2 to the risk factor. A crime that took place in the past four years contributes a value less than 2, based on how far in the past the incident occurred. A crime that took place more than four years ago has no contribution to the risk factor.
4. The CVRM risk factors are then scaled (normalized) using standard data analysis techniques to improve accuracy of the risk assessments.

Q: How does the CVRM turn the risk factors into an actual assessment of risk?

A: The risk assessment calculation consists conceptually of two main steps:

1. The core of the risk assessment is a simple weighted sum of the risk factors derived by the method described above. In other words, all of the risk factors are added together after each one has been multiplied by a number, called a *model coefficient*. This step produces a preliminary risk assessment score. The model coefficients, which are determined automatically, are shown in Table 1 for the current and previous CVRM versions.

Table 1. Model coefficients and risk factors in CVRM

Risk factor (after time-sensitivity weighting and data normalization)	Model coefficient	
	Current CVRM	*Previous CVRM*
Incidents as shooting victim	0.3071	0.2029
Age at latest arrest	0.3056	0.5152
Incidents as victim of aggravated battery or assault	0.2627	0.6567
Trend in involvement in crime incidents	0.1413	0.1466
Violent incidents as arrestee	0.1339	0.4099
Arrests for unauthorized use of a weapon	0.1330	0.1430
Narcotics arrests	*Not used*	0.4091
Affiliation with a gang	*Not used*	0.0066

2. Research has shown that an individual's risk for involvement in violent crime is influenced by that of others with whom the individual has been arrested, especially if those "co-arrests" are frequent.[1,2,3] However, the CVRM does not use such co-arrest patterns as a risk factor. Instead, the CVRM uses this information in the following way. Suppose that individuals A and B have been frequently co-arrested recently, and have participated in similar patterns of crime incidents, yet A has been shot three times before, while B has been shot only twice. The preliminary risk score from Step 1 would see the two individuals as having significantly different risks because of the difference in their numbers of shootings, but the CVRM recognizes the possibility that person A was simply less fortunate than person B, and adjusts their scores to be more similar than initially calculated.

Q: How accurate are the risk assessments provided by the CVRM?

A: The CVRM aims to measure the relative risks of individuals. One way to define accuracy of a risk model is to measure the extent to which the model identifies genuine risks. We have found that, historically, among the individuals with the highest CVRM risk scores, approximately 1 in 3 will be involved in a shooting or homicide in the next 18 months. This is an extremely high risk of a deadly outcome. For comparison, a Chicago resident with no arrests in the past four years has about a 1 in 2300 chance of being a shooting victim during the same time period. As another basis for comparison, a typical middle-aged smoker may have only about a 1 in 200 chance of developing lung cancer in an 18-month period.

TECHNICAL DESCRIPTION OF THE CVRM MODEL

Introduction

In this section, we provide a technical description of the CVRM to supplement the plain-English explanation above. For simplicity of this presentation, we provide the specific algorithm used by the CVRM model in its final form. The background for our approach can be understood by reading about our prior work in disease detection from MRI images, which inspired this work.[4]

Conditional random field

The framework used in the CVRM is a conditional random field (CRF),[5] in which the features are the individual's risk factors, as explained earlier, and the CRF provides a regularization mechanism. The model is inspired by our prior work in the detection of disease in MRI images, the relationship being a regularization method to combat the effect of noise in the estimation of risk among interconnected elements (pixels in MRI, persons in CVRM). The risk factors are the most important information for assessing risk; but, the CRF regularization helps to improve performance.

In the MRI application, it is known that both the feature vectors and the labels of two neighboring pixels tend to be statistically correlated. In prior research on violent crime, analogous correlations have been extensively demonstrated to exist among individuals who are close to one another in a graph defined by co-arrests.[1,2,3] Thus, the CVRM uses a CRF based on an undirected graph in which each node represents an individual, and the nodes representing two individuals are connected by an edge if the individuals have been co-arrested at least once during the study period. Each edge has an associated weight equal to the number of co-arrests between the two individuals.

Risk factors

Most of the risk factors (Table 1) are based on crime incidents of various types. Each risk factor of this type is computed as a weighted sum of contributions from the crime incidents of that type, with the weighting for a given incident being $c(t) \quad 1 + \exp(\ \ t/41.7)$, where t is the number of days since the incident occurred. (This functional form and its parameters were learned empirically as part of model training.) The age variable is self-explanatory. The "trend" variable is the slope of a line obtained by a least-squares fit to the individual's numbers of arrests each year for the past four years. Before running the model, each variable must be normalized. In the previous CVRM version, standard unity-based normalization was applied to all the risk factors (scaling to the range [0,1]). In the current version, the normalization method for age has been changed to a nonlinear scaling that addresses the highly skewed distribution of ages, in which hardly any individuals are in their 80's and 90's, while many are young. The new scaling uses a generalized logistic function to fit the empirical cumulative distribution function of the data. For ease of interpretation by non-experts, the scale of the age variable has been inverted so that all of the model coefficients corresponding to the risk factors are non-negative.

Symbols used

r_i = estimated risk score for individual i
S = set of individuals whose risks will be evaluated
$\boldsymbol{x}_i \in \mathbb{R}^N$ = vector containing N risk factors for individual i
$\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i \in S}$ = entire set of risk-factor data
$y_i \in \{-1,1\}$ = label for individual i
$\boldsymbol{y} = \{y_i\}_{i \in S}$ = entire set of label data
$\boldsymbol{\mu}_{il} = |\boldsymbol{x}_i - \boldsymbol{x}_l|$ = difference (vector) between two individuals' risk-factor vectors
$\boldsymbol{a} =$ vector of model coefficients (see Table 1)
$a_0 = -0.7881, (previous\ version), -0.7654\ (current\ version)$
$z(\boldsymbol{x}_i) = \boldsymbol{a}^T \boldsymbol{x}_i + a_0$ (or Platt-scaled version, to adhere to probabilistic framework)
M_i = set of individuals at geodesic distance 1 from individual i in CRF graph
N_i = set of individuals at geodesic distance 2 from individual i in CRF graph
w_{il} = edge weight (number of co-arrests between individuals i and j)
$\lambda_1 = 0.1\ (both\ versions)$
$\lambda_2 = 0.1\ (previous\ version), 0.09\ (new\ version)$
$\boldsymbol{v}_1 = [0.0885, 0.0179, 0.0116, 0.1939, 0.0645, 0.0595, 0.0578, 0.0195]^T\ (previous\ version)$,
$[.04513, .0792, 0.15, 0.012, 0.1367, 0.1874]^T\ (current\ version)$,
$\boldsymbol{v}_2 = [0.1359, 0.0794, 0.0308, 0.1539, 0.1397, 0.0318, 0.1083, 0.0676]^T\ (previous\ version)$,
$[0.7981, 0.4777, 0.7981, 0.4635, 0.1746, 0.7346]^T\ (current\ version)$,

Procedure for calculating risk scores

repeat until convergence*
{
  for all $i \in S$
  {

$$Q(y_i, \boldsymbol{x}) = \exp\left(y_i z(\boldsymbol{x}_i) + \lambda_1 \sum_{j \in M_i} y_i y_j w_{ij} \boldsymbol{v}_1^T \boldsymbol{\mu}_{ij} + \lambda_2 \sum_{k \in N_i} y_i y_k w_{ik} \boldsymbol{v}_2^T \boldsymbol{\mu}_{ik} \right)$$

$$S(\boldsymbol{x}) = \sum_{y_i} Q(y_i, \boldsymbol{x})$$

$$P(y_i, \boldsymbol{x}) = Q(y_i, \boldsymbol{x}) / S(\boldsymbol{x})$$

$$y_i \leftarrow \arg\max_{y_i} P(y_i, \boldsymbol{x})$$

  }
}
$r_i = P(y_i = 1, \boldsymbol{x})$

*Until, from one iteration to the next, the change in the average value of $P(y_i = 1, \boldsymbol{x})$ drops below 0.0015.

[1] P. Kump, D.H. Alonso, Y. Yang, J. Candella, J. Lewin, and M.N. Wernick, "Measurement of repeat effects in Chicago's criminal social network," *Appl. Comput. Informatics*, vol. 12, pp 154-160, 2016.

[2] B. Green, T. Horel, and A.V. Papachristos, "Modeling contagion through social networks to explain and predict gunshot violence in Chicago, 2006 to 2014," *JAMA Intern. Med.*, vol. 117, pp. 326-333, 2017

[3] A. V. Papachristos and C. Wildeman, "Social networks and risk of homicide victimization in an African American community," Available at SSRN 2149219, 2012.

[4] Y. Artan, M.A. Haider, D.L. Langer, A.J. Evans, Y. Yang, M.N. Wernick, and I.S. Yetik, "Prostate cancer localization with multispectral MRI using cost-sensitive support vector machines and conditional random fields," *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2444-2455, 2010.

[5] J. Lafferty, A. McCallum, and F. C. Pereira, Conditional Random Fields: Probabilistic Models For Segmenting And Labeling Sequence Data, 2001.